



Programa de formación  
**MACHINE LEARNING  
AND DATA SCIENCE MLDS**

Facultad de  
**INGENIERÍA**





# Módulo 2

# Introducción al Machine

# Learning con *Python*

Unidad 4

Aprendizaje no supervisado:  
agrupamiento

Clase sincrónica

Facultad de  
**INGENIERÍA**





## Bienvenida

# Fabio Augusto Gonzalez, PhD.

<https://dis.unal.edu.co/~fgonza/>

[fagonzalezo@unal.edu.co](mailto:fagonzalezo@unal.edu.co)



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

Departamento de Ingeniería de Sistemas e Industrial  
Facultad de Ingeniería  
Universidad Nacional de Colombia  
Sede Bogotá



## Tabla de contenidos

- 1 Aprendizaje no supervisado
- 2 Agrupamiento
  - Agrupamiento particional
  - Agrupamiento jerárquico
- 3 Algoritmo de agrupamiento K-means
- 4 Evaluación del desempeño
  - Inercia
  - Determinación del número de grupos
  - Coeficiente de silueta
  - Evaluación externa
- 5 Agrupamiento jerárquico

## Objetivos de aprendizaje



## Unidad 4 - Aprendizaje no supervisado: agrupamiento

Al finalizar la unidad usted deberá ser capaz de:

 1

Conocer los fundamentos del algoritmo de agrupamiento *k-means*.

 2

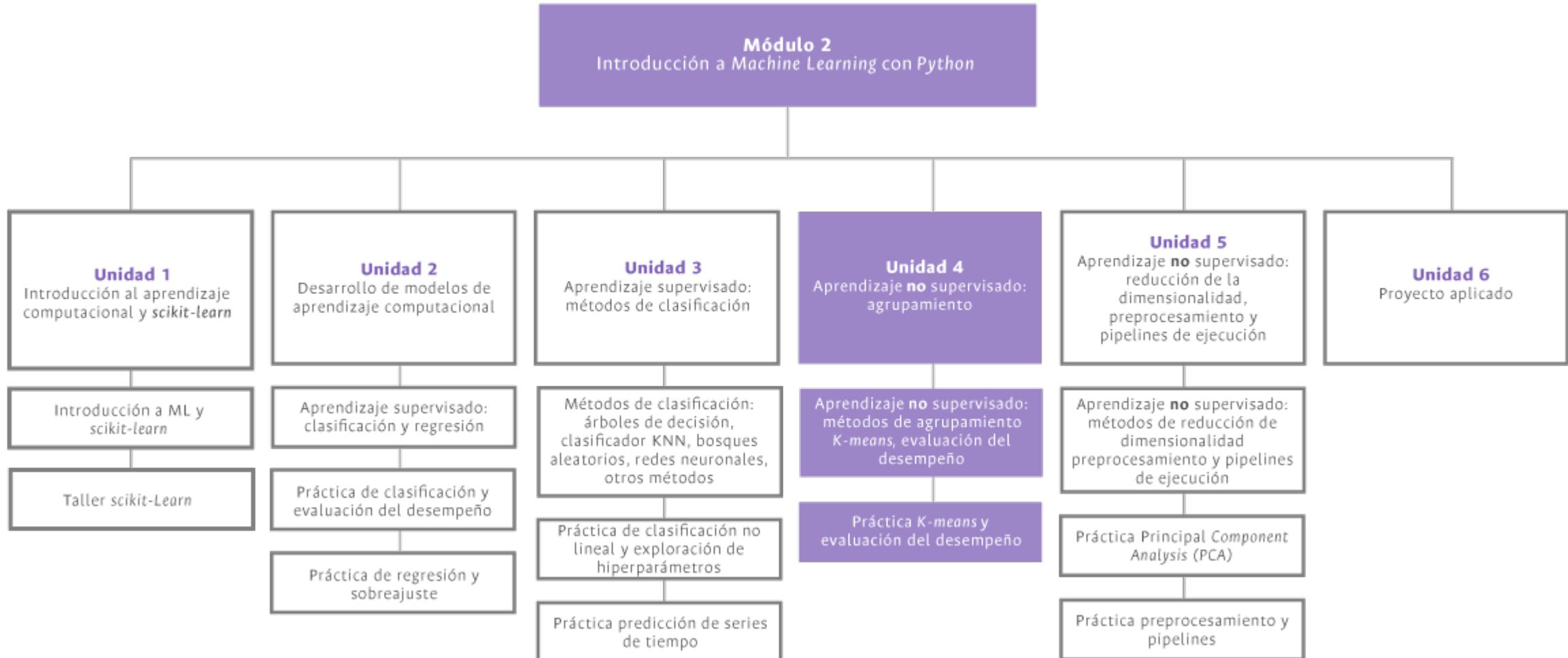
Implementar modelos de agrupamiento con ayuda de la librería *scikit-learn*.

 3

Evaluar modelos de agrupamiento mediante el uso de diferentes métricas de desempeño.

