Análisis de conglomerados

Objetivo

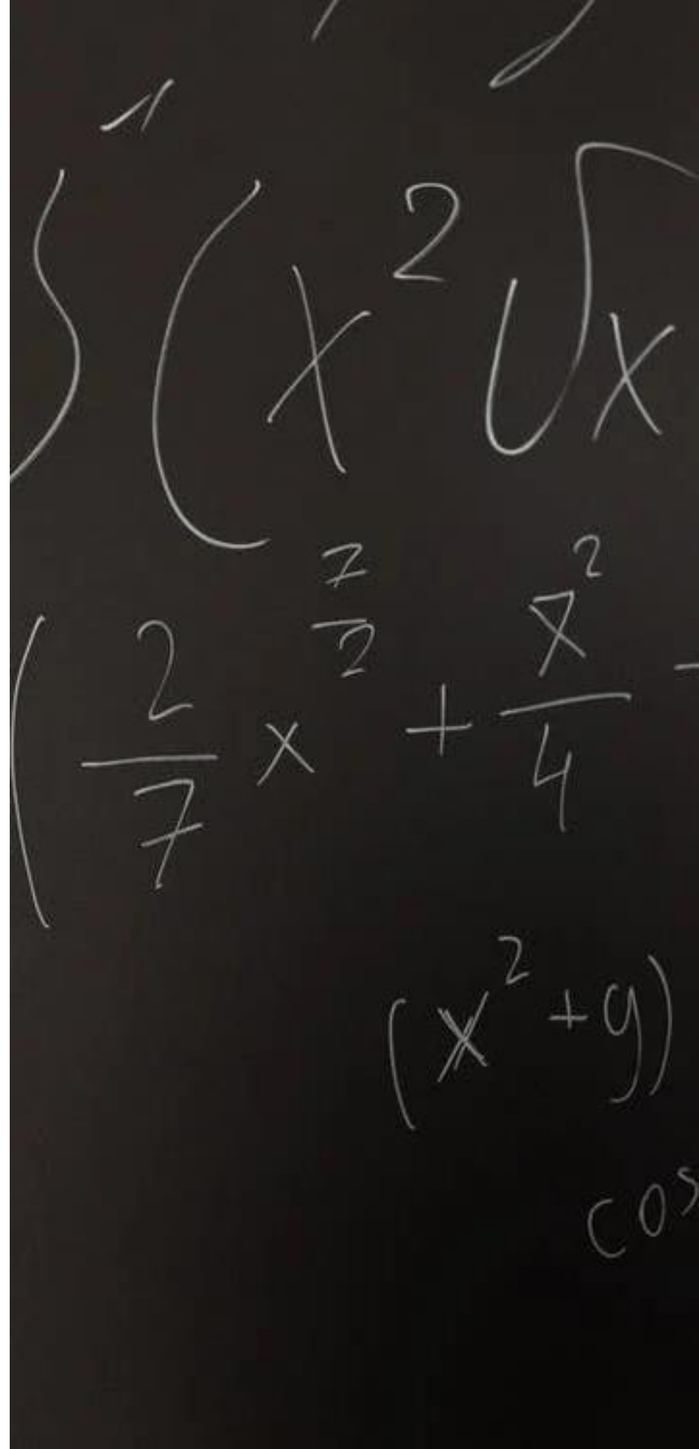
Análisis de conglomerados

- Comprender el análisis de conglomerados.
- Diferenciar conglomerados jerárquicos y no jerárquicos
- Resolver un problema utilizando python y un conjunto de datos de UCI Machine Learning Repository.

Desarrollo de conceptos

Qué es?

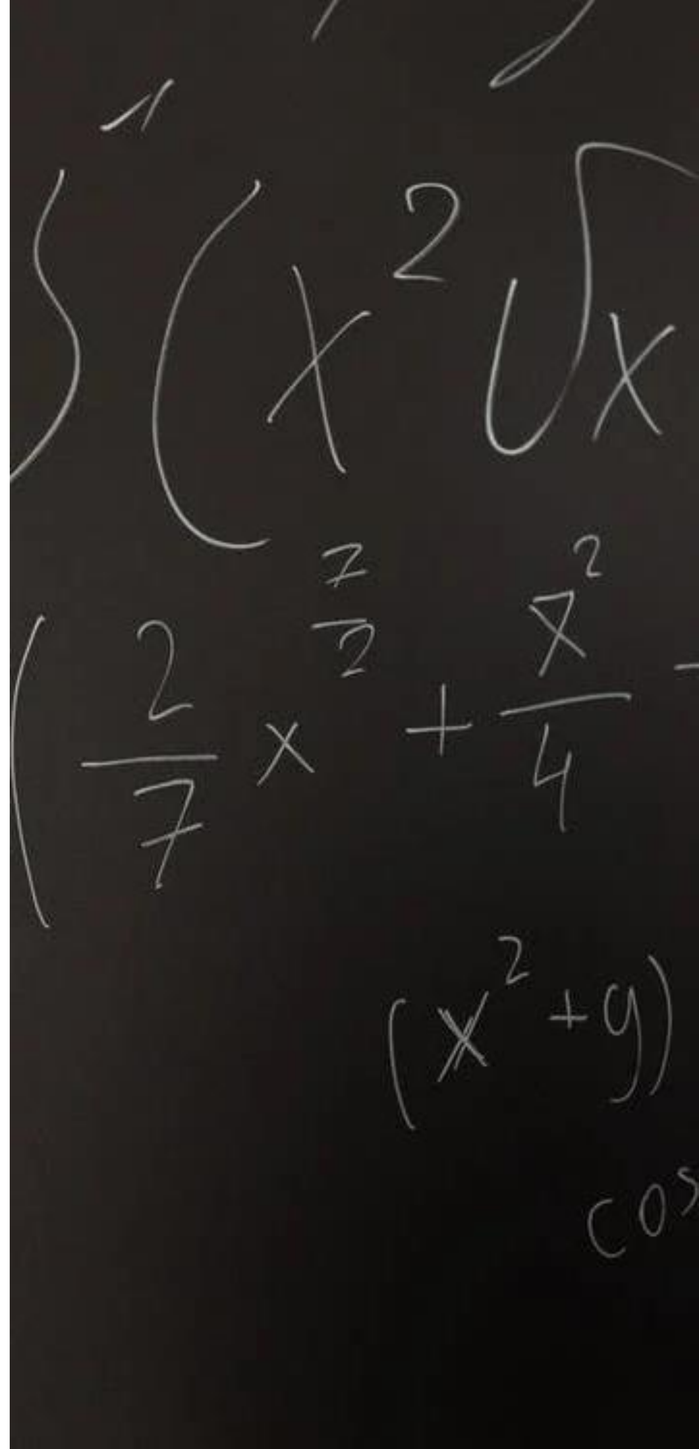
- Es el procedimiento estadístico más generalizado que considera la agrupación de objetos o casos en función a su similitud o disimilitud.
- Formar grupos de casos a partir de la proximidad que se haya establecido entre ellos.
- Países, instituciones, individuos, grupos, asociaciones, nos permite establecer posibles tipos diferenciados, en función a las características que les hacen estar próximos.



Desarrollo de conceptos

Qué es?

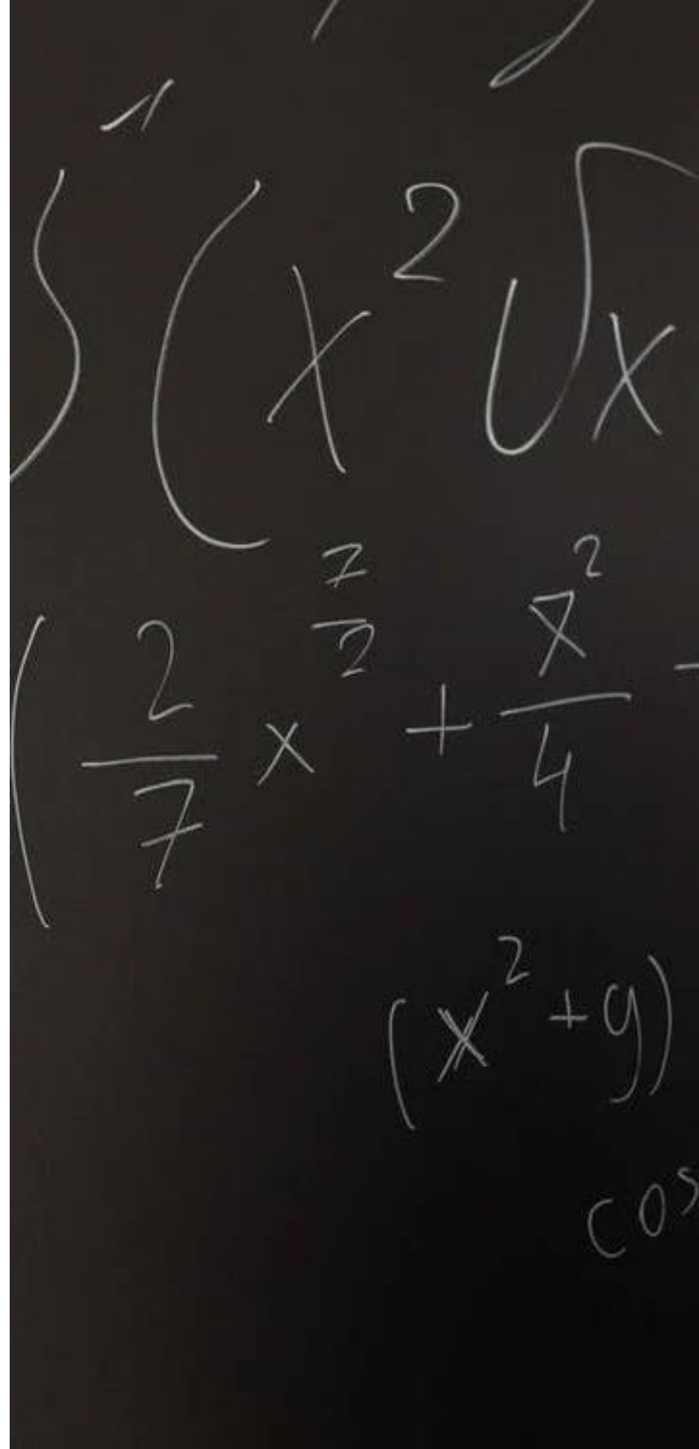
- Análisis de clúster consiste en identificar la existencia de grupos en los datos u observaciones
- Kaufman y Rousseeuw: el análisis de clústeres es el arte de encontrar grupos en los datos.
- Es de carácter exploratorio.
- Posibilidad de decidir cuántos son los grupos relevantes



Desarrollo de conceptos

En qué consiste?

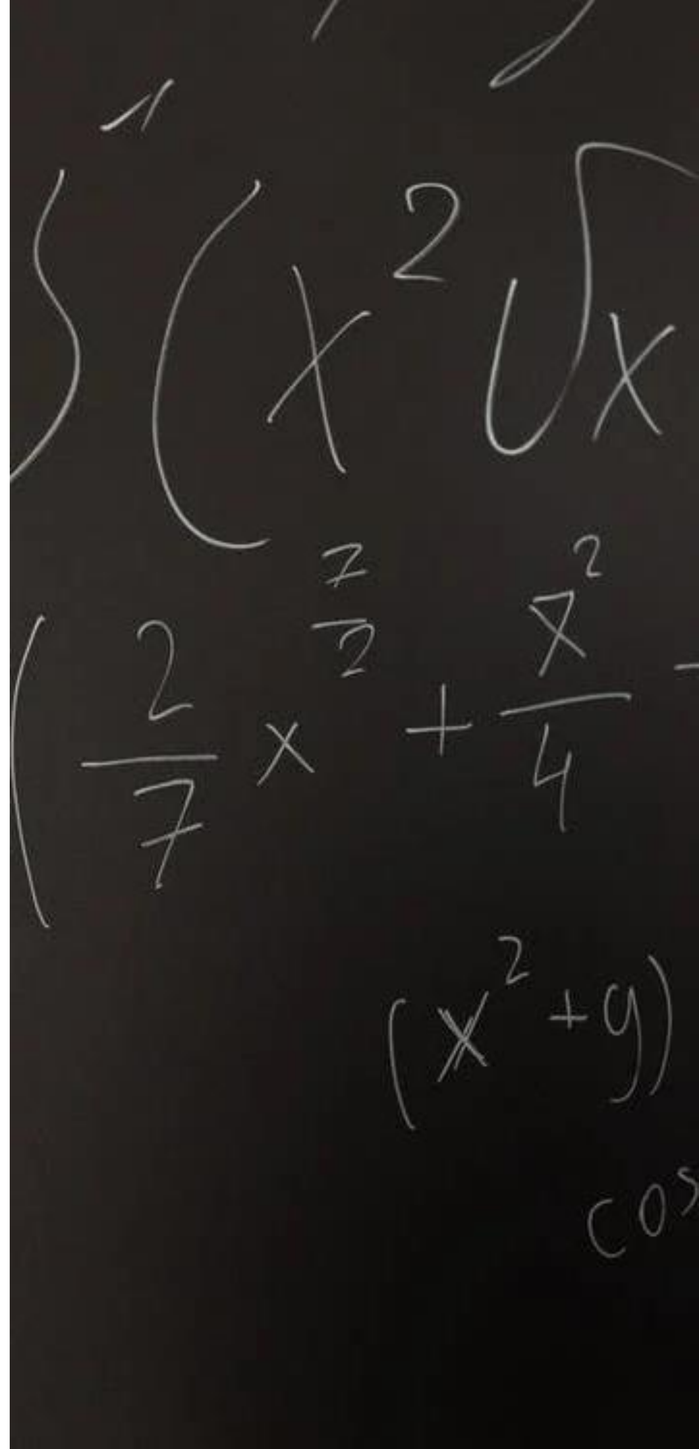
- Tomar un conjunto de casos
- Sacar las matrices de distancias entre ellos
- Determinar cuáles son más similares entre sí y cuales más diferentes.
- Medir las matrices de distancias entre casos o también entre variables.
- Los métodos de conglomerados (clúster) permiten agrupar tanto casos (por ejemplo países) como variables (por ejemplo, indicadores de desarrollo).
- Agrupar lo semejante y diferenciarse de lo diferente.



Desarrollo de conceptos

Utilidad

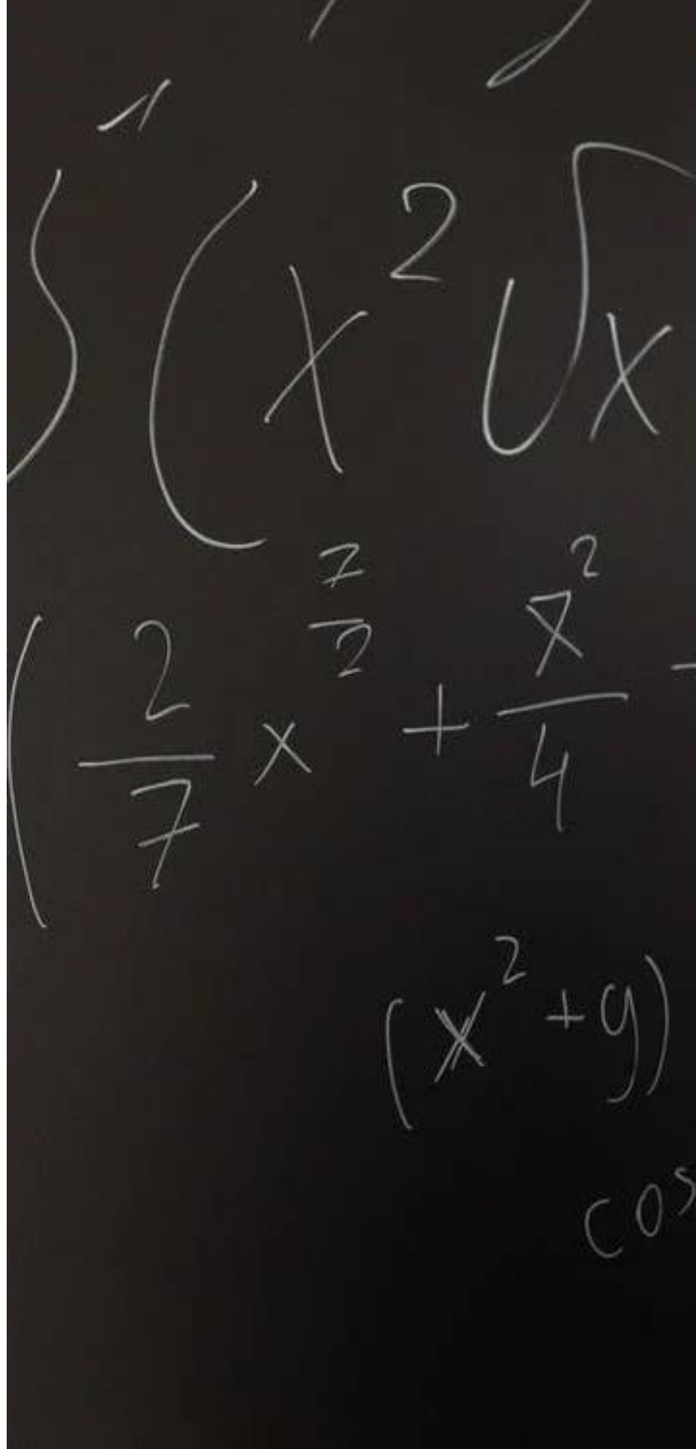
- Es exploratorio de datos.
- Los procedimientos de “minería de datos” se basan en el análisis de clústeres.
- Los métodos de clúster están pensados más para generar hipótesis que para testarlas



Desarrollo de conceptos

Supuestos:

- Puede permitir que los casos se solapen o permitir solamente la pertenencia a un clúster,
- Permitir solamente la pertenencia a un clúster es el más común.
- El número de conglomerados que obtengamos dependerán directamente de las variables que se hayan utilizado para determinar las diferencias o similitudes.



