

Píldora del módulo 2: Clustering

Cristian Yepes e Ignacio Castillo

Hoja de Ruta

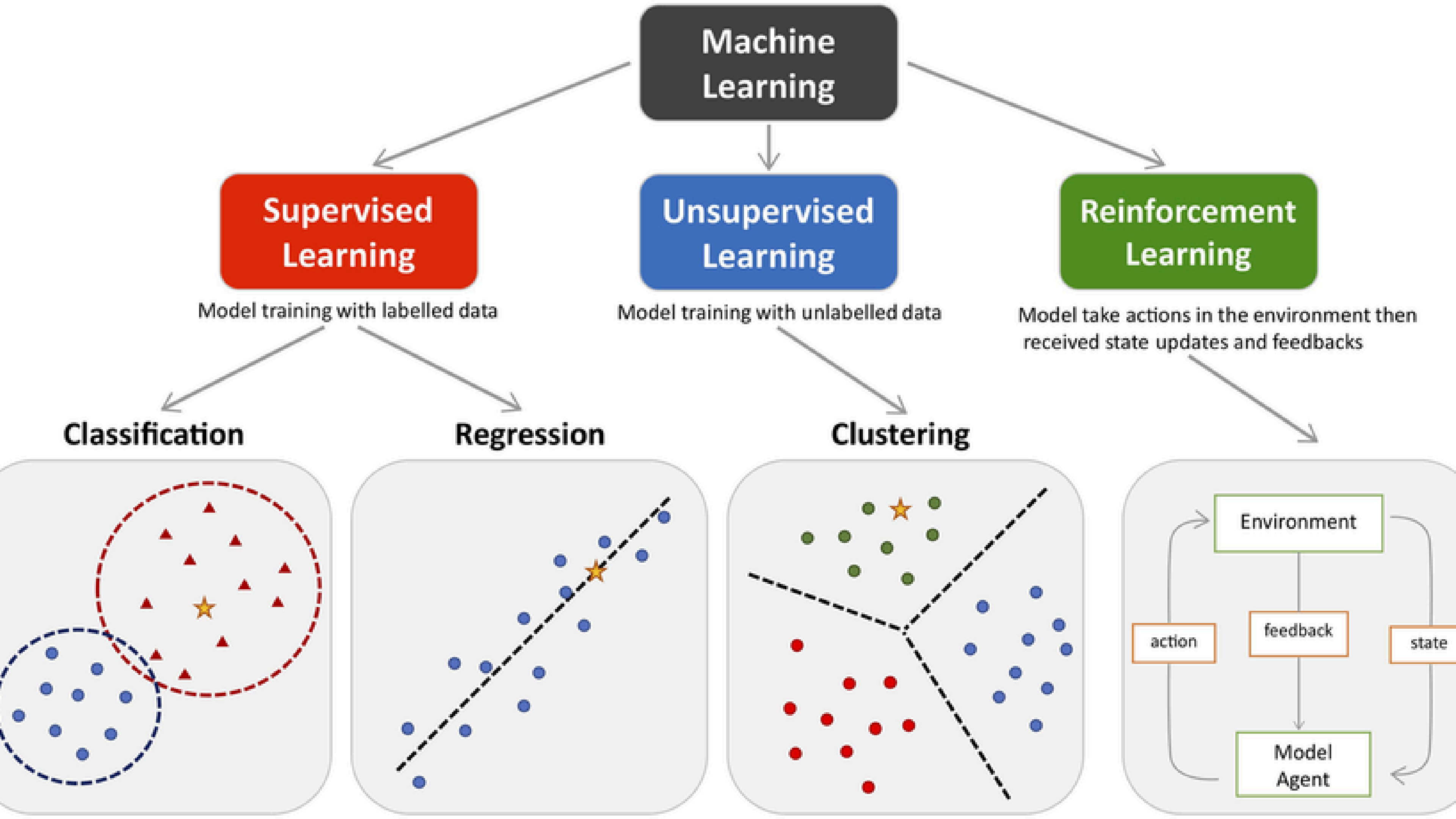
Ubicación de
Clustering dentro del
panorama del ML



comprendiendo el
Clustering a través de
un ejemplo
(participativo)

