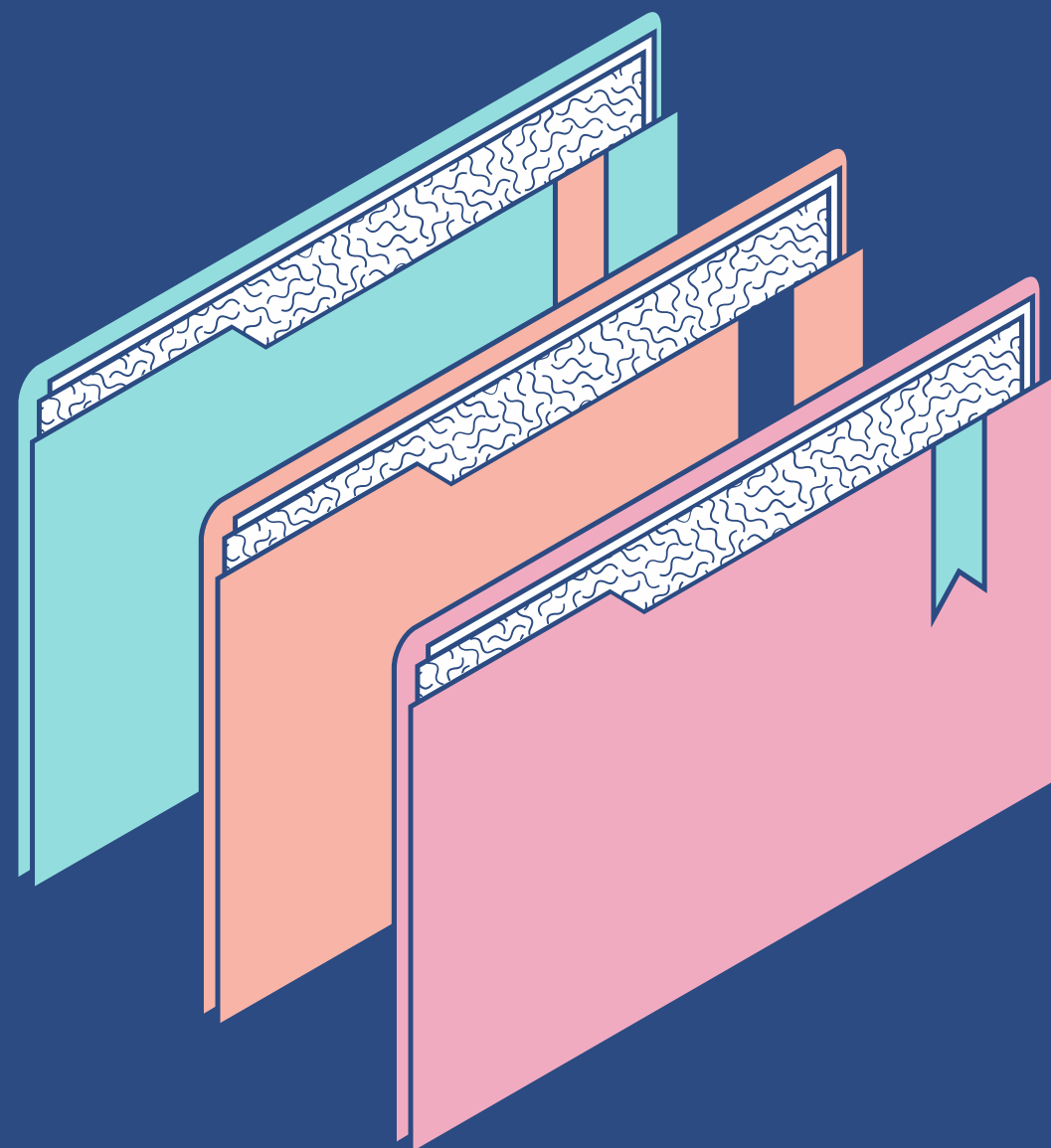


UNIVERSIDAD DE COSTA RICA  
FACULTAD DE INGENIERÍAS  
ESCUELA DE CIENCIAS DE LA  
COMPUTACIÓN E INFORMÁTICA

**Grupo ZZZ:**  
**Nathalie Alfaro Quesada - B90221**  
**José Pablo Mora Cubillo - B75044**

CI - 0124 Computabilidad y Complejidad  
II Semestre, 2024



## TAREA PROGRAMADA 2

### AVANCE 2

## ALGORITMO DE AGRUPAMIENTO EN DATOS DE CÉLULAS

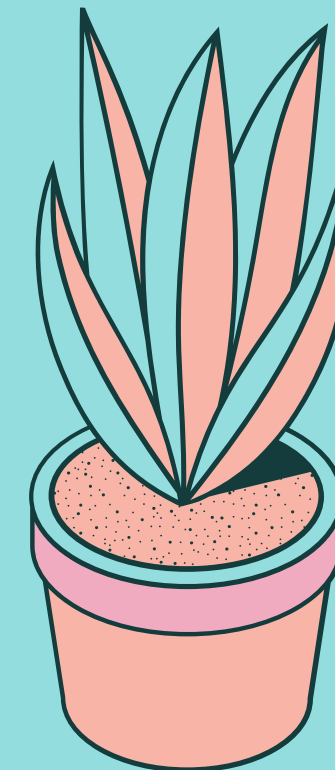
# ANEUPLOIDÍAS

Alteraciones cromosómicas que se caracterizan por la ganancia o pérdida de cromosomas completos o segmentos cromosómicos.