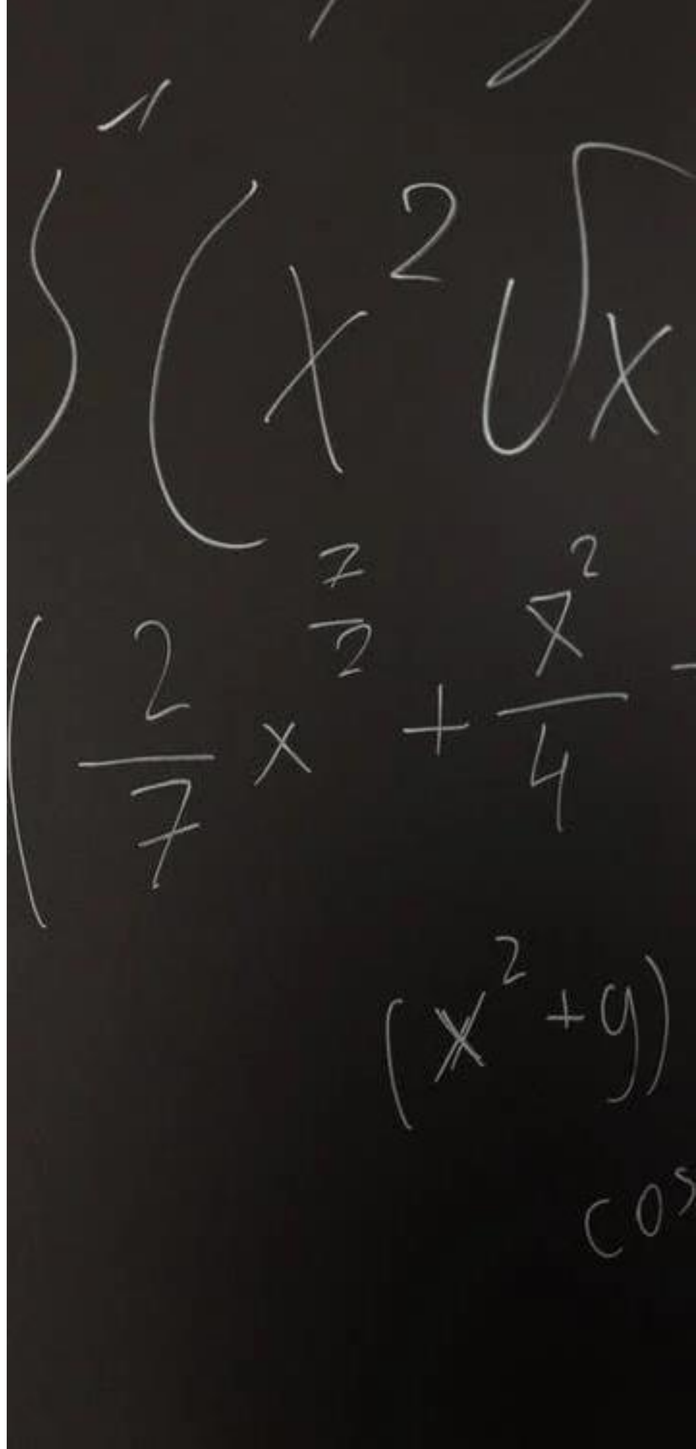
Desarrollo de conceptos

Decisiones técnicas

- Variables a seleccionar
- Medida de distancia o proximidad a utilizar
- Criterios que decidirán la adscripción de un caso a un conglomerado u otro.

Métodos

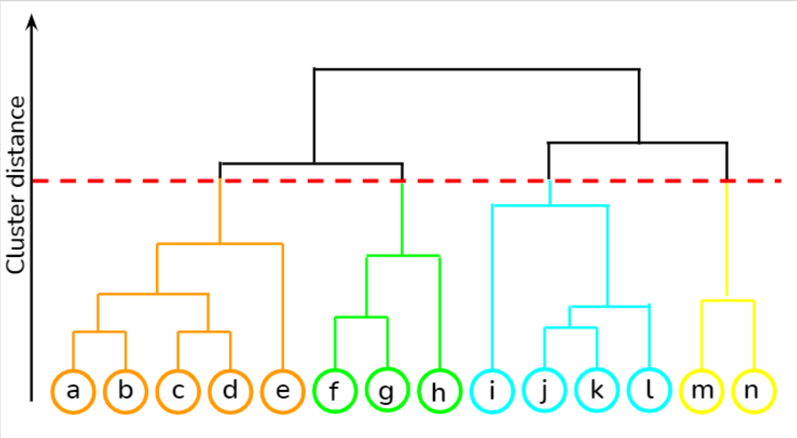
- Técnicas jerárquicas
- Técnicas no jerárquicas

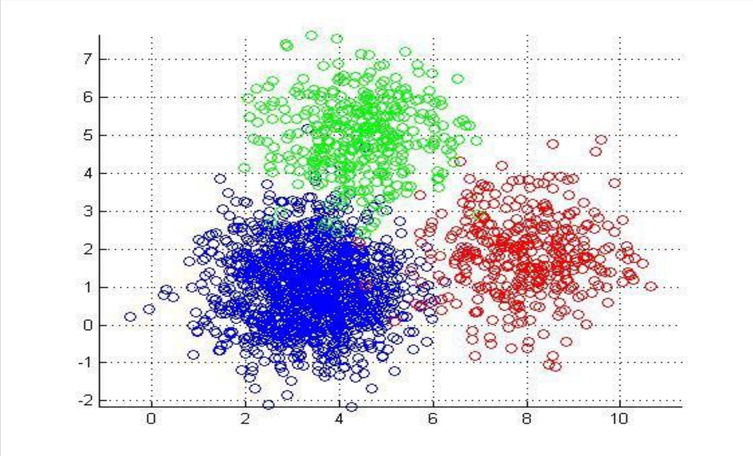


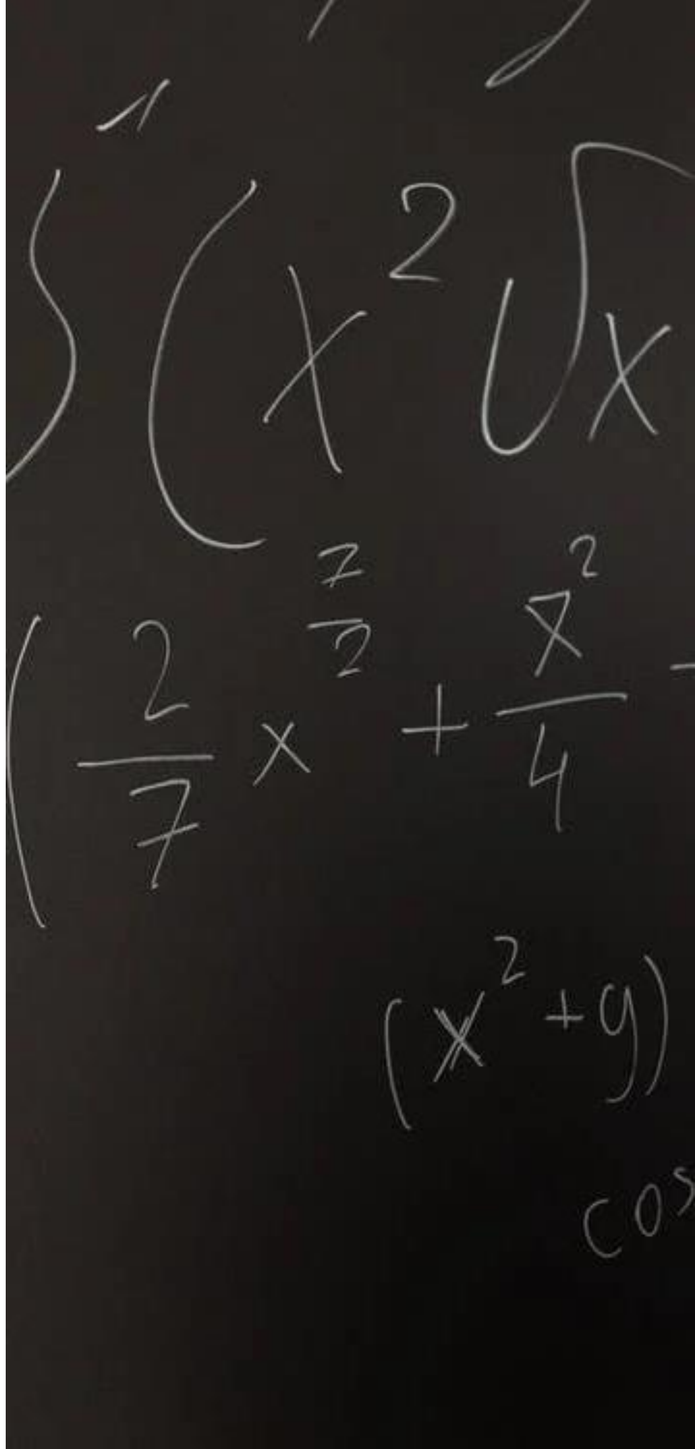
Desarrollo de conceptos

Métodos

- Jerárquicos
- No jerárquicos



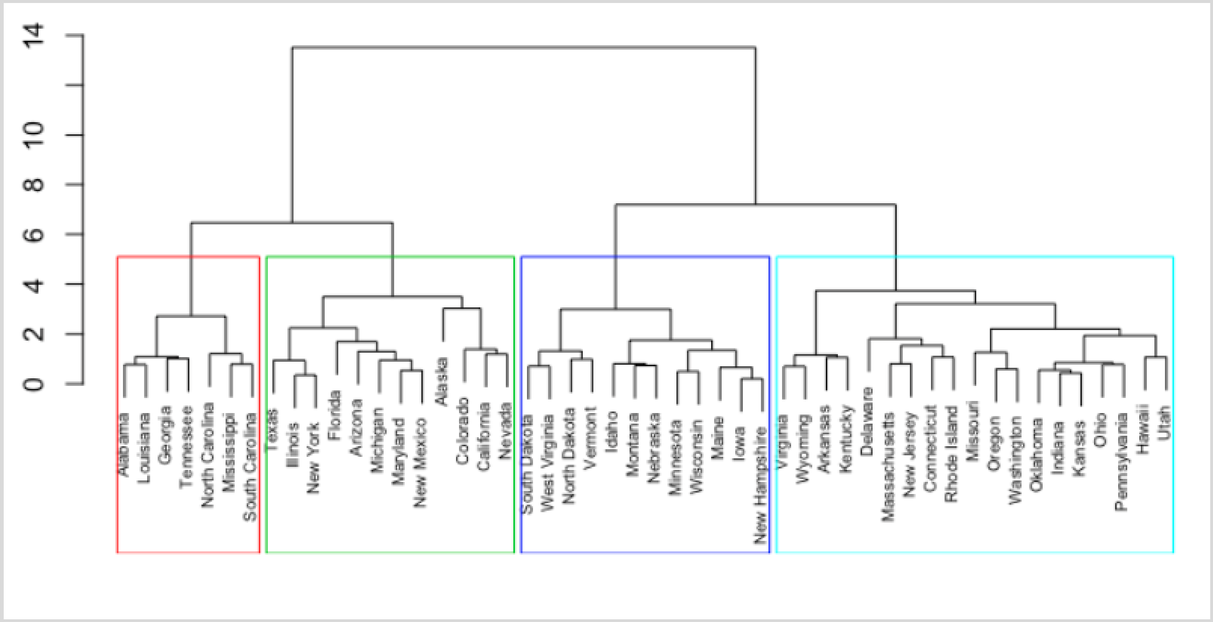


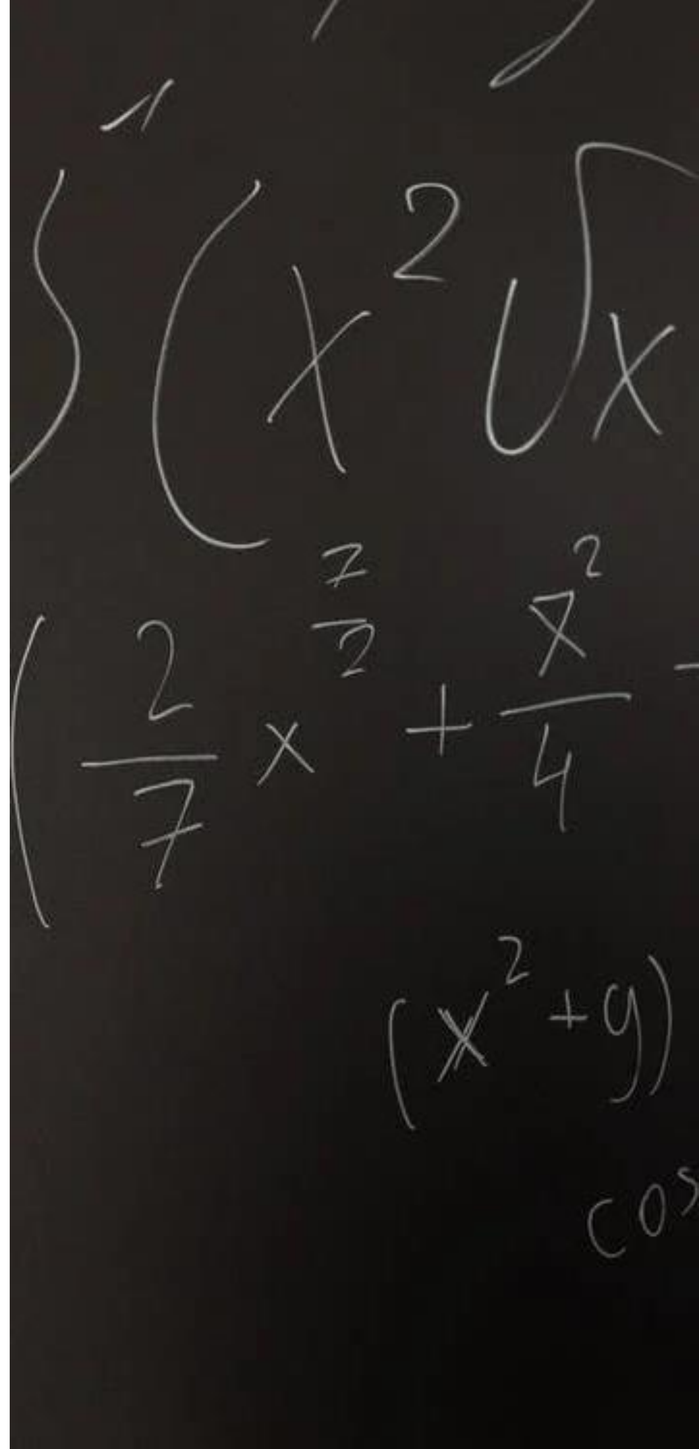


Desarrollo de conceptos

Jerárquicos

- Un clúster contiene otros clústeres, que contienen otros clústeres y así sucesivamente hasta llegar a un solo grupo que agrupa a todos

