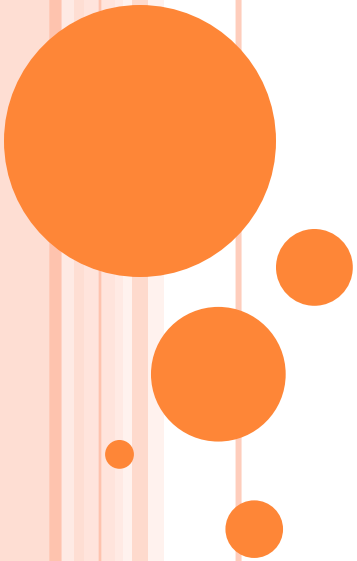


Algoritmos de recomendación

Clasificación Jerárquica


amazon


NETFLIX


Spotify®

amazon

NETFLIX

Spotify®

Premio Netflix

- Competencia abierta que premió al mejor algoritmo que permitiera predecir con anticipación la evaluación de un usuario a una película nueva utilizando únicamente las evaluaciones anteriores y ninguna información adicional respecto al usuario o la película.

US$ 1.000.000

amazon

29%

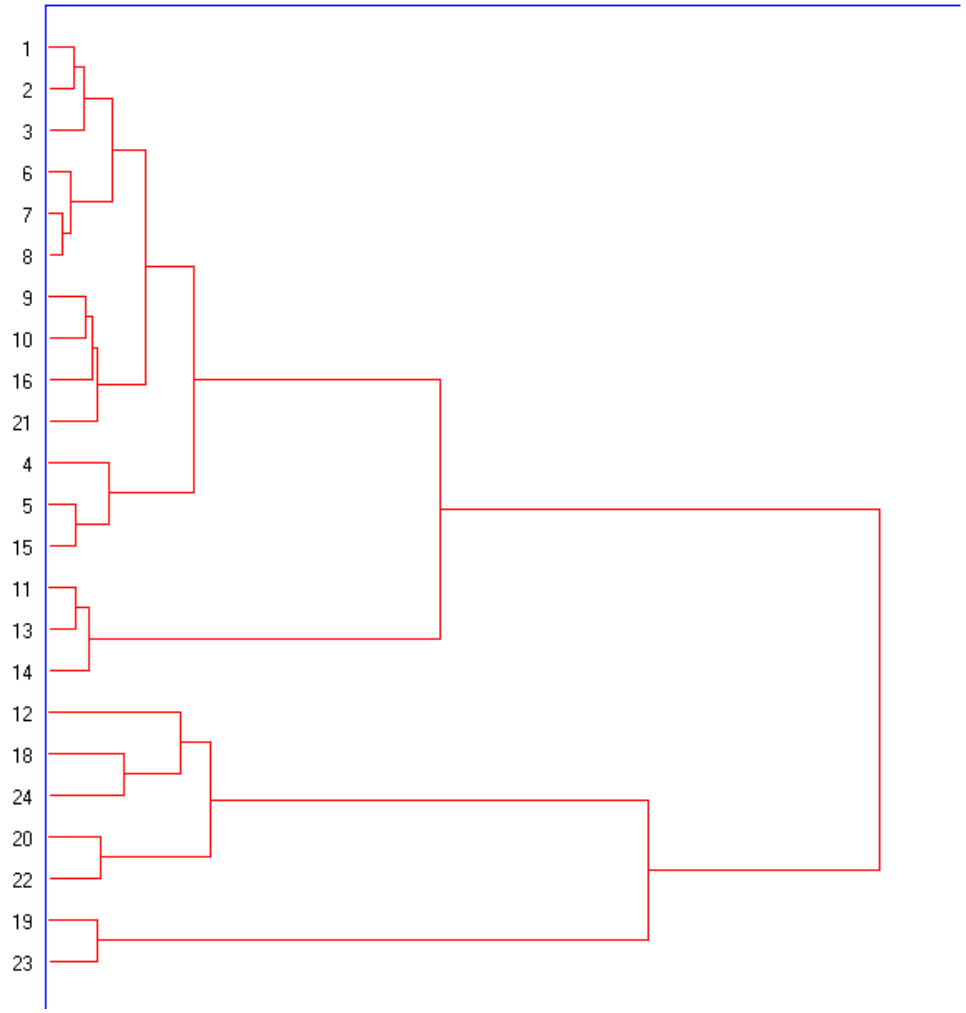
NETFLIX

10%

Spotify®

• • •

Clasificación Jerárquica



Clasificación automática

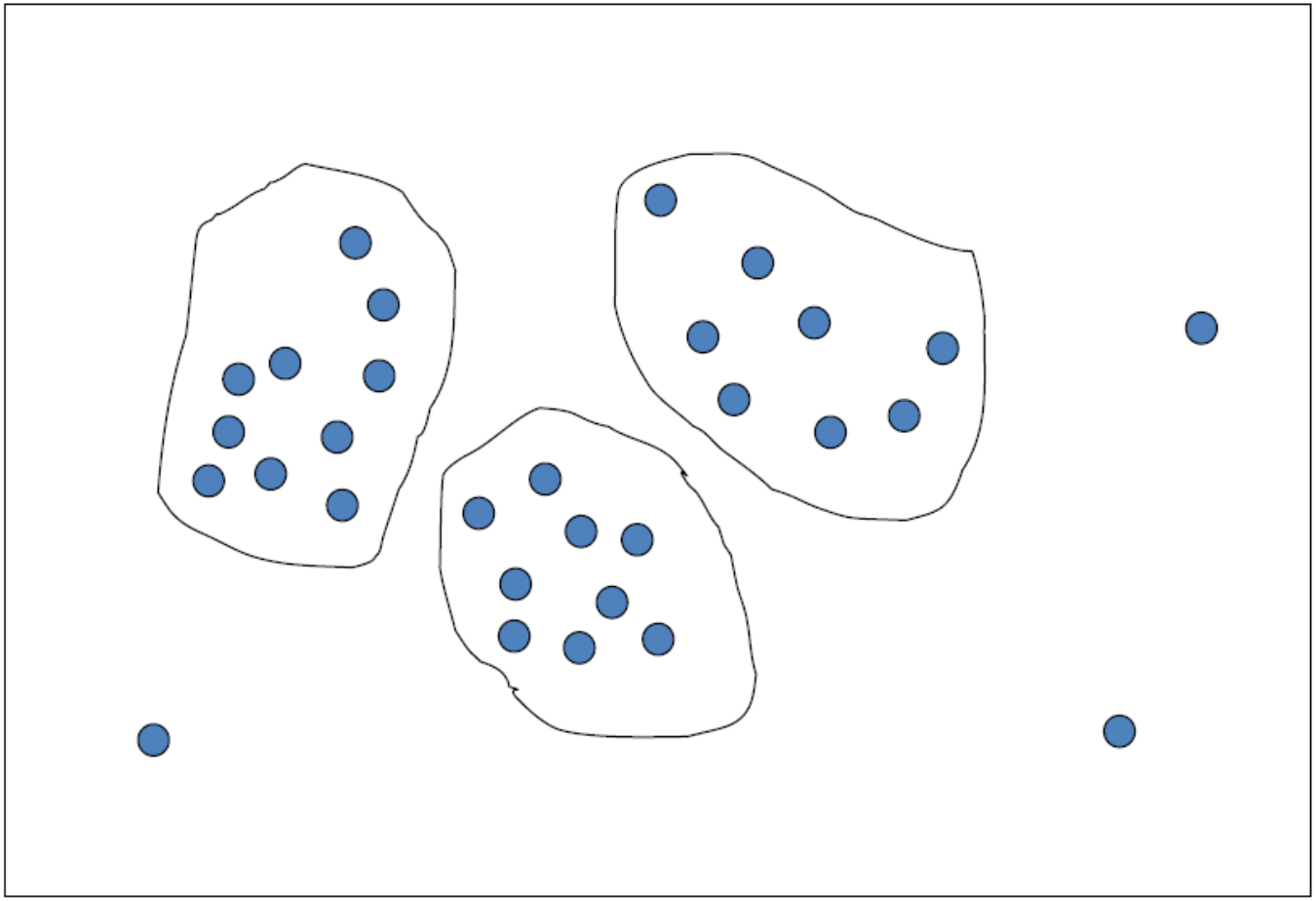
- “La clasificación automática tiene por objetivo reconocer grupos de individuos homogéneos, de tal forma que los grupos queden bien separados y bien diferenciados.”
- “Estos individuos pueden estar descritos por una tabla de datos de individuos por variables, con variables cuantitativas o cualitativas, o por una tabla de proximidades.”

Tarea de la minería de datos

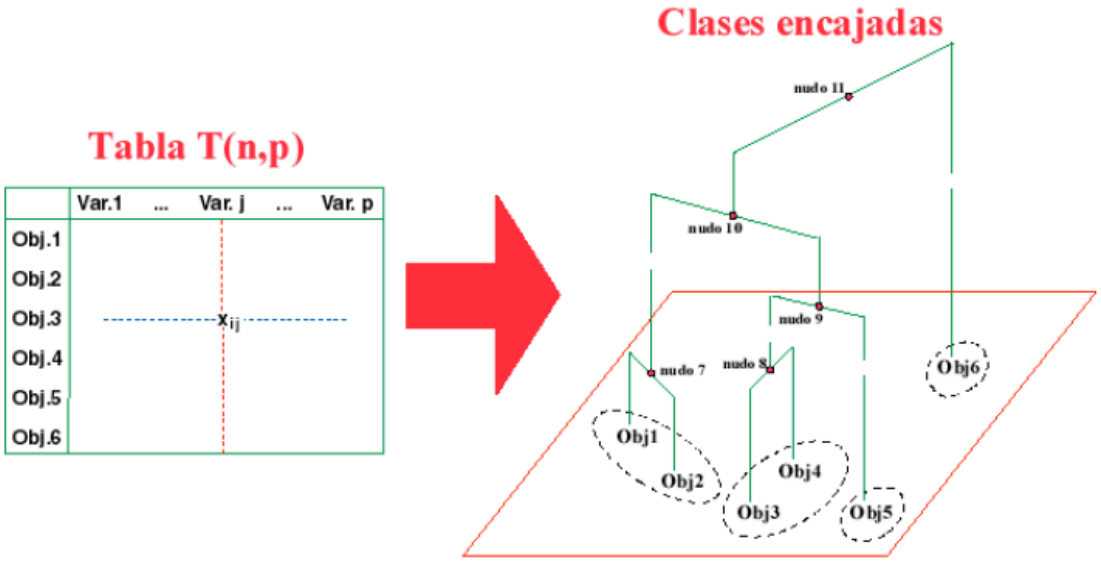
- ***"Clustering"**: Es similar a la clasificación (discriminación), excepto que los grupos no son predefinidos. El objetivo es particionar o segmentar un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similaridad de los datos o individuos en ciertas variables. Como los grupos no son dados a priori el experto debe dar una interpretación de los grupos que se forman.*
- Métodos:

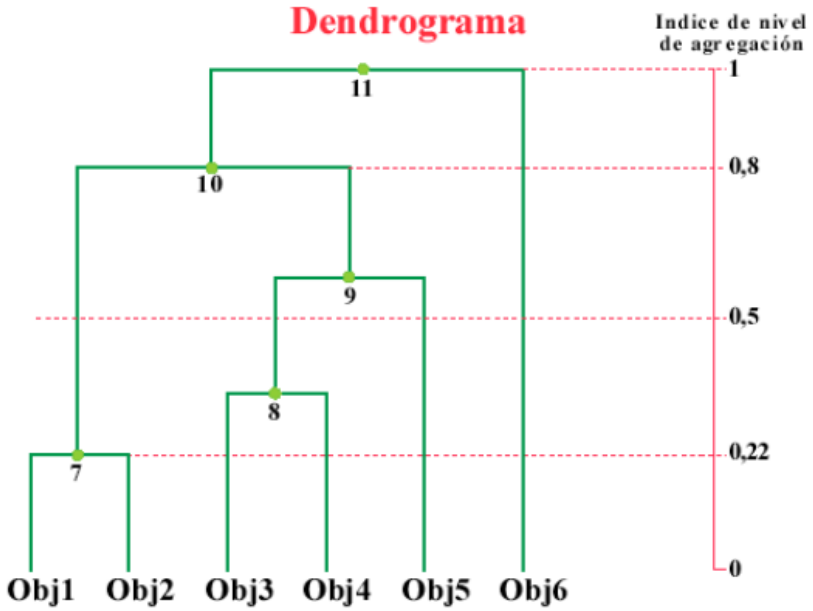
1. Clasificación Jerárquica (grupos disjuntos).
2. Nubes Dinámicas o k-means (grupos disjuntos).
3. Clasificación Piramidal (grupos NO disjuntos).

Análisis de Clúster



Clasificación Jerárquica





Definiciones básicas

Sea X la matriz de datos cuyas n filas y p columnas, forman el conjunto del cual se busca una partición. Supondremos que X es una matriz de n individuos cada uno representado por p variables.

Tabla de Datos

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

Ejemplo: Tabla de notas escolares

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Disimilitud y agregaciones

Con el propósito de encontrar una clasificación de las filas o columnas de la matriz X, el primer problema a resolver es cómo cuantificar la similitud entre esos objetos o grupos de objetos.