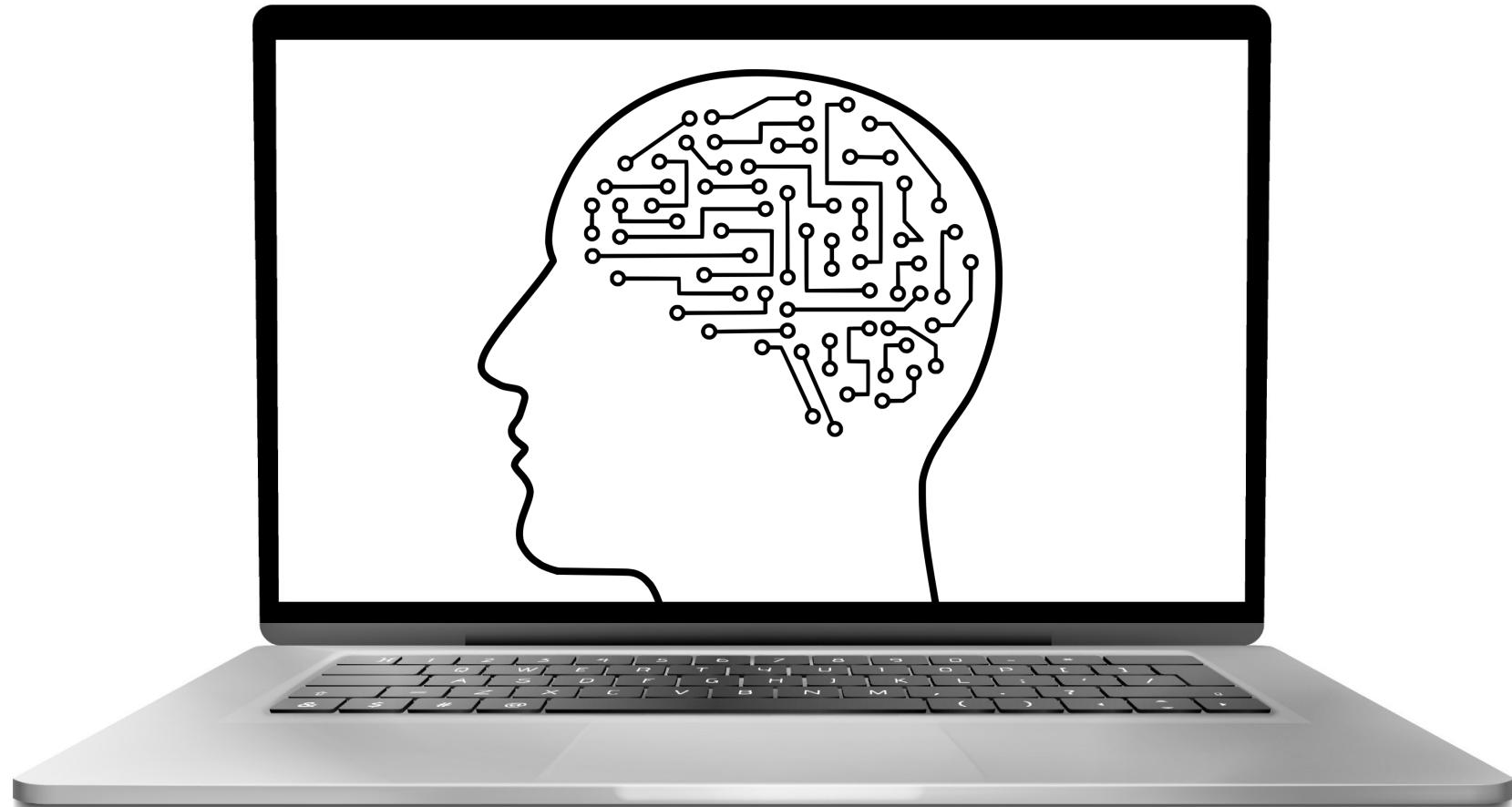


# Mapa de contenidos de la unidad



1

# Aprendizaje no Supervisado



## Aprendizaje no Supervisado



El aprendizaje no supervisado busca aprender propiedades o patrones de un conjunto de datos. En contraste con el aprendizaje supervisado, donde se cuenta con una etiqueta u objetivo en los ejemplos de entrenamiento, en el aprendizaje no supervisado no se cuenta con esta información. A continuación, podemos ver algunas de las tareas del aprendizaje no supervisado:

Agrupamiento	Encontrar grupos de ejemplos con características similares
Reducción de la dimensionalidad	Reducir el tamaño de los datos sin perder mucha información
Detección de anomalías	Se analizan los datos para encontrar ejemplos inusuales o fraudulentos
Modelos Generativos	Se estima la distribución de probabilidad generadora de los datos para crear nuevos ejemplos

2

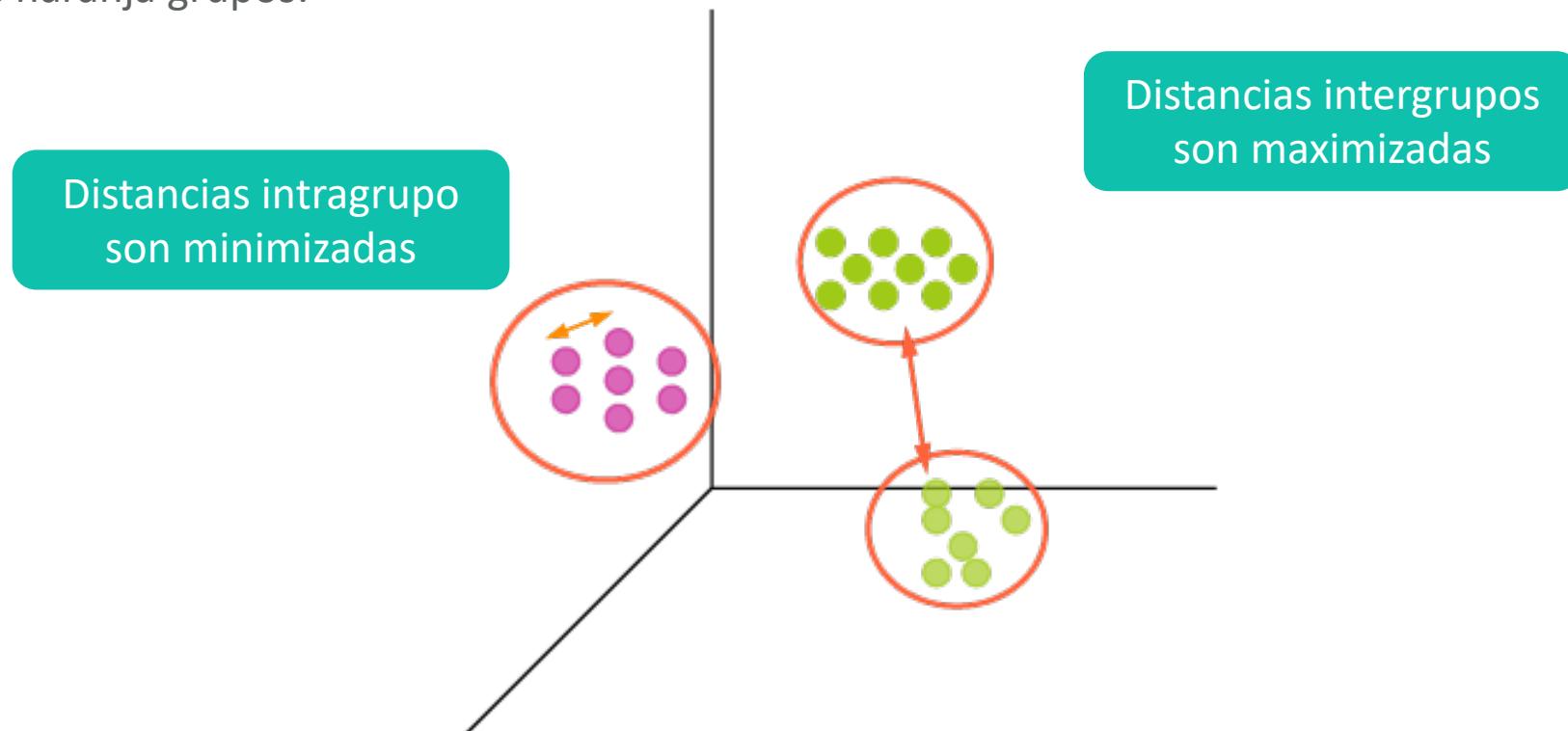
## Agrupamiento



## Agrupamiento

El **análisis de grupos** o agrupamiento busca encontrar grupos de objetos de tal forma que estos objetos sean similares o relacionados entre sí y diferentes a los objetos de otros grupos.