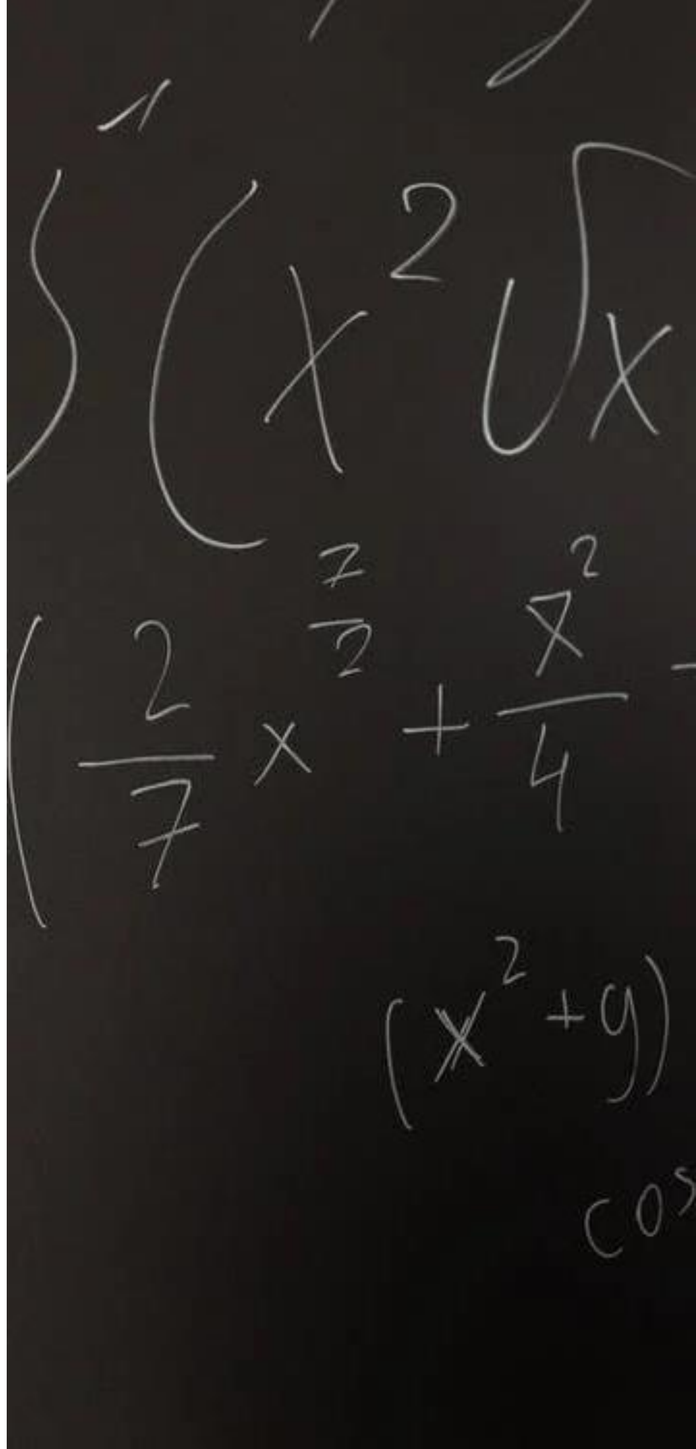




Desarrollo de conceptos

Jerárquicos

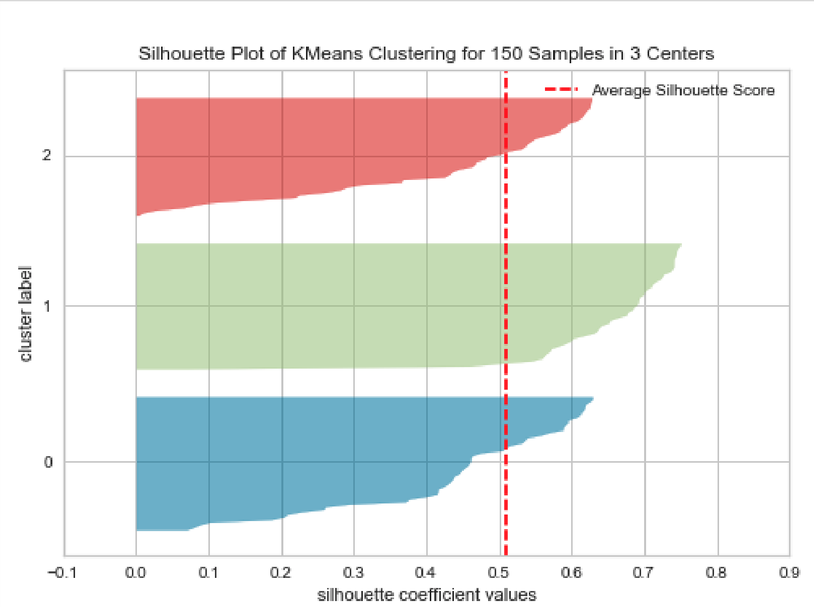
- Son exploratorios
- Y el resultado es elegir cuántos grupos existen
- Se usan procedimientos de vinculación (linkage) entre casos o variables



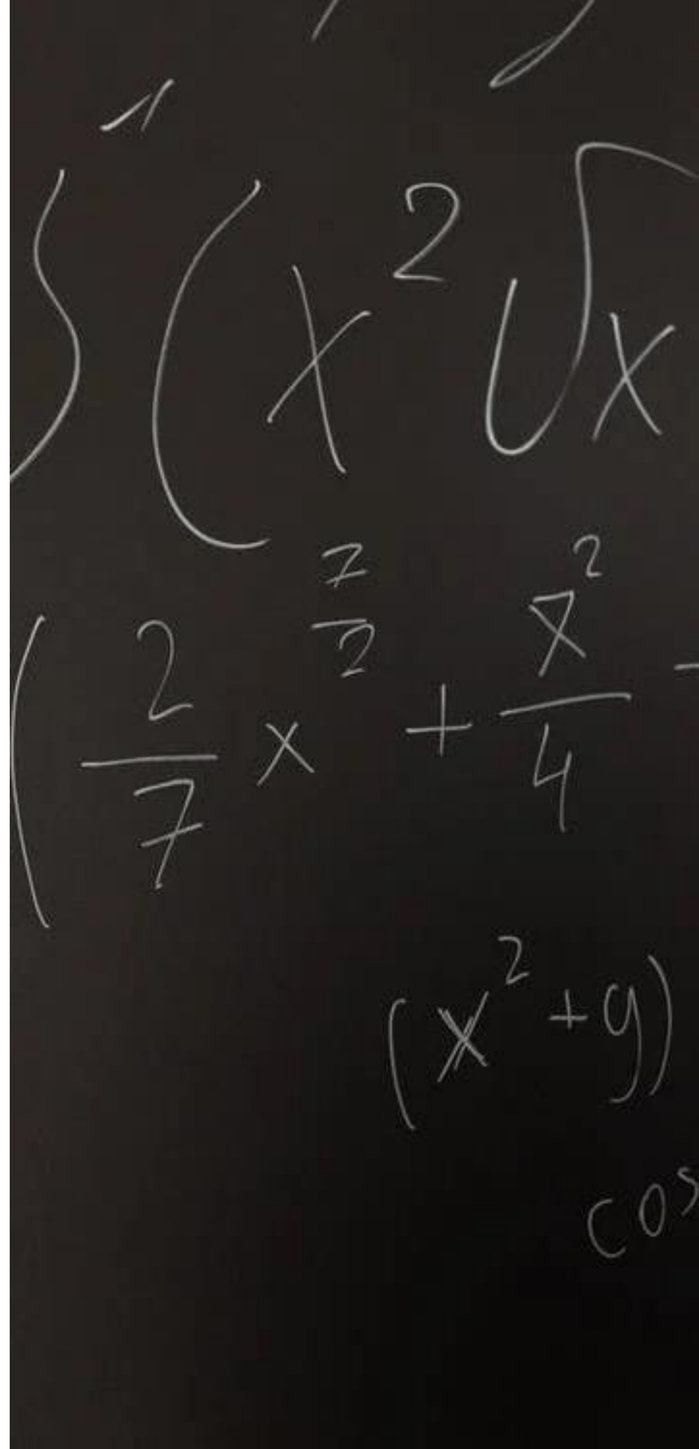
Desarrollo de conceptos

No Jerárquicos

- Los clústeres son el resultado de la partición en grupos de los casos en estudio



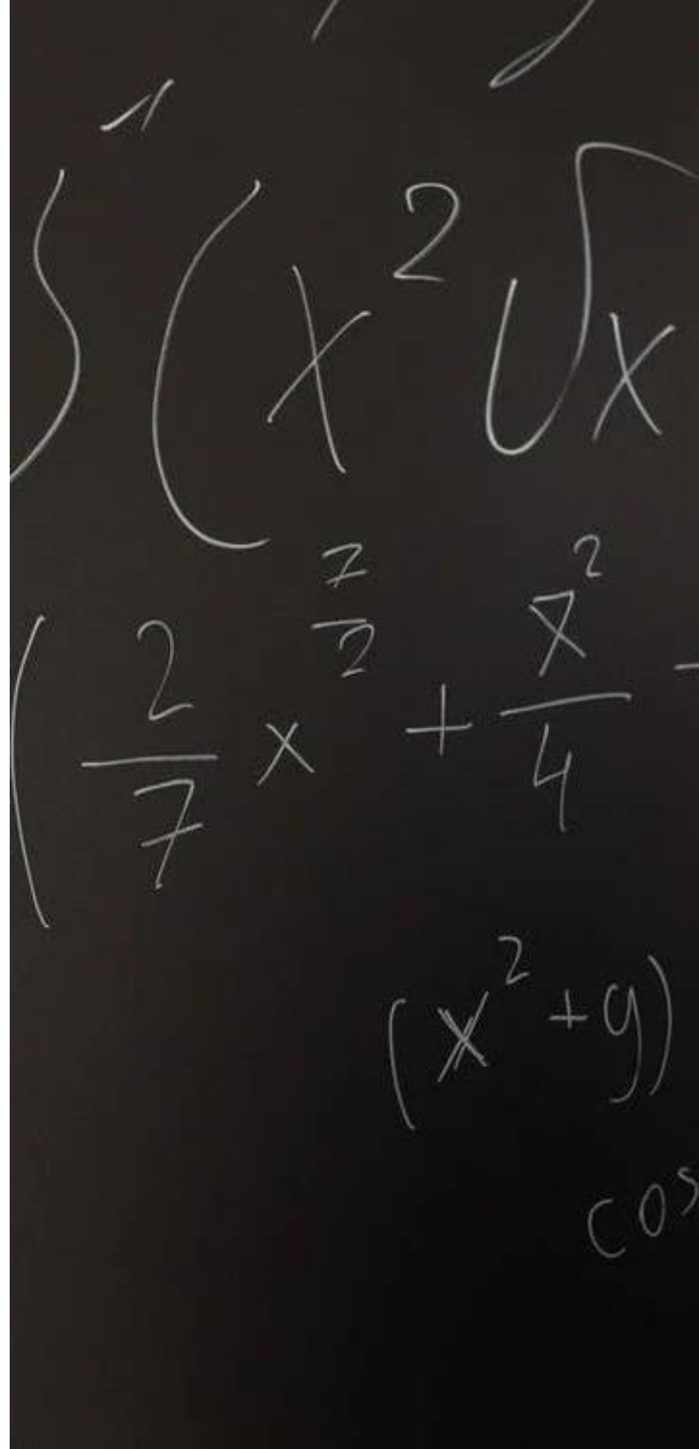
Silhouette diagram — K-Means method (Image by author)



Desarrollo de conceptos

No Jerárquicos

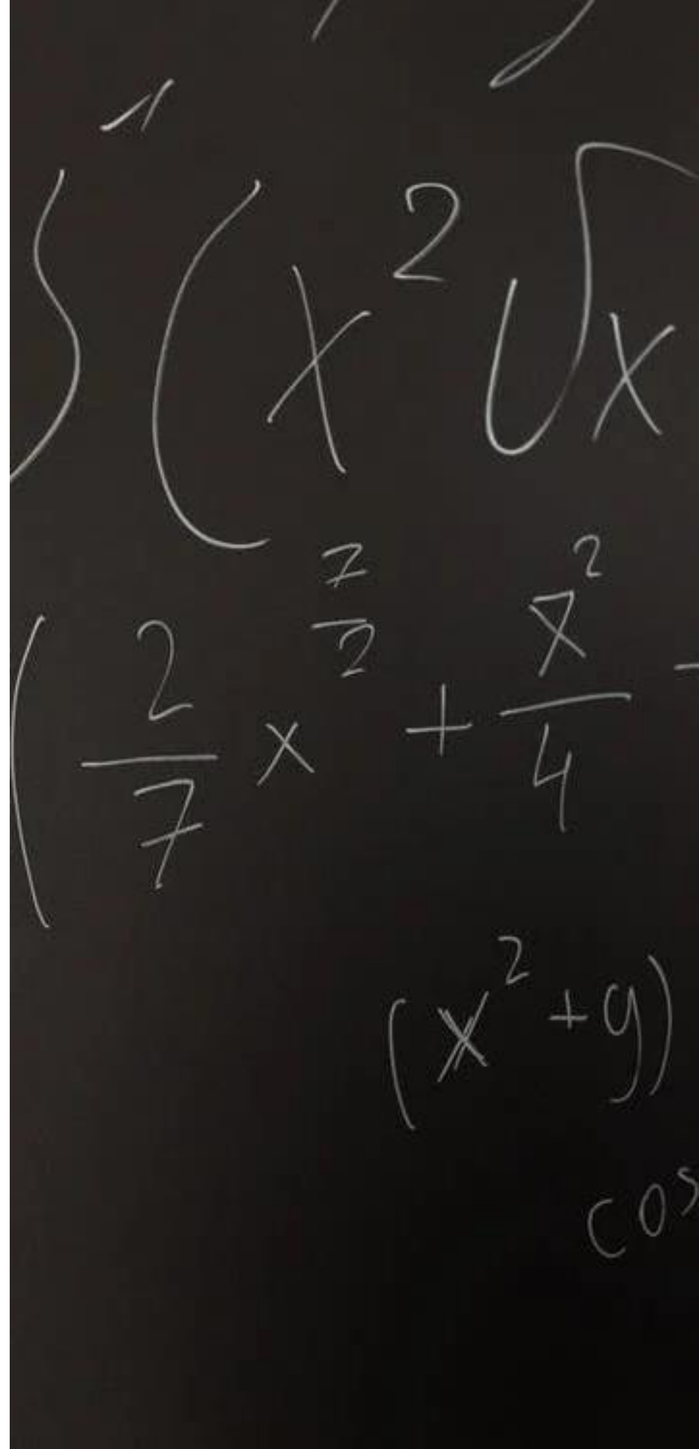
- Se indican cuántos grupos al iniciar el análisis
- Solo se establecen grupos de casos no de variables
- Requieren mayor poder de cómputo
- K-medias o k-medianas
- Se optimiza la diferencia entre los grupos y se obtiene mayor homegeneidad interna.



Desarrollo de conceptos

Criterios para un análisis de conglomerados

- Métodos de análisis (Jerárquico, No jerárquico)
- Procedimientos
  - Jerárquicos: Single linkage, Complete linkage, etc.
  - No Jerárquicos: K-medias, K-medianas
- Medida: de similitud o disimilitud
  - Euclidiana, Manhattan, Minkowski, Jaccard, Coseno.
- Transformaciones de variables (normalización)
- Reglas de finalización (número de clústers)