Algunos índices de disimilitud

Un índice de disimilitud entre objetos pertenecientes a un conjunto I, es una función d tal que:

$$d : I \times I \longrightarrow [0, +\infty[$$

y

$$d(x, y) = d(y, x) \text{ para todo } x, y \in I.$$

Ejemplos de índices

- Distancia Euclidiana: Sean $\mathbf{p} = (p_1, p_2,\ldots, p_n)$ y $\mathbf{q} = (q_1, q_2,\ldots, q_n)$ dos columnas de la tabla. La distancia Euclidiana entre ellas está definida por:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Tabla de Datos

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6.5	9.2	8.6	8
Pedro	7.5	9.4	7.3	7	7
Inés	7.6	9.2	8	8	7.5
Luis	5	6.5	6.5	7	9
Andrés	6	6	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8	6.5
Carlos	6.3	6.4	8.2	9	7.2
José	7.9	9.7	7.5	8	6
Sonía	6	6	6.5	5.5	8.7
María	6.8	7.2	8.7	9	7

Distancia Lucía-Pedro

0.25	8.41	3.61	2.56	1	3.9787

Disimilitud notas escolares

Matriz de Distancias										
	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,947	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,214	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,135782	3,265	1,69	4,77	2,53
Luis				0	3,072	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0

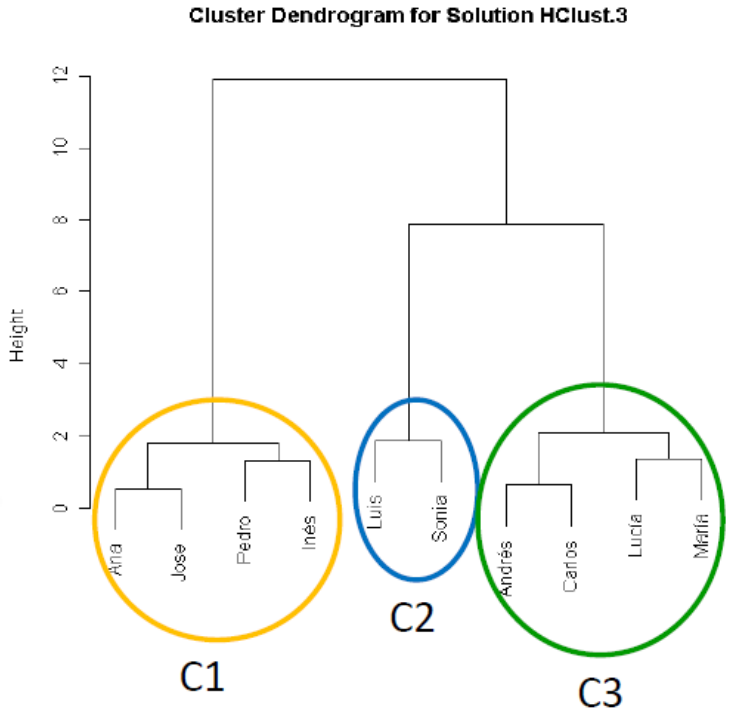
Ejemplos de índices

- Distancia Euclídia de las varianzas: cuando las variables tienen varianzas muy desiguales, la distancia entre filas puede depender más de la estructura de varianzas que de la estructura de correlaciones. Para corregir este efecto se utiliza el índice:

$$d(x_i, x_s) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sigma_j^2} (x_{ij} - x_{sj})^2}$$

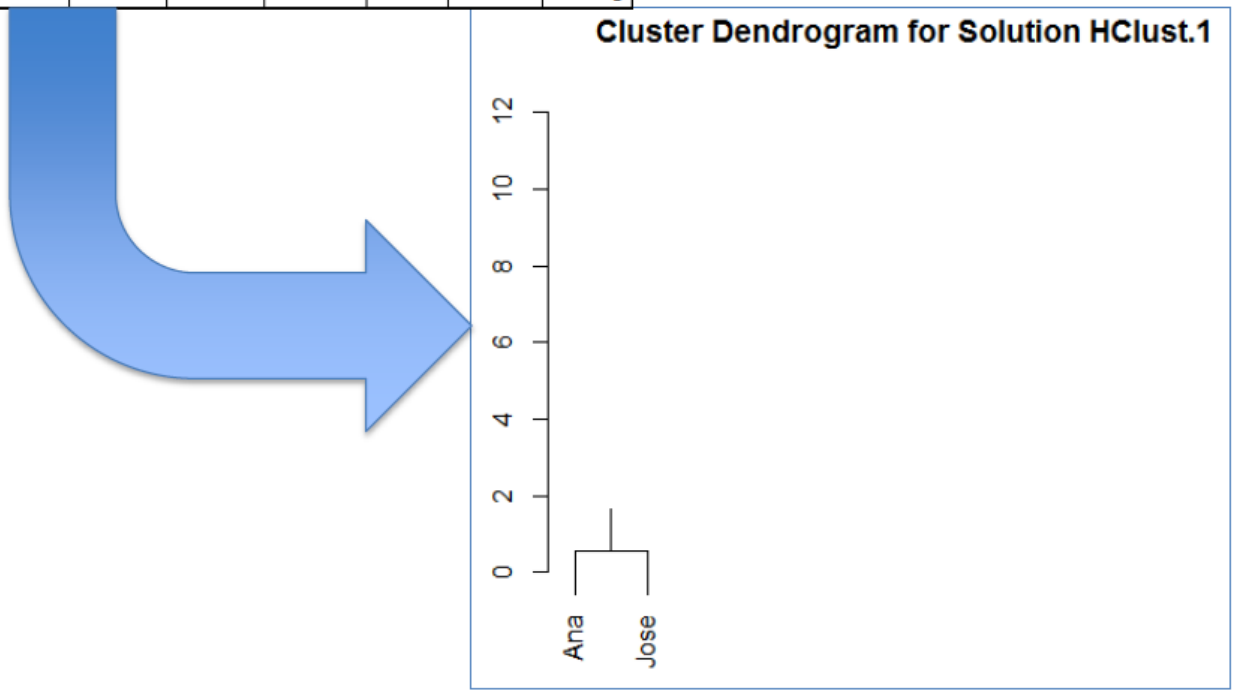
¿Cómo se construye el árbol?

Análisis de los Clústeres					
	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6,5	9,2	8,6	8
Pedro	7,5	9,4	7,3	7	7
Inés	7,6	9,2	8	8	7,5
Luis	5	6,5	6,5	7	9
Andrés	6	6	7,8	8,9	7,3
Ana	7,8	9,6	7,7	8	6,5
Carlos	6,3	6,4	8,2	9	7,2
José	7,9	9,7	7,5	8	6
Sonia	6	6	6,5	5,5	8,7
María	6,8	7,2	8,7	9	7



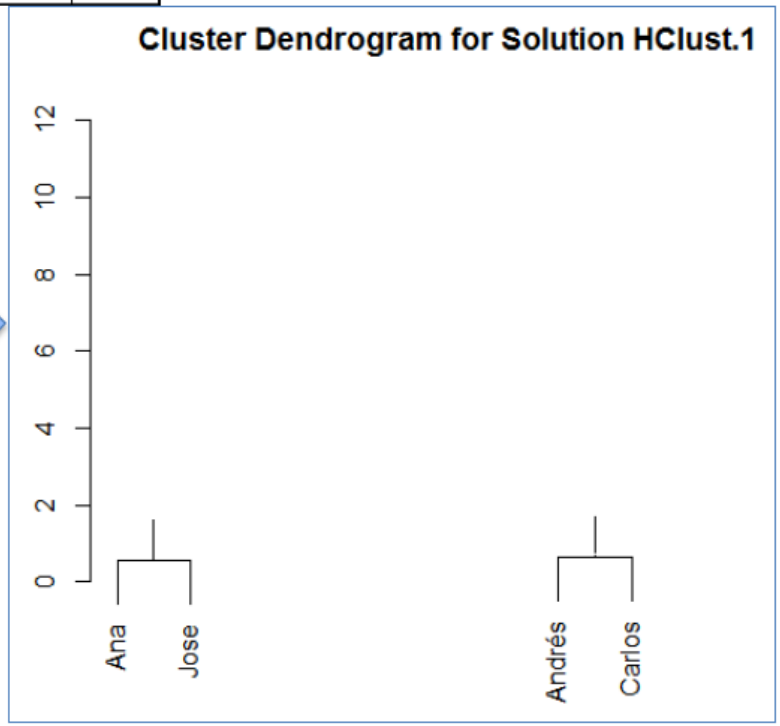
Matriz de Distancias

	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,95	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,21	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,14	3,265	1,69	4,77	2,53
Luis				0	3,07	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0



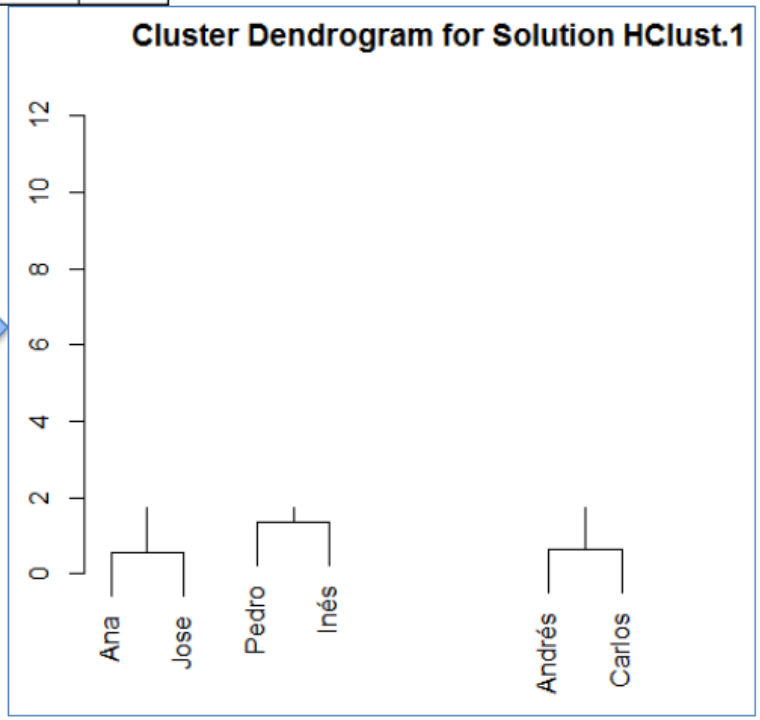
Matriz de Distancias

	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,947	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,214	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,14	3,265	1,69	4,77	2,53
Luis				0	3,072	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0



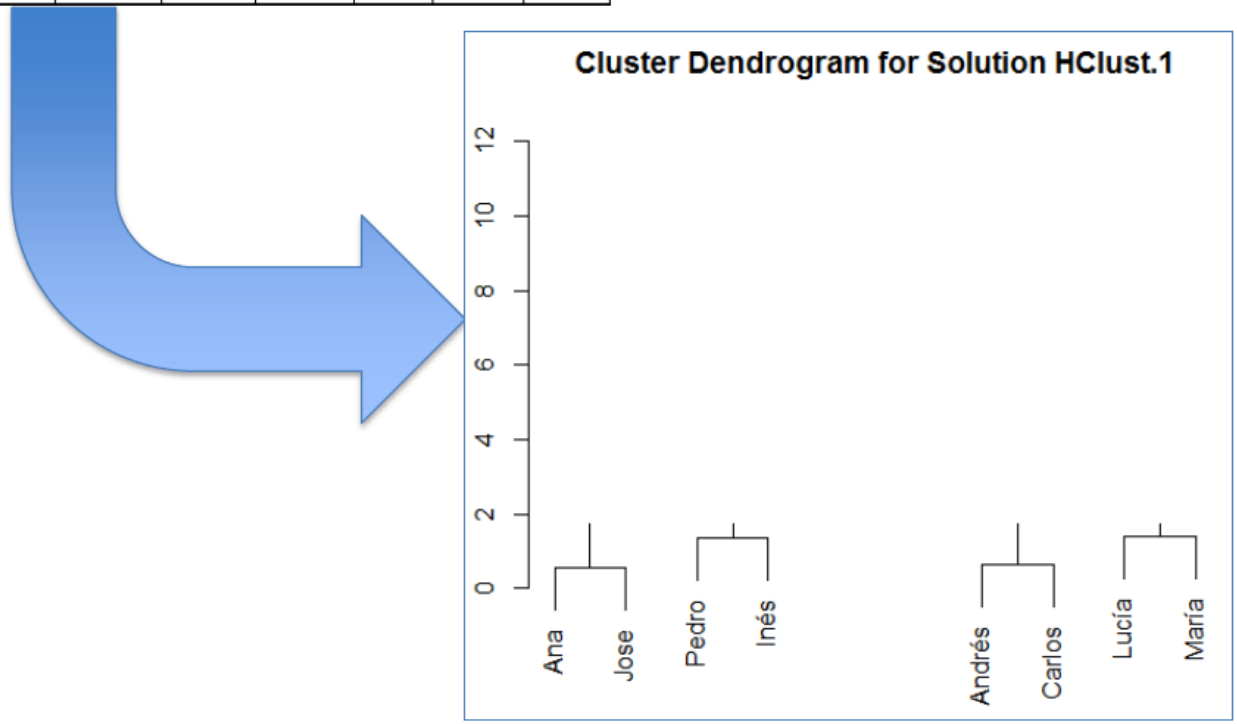
Matriz de Distancias

	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,947	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,214	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,14	3,265	1,69	4,77	2,53
Luis				0	3,072	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0