Frecuentes en el cáncer.

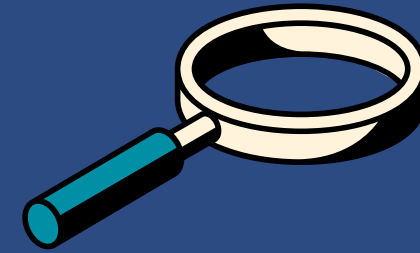
Número de copias (CN) a nivel de gen en condiciones normales este valor debería ser de 2.

Datos de CN de 718 genes implicados en cáncer para 1398 líneas celulares cancerígenas.



## OBJETIVO

Busca agrupar estas líneas celulares en grupos de acuerdo a similitudes en los valores de CN, por lo cual se va a utilizar algoritmos de agrupamiento.



## IMPORTANCIA

Se puede hipotetizar que las líneas celulares con patrones parecidos de CN tienen comportamientos similares, por ejemplo, sensibilidad al mismo tipo de fármacos.



# DATOS

symbol	ABCB1	ABI1	ABL1	ABL2	ACKR3	ACSL3	ACVR1	ACVR2A	AFDN	AFF1	...	ZNF208	ZNF331	ZNF384	ZNF429	ZNF521	ZNF626
model_id																	
SIDM00001	3.0000	2.0000	3.0000	2.0000	1.903779	2.0000	2.0000	2.5000	1.0000	2.0000	...	1.5000	1.0000	2.0000	1.5000	2.0000	1.5000
SIDM00002	3.0000	3.0000	3.0000	3.0000	3.000000	3.0000	3.0000	3.0000	2.0000	3.0000	...	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000
SIDM00003	4.0000	3.0000	3.0000	3.0000	3.000000	3.0000	3.0000	3.0000	2.0000	2.0000	...	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000
SIDM00008	5.0000	3.0000	2.0000	4.0000	4.000000	4.0000	4.0000	5.0000	3.0000	3.0000	...	3.0000	3.0000	3.0000	3.0000	4.0000	3.0000
SIDM00011	7.0000	3.0000	3.0000	4.0000	4.000000	4.0000	4.0000	4.0000	3.0000	3.0000	...	2.0000	3.0000	3.0000	2.0000	3.0000	2.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
SIDM02076	3.8307	2.0521	4.2114	3.2612	3.079400	3.0782	3.1249	3.1249	2.9867	3.1873	...	3.3713	3.7299	3.8227	3.3713	3.6534	3.3713
SIDM02077	2.9821	2.4088	2.1983	1.6317	1.636200	1.6362	1.6053	0.9731	1.6669	1.0377	...	0.9191	1.9538	2.0136	1.8329	1.9477	1.9803
SIDM02078	2.9376	1.9788	2.8560	2.0221	2.882000	2.8820	2.8694	2.8694	2.0289	2.9833	...	1.9748	3.7649	4.9985	1.9748	1.0058	1.9748
SIDM02079	2.9775	2.7970	3.8654	3.1729	1.998400	4.0418	1.9879	1.9879	2.0014	1.1020	...	1.9649	3.8902	3.6761	1.9649	3.0050	1.9649
SIDM02080	1.8506	1.9958	2.0053	2.1766	2.071200	2.0117	2.0267	2.0267	2.0100	1.0082	...	1.9783	2.0057	1.0433	1.9783	3.8952	1.9783

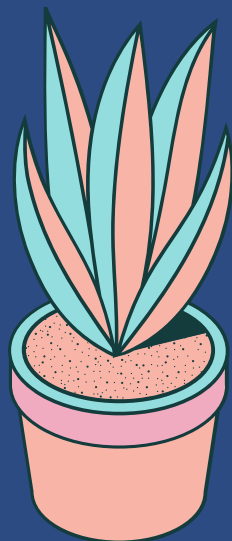
1398 rows × 718 columns

# ALGORITMOS IMPLEMENTADOS

Fuerza Bruta.

