



Instituto Politécnico Nacional  
Escuela Superior de Cómputo



# K-Means

Aprendizaje máquina e  
Inteligencia Artificial

06/06/2022

Equipo:

Alcibar Zubillaga Julián  
De Luna Ocampo Yanina

# Tabla de contenidos

01

## Introducción

¿De dónde surge este algoritmo?

02

## Principios y Algoritmo

Analizaremos el algoritmo y los principios.

03

## Programación

Veremos algunas aplicaciones.

04

## Conclusión

Por último daremos una pequeña conclusión.



01

# Introducción

—



¿Qué es el  
clustering?

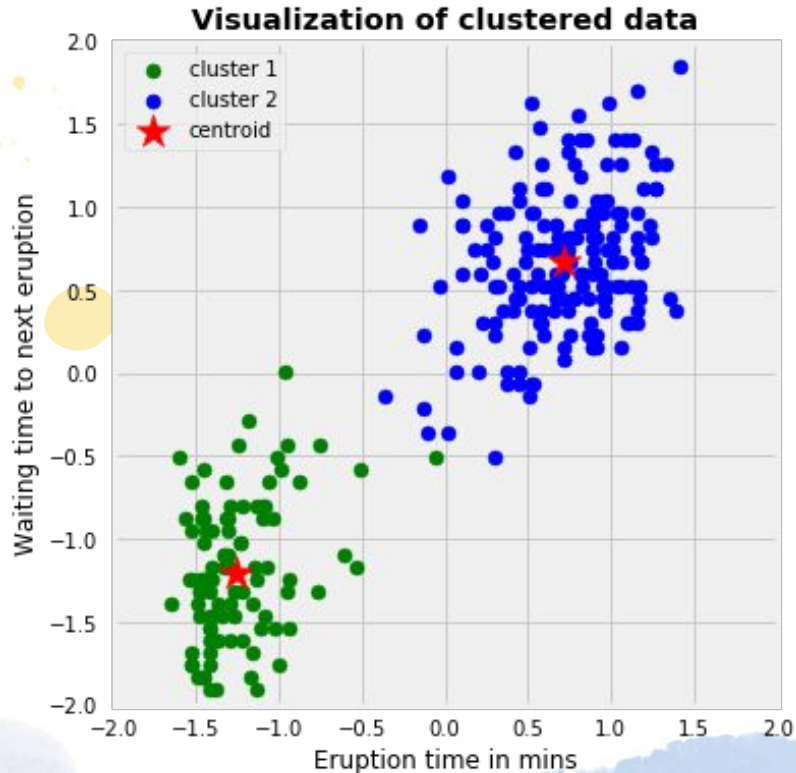


# Clustering

El objetivo es agrupar los datos que presentan ciertas semejanzas entre sus miembros, es decir “que se parezcan”

También buscamos que los datos que pertenezcan a grupos diferentes tengan rasgos lo suficientemente diferentes entre sí.

# Clustering



## 1.- Cluster

Podemos observar que en este dataset se muestra que entre más tiempo de espera entre erupciones, significa que la próxima erupción durará más.

## 2.- Cluster

En este caso, muestra que entre menos más rápidas sean esas erupciones, entonces estas durarán menos.

# Clustering - ¿Cómo se cataloga un buen clustering?



1.-

Los puntos deben tener propiedades comunes en el contexto estudiado.



2.-

Los puntos de un mismo cluster deben ser compactos y tener intersección mínima.



3.-

Los clusters deben ser identificables y de tamaño considerable.

