


# K-means

# Introducción

**Clustering** es una técnica que tiene como objetivo **organizar patrones en grupos**, de modo que los patrones que pertenecen al mismo grupo son lo suficientemente similares como para inferir que son del mismo tipo y los patrones que pertenecen a diferentes grupos son lo suficientemente diferentes como para inferir que son de otra clase.

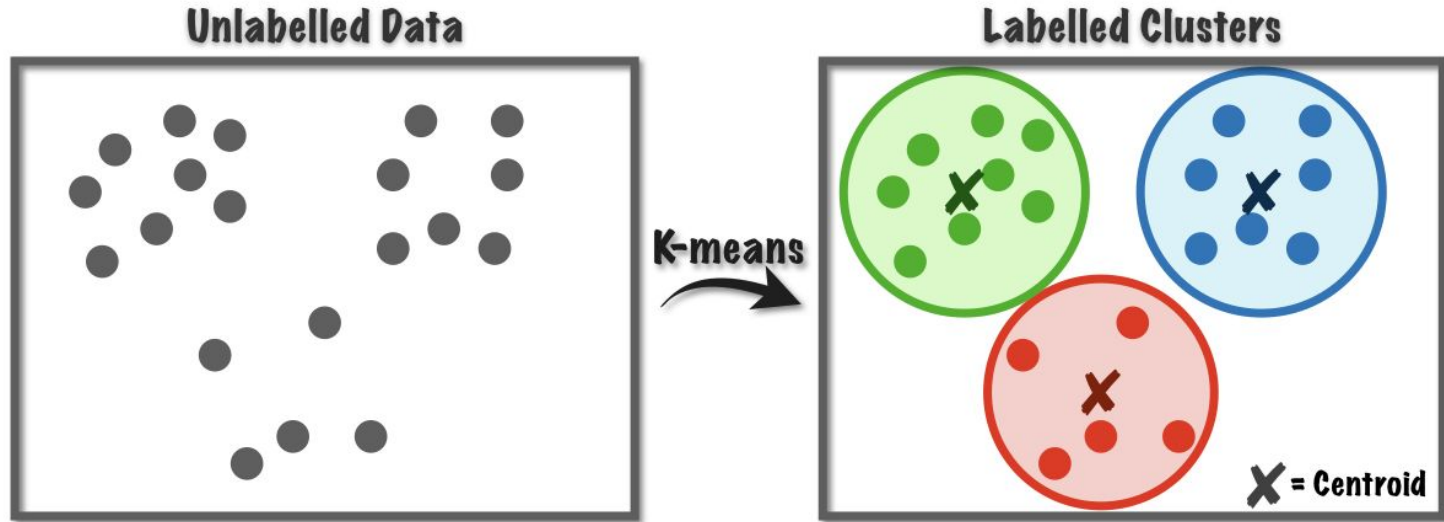


# Aplicaciones

- **Retail/ Marketing:**
    - Identificar patrones de compra de los consumidores.
    - Recomendar nuevos libros o películas a nuevos clientes.
  - **Banca:**
    - Identificar grupos de tipo de consumidores.
  - **Seguros:**
    - Detección de fraude en reclamos
    - Riesgo de seguro de los clientes
  - **Publicaciones científicas:**
    - Categorizar nuevos artículos basado en su contenido.
  - **Medicina:**
    - Caracterizar pacientes a través de su comportamiento
  - **Biología:**
    - Agrupación de marcadores genéticos para identificar familias.
- 

# K-Means

El algoritmo K-means se basa en particiones, lo que significa que cada cluster se encuentra representado por un centroide.



# Introducción

Dentro de los inconvenientes de este algoritmo es que tenemos que seleccionar el número de clusters ( $k$ ). Por lo tanto una selección inadecuada  $K$  puede llevarnos malos resultados.

Para evaluar el rendimiento del K-means tenemos el método **elbow**.