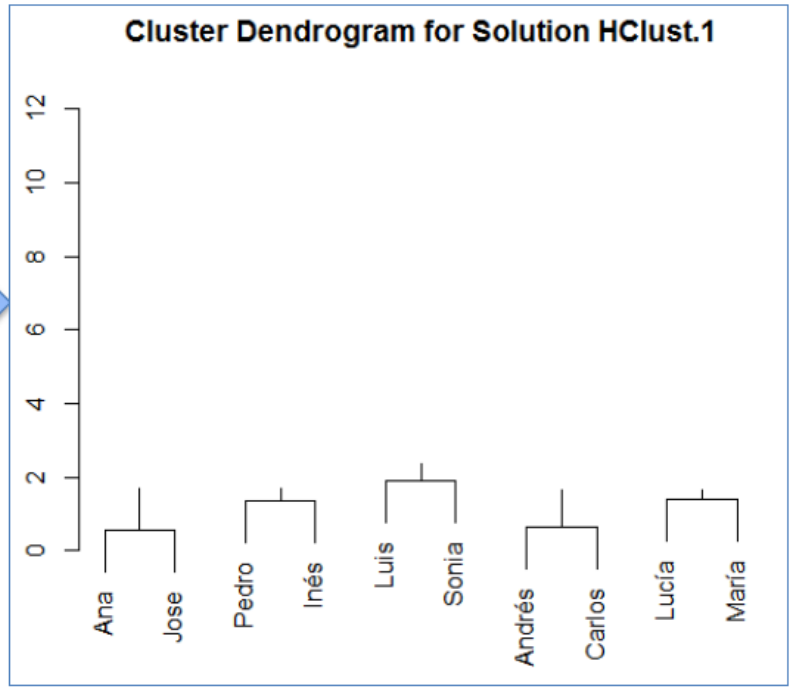
Matriz de Distancias

	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,947	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,214	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,14	3,265	1,69	4,77	2,53
Luis				0	3,072	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0



Matriz de Distancias

	Lucía	Pedro	Inés	Luis	Andrés	Ana	Carlos	José	Sonía	María
Lucía	0	3,98	3,11	3,85	1,947	3,89	1,517	4,28	4,32	1,39
Pedro		0	1,34	4,39	4,214	1,24	3,91	1,51	4,43	3,36
Inés			0	4,42	3,7	1,14	3,265	1,69	4,77	2,53
Luis				0	3,072	1,89	3,439	5,45	1,89	4,07
Andrés					0	4,2	0,656	4,46	3,9	1,73
Ana						0	3,772	0,56	5,36	3
Carlos							0	4,05	4,2	1,09
José								0	5,64	3,3
Sonía									0	4,7
María										0



Índices de agregación

- Permiten cuantificar la similitud entre grupos de objetos del conjunto a clasificar.

Una agregación es una función tal que:

$$\delta : P(I) \times P(I) \longrightarrow [0, +\infty[$$

$$\delta(x, x) = 0 \quad \forall\, x \in P(I)$$

$$\delta(x, y) = \delta(y, x),$$

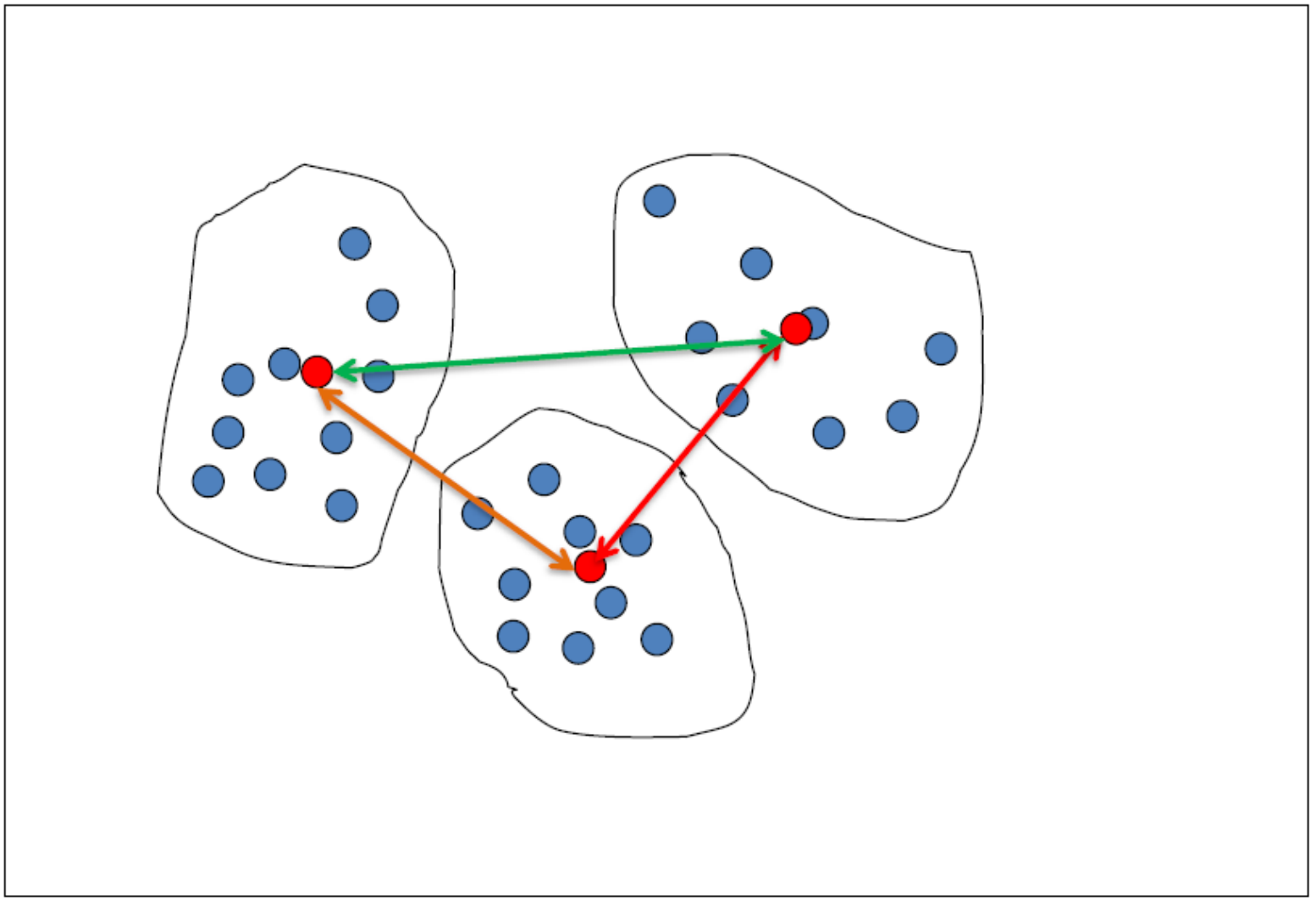
Ejemplos

- Agregación de *Ward:*

$$\delta_w(x, y) = \frac{|x| \cdot |y|}{|x| + |y|} \|g_x - g_y\|^2$$

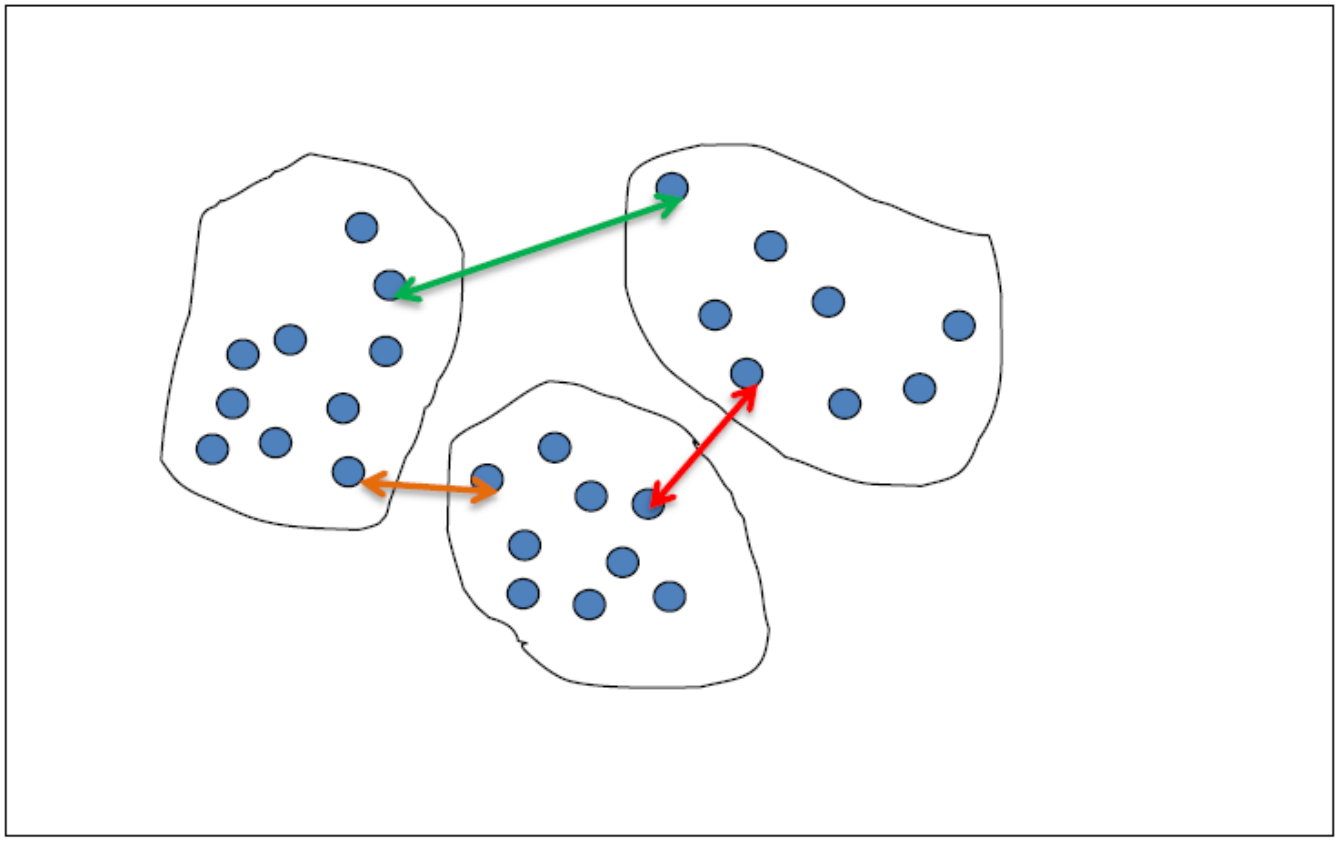
g_x es el baricentro de x.