# Aplicaciones del Clustering



Planificación  
Urbana



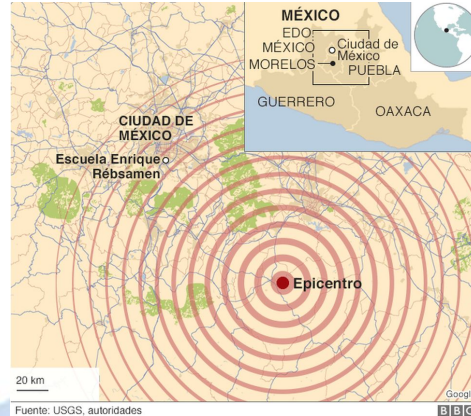
Detección de  
los epicentros



Taxonomía de  
los seres vivos

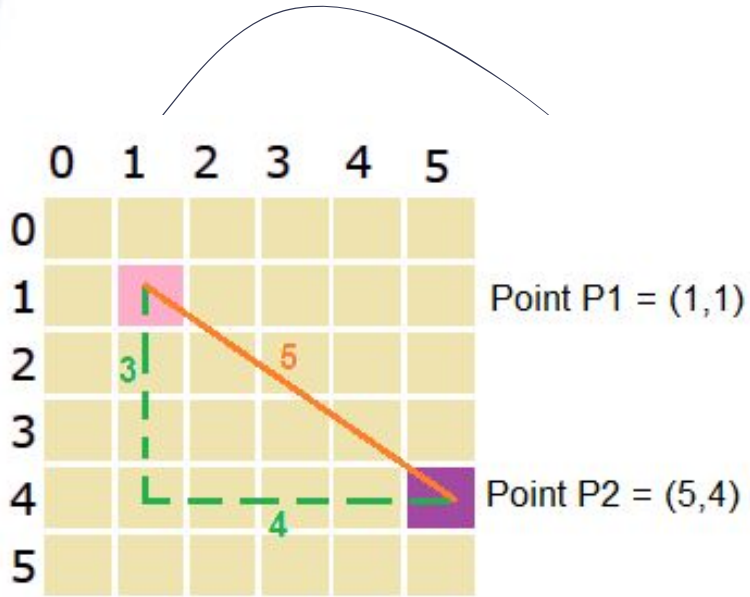


Imputar valores  
desconocidos de un  
dataset





# Tipos de Distancias



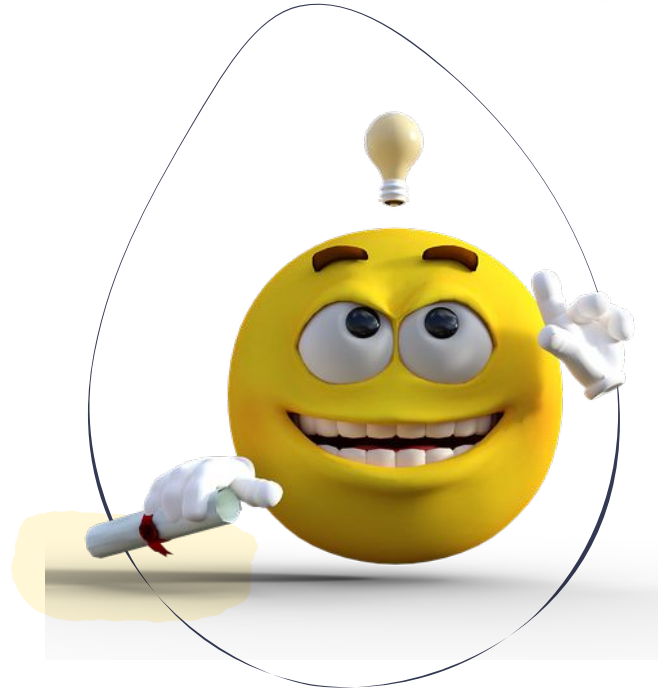
— Distancia  
Manhattan

— Distancia  
Euclidiana

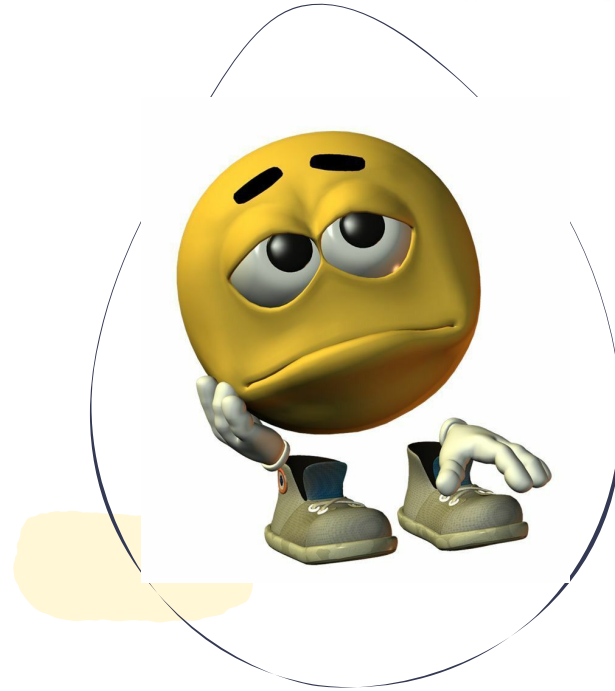
— Distancias P

# ¿Dónde surge?

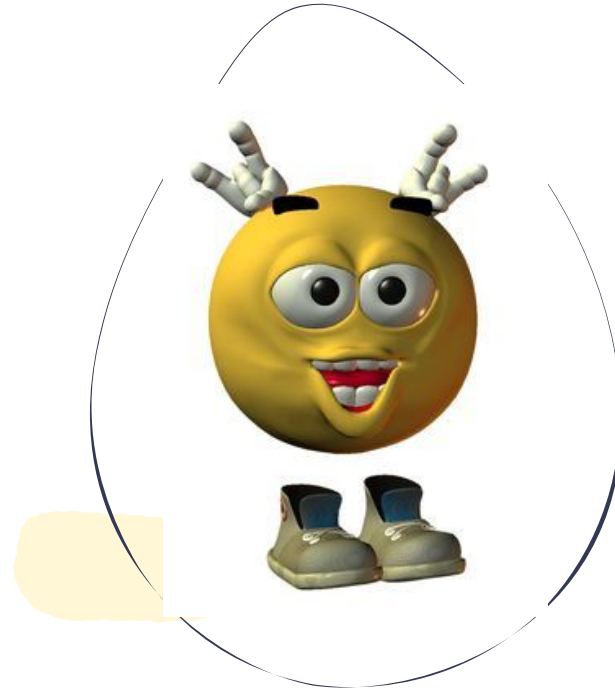
- Fue un término propuesto por James McQueen en 1967.
- Introducido por Hugo Steinhaus en 1957.
- Fue propuesto por primera vez por Stuart Lloyd en 1957 como una técnica de modelación de código de pulso.



- 