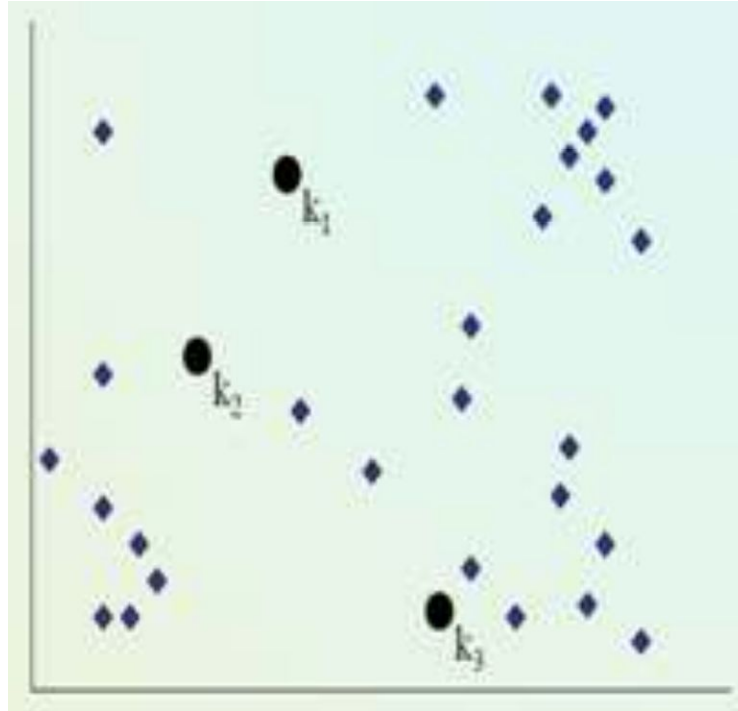


# Algoritmo K-means

1. **Elegir k centroides, y posicionarlos en el conjunto de datos en un lugar aleatorio.**
2. Calcular la distancia entre cada patrón desde los centroides.
3. Asignar a cada patrón al centroide más cercano, la clase del centroide.
4. Una vez que todos los puntos o patrones fueron asignados, recalcular la posición de los centroides a partir del centroide de los patrones asignados a cada centroide.
5. Repetir los pasos 2 al 4 hasta que los centroides se mantengan en la misma posición, llegar a un número máx. de interacciones o aceptar una tolerancia definida.



1. Elegir  $k$  centroides y posicionarlos en el conjunto de datos en un lugar aleatorio.



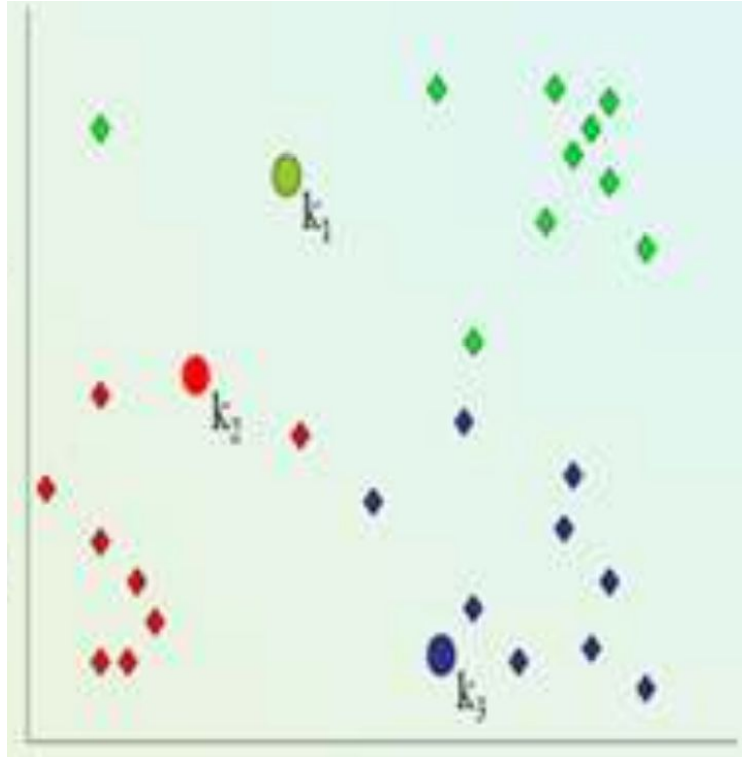
# Algoritmo K-means

1. Elegir  $k$  centroides, y posicionarlos en el conjunto de datos en un lugar aleatorio.
2. **Calcular la distancia entre cada patrón desde los centroides.**
3. **Asignar a cada patrón al centroide más cercano, la clase del centroide.**
4. Una vez que todos los puntos o patrones fueron asignados, recalcular la posición de los centroides a partir del centroide de los patrones asignados a cada centroide.
5. Repetir los pasos 2 al 4 hasta que los centroides se mantengan en la misma posición, llegar a un número máx. de interacciones o aceptar una tolerancia definida.