Agregación de Ward



Agregación del salto mínimo

$$\delta_{\min}(x, y) = \min\{d(h, k) \mid h \in x \text{ y } k \in y\}.$$



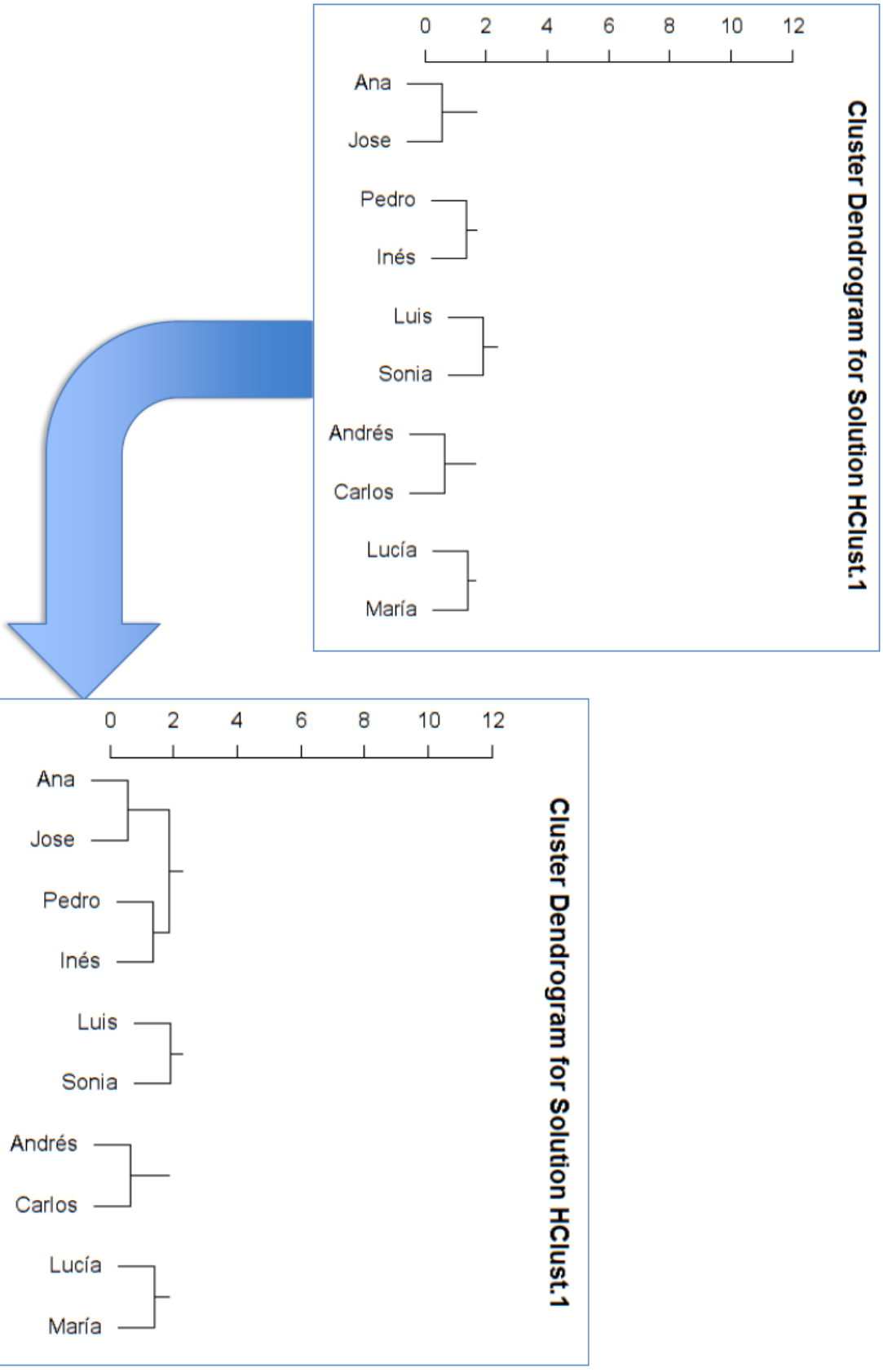
Otros ejemplos

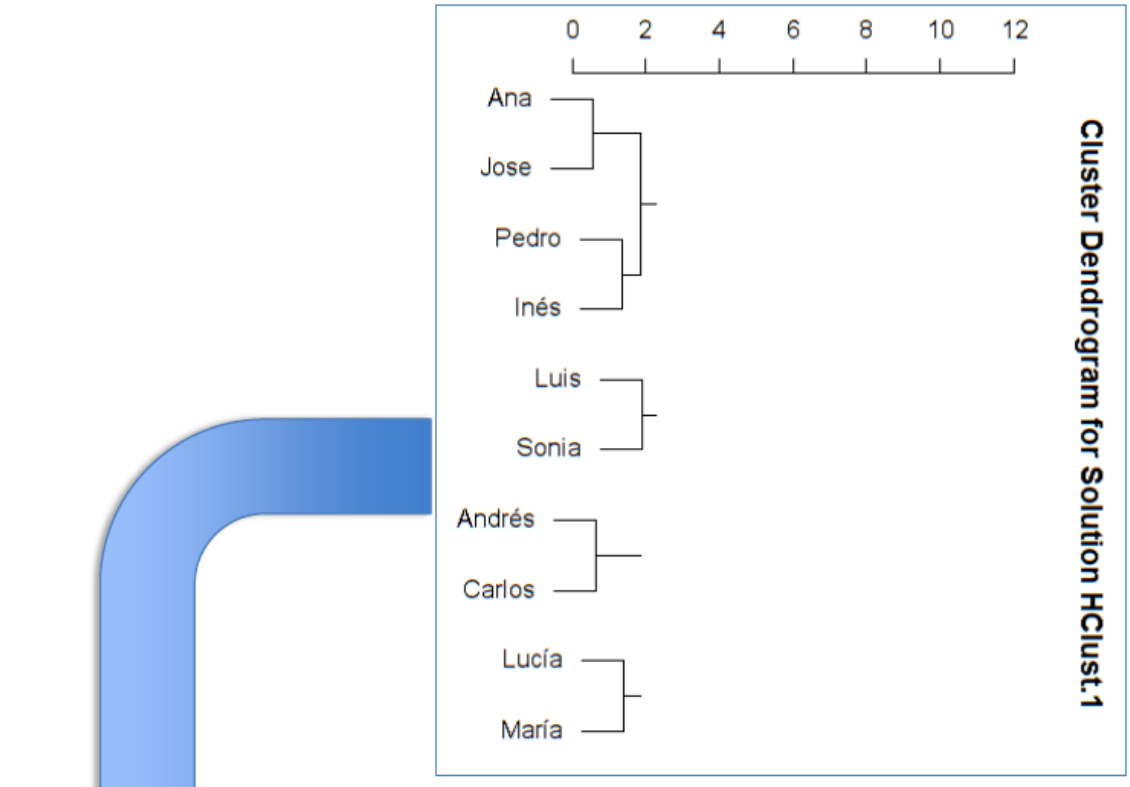
- Agregación del salto máximo:

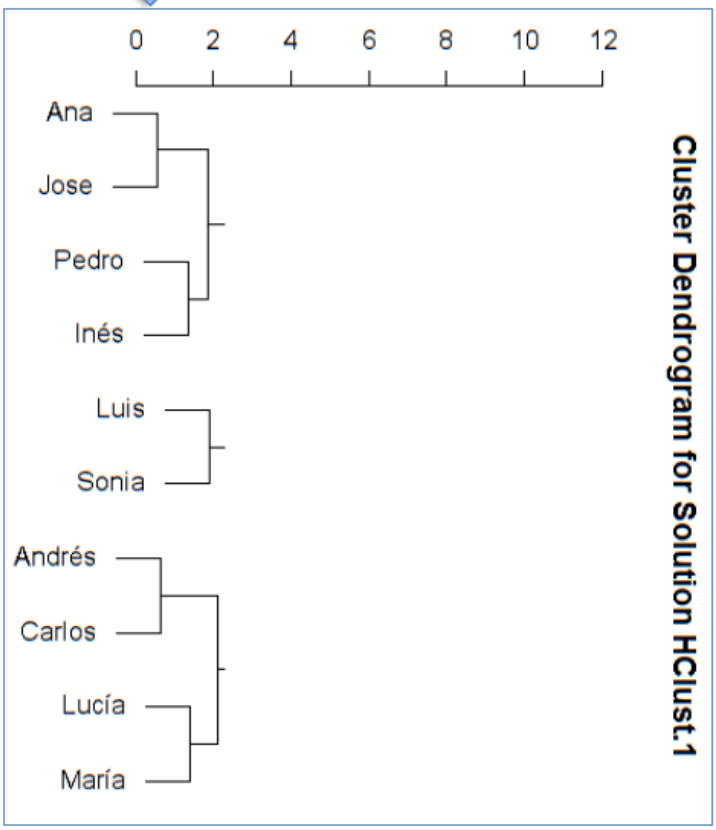
$$\delta_{\max}(x, y) = \max\{d(h, k) \mid h \in x \text{ y } k \in y\}$$

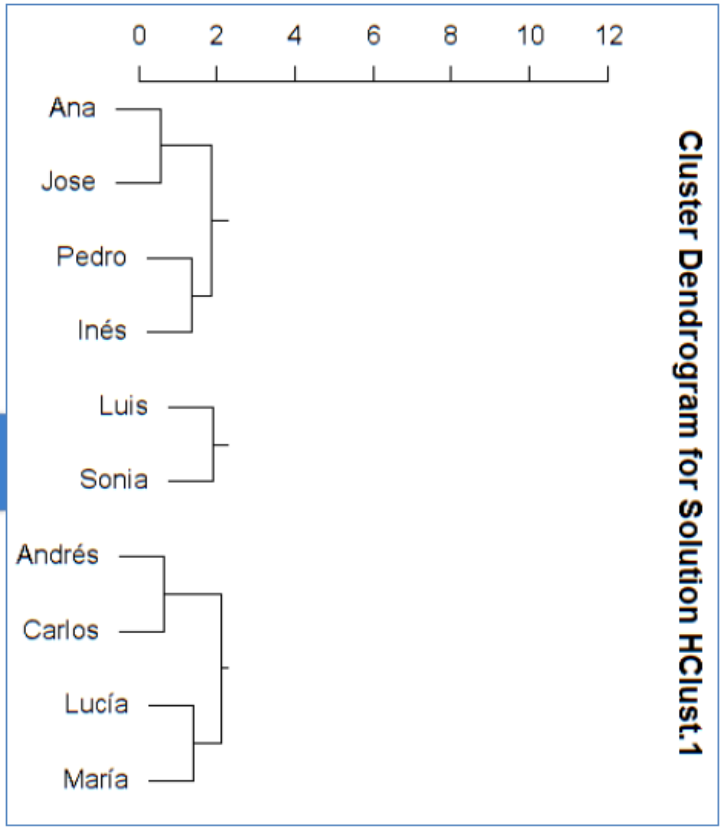
- Agregación del promedio de las disimilitudes:

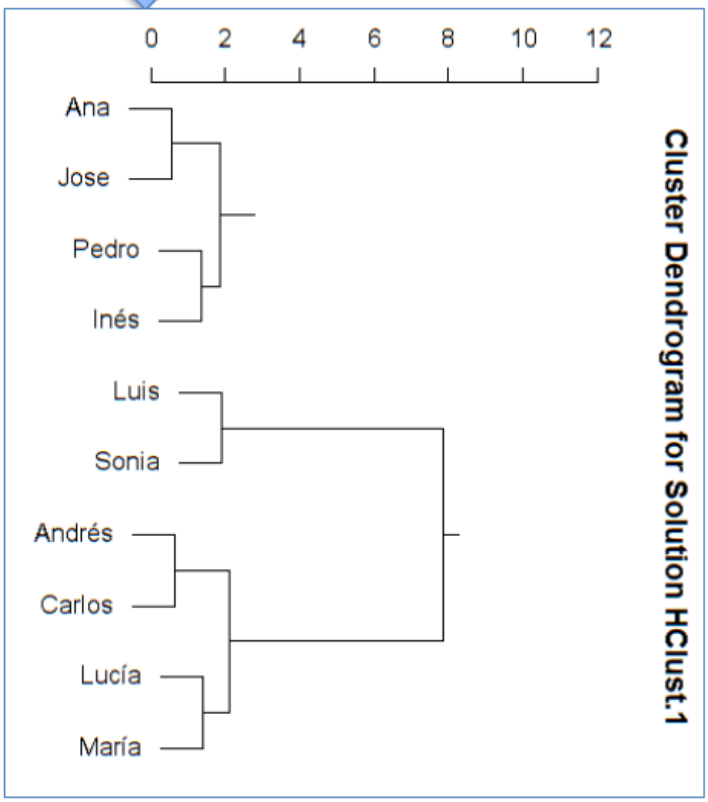
$$\delta_{\text{prom}}(x, y) = \frac{1}{|x| \quad |y|} \sum \{d(h, k) \mid h \in x \text{ y } k \in y\}$$

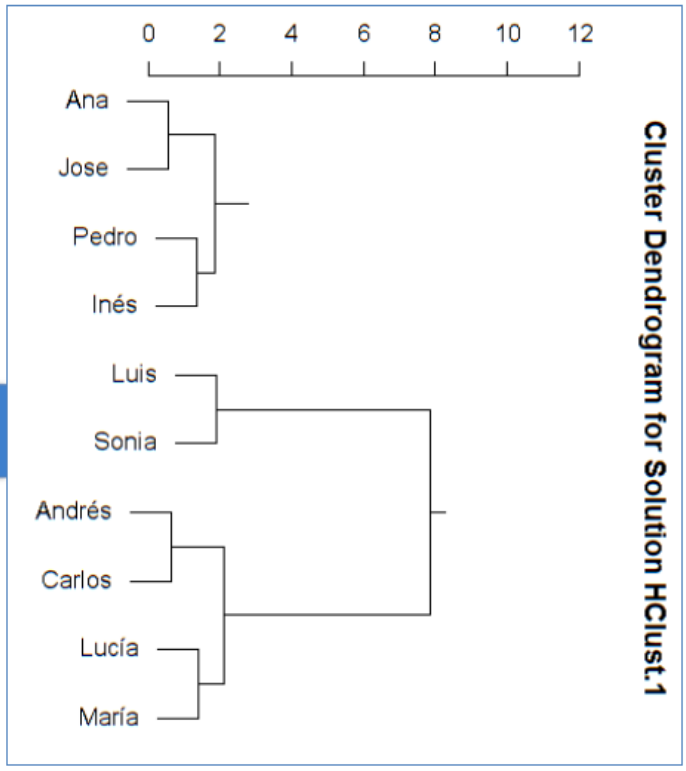

Cluster Dendrogram for Solution HClust.1
0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María
Cluster Dendrogram for Solution HClust.1
0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María