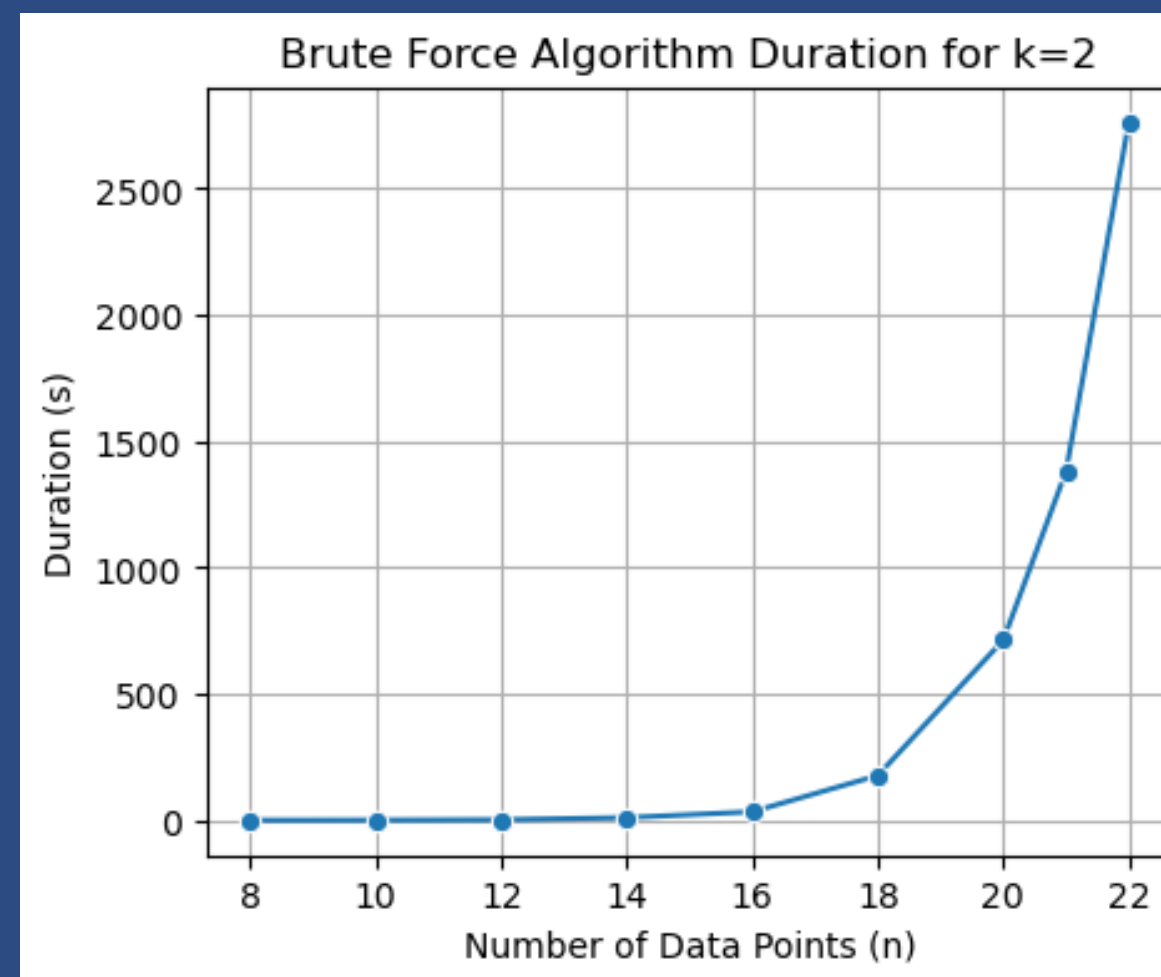
K-Means

Algoritmo genético.