- En 1965 E. W. Forgy publicó esencialmente el mismo método, por lo que a veces se le nombra Lloyd-Forgy.
  - Este no se publicó fuera de los laboratorios Bell hasta 1982.
  - Parte del aprendizaje no supervisado.



- 
- Una versión más eficiente fue propuesta y publicada en Fortran por Hartigan y Wong en 1975/1979.







02

# Principios

—



K-means es  
el algoritmo  
más famoso.

# ¿Qué es el K-means?



# ¿Qué es el K-means?



## Técnica

Técnica de análisis exploratorio de datos.



## Jerarquía

Implementa un método no jerárquico para agrupar objetos.



## Distancia

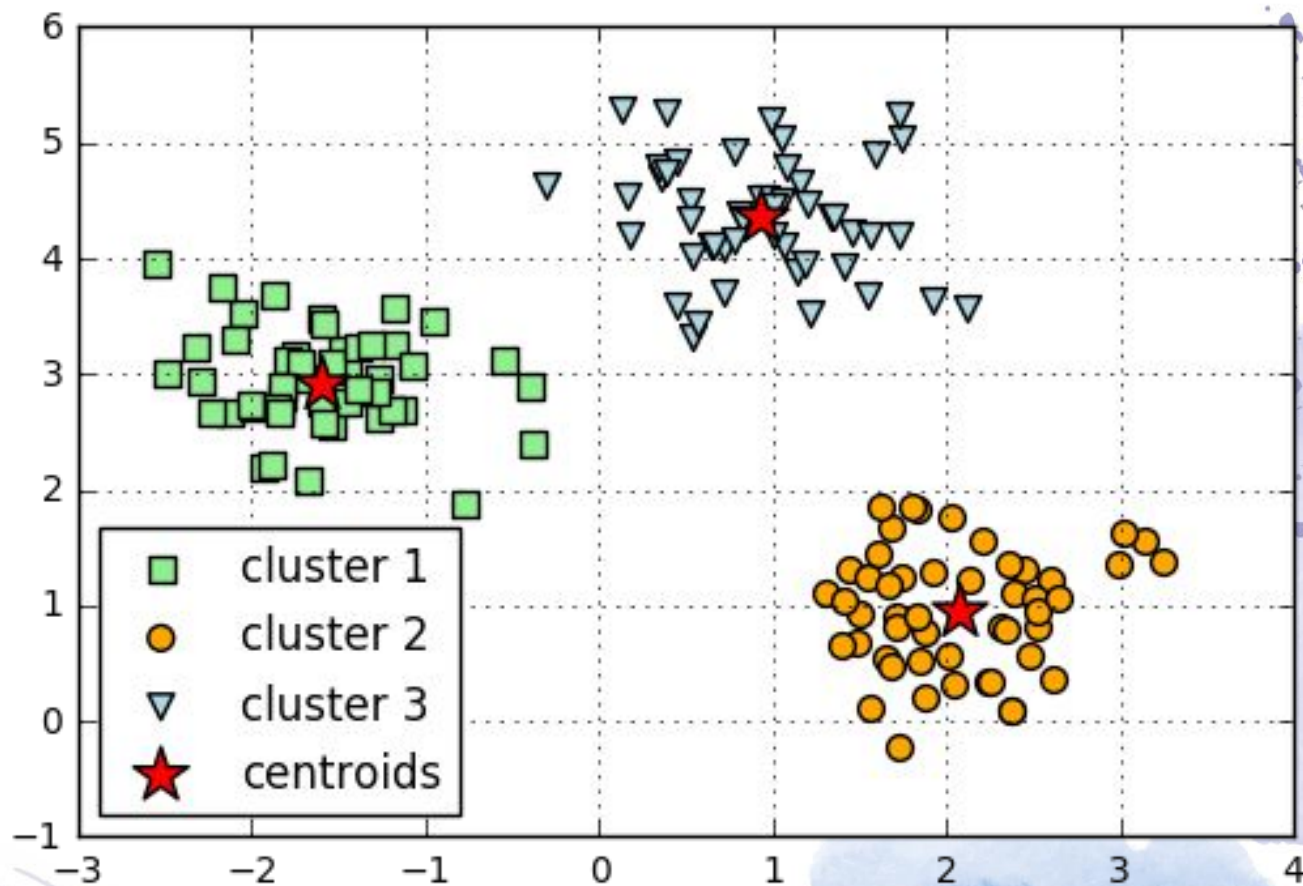
Utiliza la distancia Euclidiana para determinar el centroide para calcular las distancias.