La siguiente figura ilustra un ejemplo de agrupamiento, donde los puntos representan ejemplos y los círculos naranja grupos:



## Agrupamiento

## Aplicaciones

### Segmentación de mercado

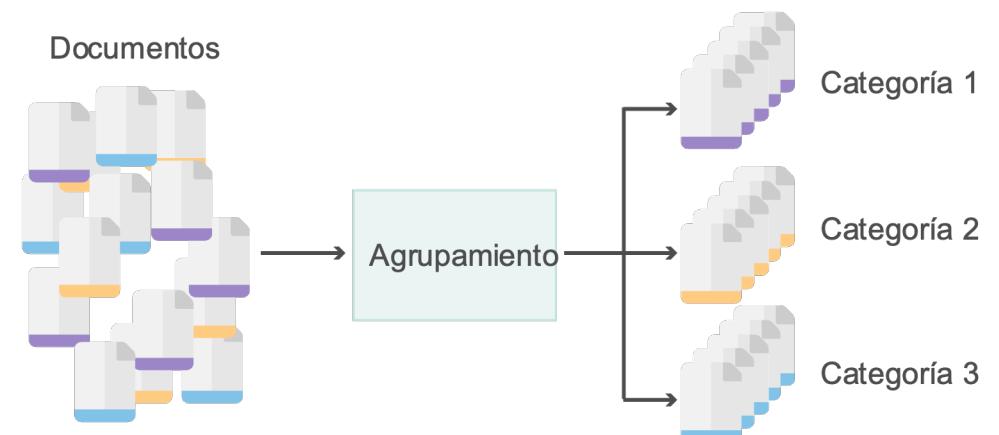
La segmentación de mercado consiste en formar grupos de tipos de potenciales clientes. La segmentación de mercado juega un rol crítico en el desarrollo de estrategias de mercado.



Fuente de figura: Subpng (s.f.)

### Agrupación de documentos

La agrupación de documentos consiste en formar grupos de documentos con temas similares. De acuerdo con Zhao & Karypis (2011) es una parte esencial de la minería de textos con aplicaciones en la recuperación de la información y manejo del conocimiento.



Fuente de figura: adaptado de IBM® Knowledge Center (s.f.)

Agrupamiento

Aplicaciones

## Segmentación de imágenes

La segmentación de imágenes consiste en formar grupos de píxeles adyacentes con contenido similar; por ejemplo, en la siguiente figura los píxeles se ven segmentados en 3 grupos que pueden ser interpretados como cuerpo de agua, vegetación y terreno. De acuerdo con Shapiro & Stockman (2001) el objetivo de segmentar es simplificar y/o cambiar la representación de una imagen en algo más significativo y fácil de analizar.

Imagen original



Imagen segmentada  
en 3 grupos



## Agrupamiento

 Tipos de Agrupamiento

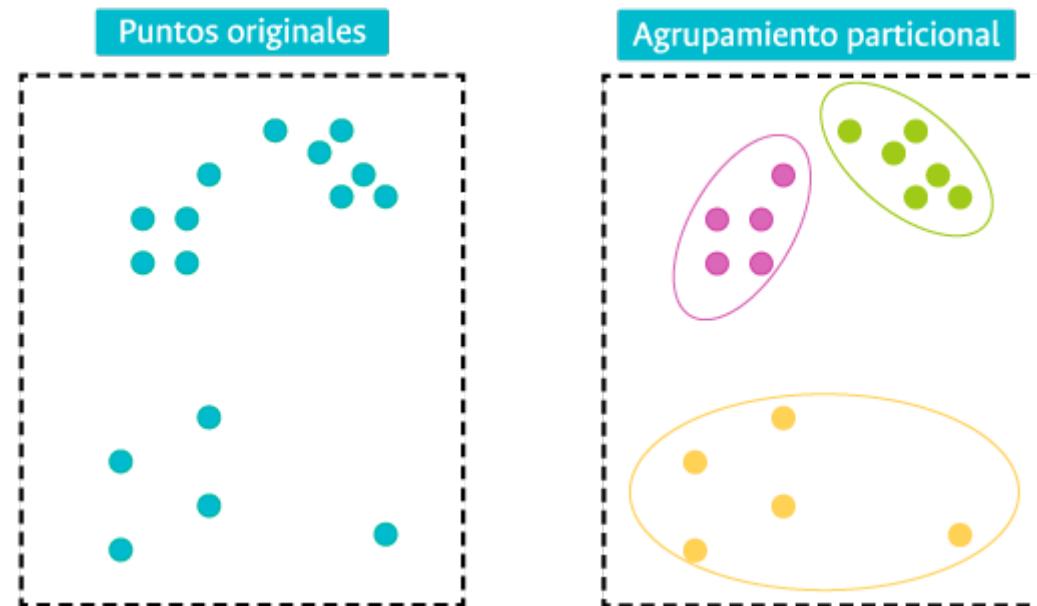
Un grupo es un conjunto de ejemplos y un agrupamiento es un conjunto de grupos. Dependiendo de la forma como se distribuyen los ejemplos en los grupos, los agrupamientos se pueden dividir en dos categorías:

**1****Agrupamiento particional****2****Agrupamiento jerárquico**

## Agrupamiento

## Agrupamiento particional

En un agrupamiento particional los elementos se dividen en subconjuntos (grupos) que no se sobreponen, de tal forma que cada elemento está exactamente en un subconjunto

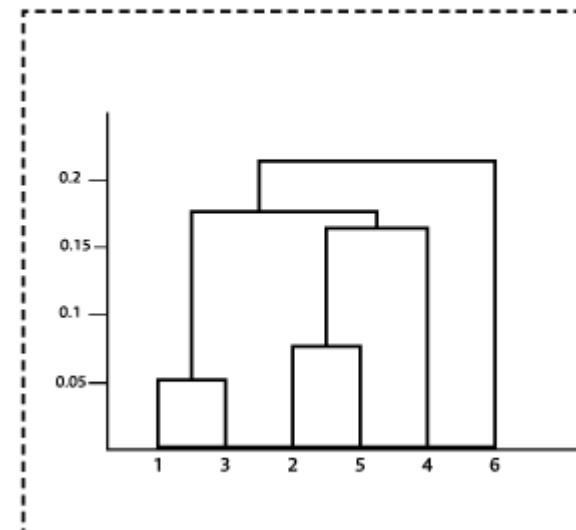


## Agrupamiento

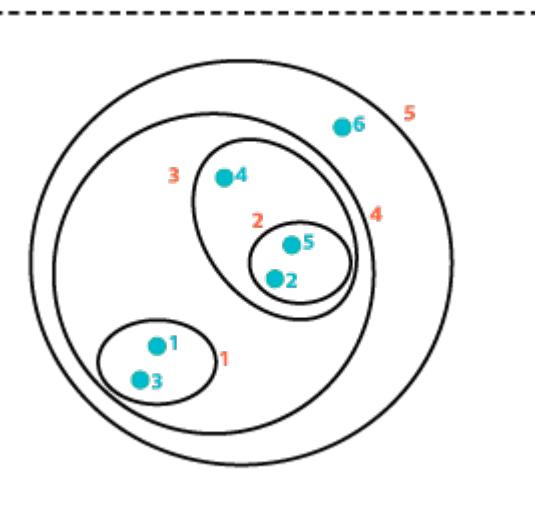
## Agrupamiento jerárquico

- Produce un conjunto de grupos anidados organizados como una jerarquía.
- Se puede visualizar como un dendrograma, que se define como un diagrama en forma de árbol que registra las secuencias de fusiones