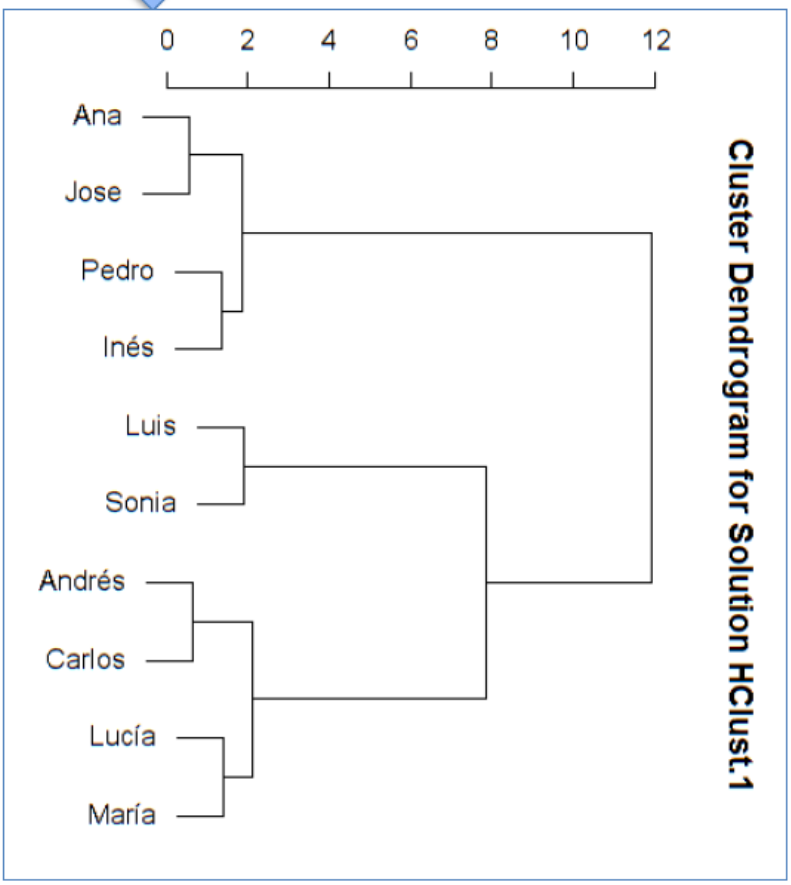

0 2 4 6 8 10 12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María
Cluster Dendrogram for Solution HClust.1


0 2 4 6 8 10 12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María
Cluster Dendrogram for Solution HClust.1


Cluster Dendrogram for Solution HClust.1
0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María


Cluster Dendrogram for Solution HClust.1
0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María


0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María
Cluster Dendrogram for Solution HClust.1


0
2
4
6
8
10
12
Ana
Jose
Pedro
Inés
Luis
Sonia
Andrés
Carlos
Lucía
María
Cluster Dendrogram for Solution HClust.1

Algoritmo

1. Inicialización: Se define P como el conjunto que contiene las clases conformadas por sólo un elemento.
2. Formación de nuevos nodos: Se funcionen los dos nodos más cercanos en sentido de la agregación elegida.
3. Actualización de P: Se agrega a P el nuevo nodo y se eliminan de él los nodos que lo conforman.
4. Test: Se detiene el algoritmo cuando el cardinal de P es mayor o igual a dos.

Ejemplo

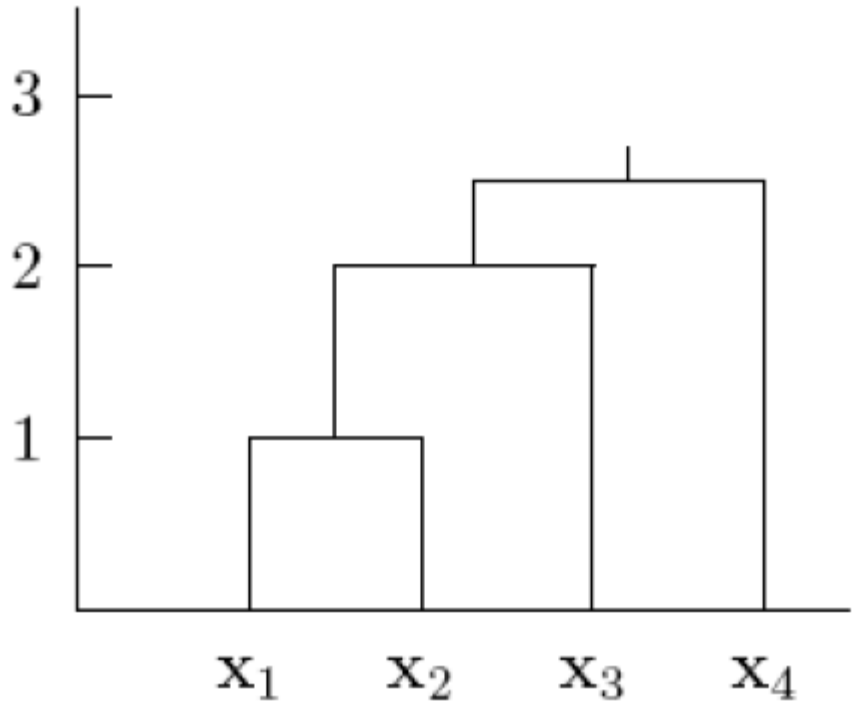
- Supongamos que tenemos los siguientes valores de disimilitud s$\Omega = \{x_1, x_2, x_3, x_4\}$:

	x_1	x_2	x_3	x_4
x_1	0	1	3	5.5
x_2		0	2	4.5
x_3			0	2.5
x_4				0

- Vemos claramente que la disimilitud mínima se alcanza para la distancia entre x1 y x2. Por lo tanto agregamos estos datos y utilizando la agregación del salto mínimo obtenemos:

	$\{x_1, x_2\}$	x_3	x_4
$\{x_1, x_2\}$	0	2	4.5
x_3		0	2.5
x_4			0

	$\{x_1, x_2, x_3\}$	x_4
$\{x_1, x_2, x_3\}$	0	2.5
x_4		0



Tarea

- Repetir el ejemplo anterior para la agregación del salto máximo.
- ¿Se obtiene el mismo dendograma?

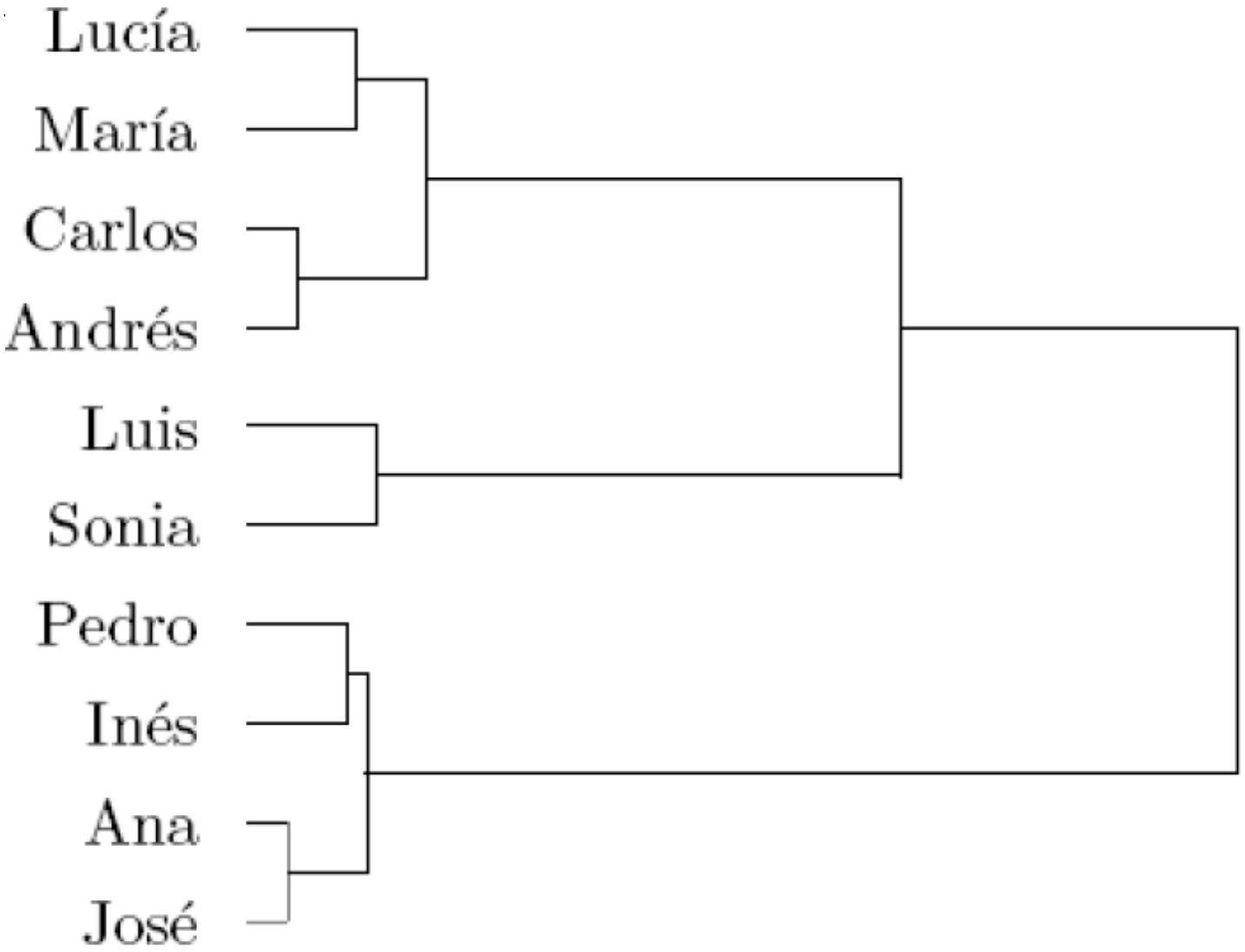
Ejemplo

- Consideremos la tabla de datos siguiente que contiene notas obtenidas por diez estudiantes en cinco materias. Todas las notas están en escala de 1 a 10.

Estudiante	Matemáticas	Ciencias	Español	Historia	Ed. Física
Lucía	7.0	6.5	9.2	8.6	8.0
Pedro	7.5	9.4	7.3	7.0	7.0
Inés	7.6	9.2	8.0	8.0	7.5
Luis	5.0	6.5	6.5	7.0	9.0
Andrés	6.0	6.0	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8.0	6.5
Carlos	6.3	6.4	8.2	9.0	7.2
José	7.9	9.7	7.5	8.0	6.0
Sonia	6.0	6.0	6.5	5.5	8.7
María	6.8	7.2	8.7	9.0	7.0

Tabla 3.1: Tabla de datos de las notas escolares.

- La clasificación jerárquica usando la agregación de *Ward* con la distancia E



Interpretación

Análisis de los Clústeres

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6.5	9.2	8.6	8
Pedro	7.5	9.4	7.3	7	7
Inés	7.6	9.2	8	8	7.5
Luis	5	6.5	6.5	7	9
Andrés	6	6	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8	6.5
Carlos	6.3	6.4	8.2	9	7.2
José	7.9	9.7	7.5	8	6
Sonia	6	6	6.5	5.5	8.7
María	6.8	7.2	8.7	9	7

Centro Gravedad C1={Pedro,Inés,Ana,José}

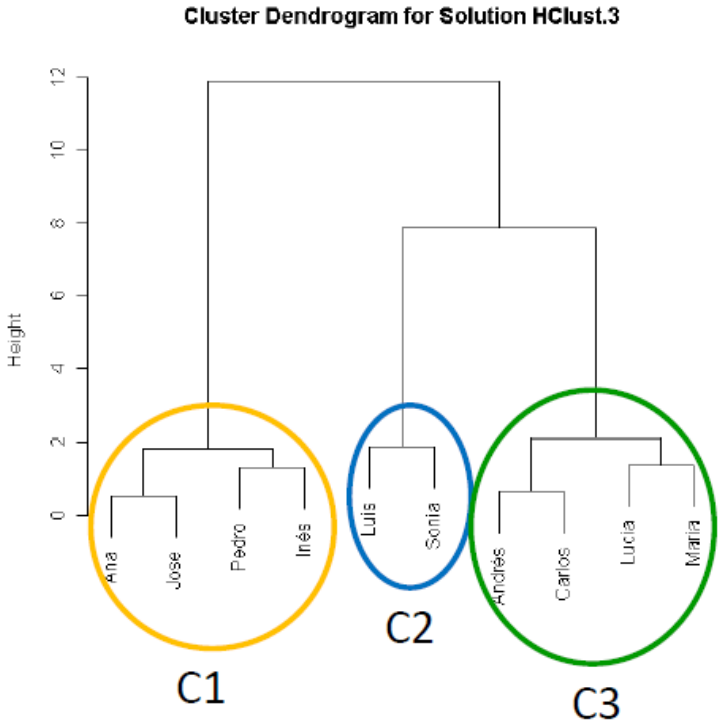
Matemáticas	Ciencias	Español	Historia	EdFísica
7.7	9.475	7.625	7.75	6.75