## Agrupar

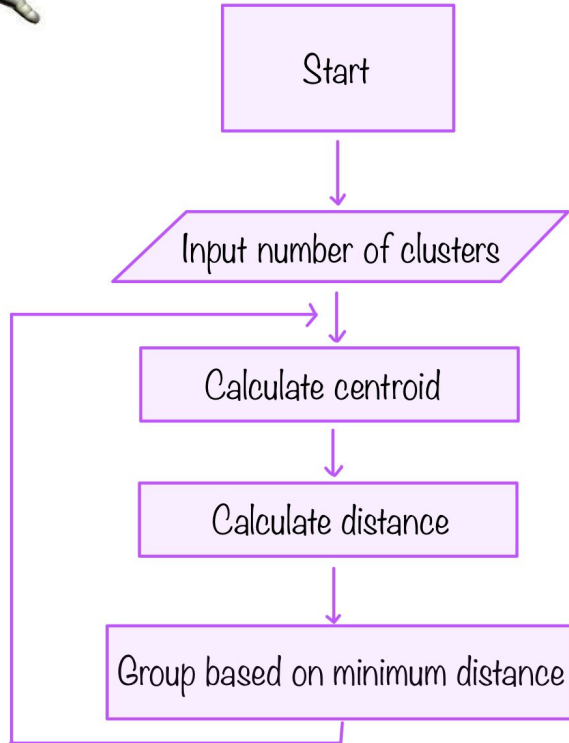
Agrupar con base en la distancia mínima.





Toma cualquier objeto aleatorio o los primeros objetos K.

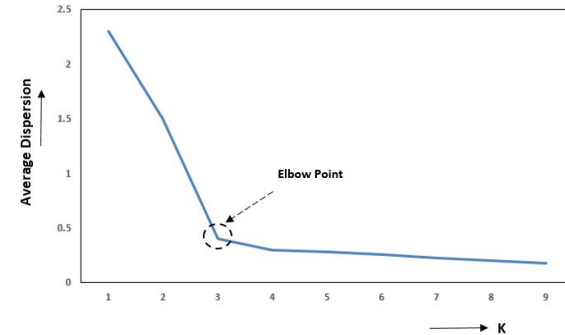
# Diagrama de flujo



## ● Elbow method

Determina el número de clústers en el conjunto de datos.

*Elbow Method for selection of optimal "K" clusters*



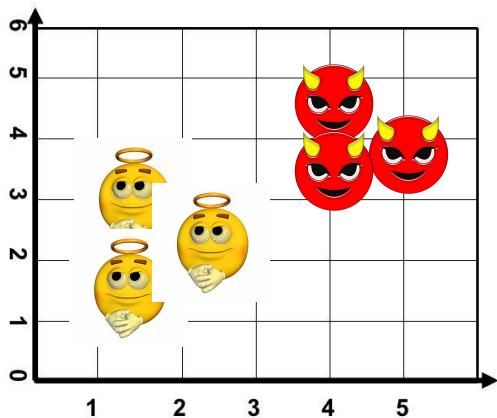
# Propiedades

## Homogeneidad:

Datos que pertenecen al mismo clúster deben ser lo más similares posible.