Dendrograma

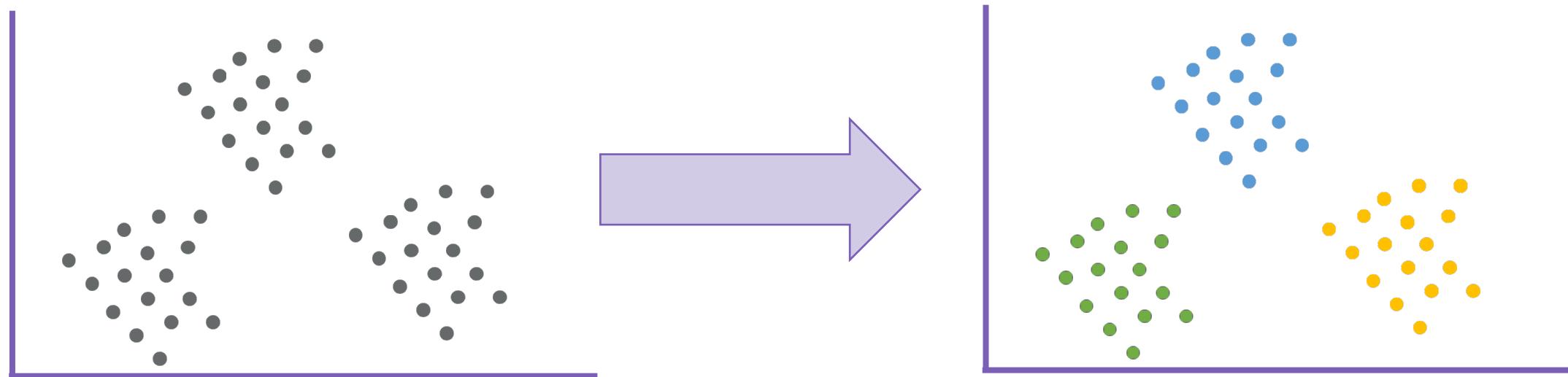


Agrupamiento jerárquico



3

## Algoritmo de Agrupamiento *K-means*



## Algoritmo de Agrupamiento *K-means*



El algoritmo *K-means* o K-medias es un método de agrupamiento particional muy popular y uno de los más sencillos.

### Ideas básicas

- Cada grupo es asociado con un **centroide**.
- Cada punto es asignado al grupo con el centroide más cercano.
- El número de grupos **K** es un hiperparámetro que debe especificarse.

## Algoritmo de Agrupamiento *K-means*



### Algoritmo

1. seleccionar  $K$  puntos como los centroides iniciales
2. Repetir
  - 2.1 formar  $K$  grupos al asignar todos los puntos al centroide más cercano
  2. 2 recalcular el centroide de cada grupo
3. Hasta que: los centroides no cambian

La clase `sklearn.cluster.KMeans` implementa este algoritmo

Algoritmo de Agrupamiento *K-means*

Animación (Video)



Algoritmo de Agrupamiento *K-means*

## Centroides

- El centroide de un grupo es la media de los puntos del grupo.
- En general los centroides no son ejemplos del conjunto de datos, aunque pertenezcan al mismo espacio.
- El algoritmo es sensible a la selección de centroides iniciales, por lo que ejecutar el algoritmo en el mismo conjunto de datos con una inicialización distinta puede producir grupos distintos.
- Normalmente la inicialización se hace de manera aleatoria.
- La implementación de *scikit-learn* ofrece por defecto la inicialización *k-means++*, la cual escoge de manera inteligente los centroides iniciales para acelerar la convergencia; *k-means++* asegura que los centroides de una inicialización aleatoria no estén muy cerca uno del otro.