Centro Gravedad C2={Luis,Sonia}

Matemáticas	Ciencias	Español	Historia	EdFísica
5.5	6.25	6.5	6.25	8.85

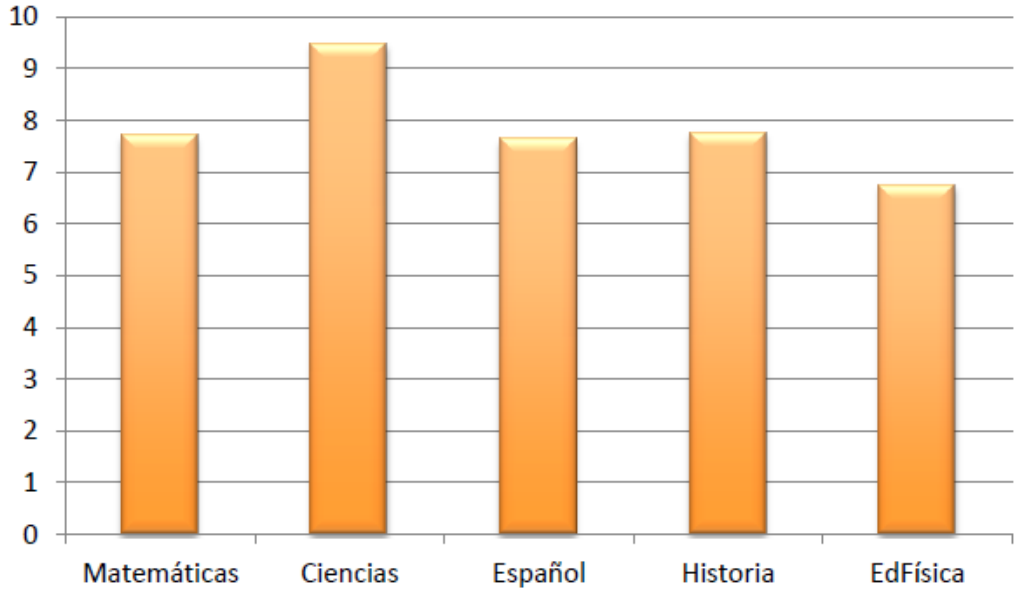
Centro Gravedad C3={Lucía,Andrés,Carlos,María}

Matemáticas	Ciencias	Español	Historia	EdFísica
6.525	6.525	8.475	8.875	7.375



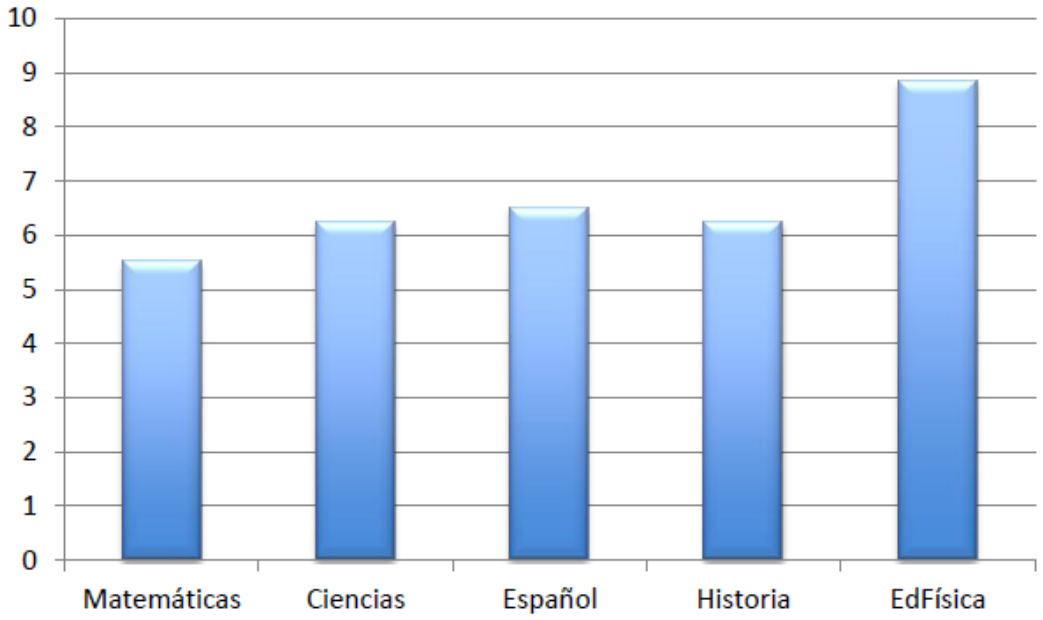
Interpretación horizontal

- El primer grupo está conformado por los estudiantes buenos en Ciencias, Matemáticas y con rendimiento promedio en las demás materias.



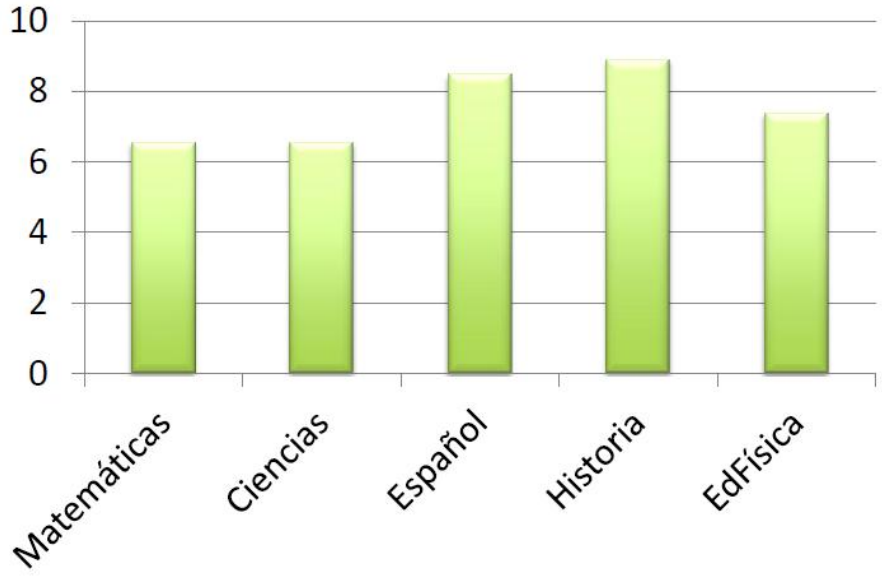
Interpretación horizontal

- El segundo grupo está conformado por los estudiantes buenos en Educación Física y que tienen un rendimiento de regular a malo en las demás materias.



Interpretación horizontal

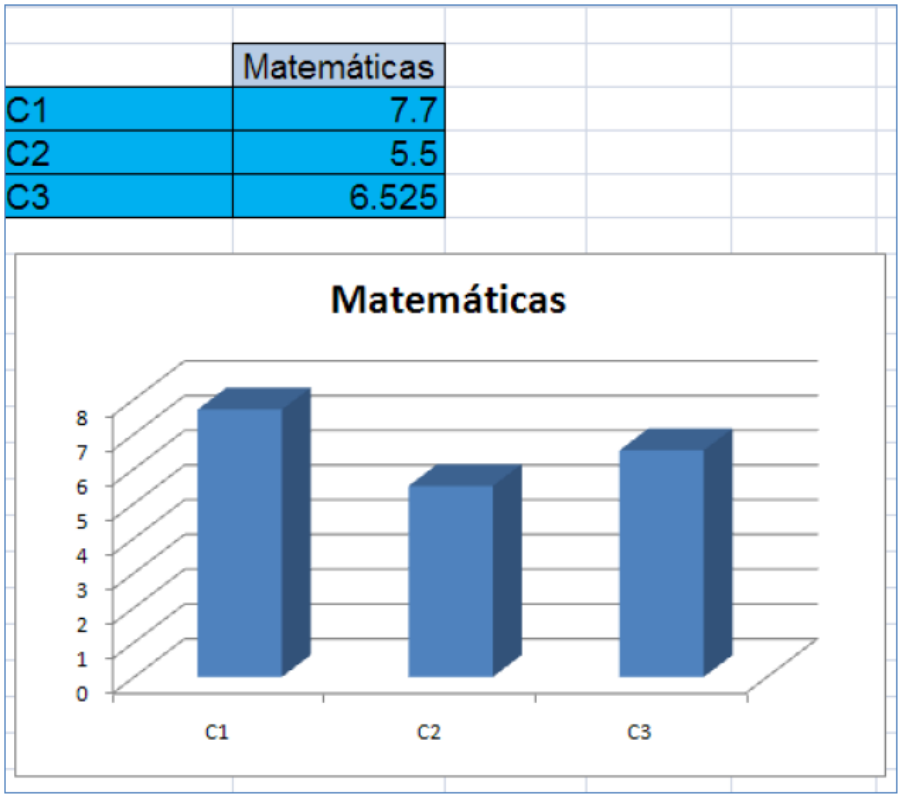
- El tercer grupo está conformado por aquellos estudiantes buenos en Español e Historia y con un rendimiento promedio en las demás asignaturas.



Interpretación vertical

- El primer grupo es el que tiene mejores resultados en matemática.

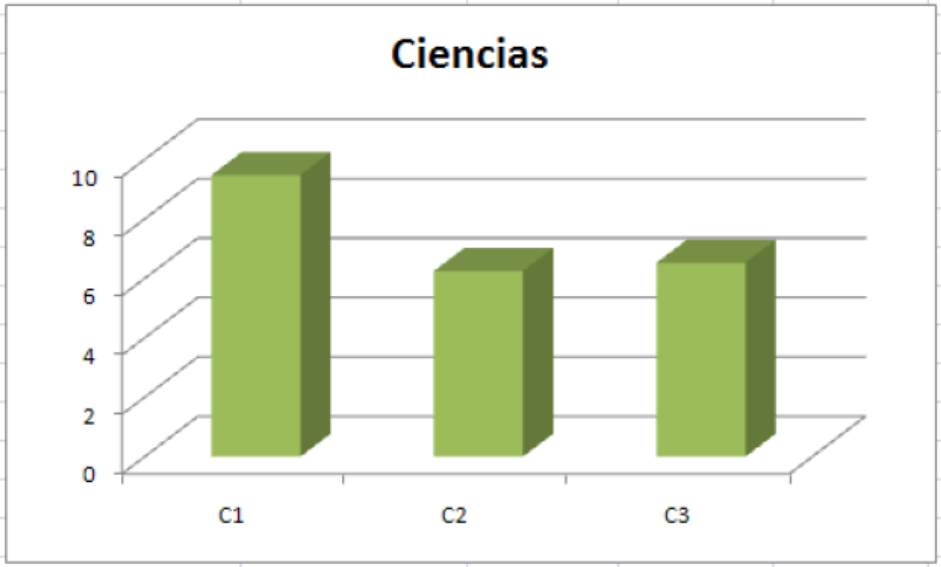
	Matemáticas
C1	7.7
C2	5.5
C3	6.525



Interpretación vertical

- El primer grupo es el con mejores resultados en Ciencias.

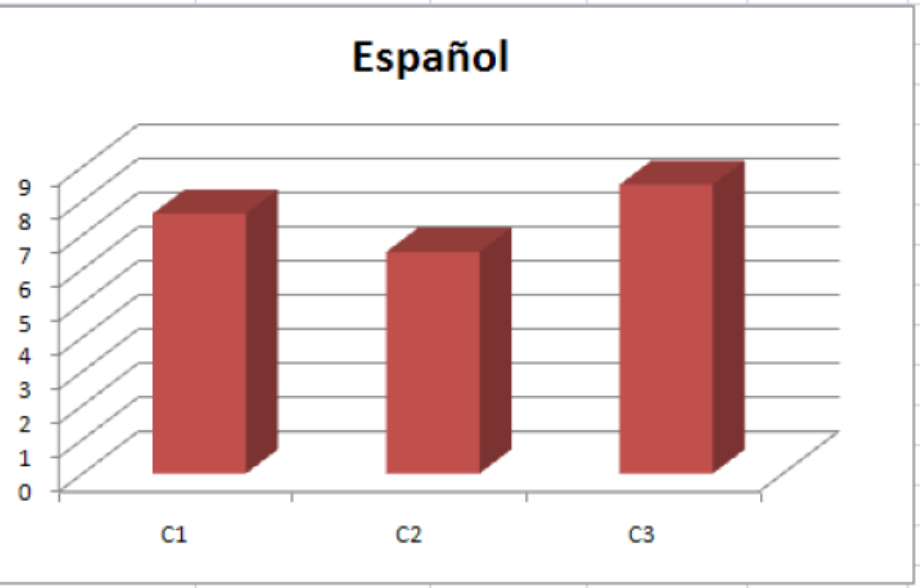
	Ciencias
C1	9.475
C2	6.25
C3	6.525



Interpretación vertical

- El tercer grupo es el mejor en Español.