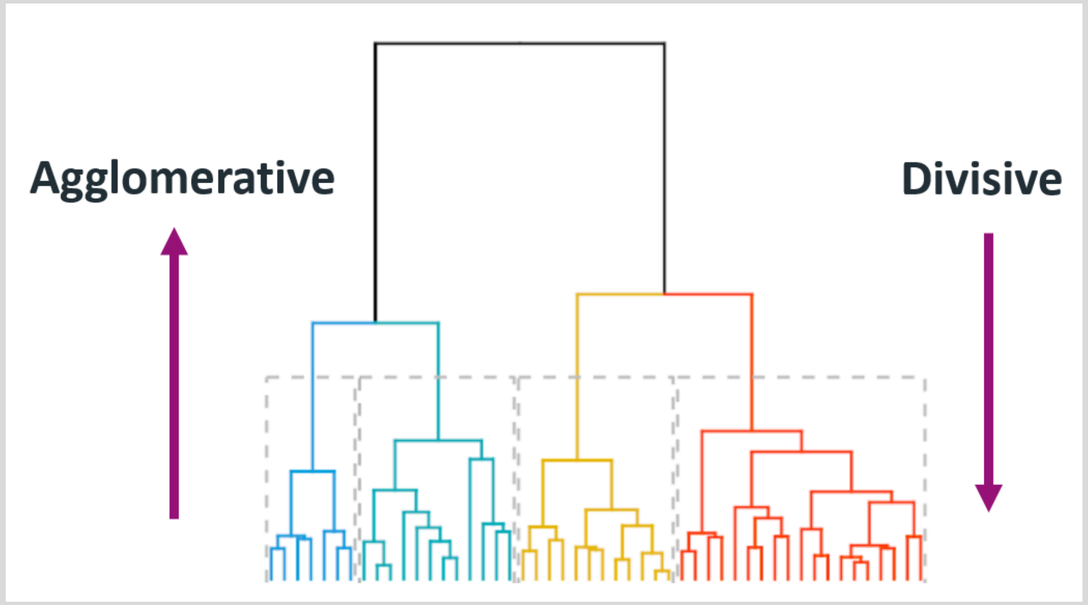
Desarrollo de conceptos

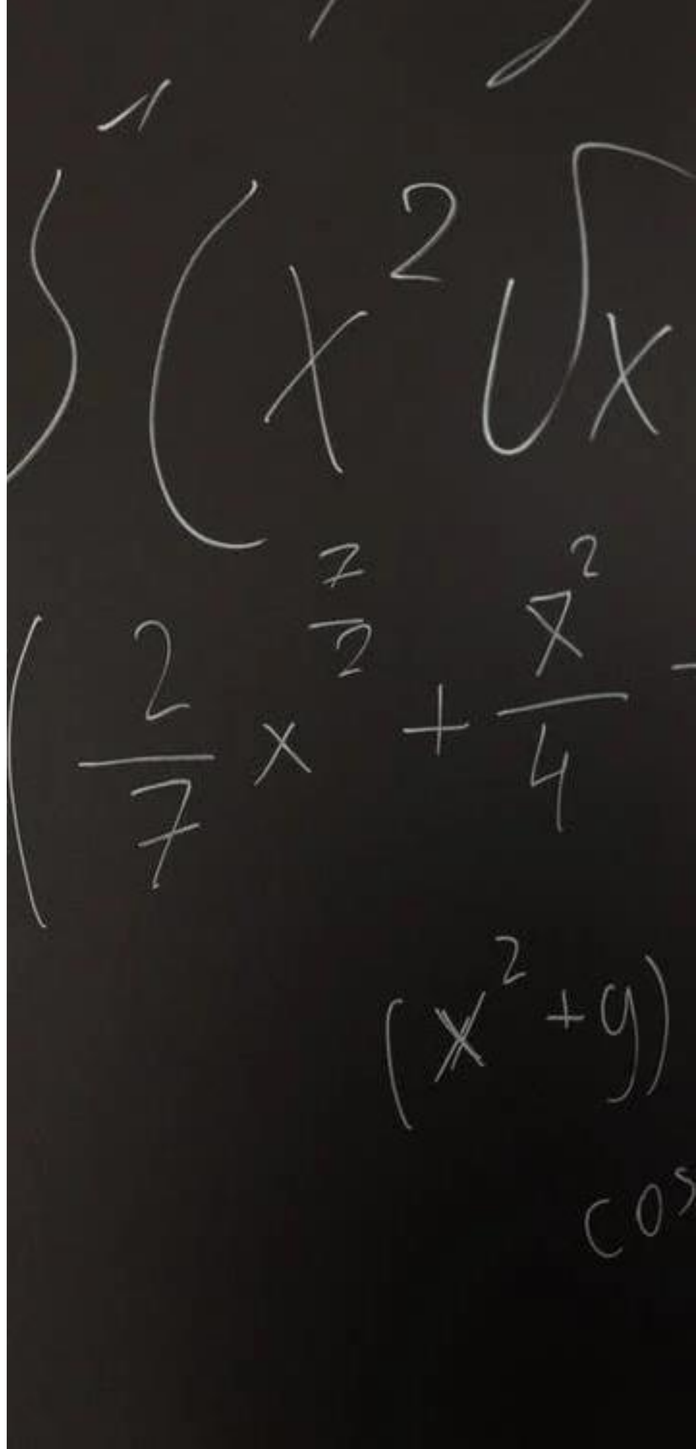
Métodos Jerárquicos

- Cuando el número de casos no es excesivo

Procedimientos

- Aglomeración
- División

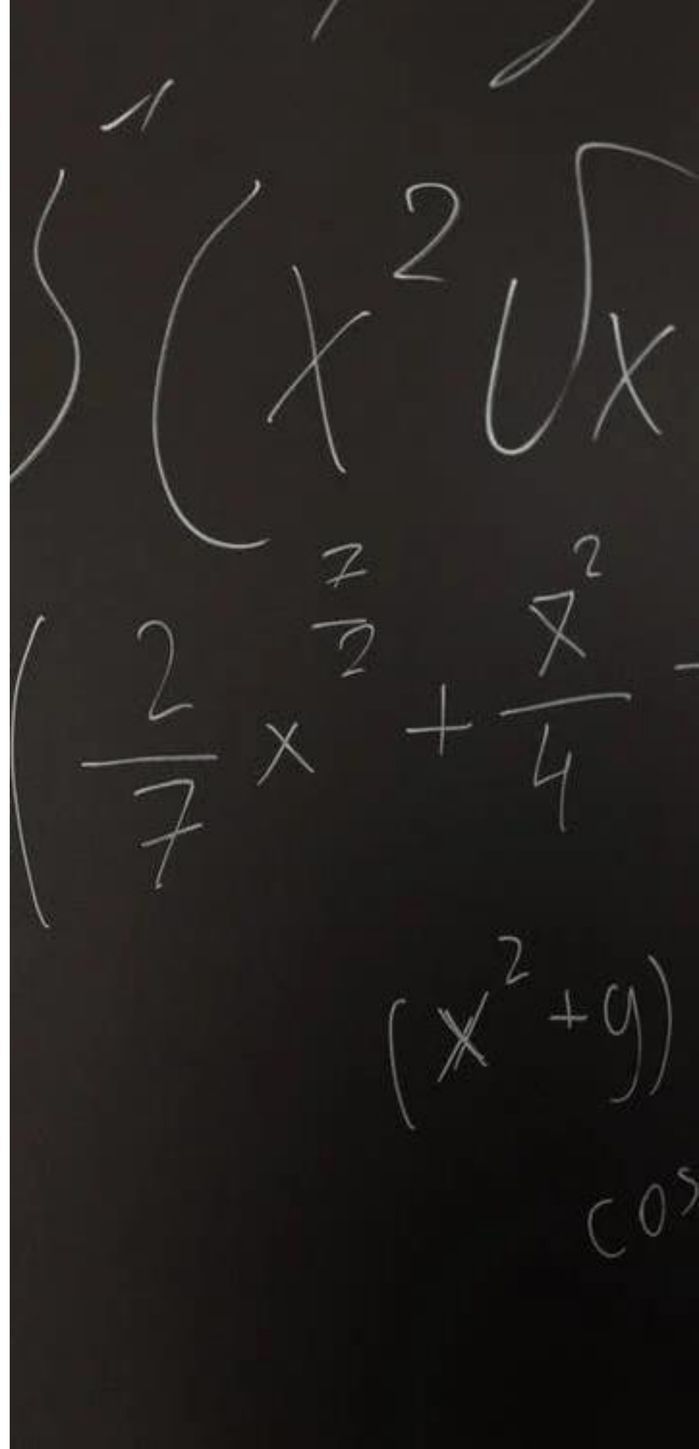




Desarrollo de conceptos

Métodos Jerárquicos - Aglomeración

- Se considera cada caso como un grupo separado
- N grupos con un tamaño de 1
- Los dos grupos (casos) más próximos se unen en un único clúster
- En ese momento existirán N-1 grupos
- Un grupo de tamaño 2 y el resto de tamaño 1
- Este procedimiento continúa hasta que todos los casos pertenecen a un único grupo.



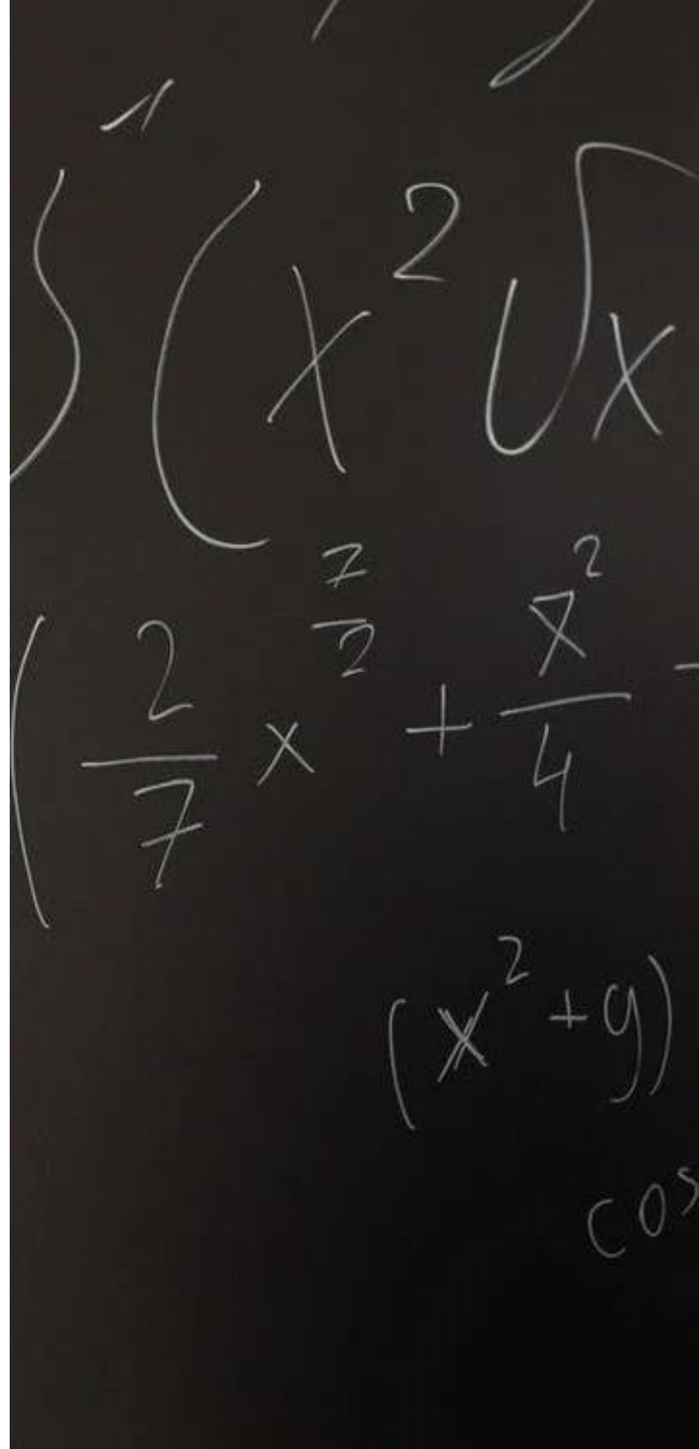
Desarrollo de conceptos

Métodos Jerárquicos - Aglomeración

a) nuevos casos se incorporan a grupos ya existentes

b) definen ellos mismos un nuevo grupo

c) se unen en un solo grupo otros grupos ya preexistentes

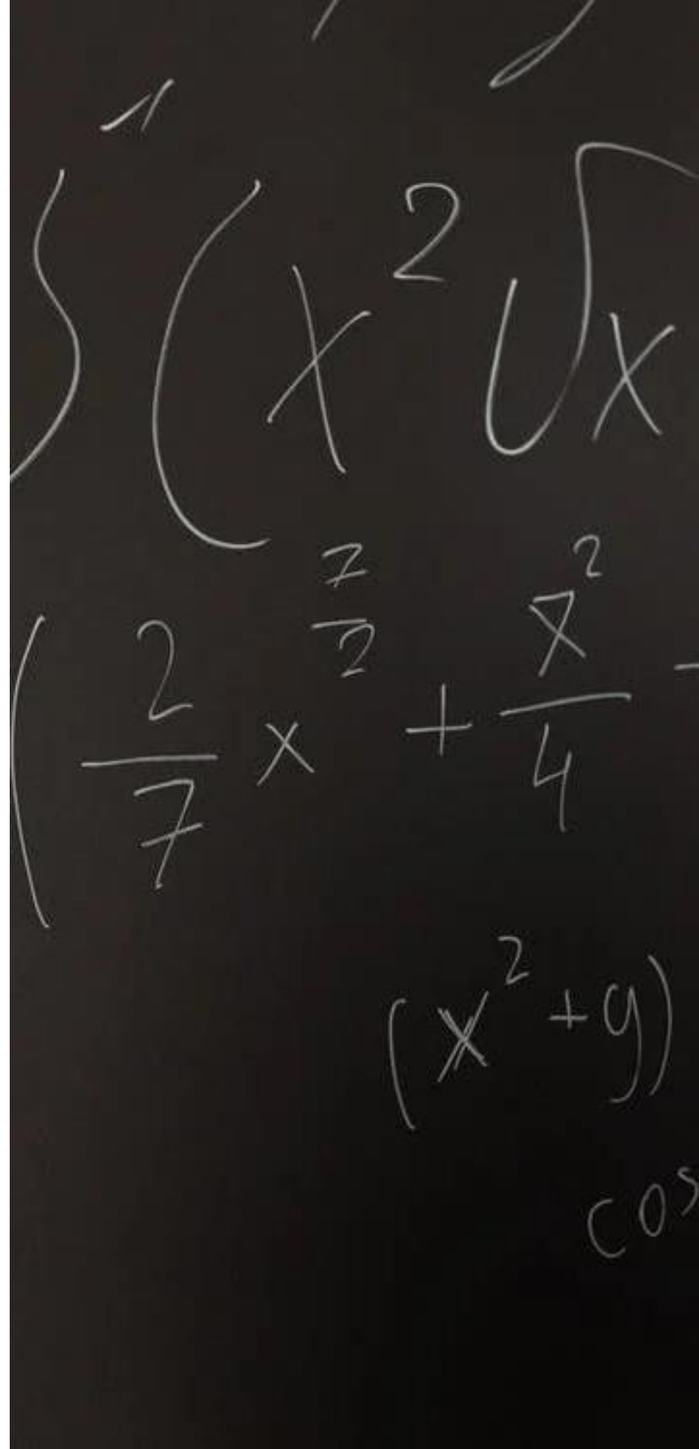


Desarrollo de conceptos

Métodos Jerárquicos - Aglomeración

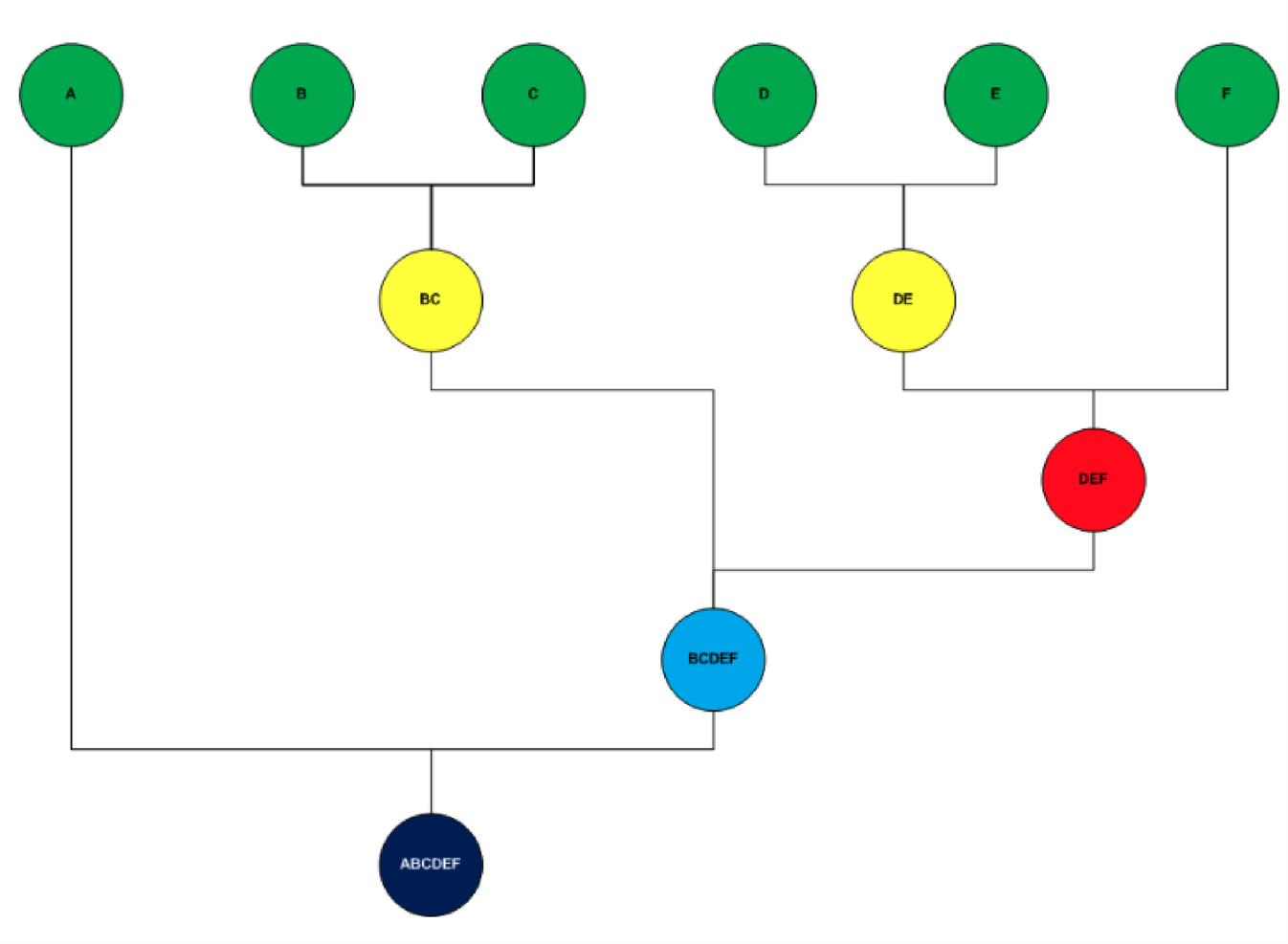
- Procedimiento

1. Asignar cada elemento a un cluster
2. Encontrar la matriz de distancias
3. Encontrar 2 clusters que tengan la distancia más corta y mezclarlos
4. Continua este proceso hasta que se forma un solo cluster grande