## Heterogeneidad:

Datos que pertenecen a diferentes agrupaciones deben ser tan diferentes como sea posible.





2.1

# Algoritmo

--



# El método de K-Means

Definimos la distancia intracluster para un cluster  $C_j$  cualquiera como:

$$SS_w(C_j) = \sum_{x \in C_j} (x - c_j)^2$$

También podemos usar la distancia intracluster normalizada para validar la eficacia del modelo:

$$S\tilde{S}_w = \sum_{j=1}^k \frac{SS_w(C_j)}{SS_T} \quad \text{donde} \quad SS_T = \sum_{i=1}^n (x_i - \bar{x})^2$$

Por lo que nuestro problema de agrupamiento se reduce a minimizar esta función:

$$SS_W(k) = \sum_{j=1}^k S_W(C_j) = \sum_{j=1}^k \sum_{x_i \in C_j} (x_i - c_j)^2$$

- donde  $k$  es el número de clusters.
- $x_i$  son los puntos que pertenecen al cluster  $j$ -ésimo.
- $c_j$  es el centroide del cluster  $j$ -ésimo.

# Algoritmo de método K-Means

El algoritmo consta de tres pasos:

1. Inicialización: una vez escogido el número de grupos,  $k$ , se establecen  $k$  centroides en el espacio de los datos, por ejemplo, escogiendo aleatoriamente.
2. Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.
3. Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo k-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

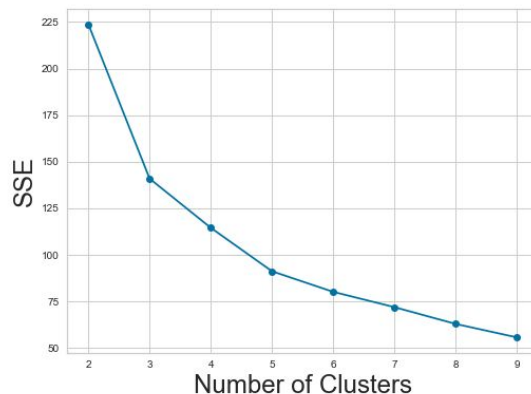


# Ajustar los parámetros del clustering



# El método del codo

Si representamos el número de clusters vs  $SSw(k)$ , la función suele presentar un codo que marca el  $k$  óptimo para el método de k-means.



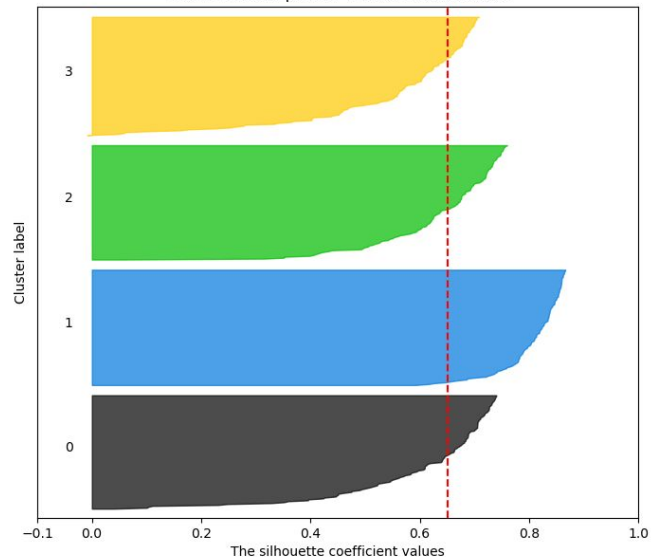
# El coeficiente de la silueta

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{si } a(i) < b(i) \\ 0 & \text{si } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{si } a(i) > b(i) \end{cases}$$