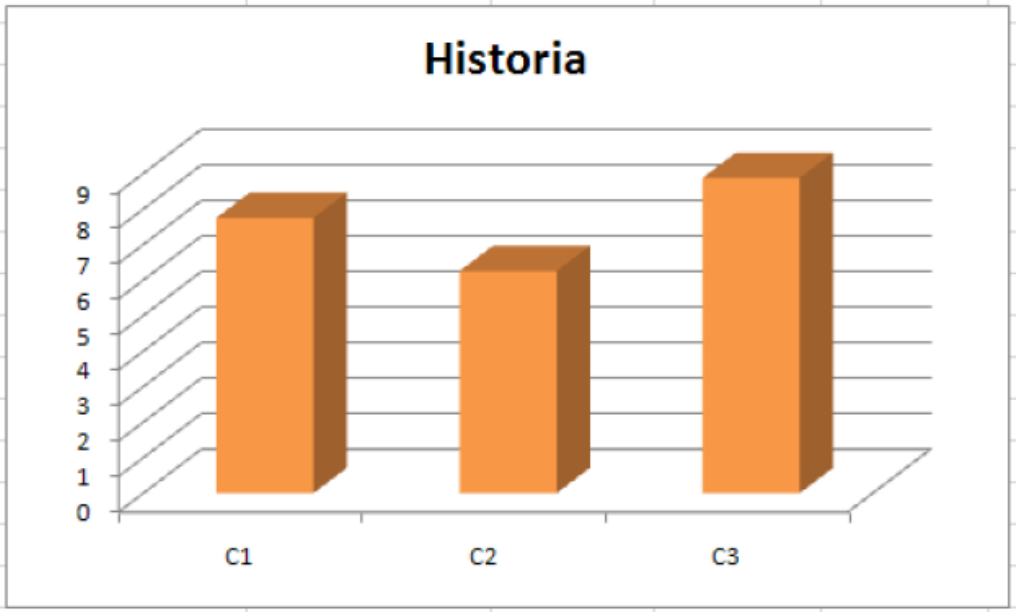
	Español
C1	7.625
C2	6.5
C3	8.475



Interpretación vertical

- El tercer grupo es el mejor en Historia.

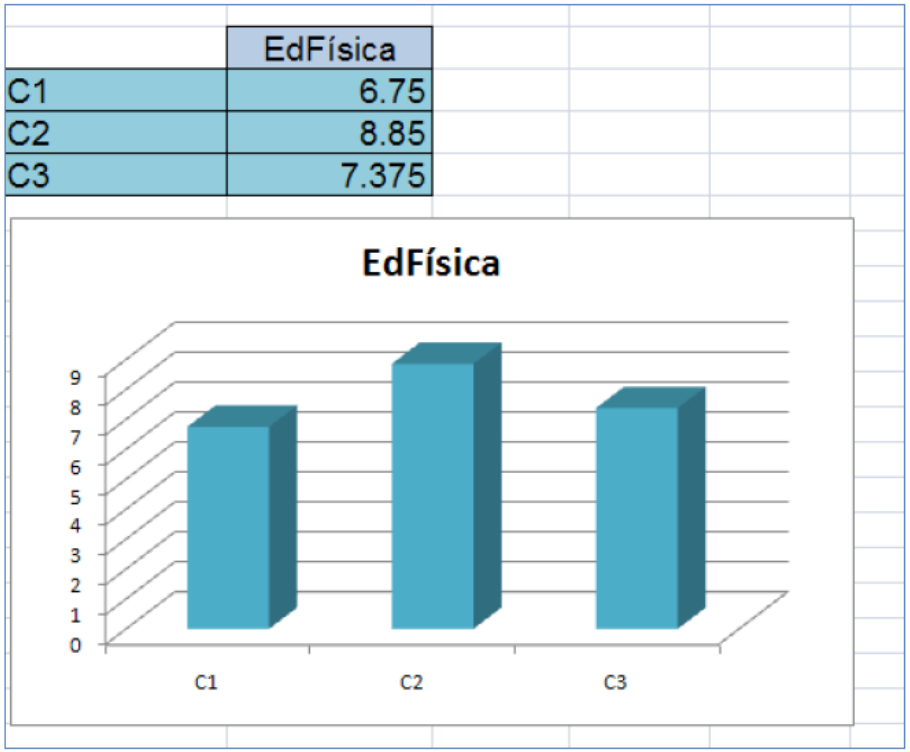
	Historia
C1	7.75
C2	6.25
C3	8.875



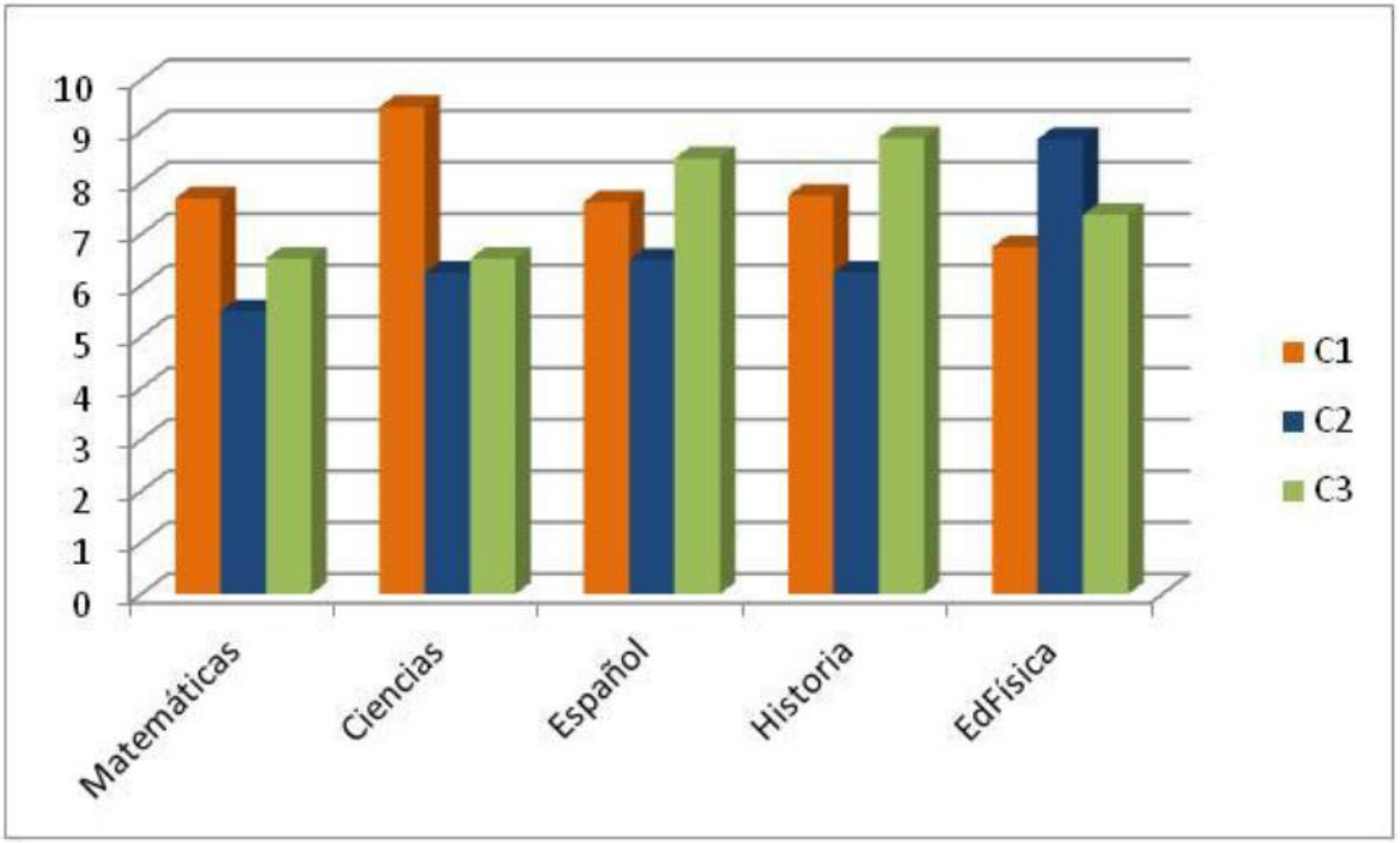
Interpretación vertical

- El segundo grupo es el mejor en Educación Física.

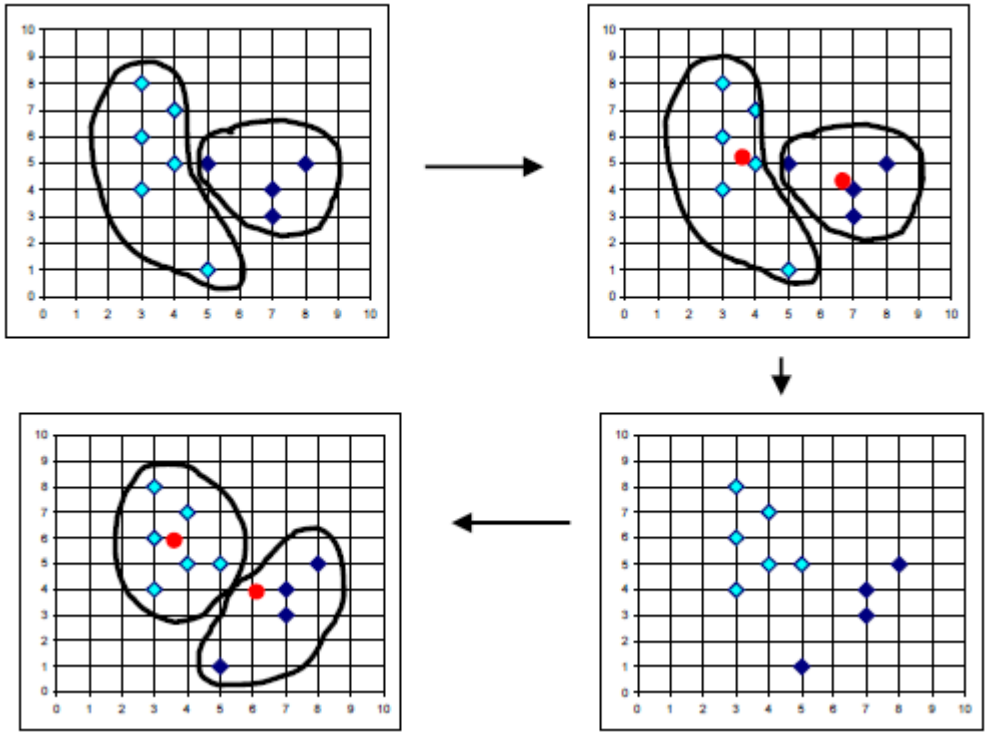
	EdFísica
C1	6.75
C2	8.85
C3	7.375



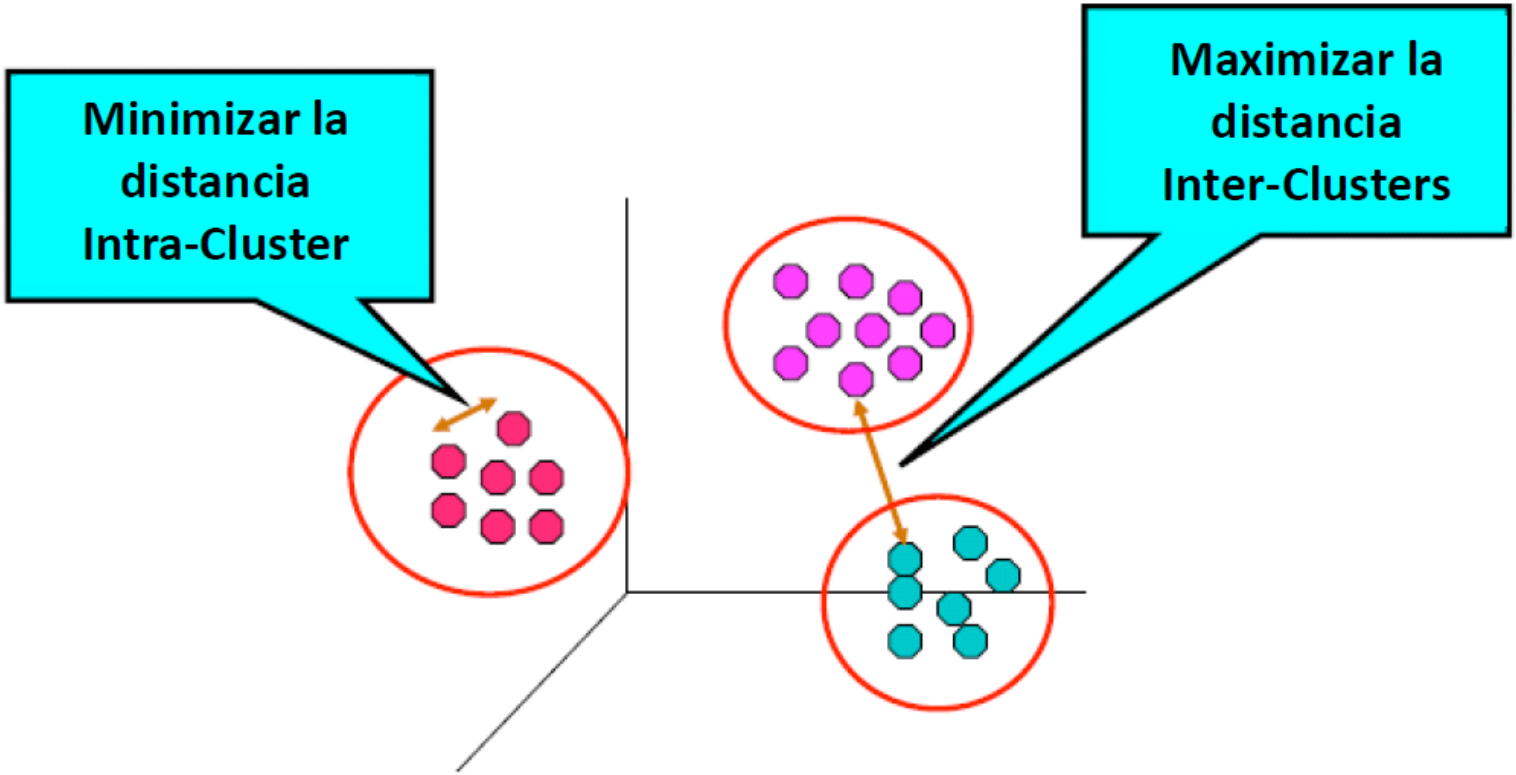
Interpretación horizontal-vertical



Método K-Means



Criterio de la inercia



Centro de gravedad

- El centro de gravedad de una clase C_K está dado por:

$$\mathbf{g}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$$

Ejemplo: Estudiantes

Ver NotasEscolaresExcelKMeans.xlsx

Análisis de los Clústeres

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6.5	9.2	8.6	8
Pedro	7.5	9.4	7.3	7	7
Inés	7.6	9.2	8	8	7.5
Luis	5	6.5	6.5	7	9
Andrés	6	6	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8	6.5
Carlos	6.3	6.4	8.2	9	7.2
José	7.9	9.7	7.5	8	6
Sonía	6	6	6.5	5.5	8.7
María	6.8	7.2	8.7	9	7