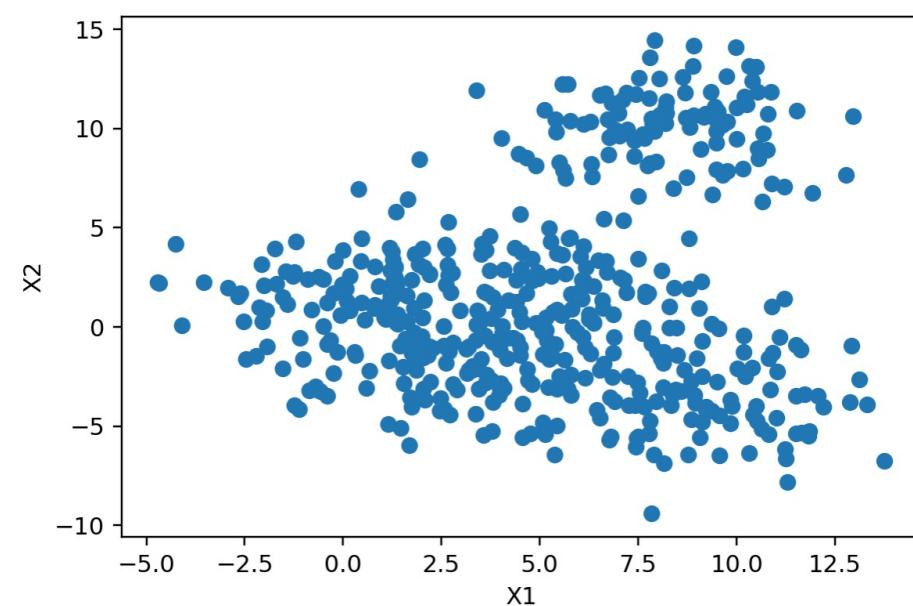
Algoritmo de Agrupamiento *K-means*

## Medida de cercanía

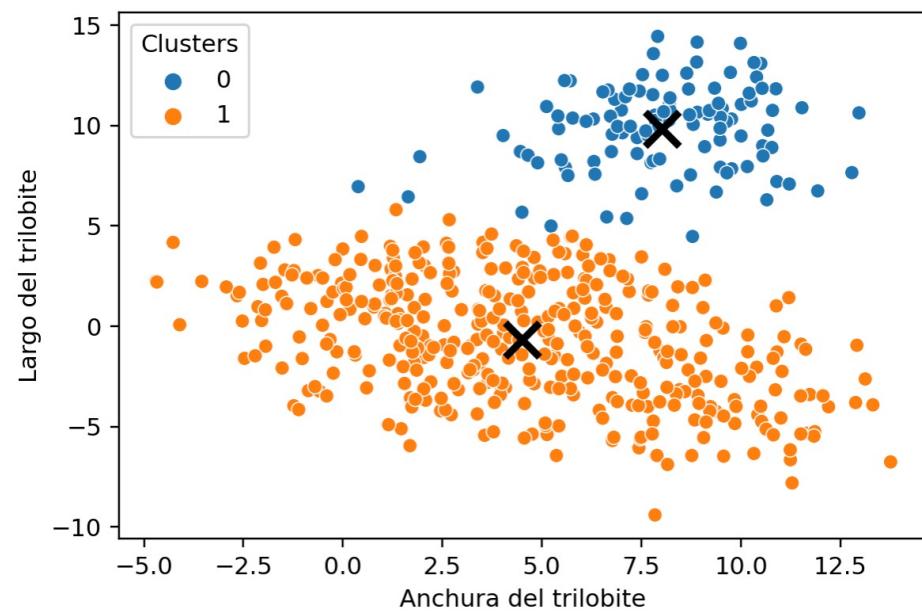
- Generalmente la cercanía se mide con la distancia euclíadiana, aunque se pueden usar otras métricas.
- La implementación de *K-means* en *scikit-learn* solo soporta la distancia euclidean.
- *K-means* siempre converge usando la distancia euclíadiana, pero el criterio de convergencia puede ser reemplazado por “hasta que relativamente pocos puntos cambien de grupo”.

Algoritmo de Agrupamiento *K-means*

## K-means en scikit learn

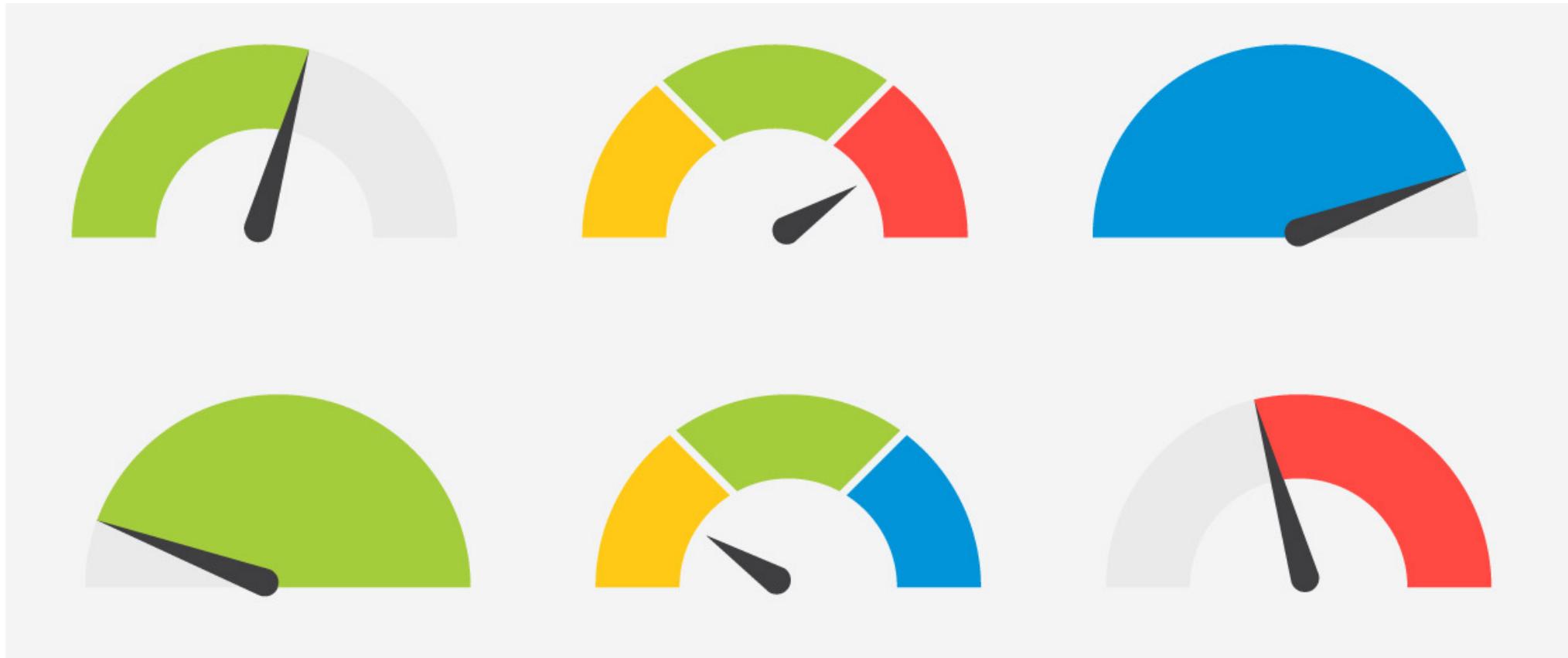


```
1 n = 2
2
3 km = KMeans(n_clusters = n)
4 km.fit(X_cluster)
5
6 y = km.predict(X_cluster)
```





## Evaluación del desempeño



## Evaluación del desempeño



- La evaluación del desempeño de un agrupamiento se puede realizar con o sin datos adicionales.
- Dependiendo del uso o no de datos adicionales, las medidas de evaluación se pueden clasificar en:
  - **Evaluación interna:** se usa información intrínseca para determinar que tan compactos y/o separados son los grupos.
  - **Evaluación externa:** se usa información externa como etiquetas para determinar qué tan homogéneos son los grupos resultantes.

## Evaluación del desempeño



## Inercia

La **inercia** o la **suma de errores cuadráticos intragrupo** corresponde a la suma de las distancias al cuadrado de cada ejemplo al centroide más cercano. La inercia es una medida de desempeño interna. Intuitivamente, la inercia estima que tan compactos son los grupos de un agrupamiento.

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

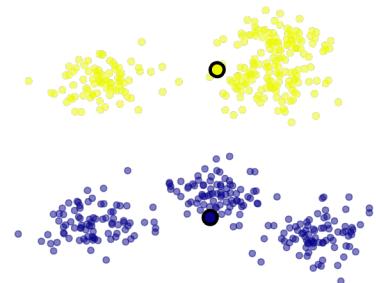
El algoritmo *K-means* está diseñado para minimizar esta medida.

## Evaluación del desempeño

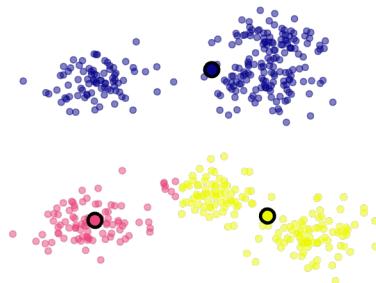


A continuación, podemos observar varios agrupamientos y su respectiva inercia.