2-3. Calcular la distancia entre cada patrón y los centroides y asignar a cada patrón la clase del centroide más cercano

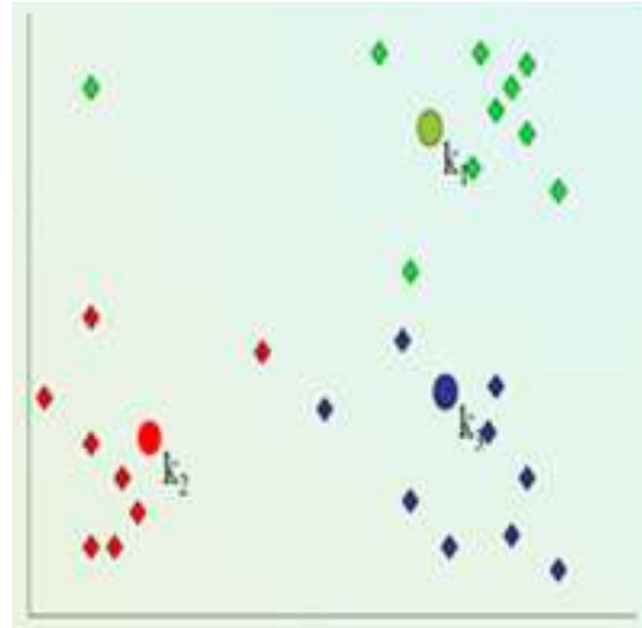
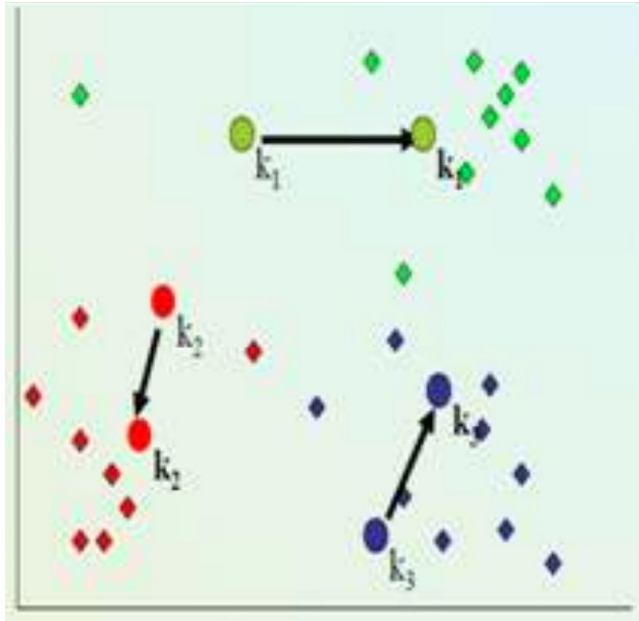