Centro Gravedad Total de la Nube de Puntos

Matemáticas	Ciencias	Español	Historia	EdFísica
6.79	7.65	7.74	7.9	7.42

Centro Gravedad C1={Pedro,Inés,Ana,José}

Matemáticas	Ciencias	Español	Historia	EdFísica
7.7	9.475	7.625	7.75	6.75

Centro Gravedad C2={Luis,Sonia}

Matemáticas	Ciencias	Español	Historia	EdFísica
5.5	6.25	6.5	6.25	8.85

Centro Gravedad C3={Lucía,Andrés,Carlos,María}

Matemáticas	Ciencias	Español	Historia	EdFísica
6.525	6.525	8.475	8.875	7.375

Inercias

- Inercia total:

$$I = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{g}||^2$$

- Inercia inter-clases: Es la inercia de los centros de gravedad de cada clase respecto al centro de gravedad total.

$$B(P) = \sum_{k=1}^{K} \frac{|C_k|}{n} ||\mathbf{g}_k - \mathbf{g}||^2$$

Inercias

- Inercia intra-clases: Es la inercia de los individuos de cada clase con respecto al centro de gravedad de la clase.

$$W(P) = \sum_{k=1}^{K} I(C_k) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$$

Teorema de Fisher

$$I=B(P)+W(P)$$

Ejercicio: ejemplo estudiantes

Análisis de los Clústeres

	Matemáticas	Ciencias	Español	Historia	EdFísica
Lucía	7	6.5	9.2	8.6	8
Pedro	7.5	9.4	7.3	7	7
Inés	7.6	9.2	8	8	7.5
Luis	5	6.5	6.5	7	9
Andrés	6	6	7.8	8.9	7.3
Ana	7.8	9.6	7.7	8	6.5
Carlos	6.3	6.4	8.2	9	7.2
José	7.9	9.7	7.5	8	6
Sonia	6	6	6.5	5.5	8.7
María	6.8	7.2	8.7	9	7

Cálculo de I=Inercia Total

	4.3246
	4.7466
	3.1426
	9.3706
	4.3646
	5.6806
	3.2726
	7.5186
	12.283
	2.5106
I=	*5.7214*

Cálculo de B(P)=Inercia Inter-Clases

	4.6434
	9.9291
	2.8287
B(P)=	4.9747

Cálculo de W(P)=Inercia Intra-Clases

1.2181	0.7763	0.8975
1.0131	0.8513	0.8975
0.1881	0.1563	1.795
0.7381	0.7313	
3.1575	2.515	

W(P)= *0.7468*

Centro Gravedad Total de la Nube de Puntos

Matemáticas	Ciencias	Español	Historia	EdFísica
6.79	7.65	7.74	7.9	7.42

I=B(P)+W(P) *5.7214*

Centro Gravedad C1={Pedro,Inés,Ana,José}

Matemáticas	Ciencias	Español	Historia	EdFísica
7.7	9.475	7.625	7.75	6.75

Centro Gravedad C2={Luis,Sonia}

Matemáticas	Ciencias	Español	Historia	EdFísica
5.5	6.25	6.5	6.25	8.85

Centro Gravedad C3={Lucía,Andrés,Carlos,María}

Matemáticas	Ciencias	Español	Historia	EdFísica
6.525	6.525	8.475	8.875	7.375

Objetivo

- Recordemos que la idea es generar clases lo más homogéneas posibles. Esto coincide con minimizar la inercia inter-clases $W(P)$ y maximizar la inercia inter-clases $B(P)$.
- Gracias al Teorema de Fisher, dado que la inercia total es constante, la optimización de cualquiera de estas inercias inmediatamente optimiza la otra.

Ejercicio combinatorio

- Calcule el número de particiones en dos clases de un conjunto de 60 elementos.
- Repita lo anterior para 100 elementos.

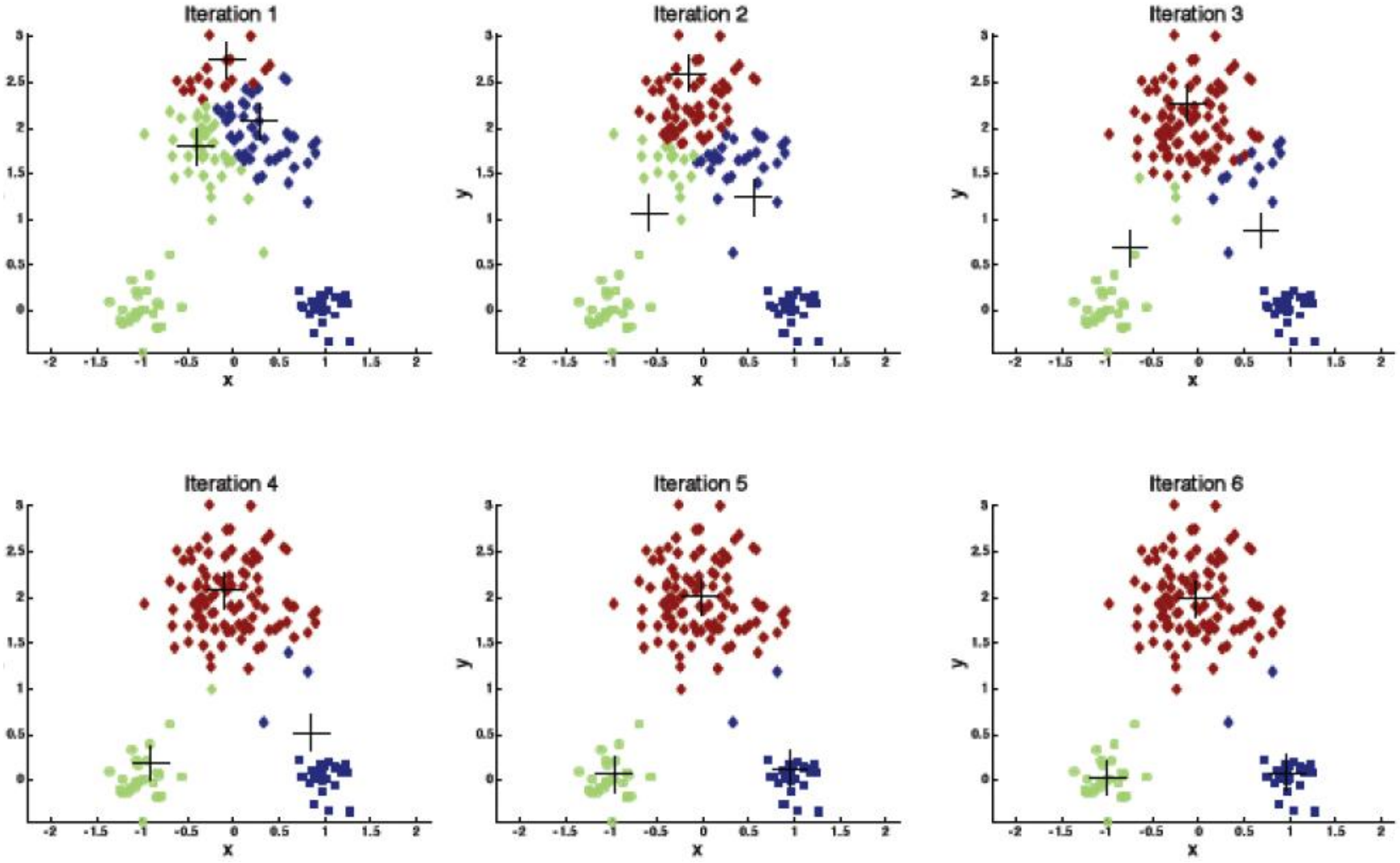
Método K-Means

Objetivo: Encontrar una partición P del conjunto de puntos y representantes de cada una de las clases de modo que $W(P)$ sea mínima.

Método de Forgy

- Consiste en un método de reasignación-recentraje que itera sucesivamente las dos operaciones siguientes hasta lograr convergencia:

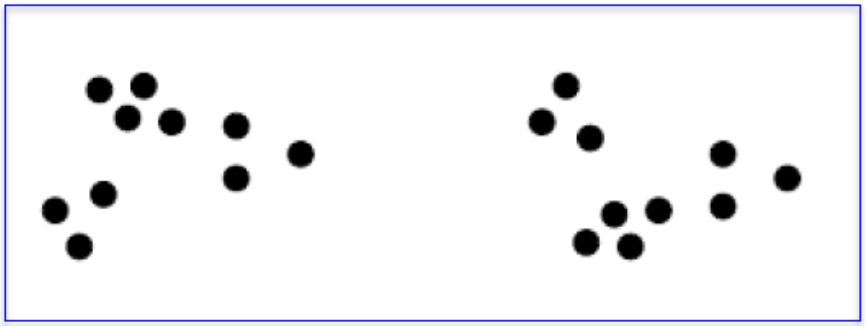
1. Representar una clase por su centro de gravedad;
2. Asignar los objetos a la clase del centro de gravedad más cercano.


Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y

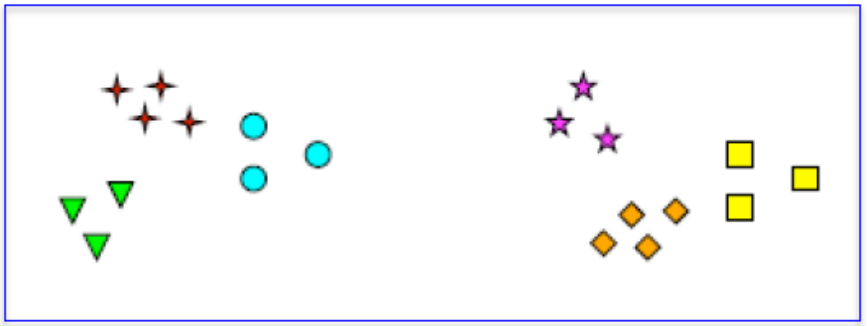
Método de McQueen

- Tal y como en el método de Forgy, las clases son representadas por su centro de gravedad y se examina a cada individuo de modo de asignarlo a la clase más cercana.
- La diferencia con el método de Forgy radica en que luego de asignar un individuo a una clase, el centro de ésta es re-calculado inmediatamente.

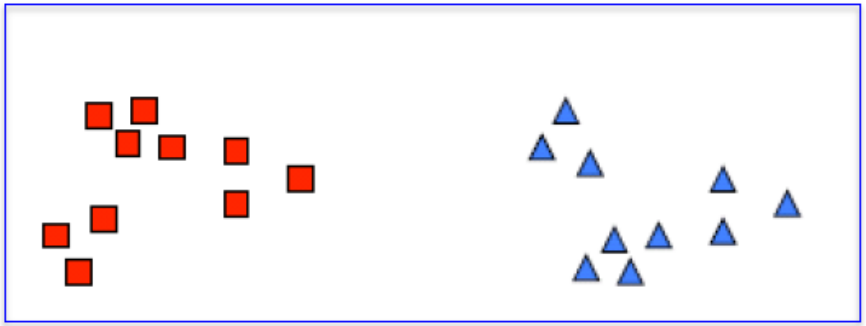
¿Cuántas clases?



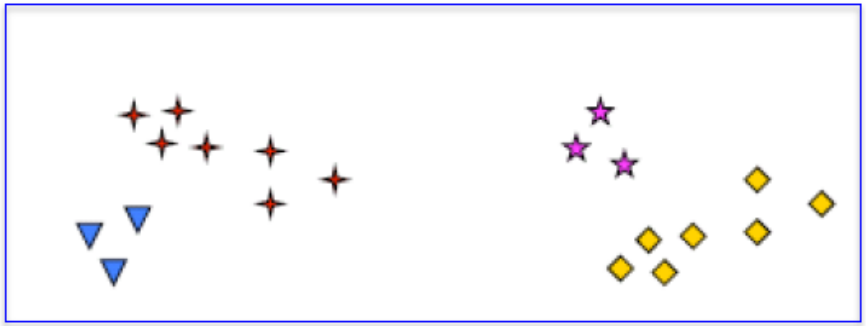
Datos originales



6 clústeres



2 clústeres



4 clústeres

Idea

- Graficar la inercia intra-clases versus el número de clases. A esto se le llama “codo”.