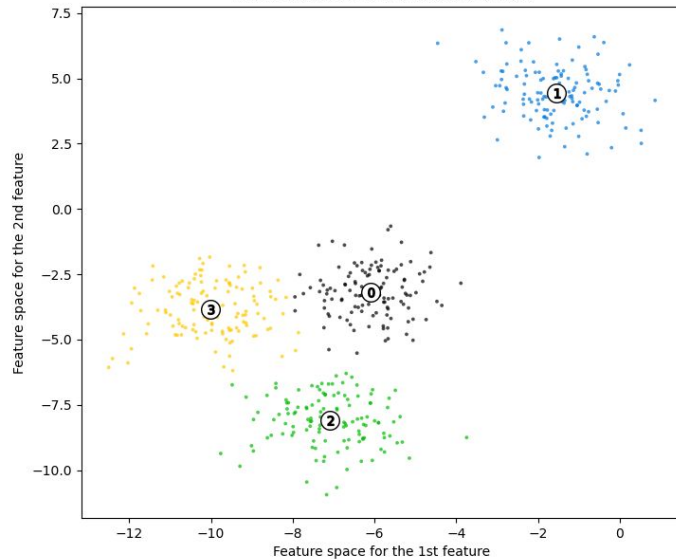
- Cuando  $S(i)$  tiende a 1, entonces  $a(i) \ll b(i)$  y por tanto el punto está muy bien clasificado
- Cuando  $S(i)$  tiende a **-1**, entonces  $a(i) \gg b(i)$  y por tanto, el punto estaría mejor en su cluster vecino.
- El promedio de  $S(i)$  sobre todos los puntos de un cluster nos informa de cómo de bien agrupados están.

### Silhouette analysis for KMeans clustering on sample data with $n\_clusters = 4$

The silhouette plot for the various clusters.



The visualization of the clustered data.



## sklearn.cluster.KMeans

```
class sklearn.cluster.KMeans(n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0,  
random_state=None, copy_x=True, algorithm='lloyd')
```

[\[source\]](#)

### Parámetros de Scikit-Learn

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^k \sum_{\ell \in C_i} d(\ell, z_i) \rightarrow \min_{\mathcal{C}, \mathcal{Z}}$$

Fórmula General para las distancias en el  
K-Means



# Mejoramiento del algoritmo K-Means



## K - Medoids

En contraste con el algoritmo k-means, **k-medoids** escoge datapoints como centros y trabaja con una métrica arbitraria de distancias.



## Método de Otsu

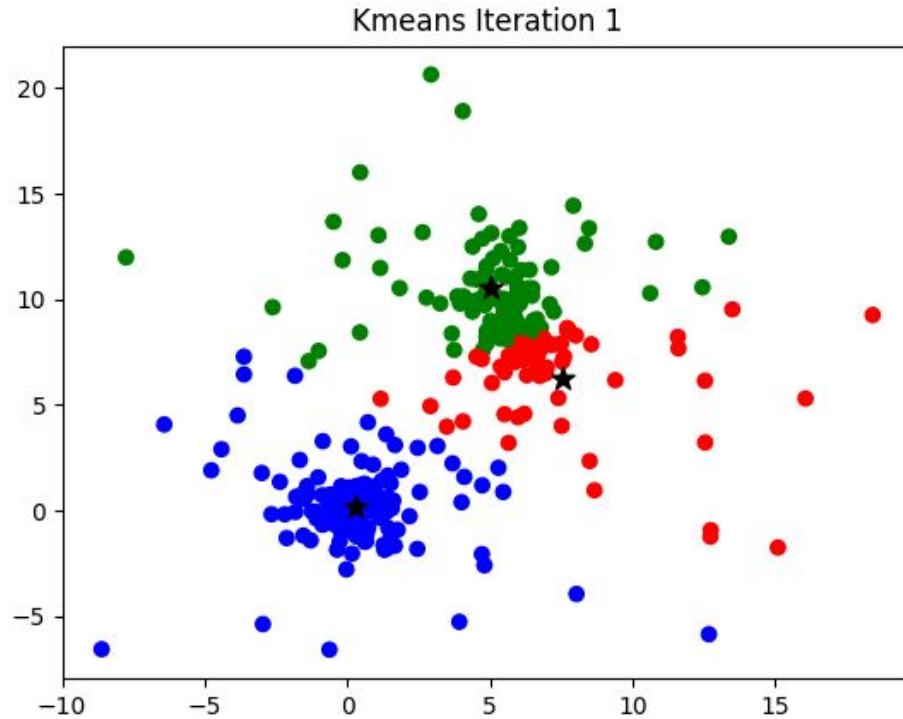
En concreto, se utiliza la varianza, que es una medida de la dispersión de valores – en este caso se trata de la dispersión de los niveles de gris.



## Fuzzy clustering

El agrupamiento difuso en el que cada punto de datos puede pertenecer a más de un grupo.