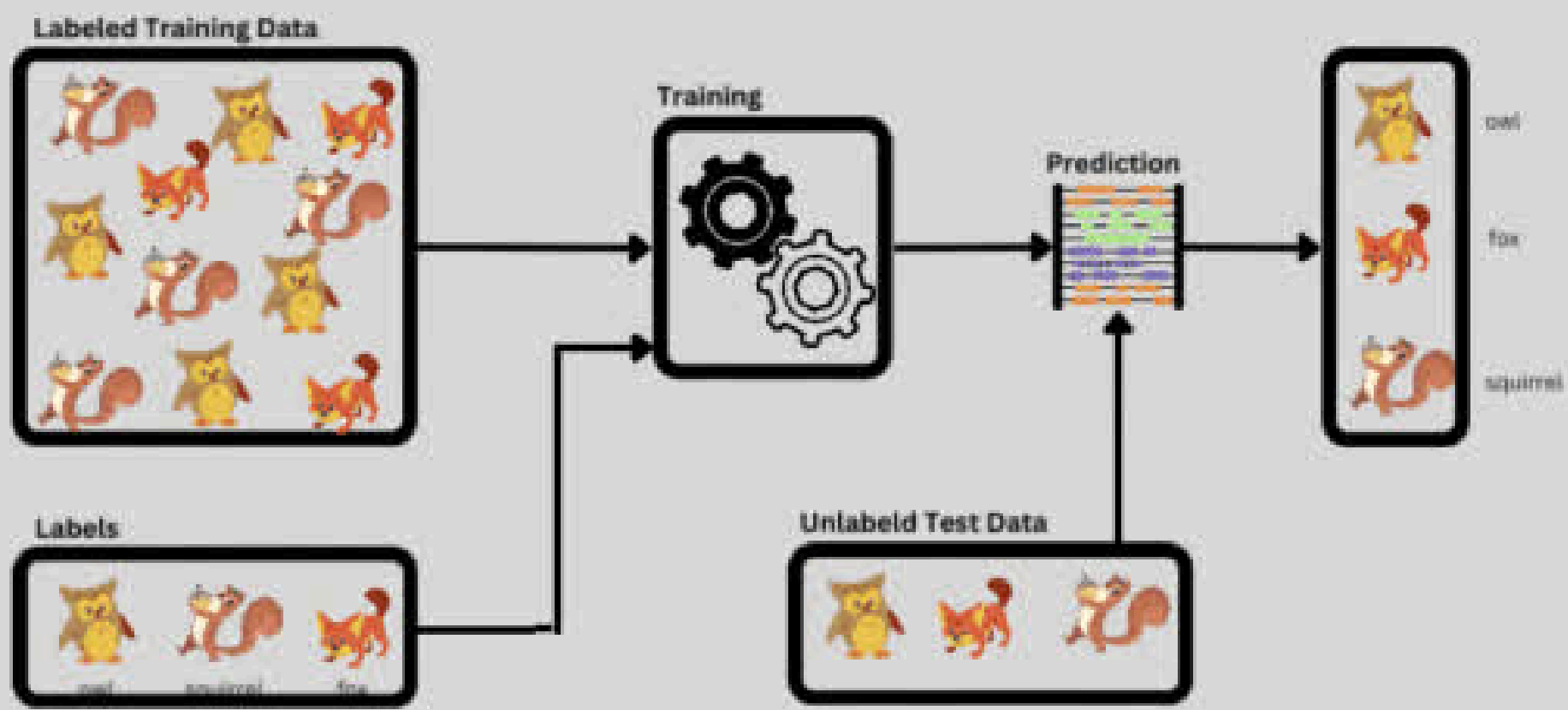


¿cómo evaluamos un
modelo Clustering?

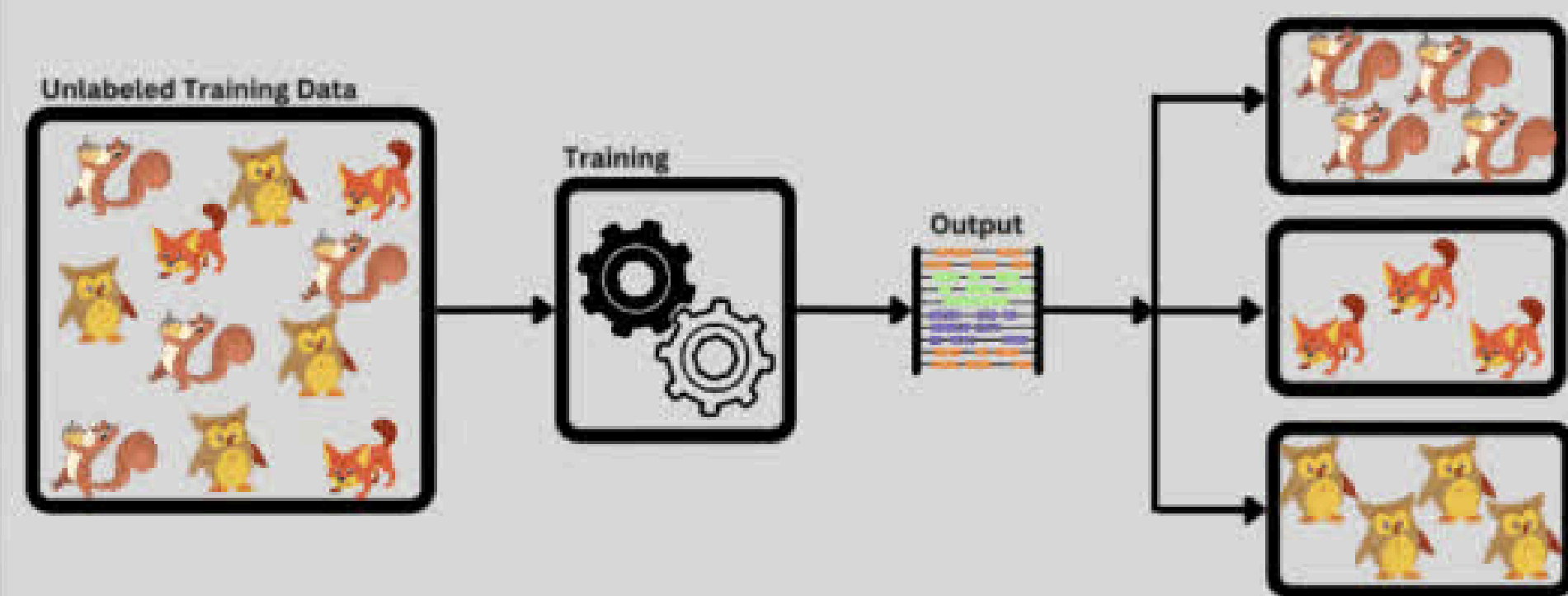


Machine Learning

Supervised Learning



Unsupervised Learning



Datamapu

La principal distinción entre ambos enfoques radica en el **uso de datos etiquetados** (el aprendizaje supervisado utiliza datos etiquetados -requieren intervención humana previa para etiquetarlos-, mientras que un algoritmo de aprendizaje no supervisado no)

En el aprendizaje supervisado, el algoritmo "aprende" del conjunto de datos de entrenamiento realizando predicciones iterativas sobre los datos y ajustándolos para obtener la respuesta correcta.