Desarrollo de conceptos

Métodos Jerárquicos - Aglomeración



Desarrollo de conceptos

Métodos Jerárquicos - División

a) Todos los casos forman un único grupo

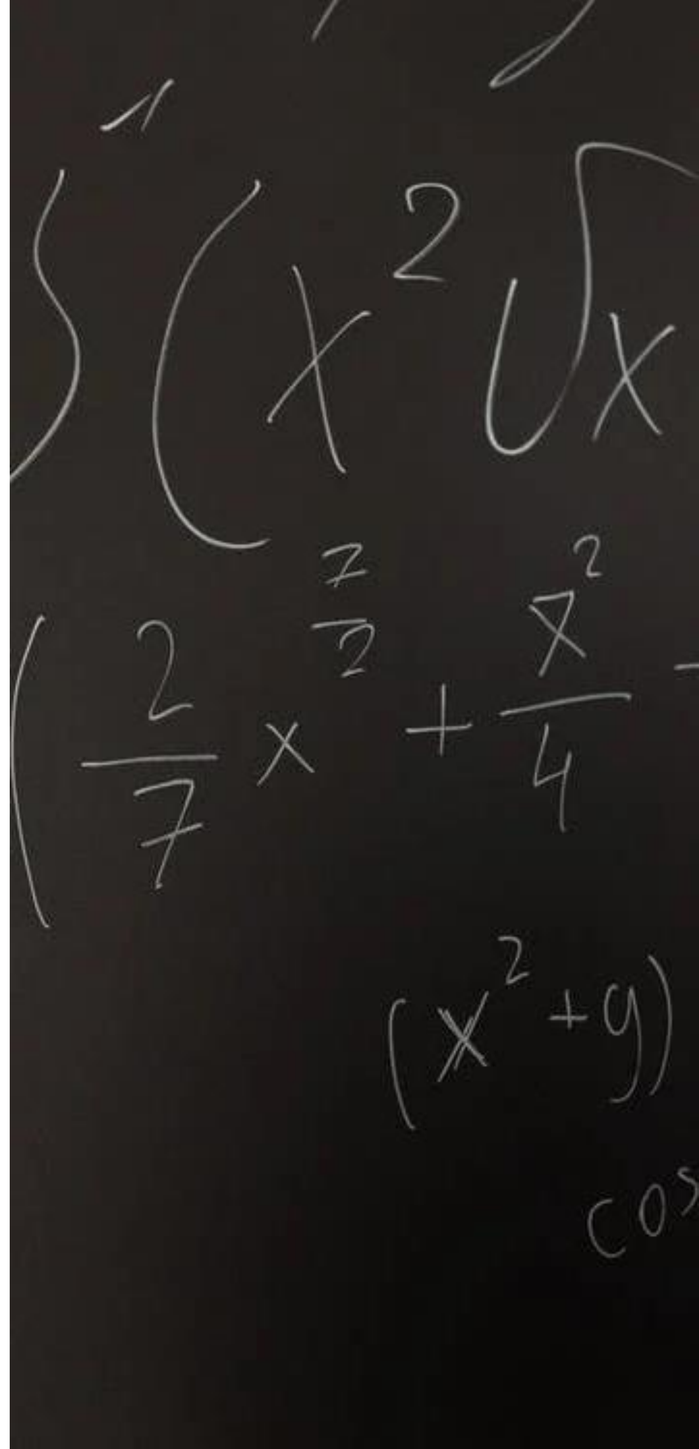
b) Este grupo se divide para crear dos grupos

c) Uno de esos dos grupos se divide en otros dos, se generan tres

d) Uno de los tres grupos se subdivide para formar otros dos, produciendo un total de cuatro grupos

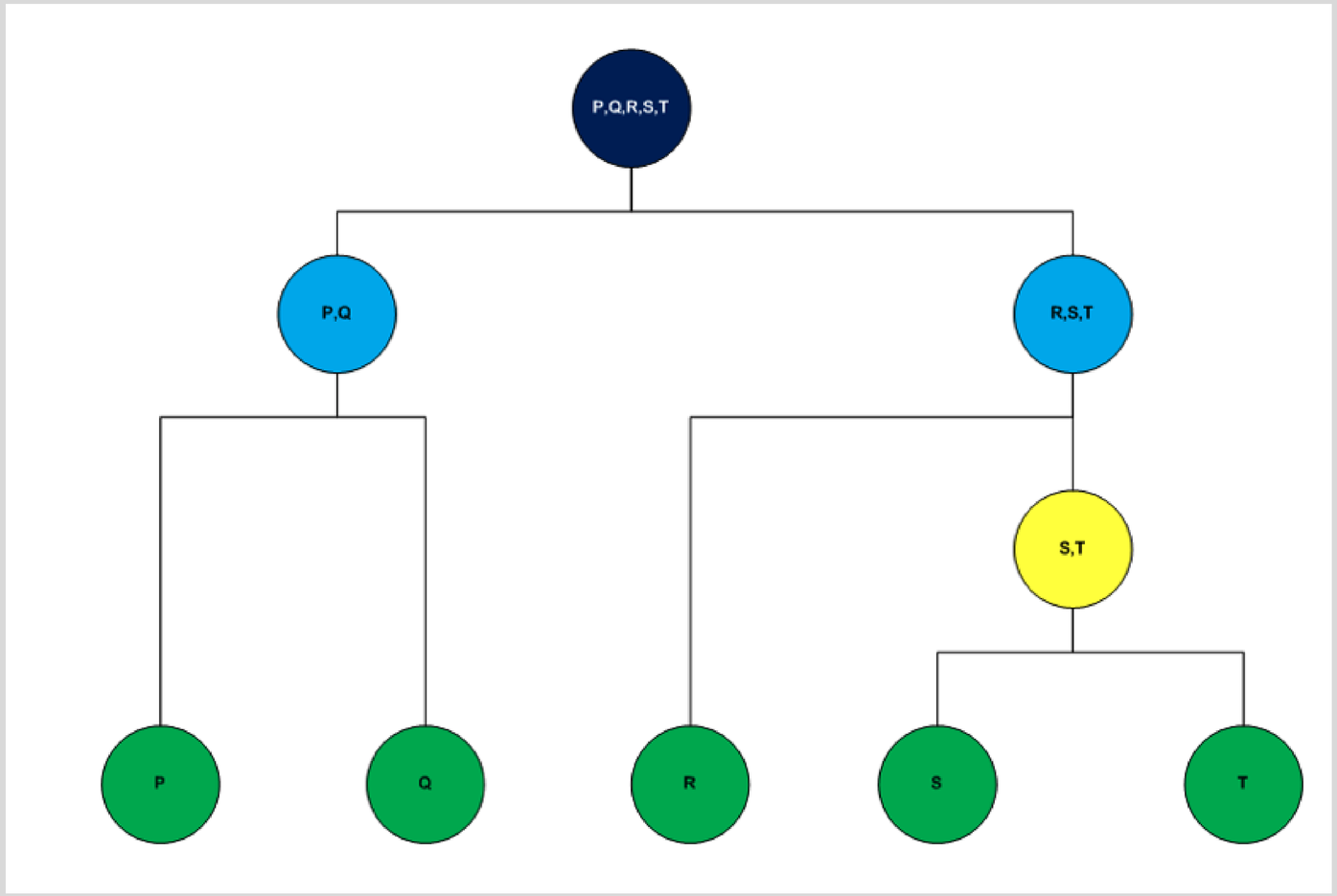
e) Se continúa hasta que finalmente hay tantos grupos como casos

*son bastante infrecuentes



Desarrollo de conceptos

Métodos Jerárquicos - División

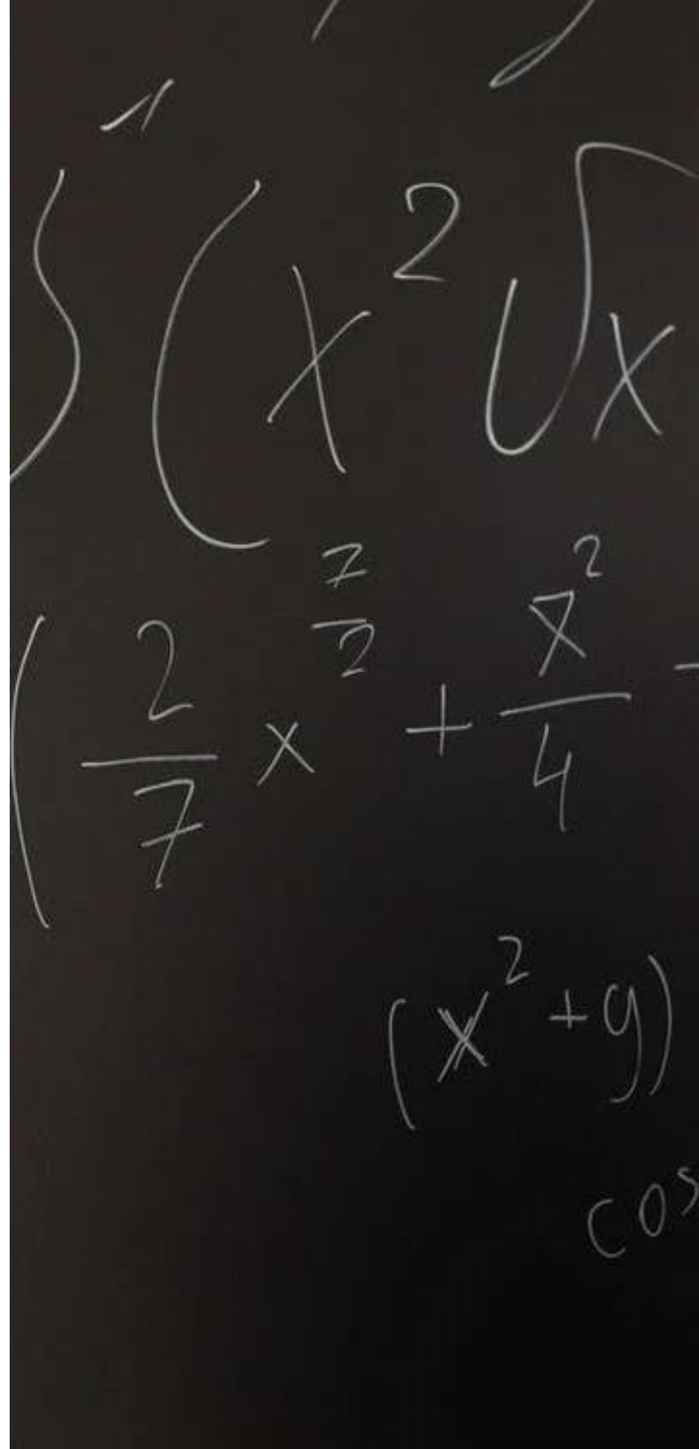


Desarrollo de conceptos

Métodos Jerárquicos - Aglomeración

- Métodos (Distancia o disimilaridad)

Método de vinculación (linkage) $d_{(R,P+Q)} =$	α_1	α_2	β	γ
Single	1/2	1/2	0	–1/2
Complete	1/2	1/2	0	1/2
Average	$n_P / (n_P + n_Q)$	$n_Q / (n_P + n_Q)$	0	0
Weighted	1/2	1/2	0	0
Centroid	$n_P / (n_P + n_Q)$	$n_Q / (n_P + n_Q)$	$-(n_P n_Q / (n_P + n_Q)^2$	0
Median	1/2	1/2	–1/4	0
Ward	$(n_R + n_P) / (n_R + n_P + n_Q)$	$(n_R + n_Q) / (n_R + n_P + n_Q)$	$- n_R / (n_R + n_P + n_Q)$	0
Flexibeta	$(1 - \beta) / 2$	$(1 - \beta) / 2$	β	0



Solución

- Conglomerado Jerárquico
- Dentro de Colab
  - https://colab.research.google.com

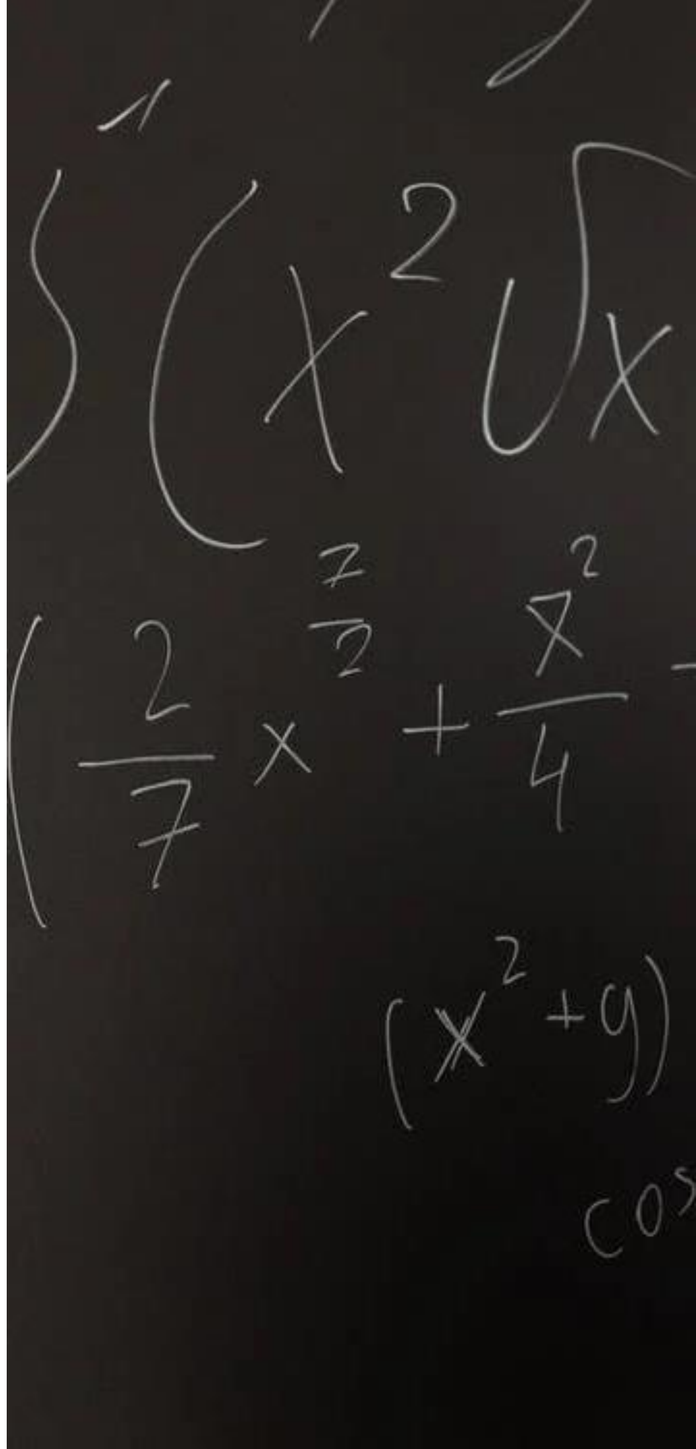
Desarrollo de conceptos

Métodos No Jerárquicos

- Cuando consideramos miles de casos

Procedimientos

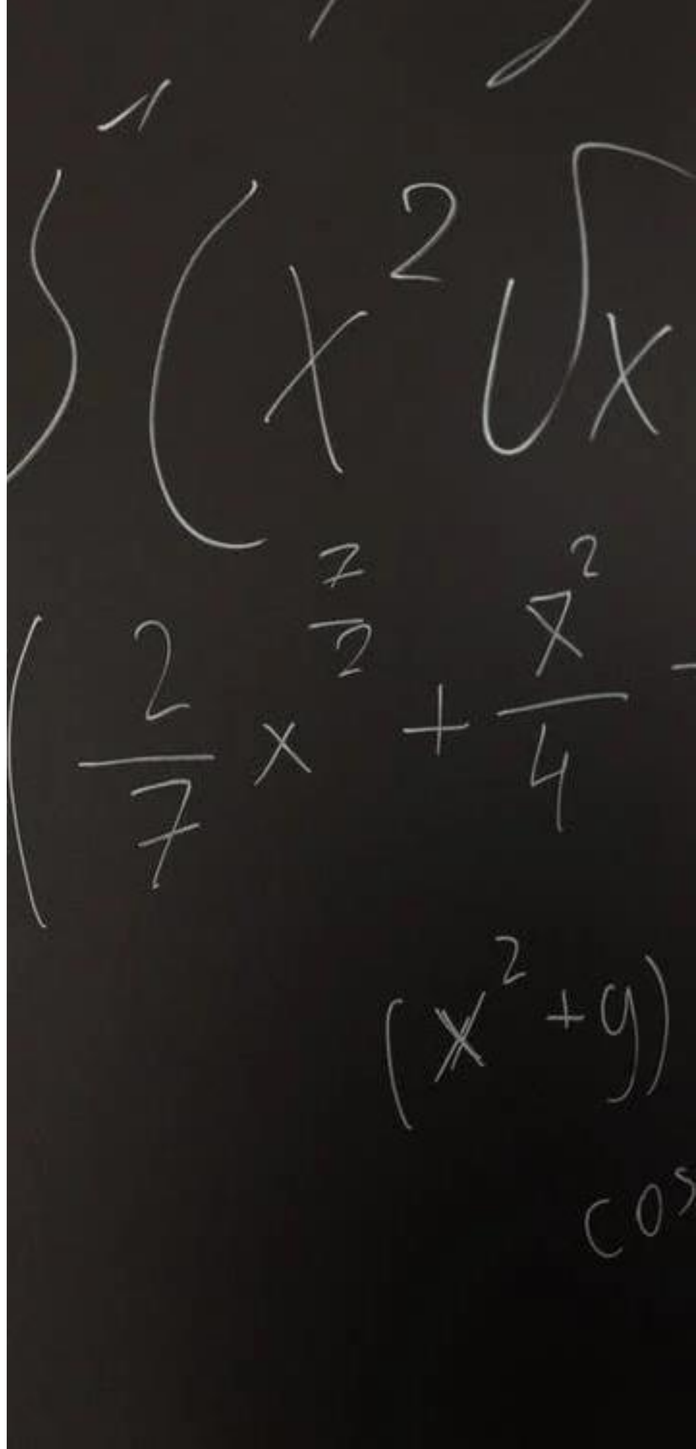
- K-medias (se debe indicar el número de clústers)
- K-medianas - two steps cluster analysis