# Ejemplo:



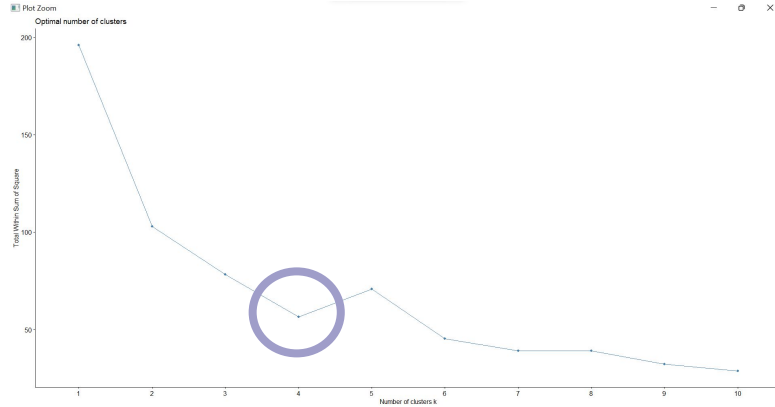


03

# Programación e Implementación

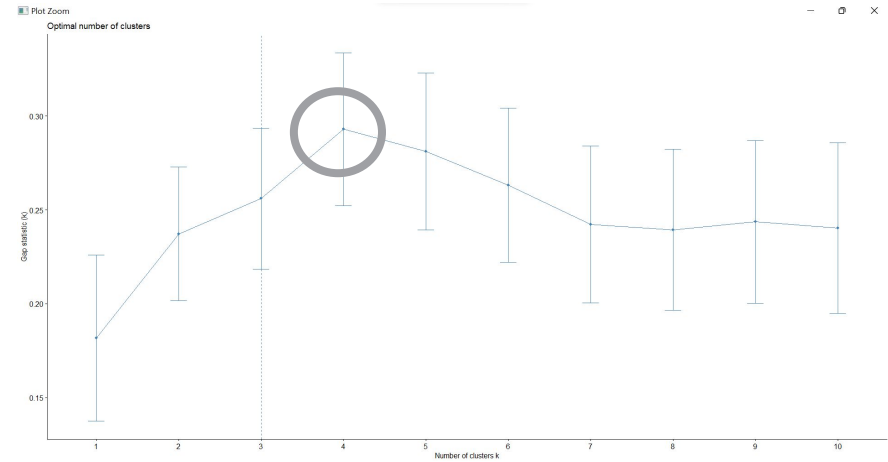
Python y R

# R



## Gap Statistics

## Elbow Method





# R

> KML  
K-means clustering with 4 clusters of sizes 13, 13, 16, 8

```
# 13 states were assigned to the first cluster  
# 13 states were assigned to the second cluster  
# 16 states were assigned to the third cluster  
# 8 states were assigned to the fourth cluster
```



# R

```
> aggregate(USArrests, by=list(cluster=km$cluster), mean)
```

	cluster	Murder	Assault	UrbanPop	Rape
1	1	3.60000	78.53846	52.07692	12.17692
2	2	10.81538	257.38462	76.00000	33.19231
3	3	5.65625	138.87500	73.87500	18.78125
4	4	13.93750	243.62500	53.75000	21.41250

```
# The mean number of murders per 100,000 citizens among the states in cluster 1 is 3.6.  
# The mean number of assaults per 100,000 citizens among the states in cluster 1 is 78.5.  
# The mean percentage of residents living in an urban area among the states in cluster 1 is 52.1%.  
# The mean number of rapes per 100,000 citizens among the states in cluster 1 is 12.2.
```

	Murder	Assault	UrbanPop	Rape	cluster
Alabama	13.2	236	58	21.2	4
Alaska	10.0	263	48	44.5	2
Arizona	8.1	294	80	31.0	2
Arkansas	8.8	190	50	19.5	4
California	9.0	276	91	40.6	2
Colorado	7.9	204	78	38.7	2