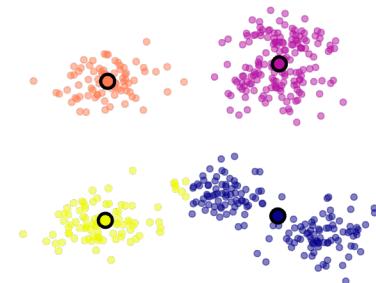
k=2, Inercia=14885.417



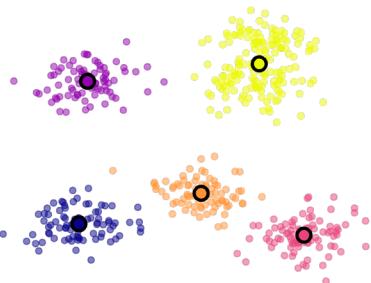
k=3, Inercia=9701.574



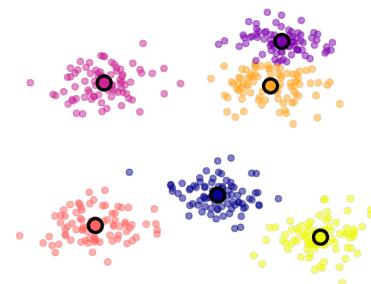
k=4, Inercia=4639.264



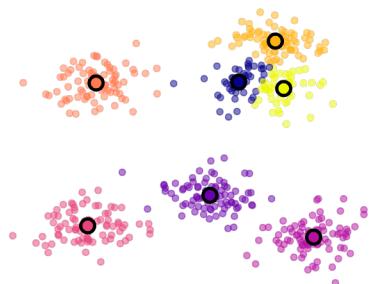
k=5, Inercia=2710.072



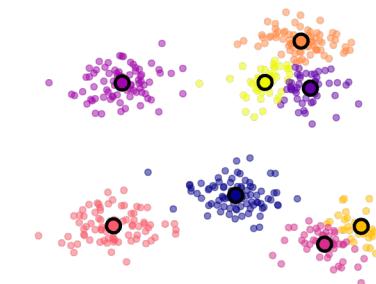
k=6, Inercia=1929.003



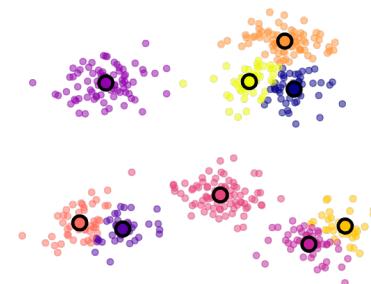
k=7, Inercia=1785.968



k=8, Inercia=1644.362



k=9, Inercia=1523.512



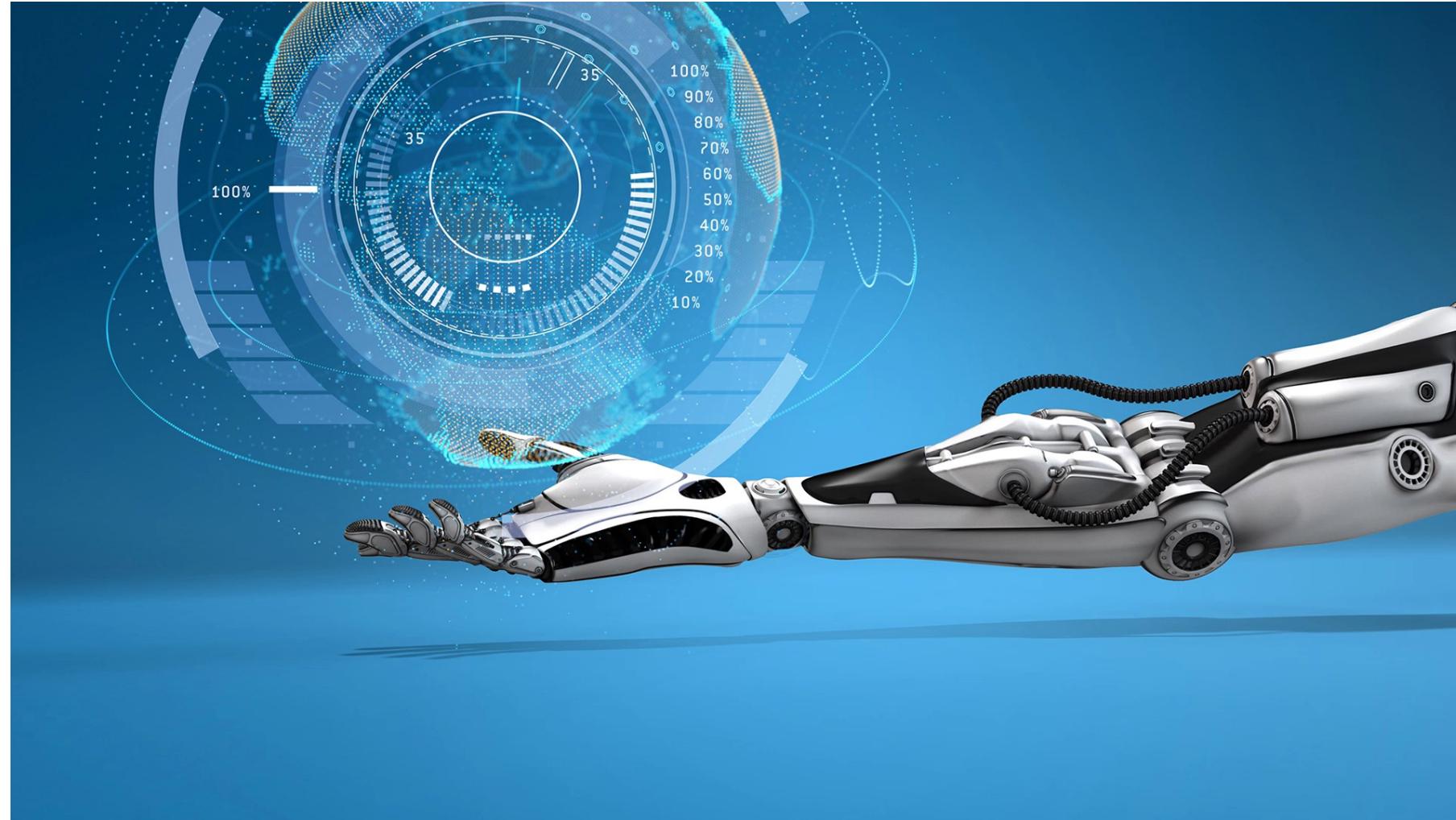
## Evaluación del desempeño

 Determinación del número de grupos

- **El método del codo** es una heurística para seleccionar el hiperparámetro **K**.
- En este método se ejecuta *K-means* con varios valores, por ejemplo, entre 2 y 10.
- Luego, se grafican los valores de K contra los respectivos valores de inercia.
- El valor de K óptimo, de acuerdo con esta heurística, es el codo de la gráfica (el punto de donde al aumentar el valor de *K* no hay una reducción significativa de la inercia).
- La intuición detrás de esta estrategia es que agregar más grupos no mejora sustancialmente el desempeño del agrupamiento, de forma similar al sobreajuste.
- En la práctica, puede que el “codo” no sea muy prominente.