Desarrollo de conceptos

Métodos No Jerárquicos – K-medias (K-Means)

- los grupos que se construyen son excluyentes entre sí desde el inicio
- Intenta maximizar las diferencias entre grupos
- Busca máximizar homogeneidad interna dentro de los grupos.

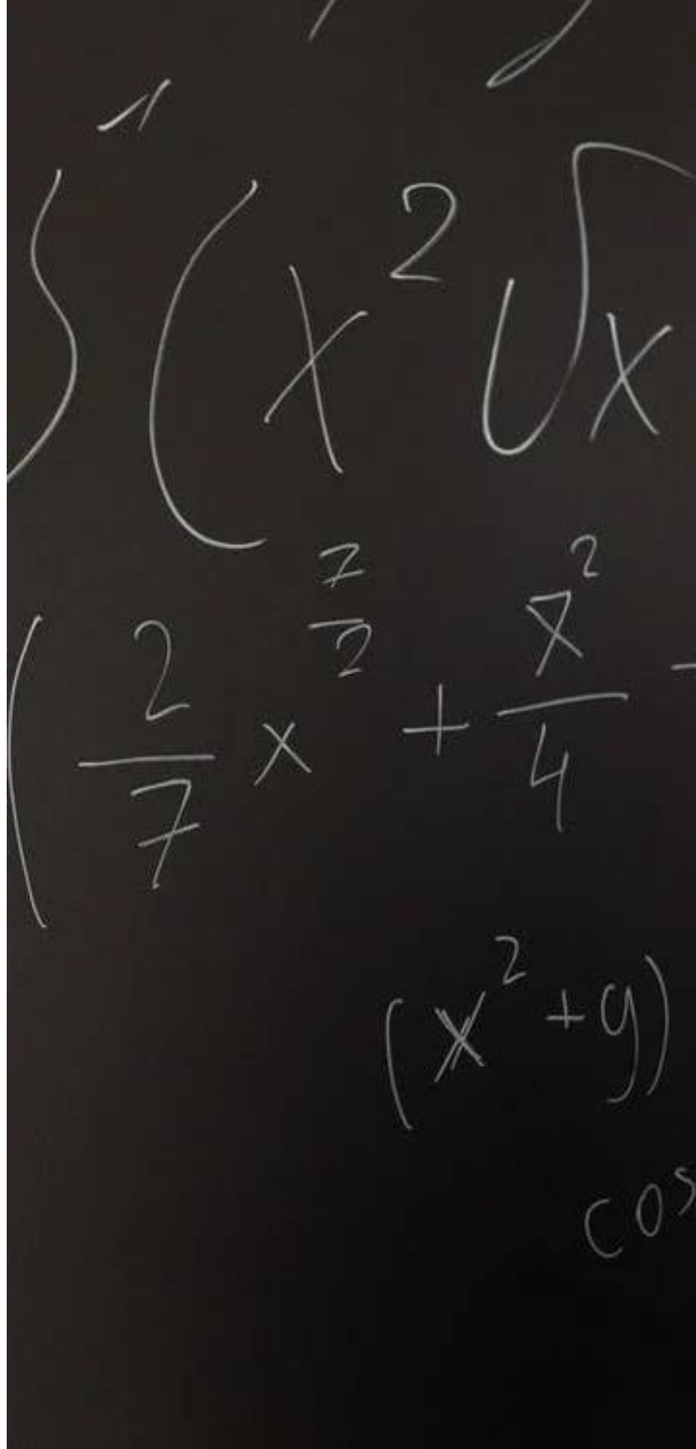


Desarrollo de conceptos

Métodos No Jerárquicos – K-medias (K-Means)

Método

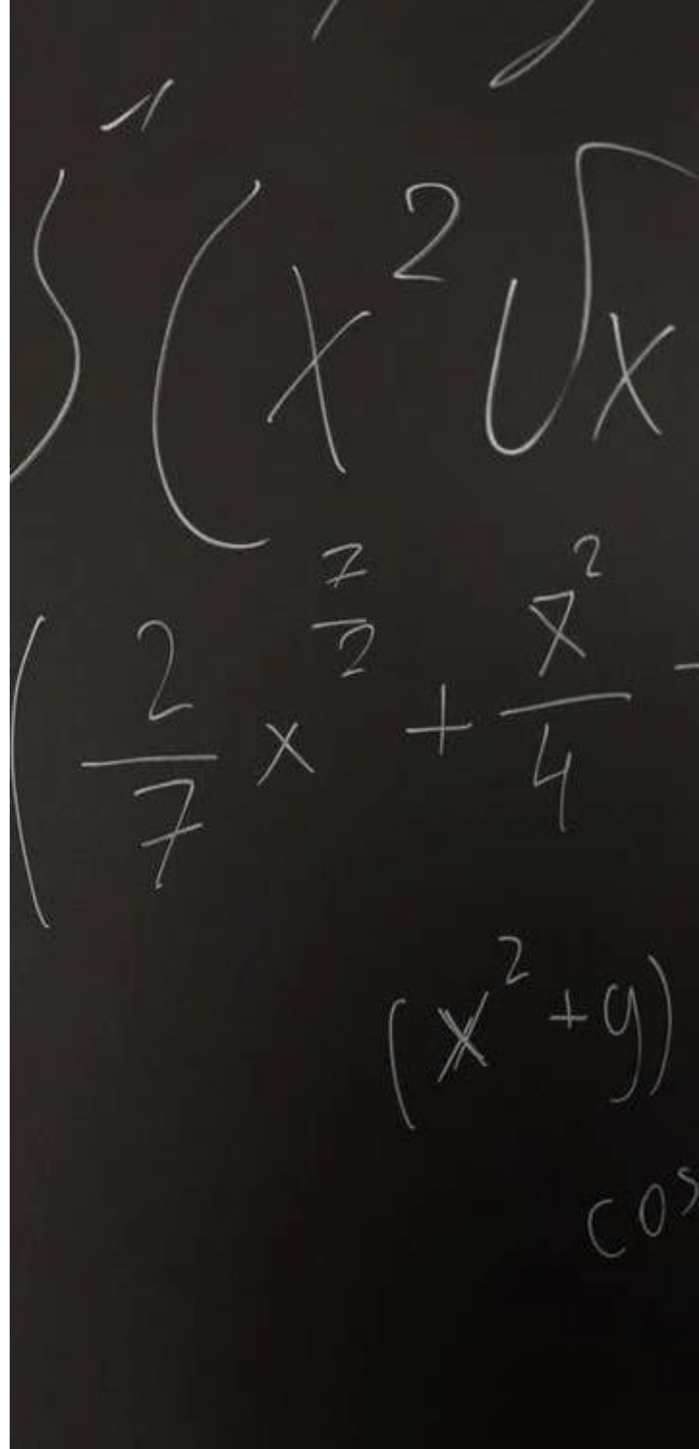
- Se indica el número de clústers
- Se selecciona un caso para cada cluster, que estén lo más separados del centro de todos los casos
- Cada caso es asignado al grupo de cuyo centro se encuentra más próximo
- Asignar cada caso a un cluster, reduciendo la suma de cuadrados intragrupos *
- Se continúa hasta que la suma de cuadrados intragrupos no se puede reducir más



Desarrollo de conceptos

Métodos No Jerárquicos – K-medias (K-Means)

- El procedimiento por el cual se van construyendo los clústeres varía dependiendo de si se conoce el valor del centro de los grupos, o si por el contrario los centros deben de ser estimados de forma iterativa, eso sí, siempre partiendo de un número prefijado de clústeres
- Se recomienda tomar una muestra y aplicar una análisis exploratorio *Método jerárquico

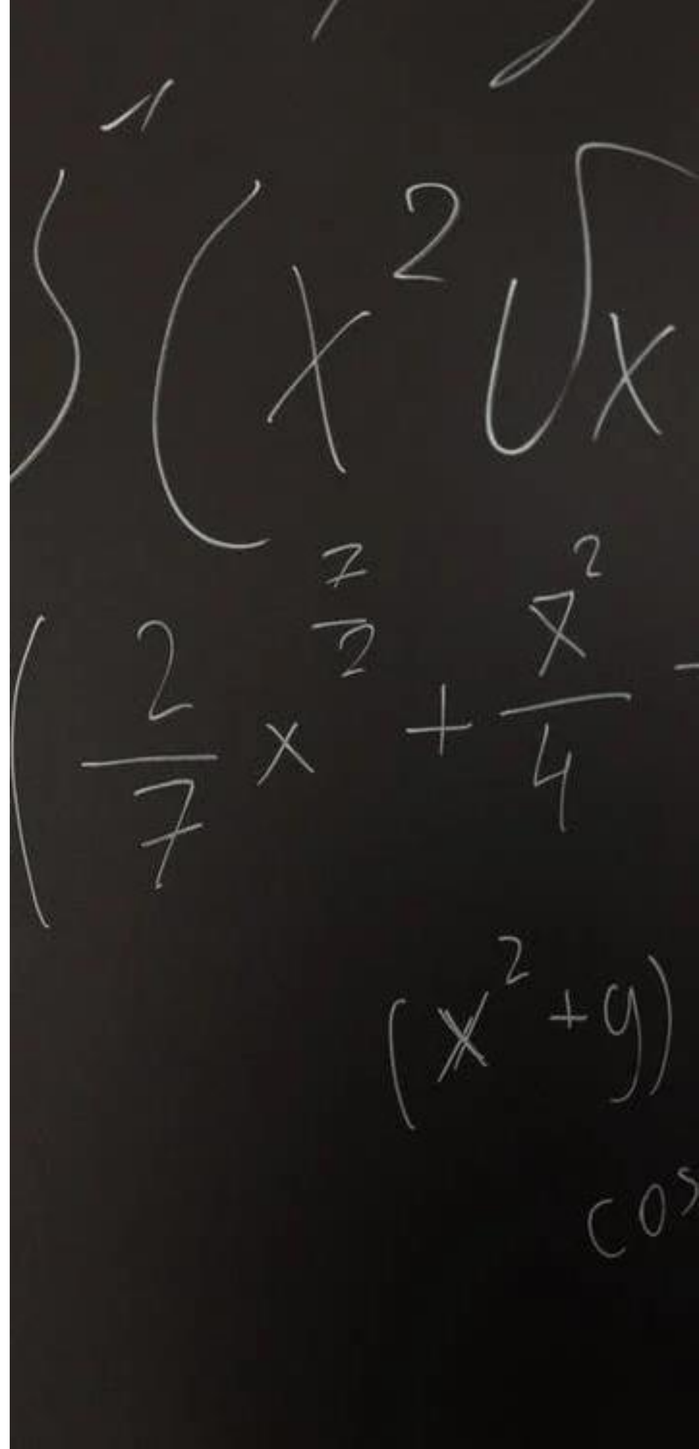


Desarrollo de conceptos

Métodos No Jerárquicos – K-medias (K-Means)

Estrategias

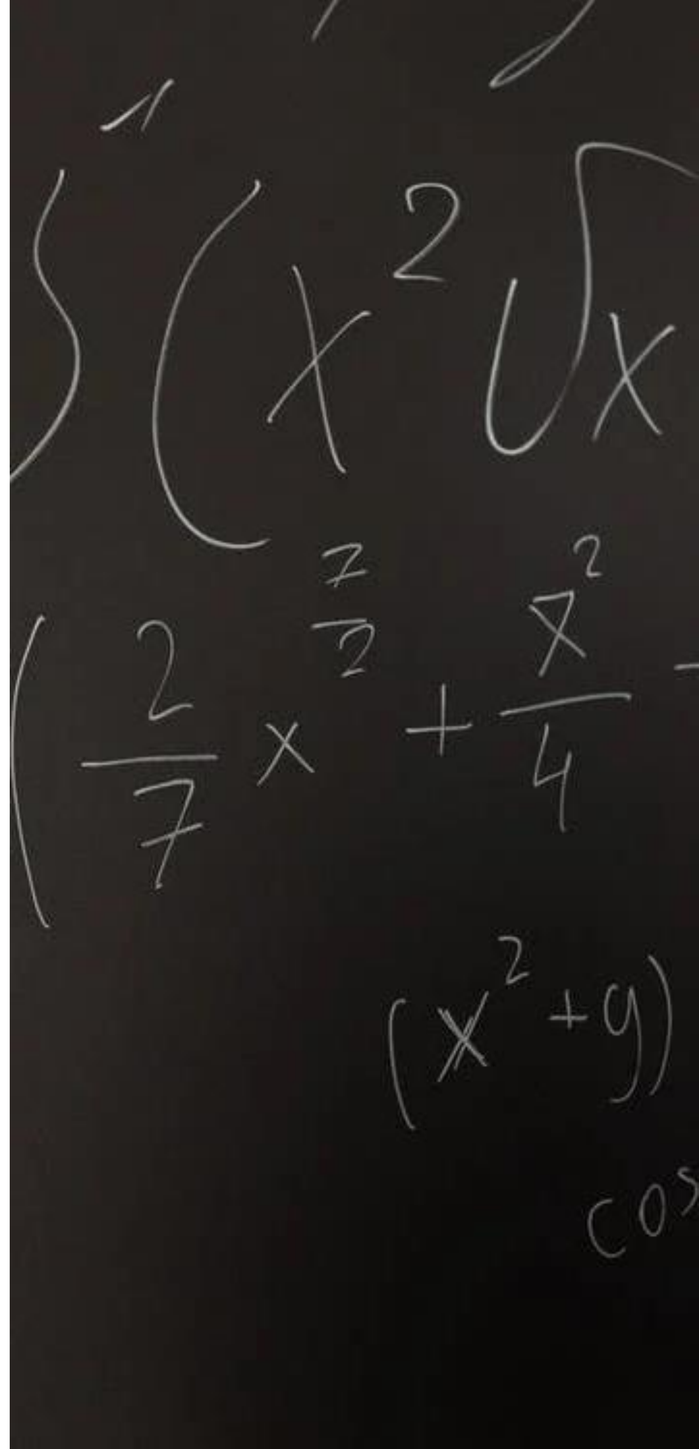
- Partir de los casos con una mayor distancia entre ellos y tomarlos como una estimación de los centros de los futuros clústeres.
- Tomar los k primeros casos como centros iniciales para los grupos
- Tomar los últimos k casos
- Decidir de forma aleatoria los centros