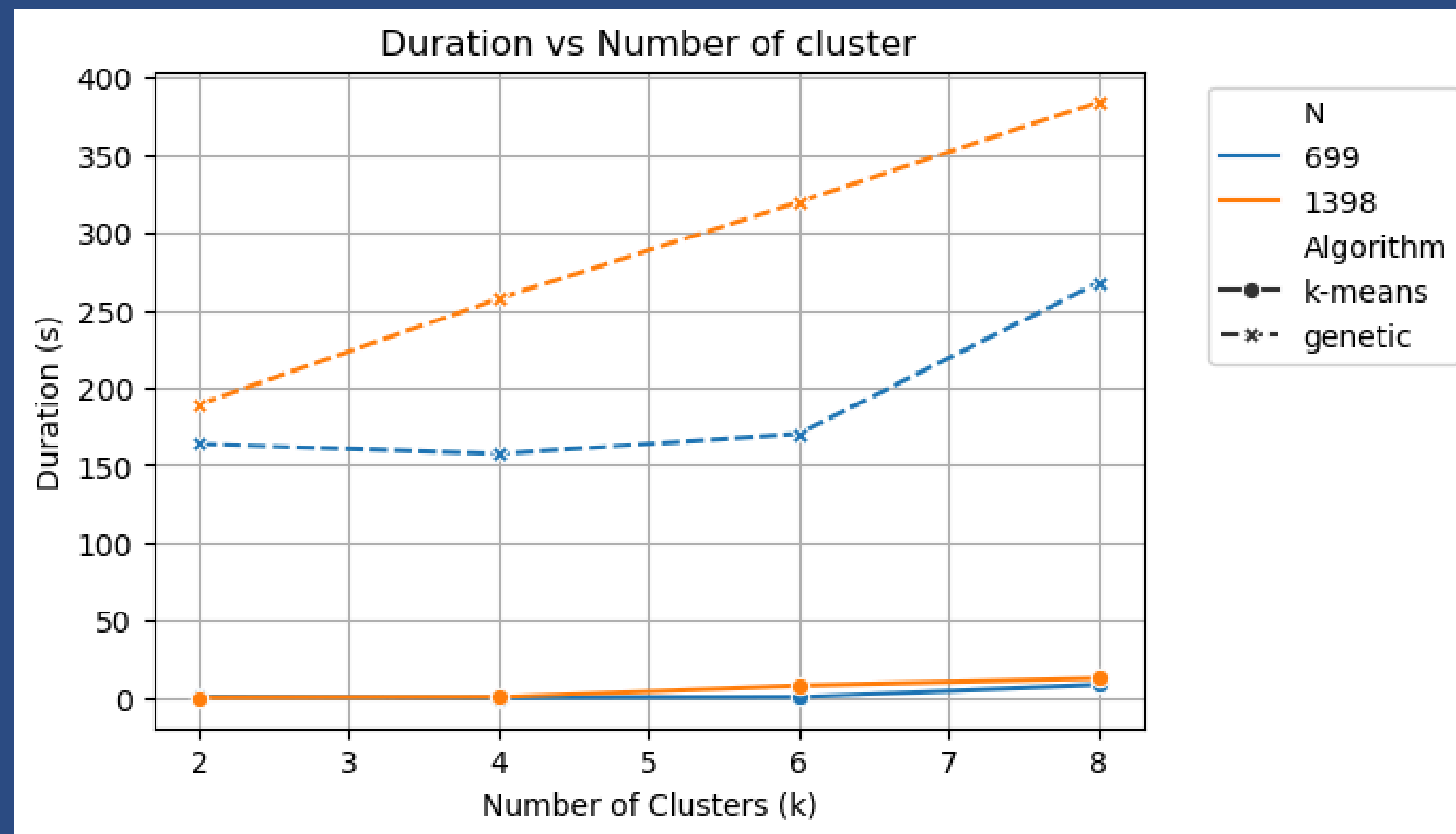
# RENDIMIENTO DE FUERZA BRUTA

- Debe realizar  $k^n$  comparaciones, donde  $k$  es el número de clusters y  $n$  la cantidad de datos.
- No fue posible ejecutarlo para el set de datos completos ( $n = 1398$ )



# RENDIMIENTO DE K-MEANS Y ALG. GENÉTICO (1)

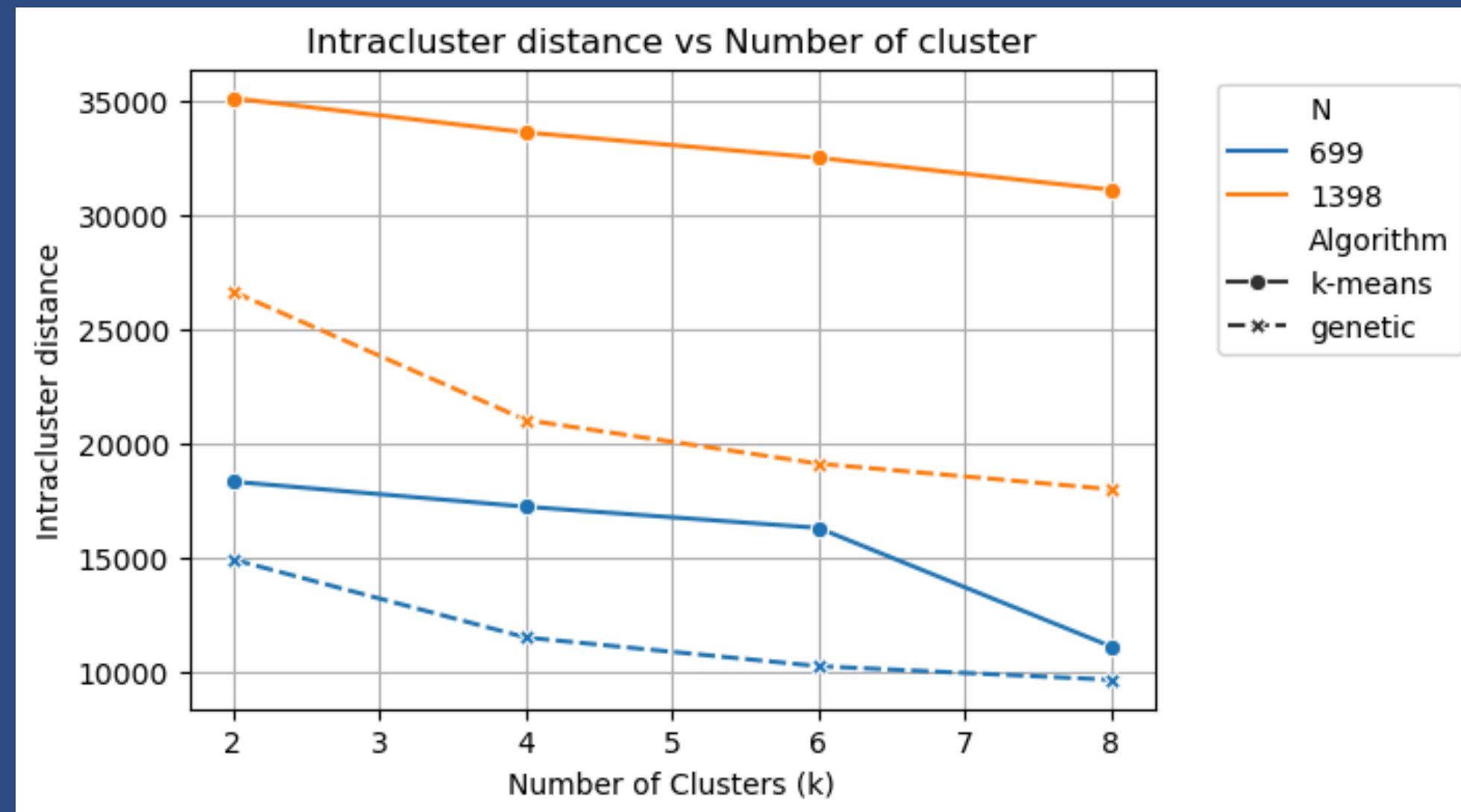
- Duración incrementa con el número de clusters.