Evaluación del desempeño

## Determinación del número de grupos (Video)



## Evaluación del desempeño

 Coeficiente de Silueta

El coeficiente de silueta es una medida de desempeño interna definida como una métrica compuesta para cada ejemplo, de la siguiente manera:

$$s = \frac{b - a}{\max(a, b)}$$

Donde:

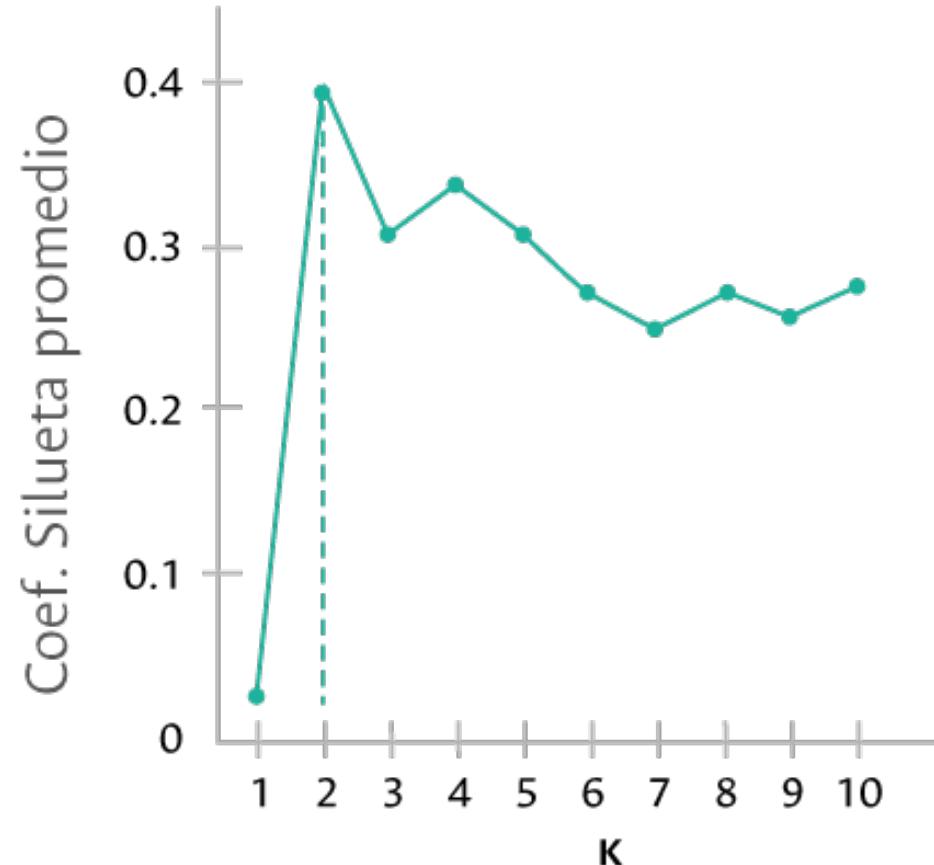
- **a**: la distancia media entre el ejemplo y el resto de puntos en su mismo grupo.
- **b**: la distancia media entre el ejemplo y el resto de puntos en el siguiente grupo más cercano

Para un conjunto de datos se calcula el coeficiente de silueta promedio; este cálculo puede ser, a nivel computacional, bastante costoso.

En esencia, el coeficiente de silueta es una diferencia de distancias normalizada, en el cual se captura que tan cerca está un ejemplo a su propio grupo comparado con el *siguiente grupo más cercano*.

Una ventaja de del coeficiente de silueta es que sus valores están acotados entre -1 y 1. El valor óptimo es 1, pero en la práctica se buscan valores que estén por encima de 0.

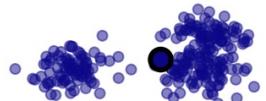
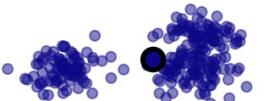
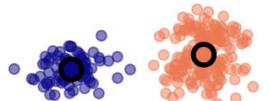
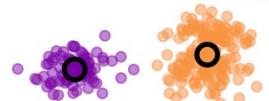
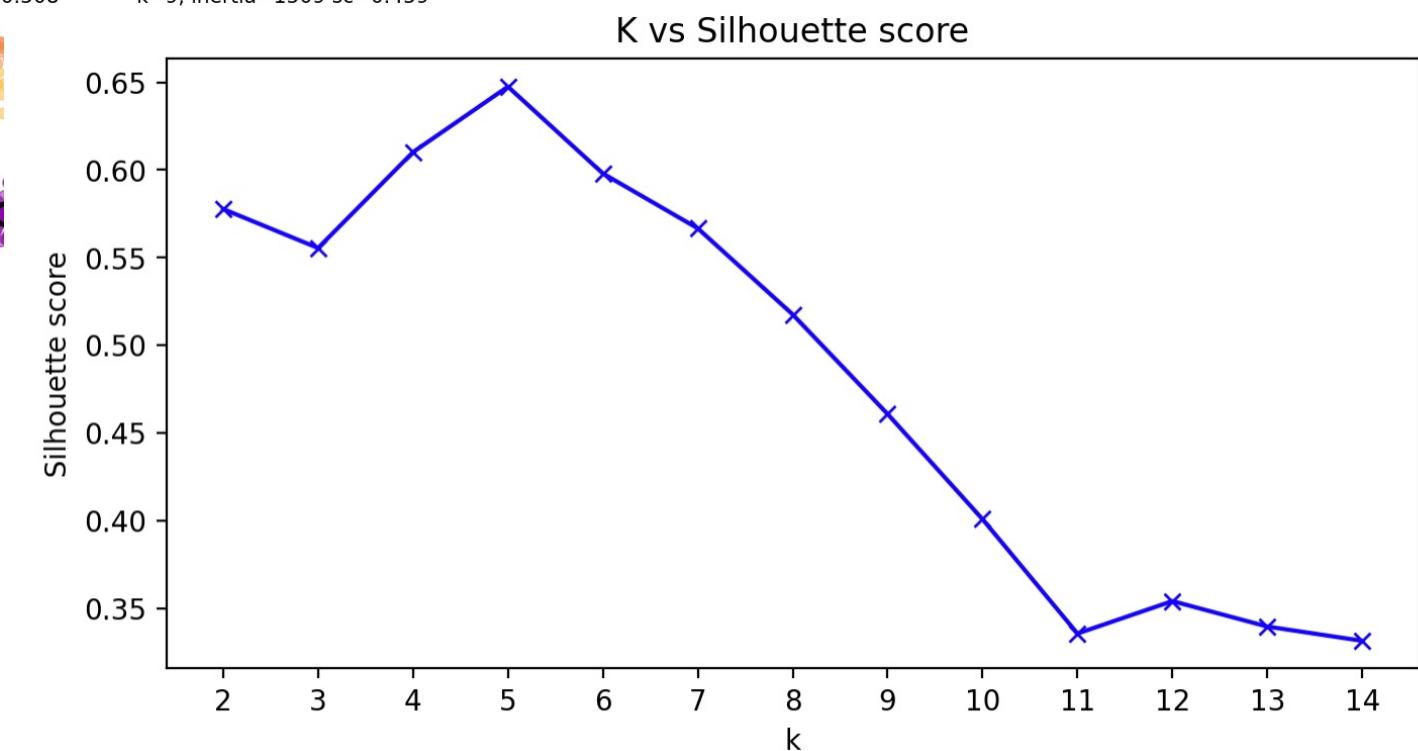
## Evaluación del desempeño

 Coeficiente de Silueta

En la figura se muestra el coeficiente de silueta para varios valores de K. El agrupamiento que es seleccionado es aquel con mayor coeficiente de silueta.