Los modelos de aprendizaje no supervisado, en cambio, trabajan por sí solos para descubrir la estructura inherente de los datos no etiquetados. **En el aprendizaje supervisado, el objetivo es predecir los resultados de los nuevos datos.** Con un algoritmo de **aprendizaje no supervisado, el objetivo es obtener información valiosa de grandes volúmenes de datos nuevos.** El propio aprendizaje automático determina qué es diferente o interesante del conjunto de datos. El aprendizaje supervisado es un método simple, que generalmente se calcula mediante programas como R o Python. En el aprendizaje no supervisado, se necesitan herramientas potentes para trabajar con grandes cantidades de datos sin clasificar. Los modelos de aprendizaje no supervisado son computacionalmente complejos porque requieren un gran conjunto de entrenamiento para producir los resultados esperados.

Unsupervised Machine Learning

Clustering

- Groups unlabeled data
- Types: Exclusive (K-Means), Overlapping, Hierarchical, Probabilistic (GMM)

Association Rules

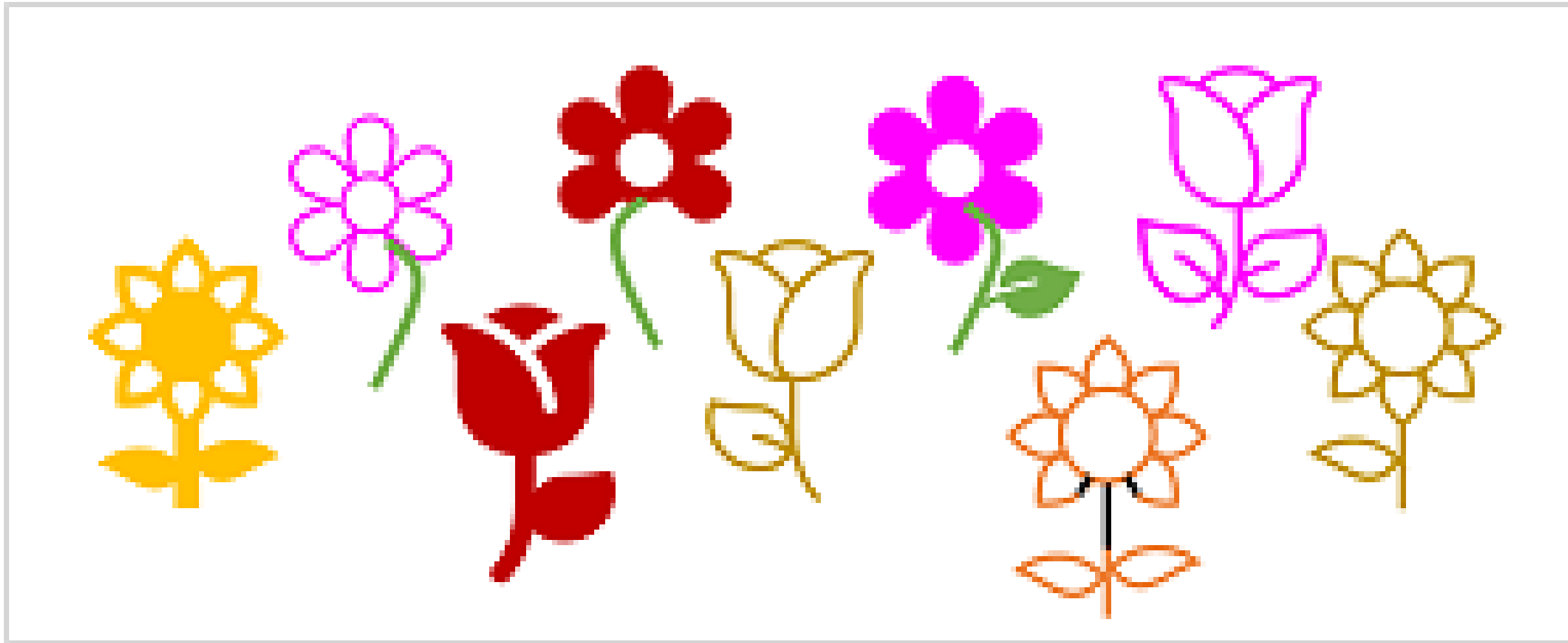
- Finds variable relationships
- Used in recommendation engines
- Algorithms: Apriori, Eclat, FP-Growth

Dimensionality Reduction

- used when the number of features, or dimensions, in a given dataset is too high
- Methods: PCA, SVD, Autoencoders


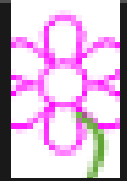






Clustering (agrupación) es una forma de aprendizaje automático no supervisado en la que las observaciones se agrupan en clústeres según las similitudes en sus características. Este tipo de aprendizaje automático se considera no supervisado porque no utiliza etiquetas previamente conocidas para entrenar un modelo sino que la etiqueta es el clúster al que se asigna la observación, basándose únicamente en sus características.

Por ejemplo, supongamos que un botánico observa una muestra de flores y registra el número de pétalos de cada flor:

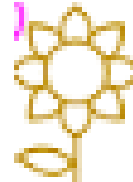
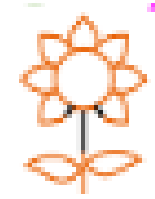
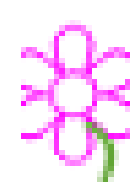
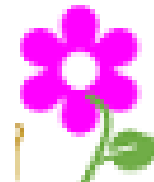
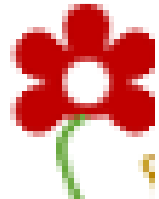


No hay etiquetas conocidas en el conjunto de datos, solo **UNA CARACTERÍSTICA**. El objetivo no es identificar los diferentes tipos (especies) de flores, sino agrupar flores similares según el número de pétalos.