Desarrollo de conceptos

Algoritmo

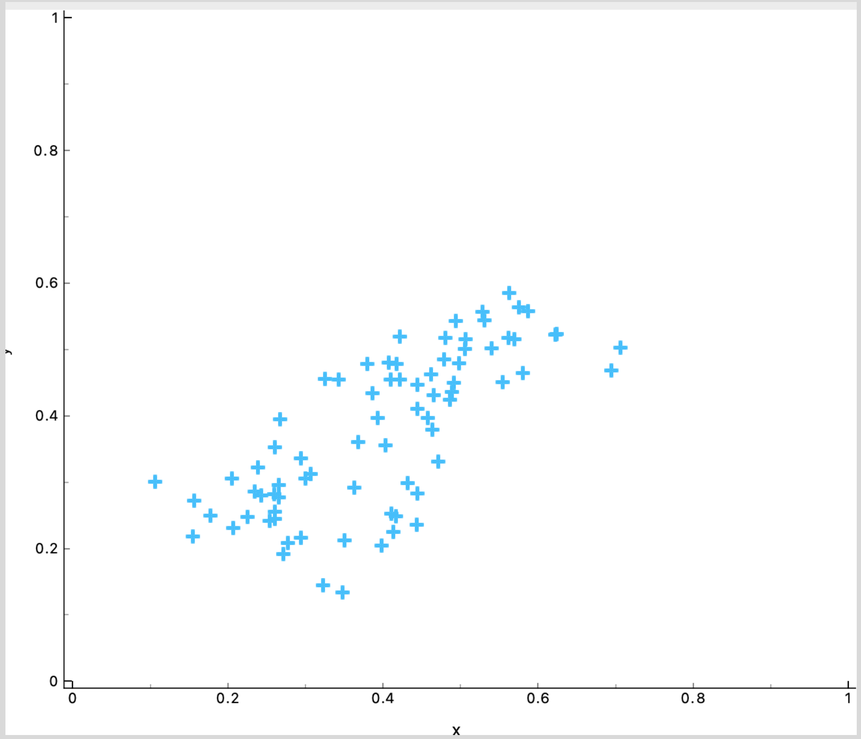
1. Definir el número de clusters **n**
2. Generar aleatoriamente **n** (centroides)
3. Asignar cada elemento del conjunto de datos al centroide más cercano para formar **n** grupos
4. Reasignar la posición de cada centroide
5. Reasignar los elementos de datos al centroide más cercano
    1. Si hubo elementos que se asignaron a un centroide distinto al original, regresar al paso 4, de lo contrario, el proceso ha terminado