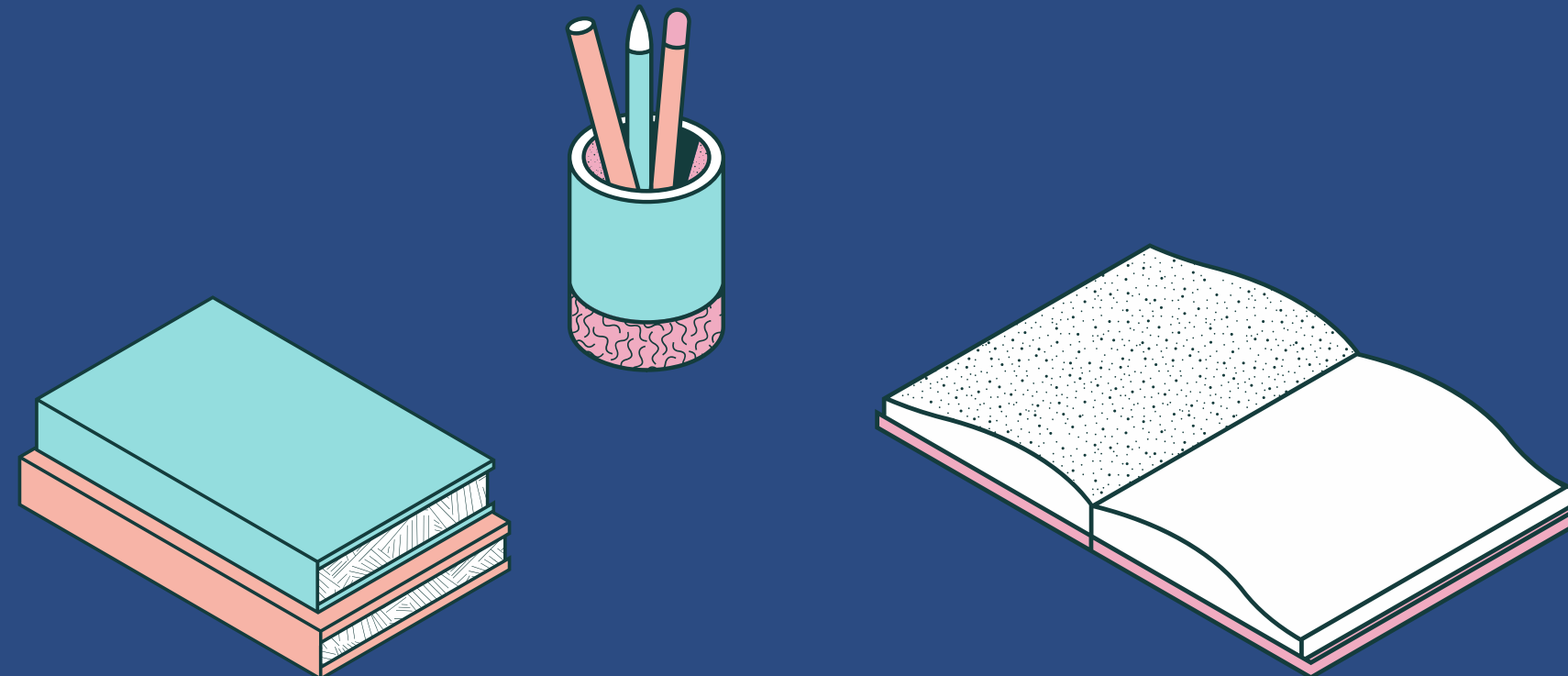
# RENDIMIENTO DE K-MEANS Y ALG. GENÉTICO (2)