Petal Count		Petals (x_2)
		5
		6
		3
		3
		6
		8
		3
		7
		8
		8

K - means



0

1

2

3

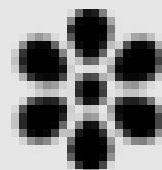
4

5

6

7





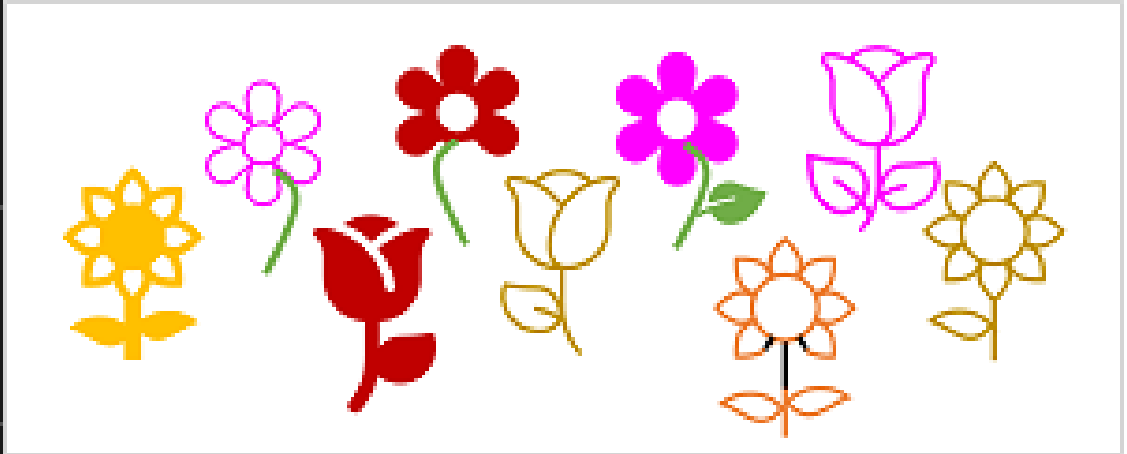
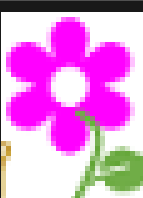
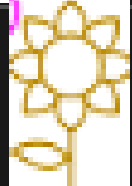