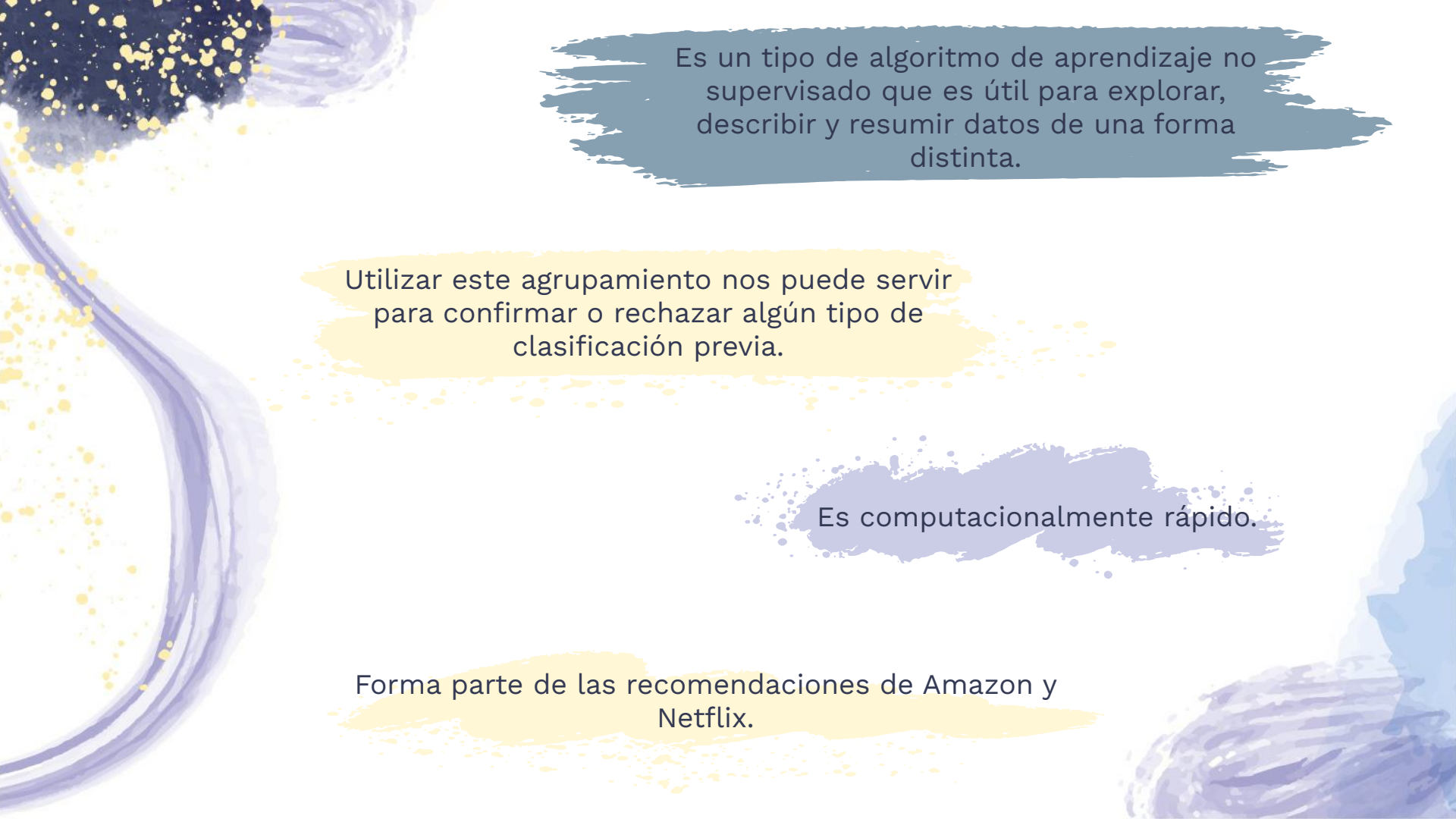




04

# Conclusión

--



Es un tipo de algoritmo de aprendizaje no supervisado que es útil para explorar, describir y resumir datos de una forma distinta.

Utilizar este agrupamiento nos puede servir para confirmar o rechazar algún tipo de clasificación previa.

Es computacionalmente rápido.

Forma parte de las recomendaciones de Amazon y Netflix.



# Referencias

- Anuradha Bhatia. K-Mean Clustering. (13 de mayo de 2017). Accedido el 6 de junio de 2022. [Video en línea]. Disponible: <https://www.youtube.com/watch?v=wt-X61BnUCA>
- Y. S.Thakare y S. B. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", *International Journal of Computer Applications*, vol. 110, n.º 11, pp. 12–16, enero de 2015. Accedido el 6 de junio de 2022. [En línea]. Disponible: <https://doi.org/10.5120/19360-0929>
- K. S. Kadam and S. B. Bagal, —Fuzzy Hyperline Segment Neural Network Pattern Classifier with Different Distance MetricsII, *International Journal of Computer Applications* 95(8):6-11, June 2014.



**THANKS!**