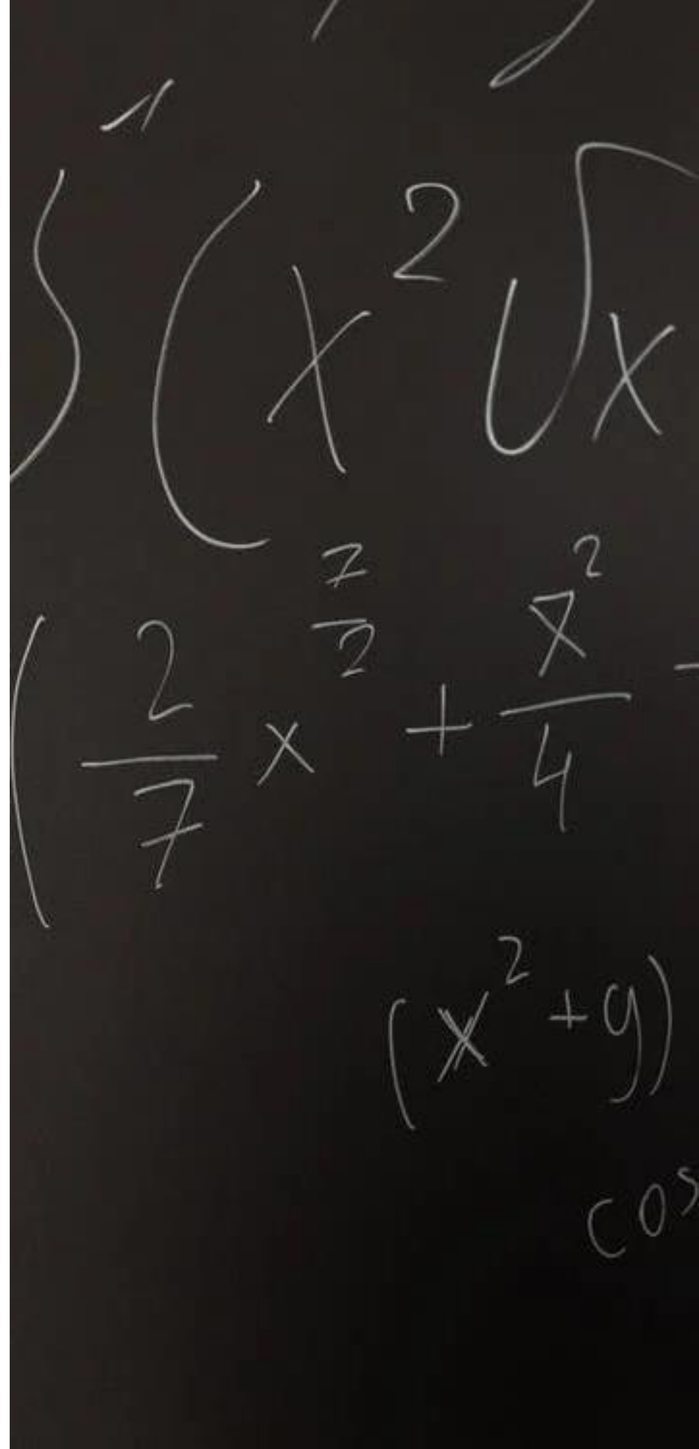


Desarrollo de conceptos

Algoritmo

1. Definir el número de clusters, **2 clusters**

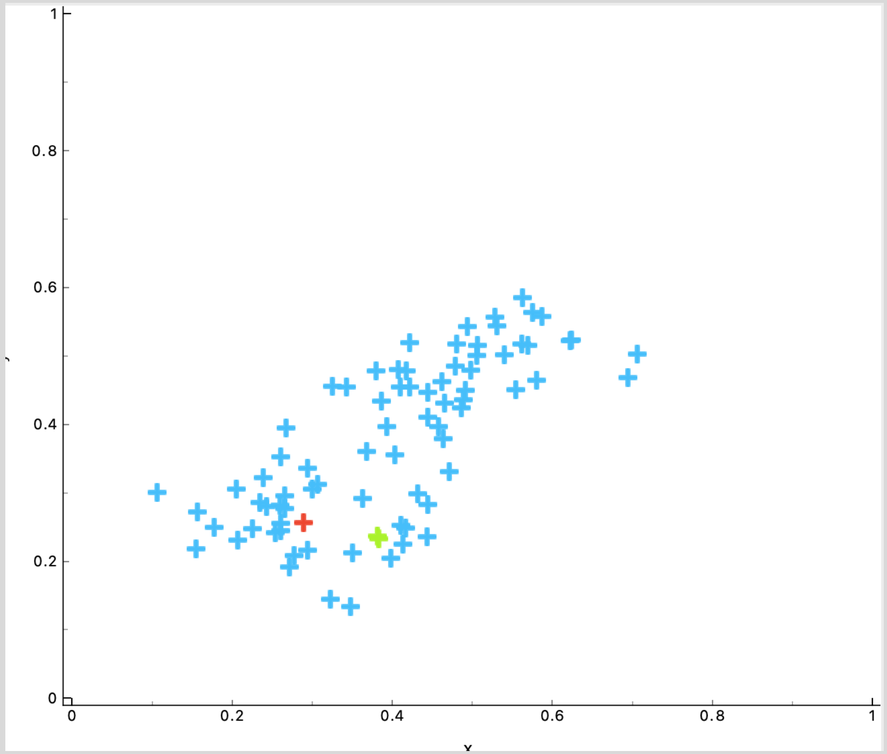


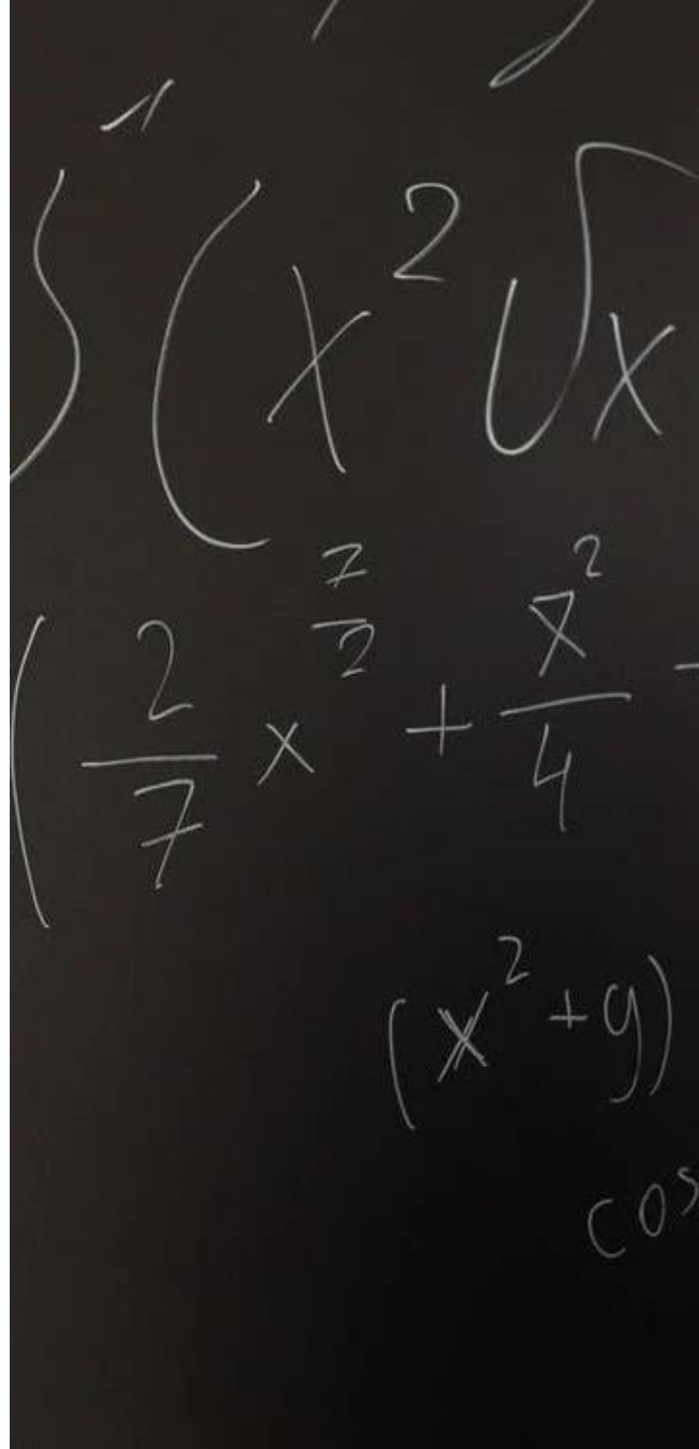


Desarrollo de conceptos

Algoritmo

2. Generar aleatoriamente **n** (centroides)

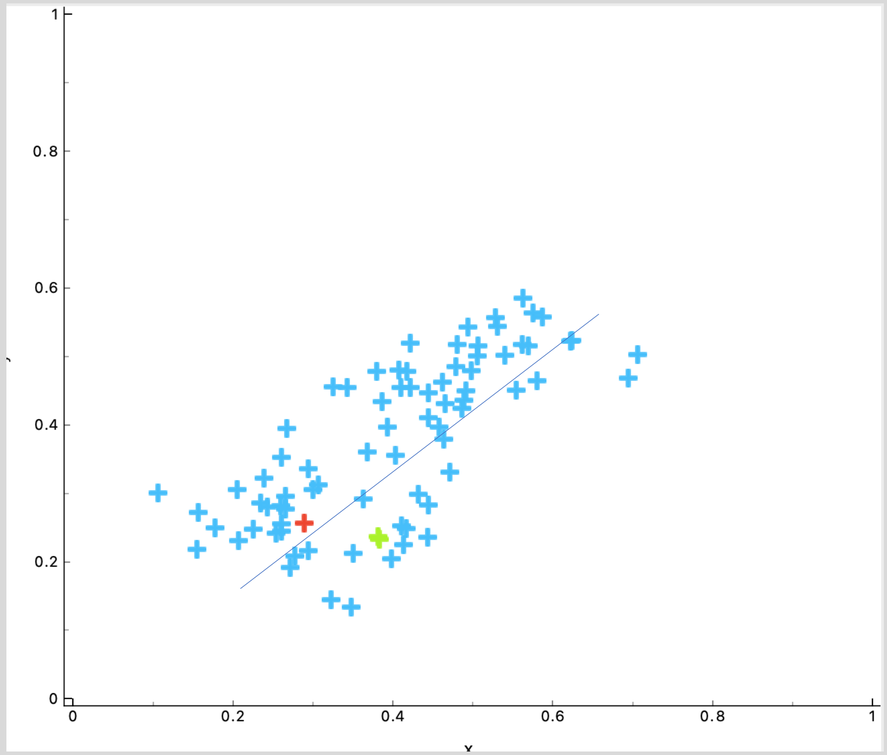


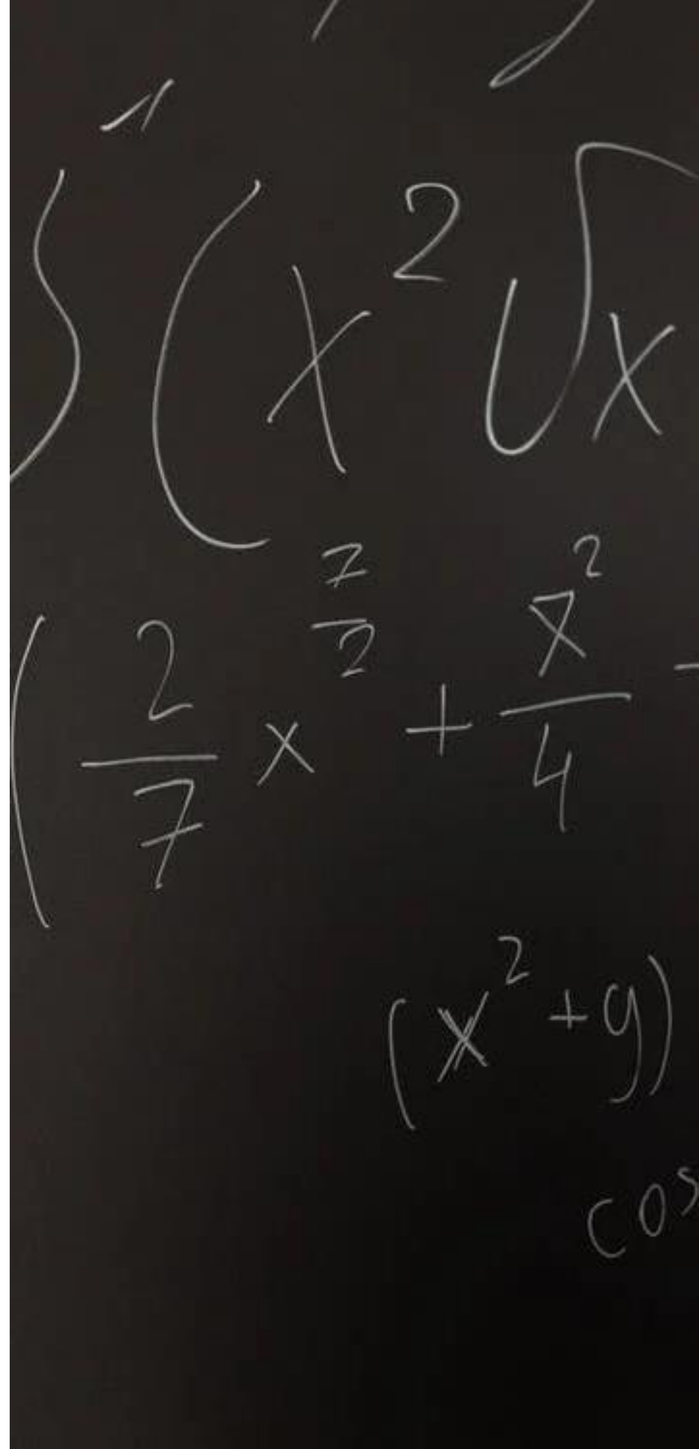


Desarrollo de conceptos

Algoritmo

3. Asignar cada elemento del conjunto de datos al centroide más cercano para formar **n** grupos

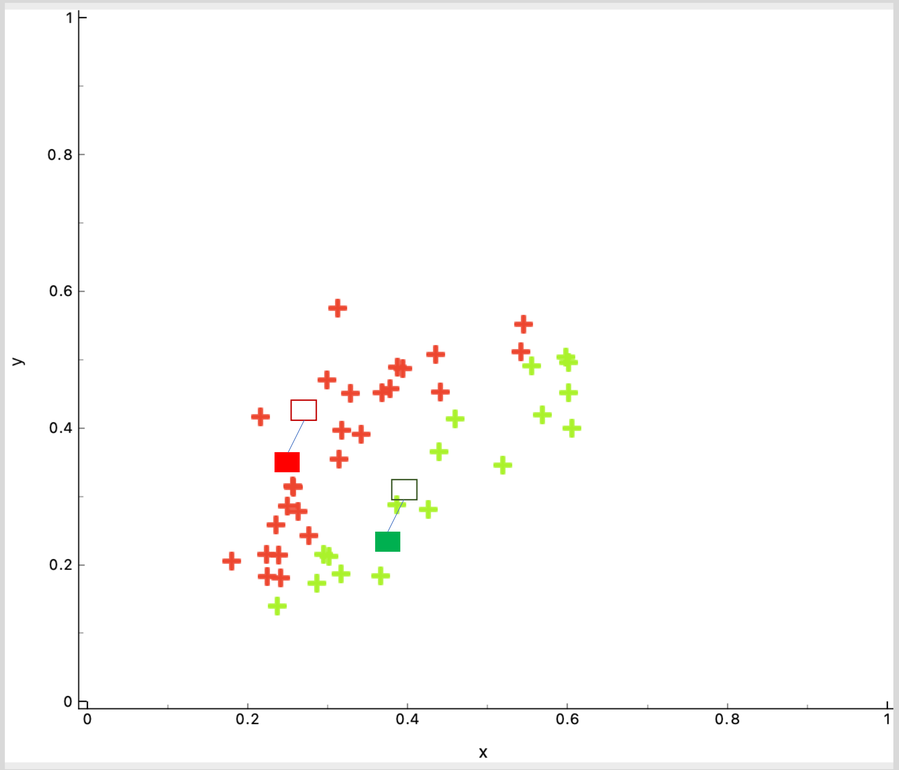


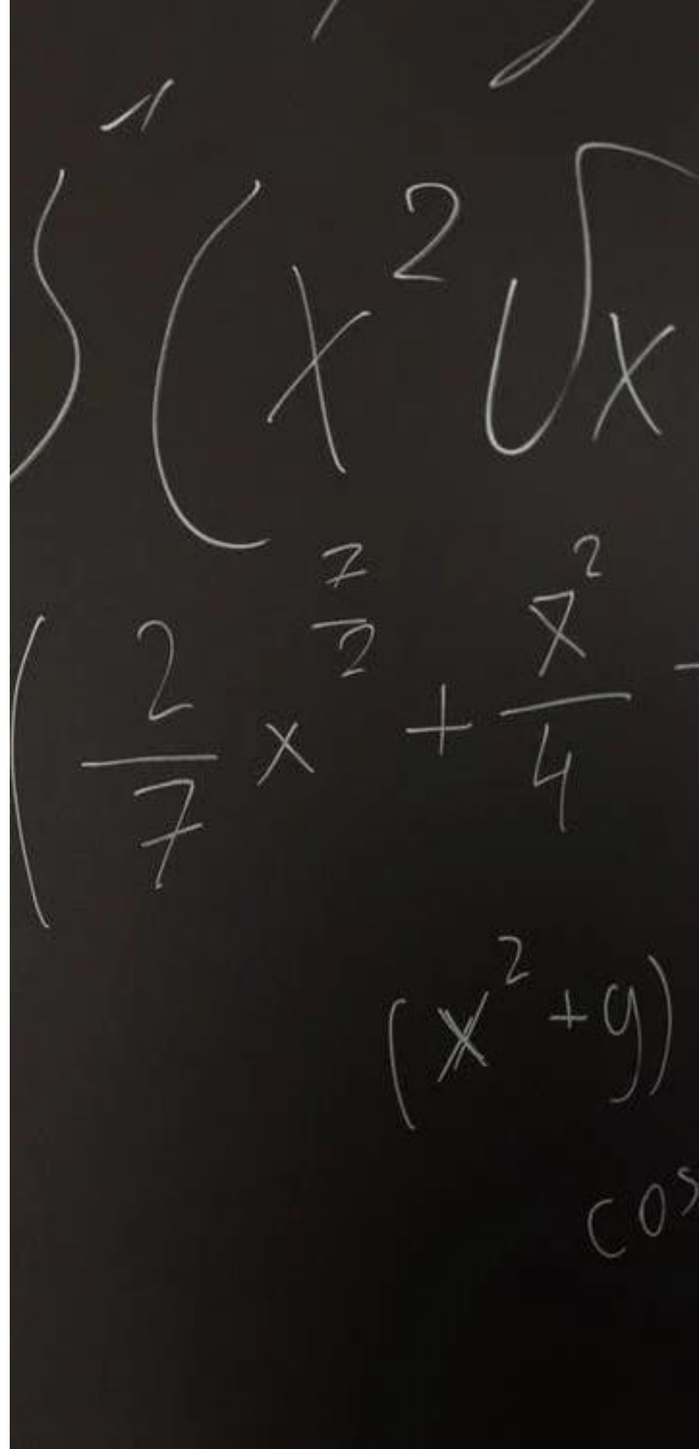


Desarrollo de conceptos

Algoritmo

4. Reasignar la posición de cada centroide

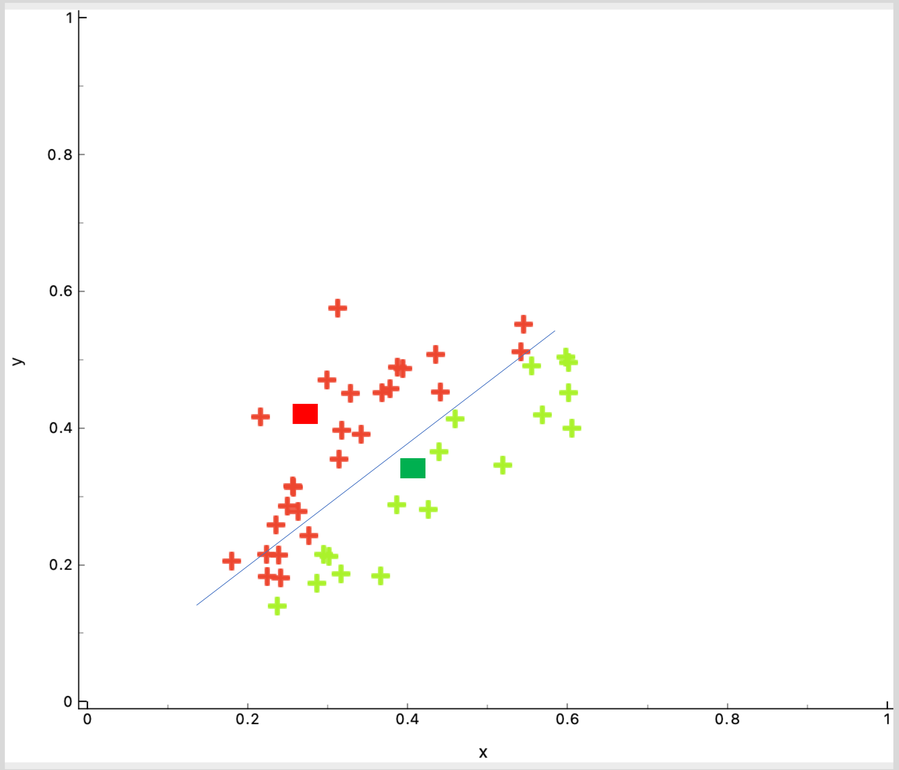


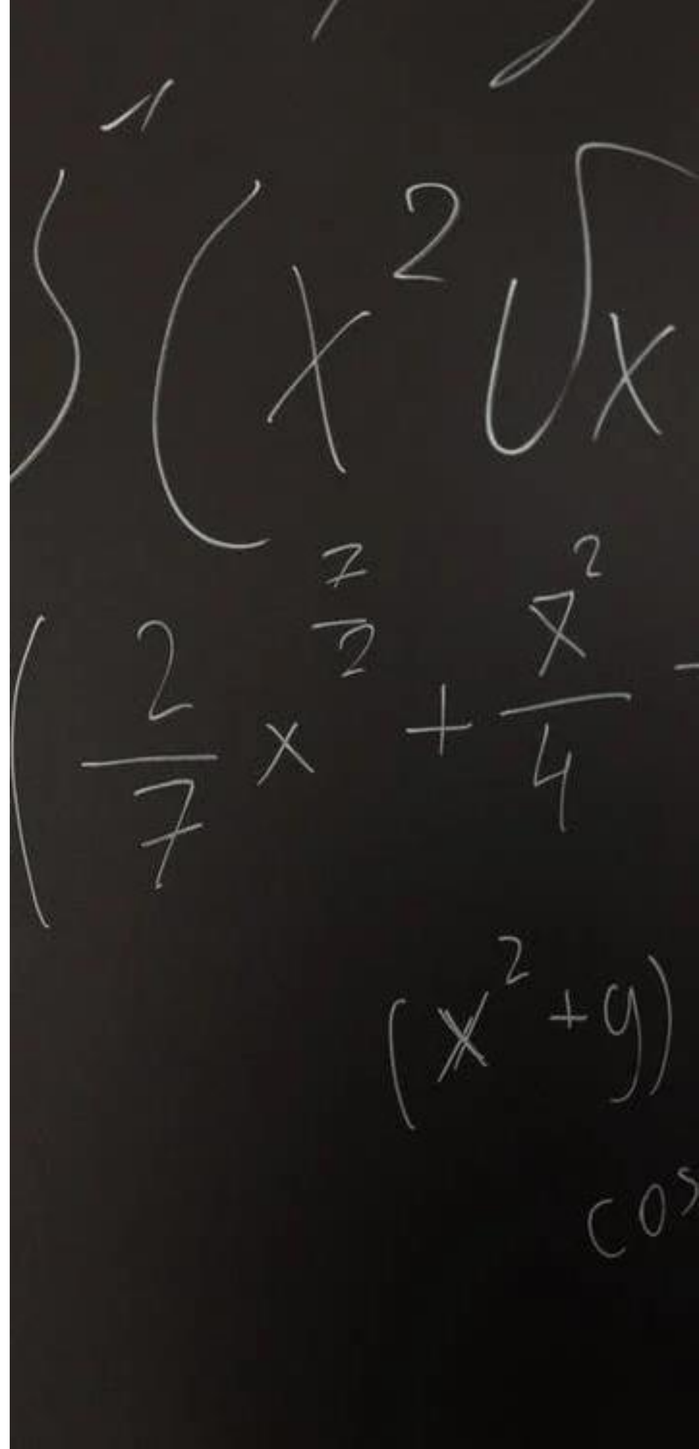


Desarrollo de conceptos

Algoritmo

5. Reasignar los elementos de datos al centroide más cercano

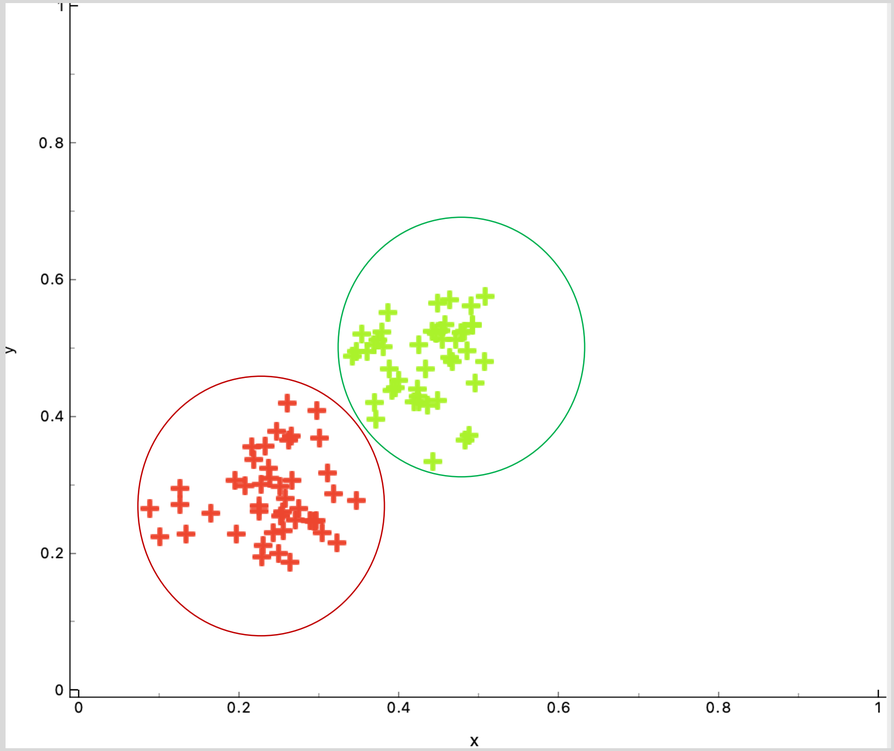


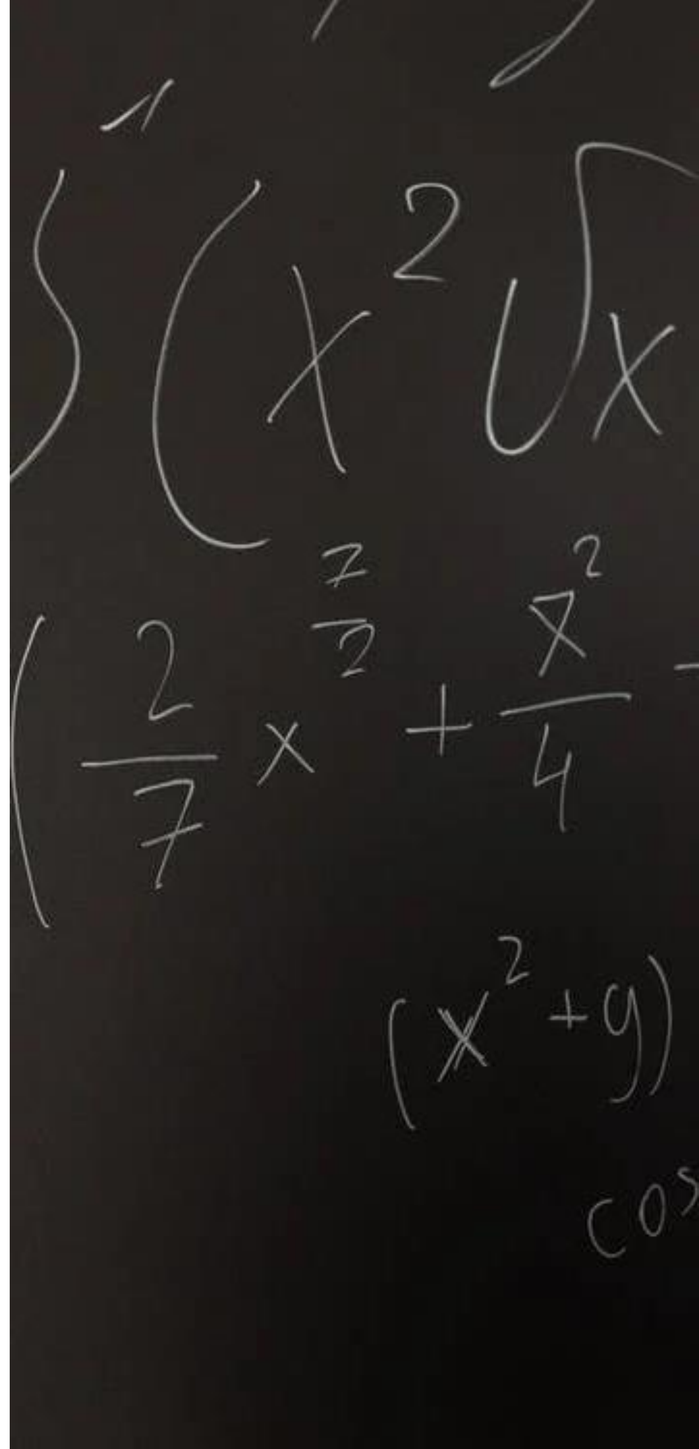


Desarrollo de conceptos

Algoritmo

Proceso terminado





Solución

- K-means
- Dentro de Colab
  - https://colab.research.google.com

Recursos

Bibliográficos

- Alpaydin, Ethem, Introduction to Machine Learning, third edition, MIT Press, 2014, Chapter 7 Clustering
- Alaminos Chica, Antonio, Análisis multivariante para las Ciencias Sociales I
- Hair, Anderson, tatham & Black, Análisis Multivariante, Prentihce Hall, 1999
- Peña, Daniel, Análisis de Datos Multivariantes, 2022
- Himanshu Sharma, Hierarchical Clustering, https://harshsharma1091996.medium.com/hierarchical-clustering-996745fe656b
- Deinda Afiya Pangestu, The Difference Between Hierarchical and Non-Hierarchical Clustering, https://medium.com/@deindaafiya/the-diffence-between-hierarchical-and-non-hierarchical-clustering-2599e45c395