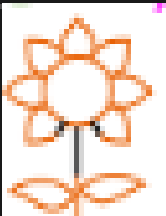

8



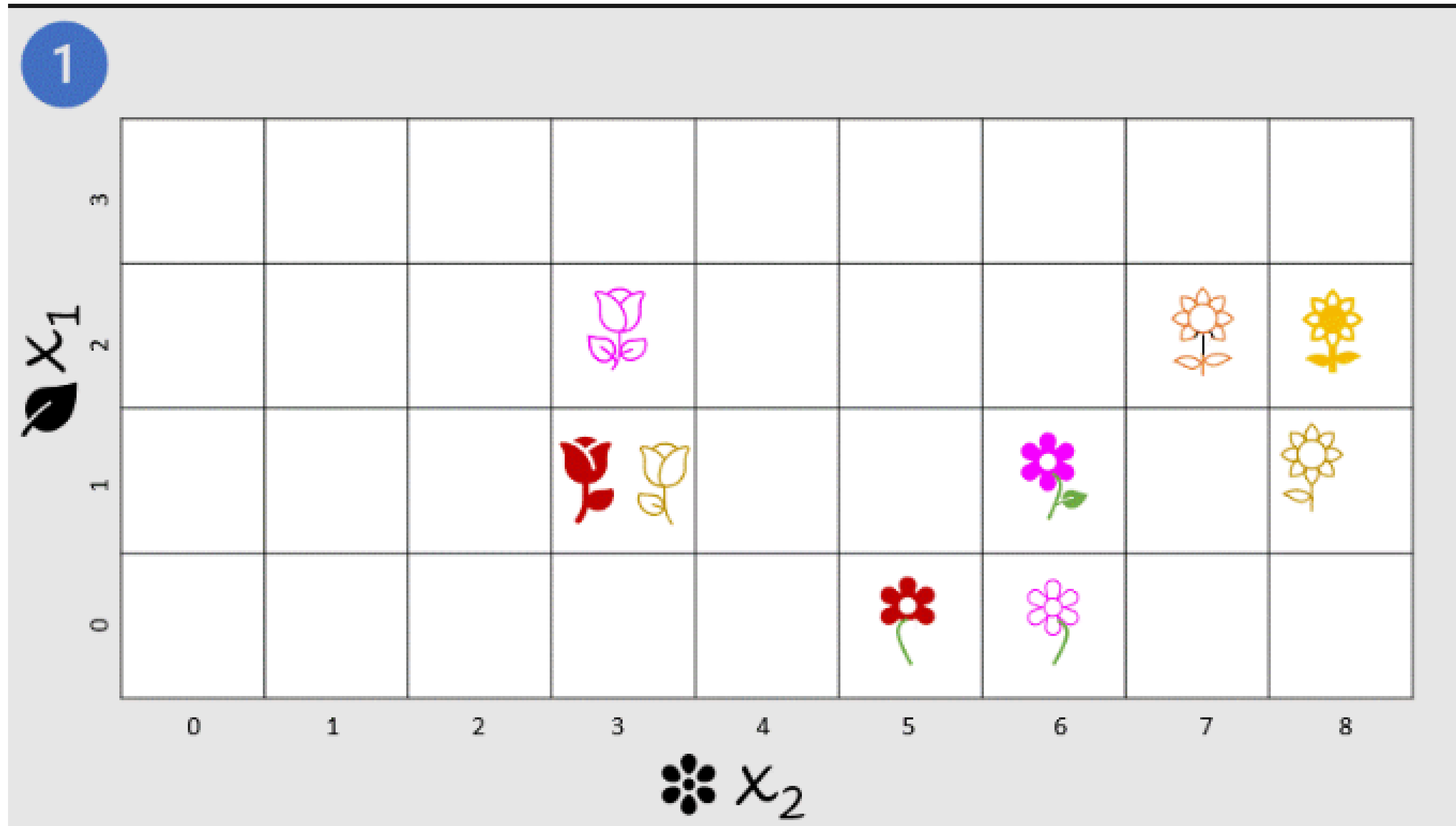
X

Entrenamiento de un modelo de Clustering

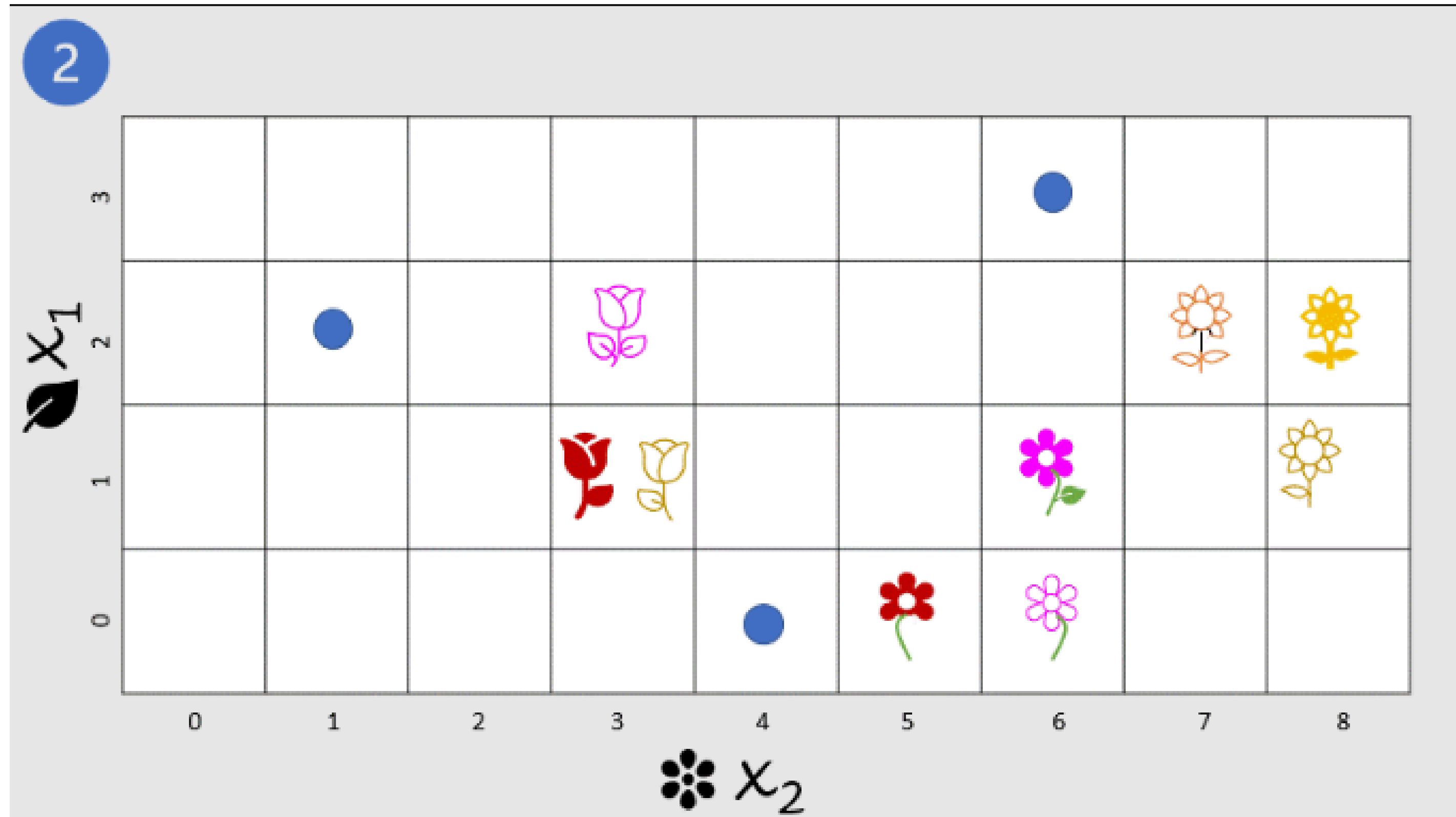
Existen múltiples algoritmos que se pueden utilizar para el agrupamiento. Uno de los más utilizados es el agrupamiento **K-Means**, que consta de los siguientes pasos:

Leaves (x_1)		Petals (x_2)	
0		5	
0		6	
1		3	
1		3	
1		6	
1		8	
2		3	
2		7	
2		8	

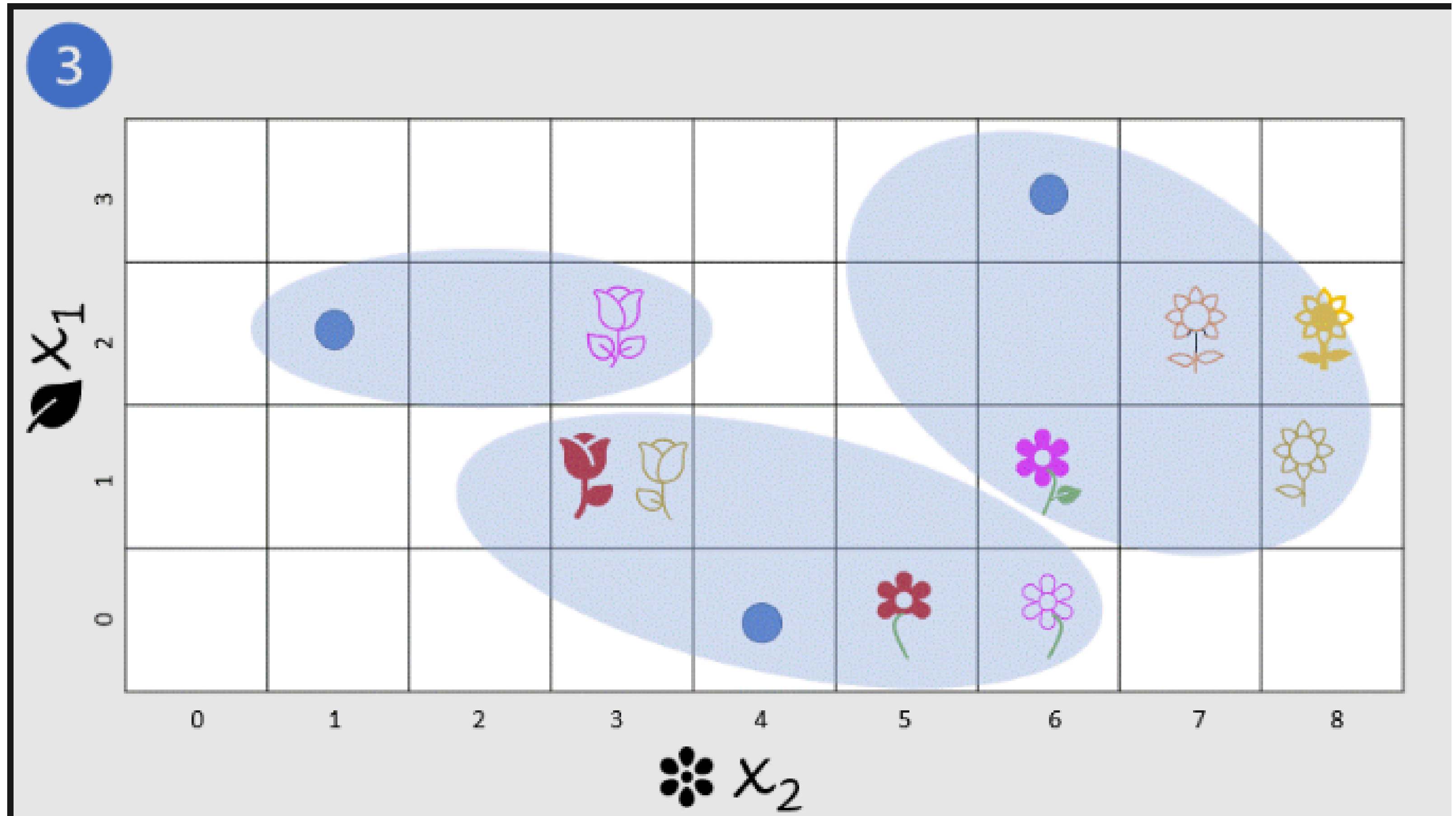
1. Los valores de la característica (x) se vectorizan para definir coordenadas n-dimensionales (donde n es el número de características). En el ejemplo de la flor, tenemos dos características: número de hojas (x1) y número de pétalos (x2). Por lo tanto, el **vector de características** tiene dos coordenadas que podemos usar para representar conceptualmente los puntos **de datos** en un espacio bidimensional ([x1,x2]).



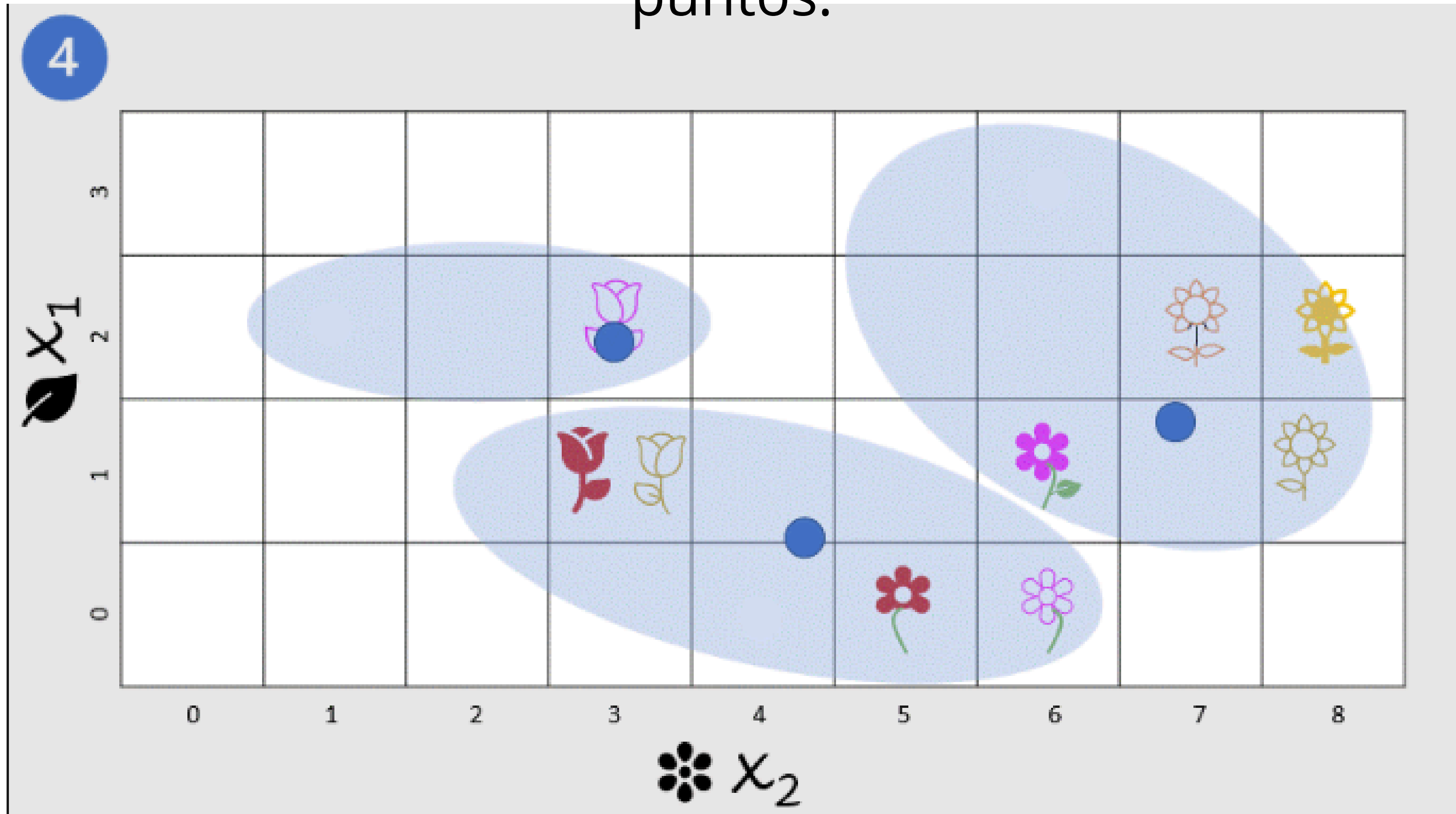
2. **Decide** cuántos grupos quieres usar para agrupar las flores (**hiperparámetro K**. Por ejemplo, para crear tres grupos, usarías un valor $k=3$). Luego, se trazan **k puntos en coordenadas aleatorias**. Estos puntos se convierten en los puntos centrales de cada grupo, por lo que se llaman **centroides**.