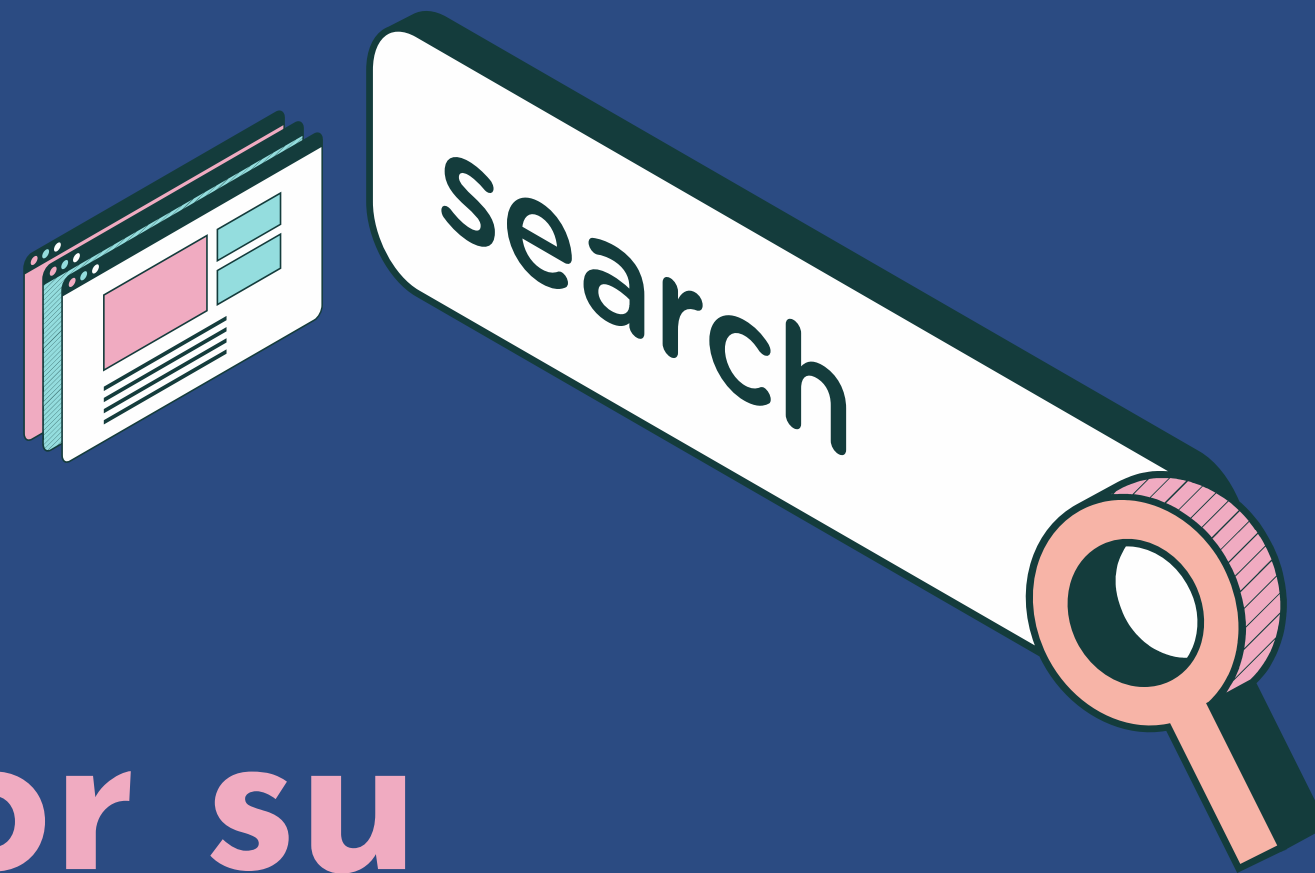
- El algoritmo genético tiene una duración mayor, pero tiende a dar mejores resultados. K-means se queda atrapado en óptimos locales.



# REFERENCIAS

Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern recognition*, 33(9), 1455-1465.