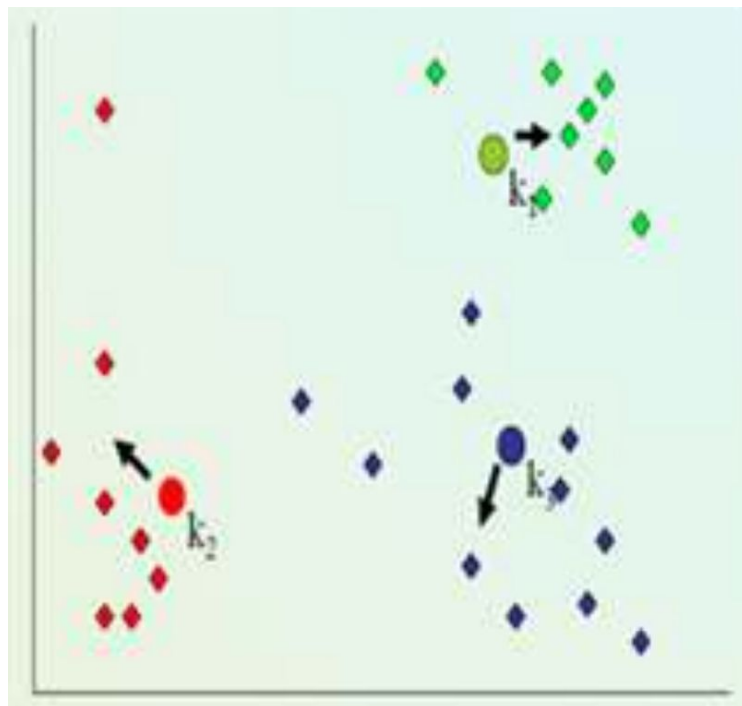
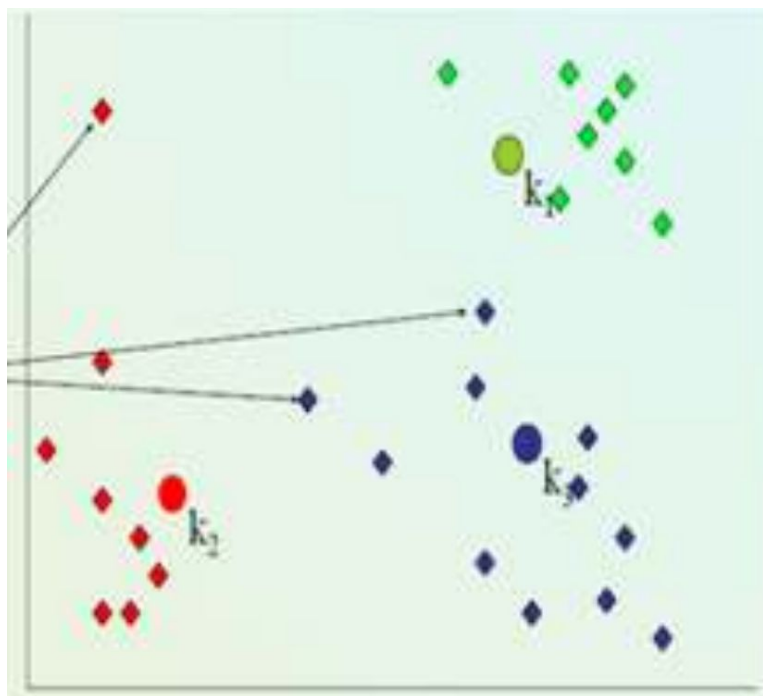
# Algoritmo K-means

1. Elegir  $k$  centroides, y posicionarlos en el conjunto de datos en un lugar aleatorio.
2. Calcular la distancia entre cada patrón desde los centroides.
3. Asignar a cada patrón al centroide más cercano, la clase del centroide.
4. **Una vez que todos los puntos o patrones fueron asignados, recalcular la posición de los centroides a partir del centroide de los patrones asignados a cada centroide.**
5. Repetir los pasos 2 al 4 hasta que los centroides se mantengan en la misma posición, llegar a un número máx. de interacciones o aceptar una tolerancia definida.



4. Una vez que todos los puntos o patrones fueron asignados, recalcular la posición de los centroides a partir del centroide de los patrones asignados a cada centroide.





# Introducción

- Cuando aplicamos k-means a datos reales es importante transformar los atributos del dataset a una escala min-max u otra, para que todos los atributos se encuentren en la misma escala
- En el caso de K-means los clusters nunca se van a traslapar (mezclar)



# Inercia del cluster

La inercia del cluster está dada por la siguiente ecuación:

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \|\mathbf{x}^i - \mu^j\|^2$$

Donde  $\mu$  es el centroide para el grupo  $j$  y  $w$  es igual a 1 si la muestra  $\mathbf{x}$  está en el grupo  $j$ ; sino es igual a 0.

- A través de la inercia podemos medir la semejanza entre patrones de un cluster.



# Elección de la $K$

Dentro de los inconvenientes de este algoritmo es que tenemos que seleccionar el número de clusters ( $k$ ). Por lo tanto una selección inadecuada  $K$  puede llevarnos malos resultados.

Para evaluar el rendimiento del K-means tenemos el método **elbow**.