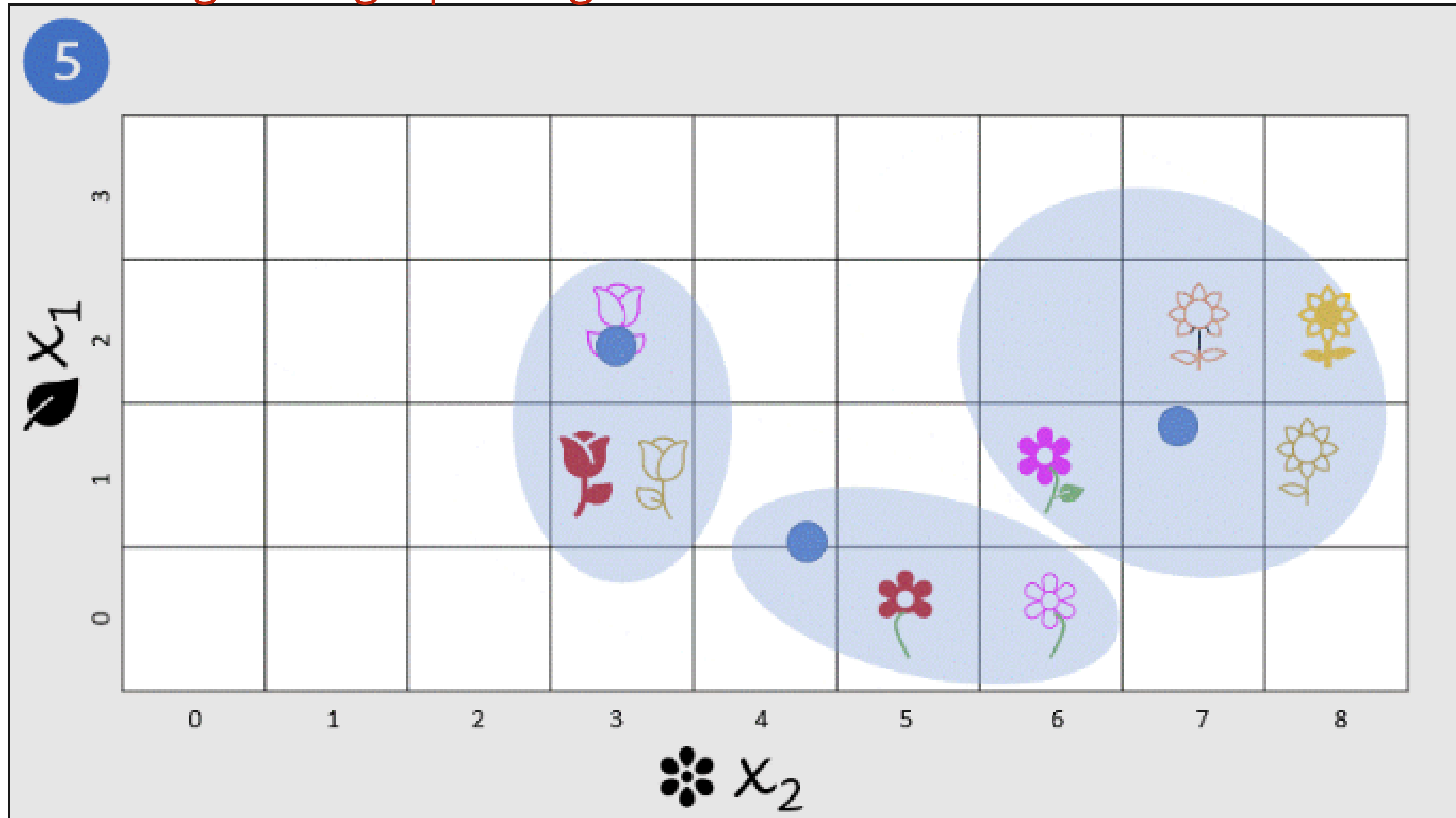
3. A cada punto de datos (en este caso una flor) se le **asigna su centroide más cercano.**