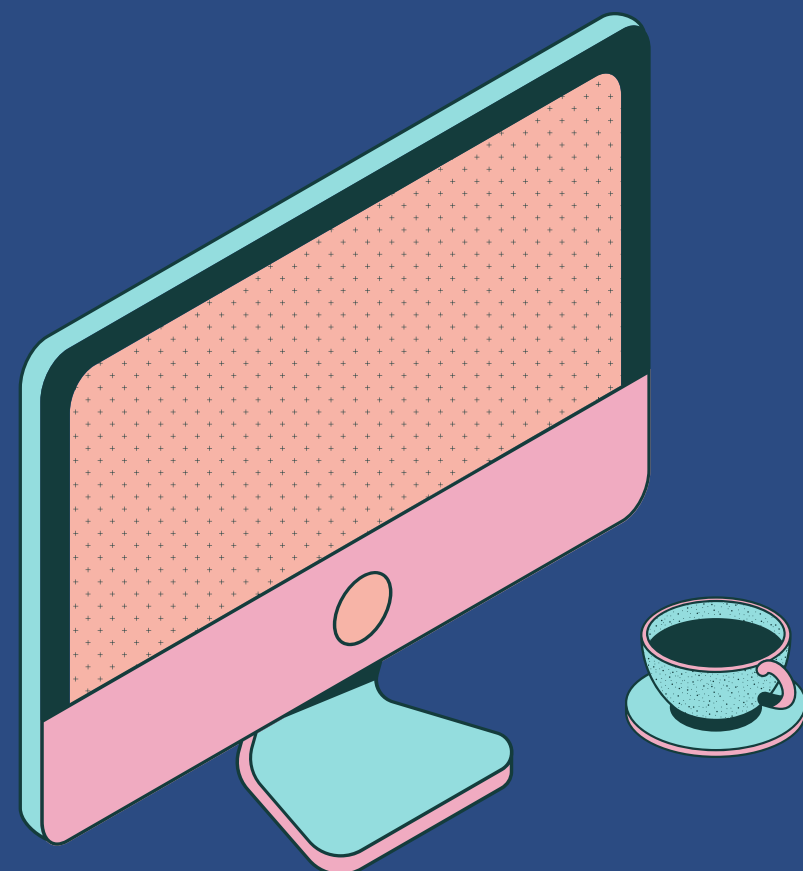




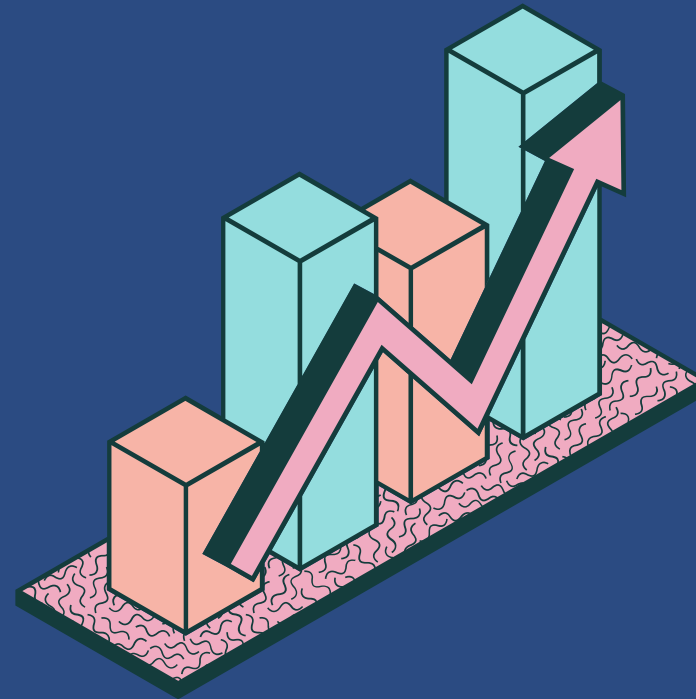
Gracias por su  
atención



# Anexos



# CLUSTERING



Técnica utilizada en el análisis de datos para agrupar un conjunto de objetos en grupos (o clusters) de manera que los objetos dentro de un mismo grupo sean más similares entre sí que con los objetos de otros grupos.

Considerados *NP-Hard*.

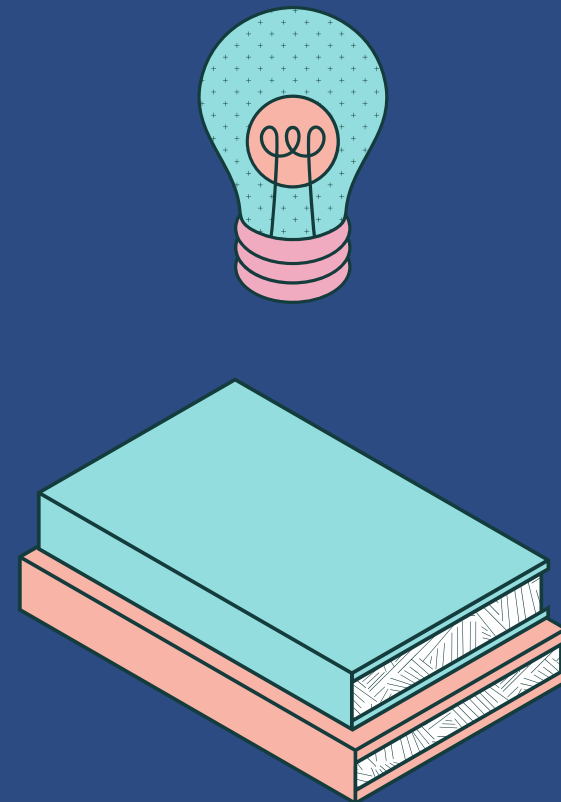
# HEURÍSTICA

Algoritmo k-means (también llamado algoritmo de Lloyd).