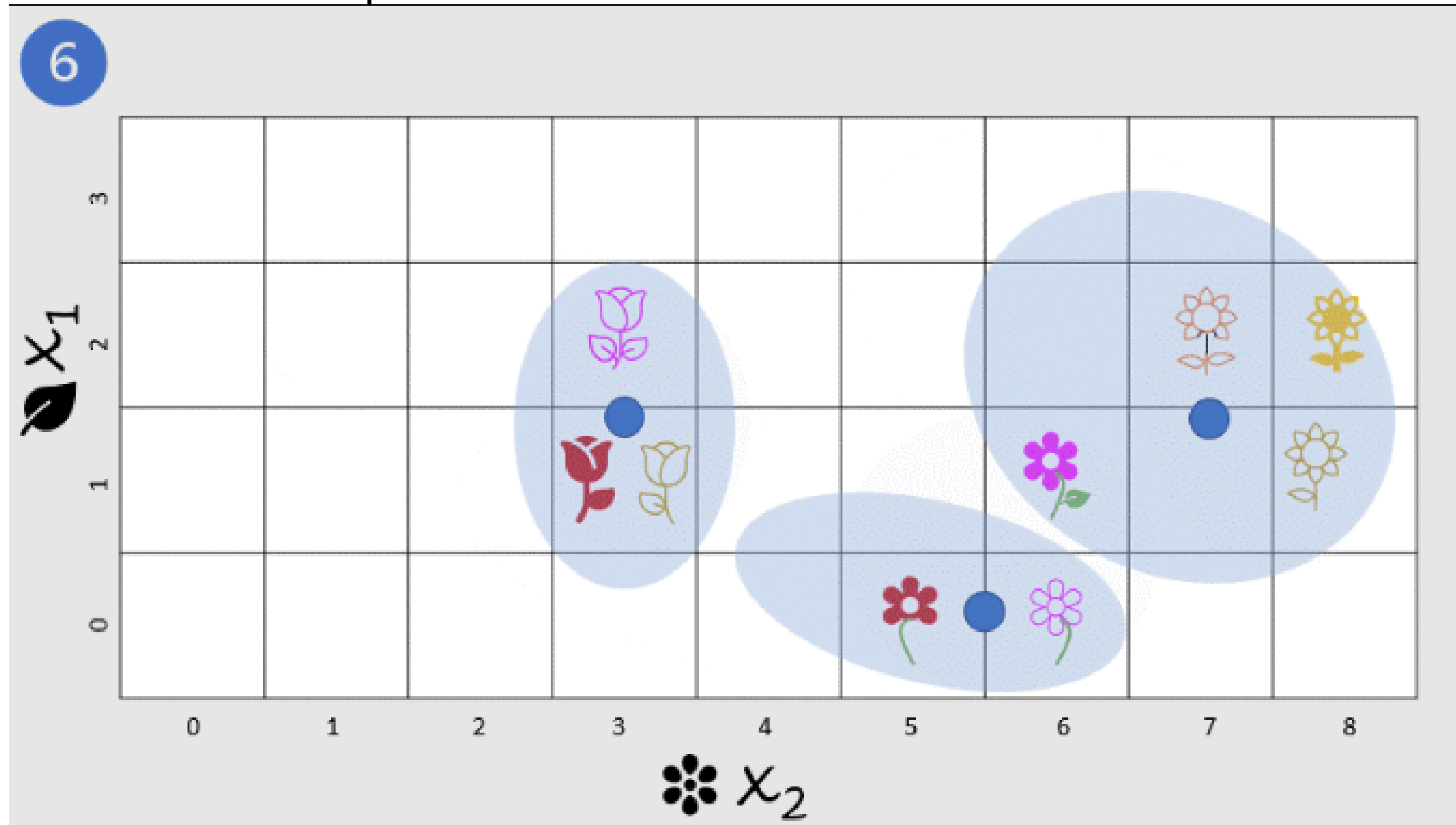
4. Cada **centroide** se mueve al **centro de los puntos de datos asignados a él** en función de la distancia media entre dichos puntos.



5. Después de mover el centroide, los puntos de datos pueden estar ahora más cerca de un centroide diferente, por lo que los puntos de datos **se reasignan a grupos según el nuevo centroide más cercano.**



6. Los pasos de movimiento del centroide y reasignación de grupos **se repiten** hasta que los grupos **se vuelven estables** o se alcanza un número máximo predeterminado de iteraciones.



Evaluación de un modelo de agrupamiento

Dado que no existe una etiqueta conocida para comparar las asignaciones de agrupamiento previstas, la evaluación de un modelo de agrupamiento se basa en la separación de los agrupamientos resultantes entre sí. Existen múltiples métricas que se pueden utilizar para evaluar la separación de agrupamientos, entre ellas:

- **Distancia promedio al centro del agrupamiento:** Cuán cerca, en promedio, está cada punto del agrupamiento del centroide del mismo.
- **Distancia promedio al otro centro:** Cuán cerca, en promedio, está cada punto del agrupamiento del centroide de todos los demás agrupamientos.
- **Distancia máxima al centro del agrupamiento:** La distancia más lejana entre un punto del agrupamiento y su centroide.
- **Silüeta:** Un valor entre -1 y 1 que resume la relación entre la distancia entre los puntos del mismo agrupamiento y los puntos de agrupamientos diferentes (cuanto más cercano a 1, mejor es la separación de agrupamientos).