Algoritmo iterativo.

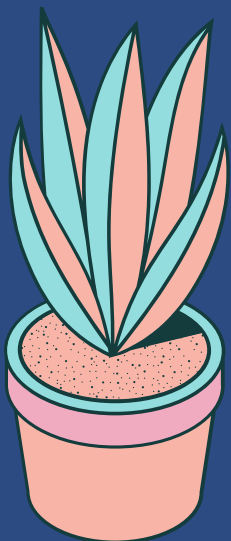
Problemas de agrupamiento y que se puede quedar atrapado en óptimos locales.

Se reciben los datos y un valor  $k$  que representa la cantidad de grupos a generar y la primera iteración  $k$  centroides aleatorios y cada entrada de los datos se asigna al cluster cuyo centroide se encuentra más cerca según una distancia euclidiana.



# K-MEANS

1. Inicializar los centroides, ya sea aleatoriamente o seleccionando puntos al azar del conjunto de datos.
2. Asignar cada punto al cluster cuyo centroide esté más cercano, utilizando la distancia euclidiana.
3. Recalcular los centroides de los clusters, donde el nuevo centroide estará formado por los promedios por dimensión de cada entrada presente en el cluster.
4. Repetir los pasos de (2) y (3) hasta que las asignaciones de los puntos de datos ya no cambien significativamente.



# DISTANCIA EUCLIDIANA

$$d = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + \dots + (Xn - Yn)^2}$$

Donde  $X_i$  corresponde al valor del centroide en la dimensión  $i$  y  $Y_i$  corresponde al valor de la entrada en la dimensión  $i$ . Si tenemos  $K = 4$  y 3 dimensiones (a veces también llamadas *features*), los cluster van a tener la siguiente forma:

$$C1 = (X1, Y1, Z1); C2 = (X2, Y2, Z2); C3 = (X3, Y3, Z3); C4 = (X4, Y4, Z4)$$



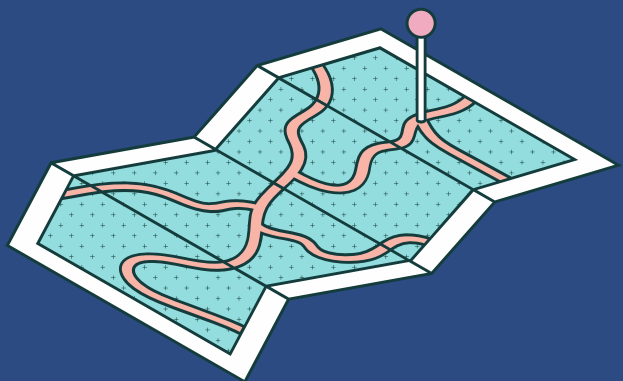
# METAHEURÍSTICA

Algoritmos genéticos para clustering, basándonos en Maulik y Bandyopadhyay (2000).

Se asume que el usuario va a brindar la cantidad  $k$  de clusters en los que se deben agrupar los datos.

Empleando una distancia euclidiana.

Espacio de búsqueda consiste en todos los posibles valores que van a tomar los  $k$  centroides, el cromosoma va a estar formado por estos mismos valores.



# CASO DE K = 4 Y 3 DIMENSIONES

Posible cromosoma sería:

$[(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), (X_3, Y_3, Z_3), (X_4, Y_4, Z_4)]$

De manera que para la mezcla de la etapa de cruce, se pueden intercambiar clusters y/o valores dentro de los clusters.

Para calcular el valor de *fitness* de cada cromosoma se toma el inverso de la sumatoria de las distancias euclidianas para cada muestra dentro de cada cluster:

Donde  $N_k$  es el número de clusters,  $E_i$  es el número de entradas en el cluster  $i$  y  $d_{kj}$  es la distancia euclidiana de la entrada  $j$  en el cluster  $i$  con el centroide  $i$ .

$$Fitness = \frac{1}{\sum_{i=1}^{N_k} \sum_{j=1}^{E_i} d_{ij}}$$

# FUERZA BRUTA

Probar todos los valores que pueden tomar los centroides.

Probar todas las posibles combinaciones que se pueden realizar entre todos los datos y los  $k$  clusters.