## Evaluación del desempeño

## Coeficiente de Silueta

 $k=2$ , inertia=14885 sc=0.578 $k=3$ , inertia=9706 sc=0.554 $k=4$ , inertia=4639 sc=0.610 $k=5$ , inertia=2710 sc=0.648 $k=6$ , inertia=1929 sc=0.598 $k=7$ , inertia=1786 sc=0.566 $k=8$ , inertia=1650 sc=0.508 $k=9$ , inertia=1509 sc=0.459

## Evaluación del desempeño



### Evaluación externa

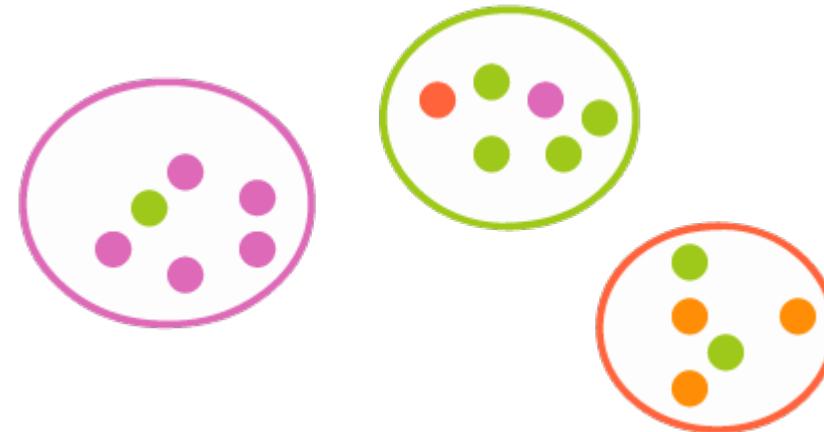
En la evaluación externa se usan datos adicionales que no estaban disponibles durante el entrenamiento al algoritmo de agrupamiento. Por ejemplo, las categorías reales de los ejemplos.

#### Homogeneidad

La homogeneidad o pureza es una métrica de evaluación externa. Para calcular esta pureza, cada grupo es asociado con la clase mayoritaria; luego, La métrica se evalúa contando el número de ejemplos clasificados correctamente y dividiendo por N (Manning, Raghavan & Schütze, 2008).

Otros ejemplos de métricas de evaluación externa son la métrica F (F1-score), el índice de Rand (Rand Index) y métricas basadas en la información mutua.

## Evaluación del desempeño

 Ejemplo cálculo de la homogeneidad o pureza

Los colores indican la clase de los ejemplos. Esta información no se tiene durante la aplicación del algoritmos de agrupamiento.

Para cada grupo se calcula la clase mayoritaria. En el diagrama se muestra con el color correspondiente.

Se cuentan el número de elementos en cada grupo que pertenecen a la clase mayoritaria: Grupo 1 (violeta): 5, Grupo 2 (verde): 4, Grupo 3 (naranja): 3.

Se suman estos valores y se dividen por el total de elementos: **pureza = (1/17) x (5+4+3) ≈ 0.71.**

## Evaluación del desempeño

## Matriz de contingencia

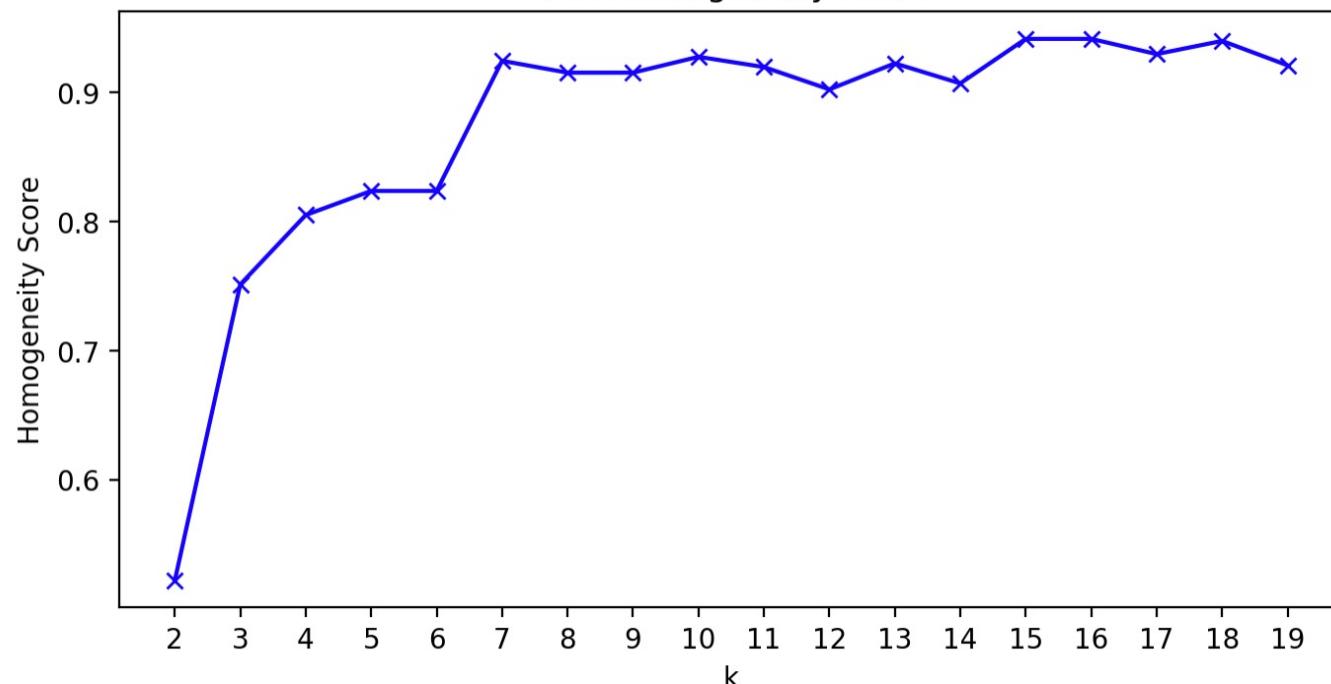
**Cluster 0 Cluster 1 Cluster 2**

<b>setosa</b>	0	50	0
<b>versicolor</b>	2	0	48
<b>virginica</b>	36	0	14

**Cluster 0 Cluster 1 Cluster 2 Cluster 3**

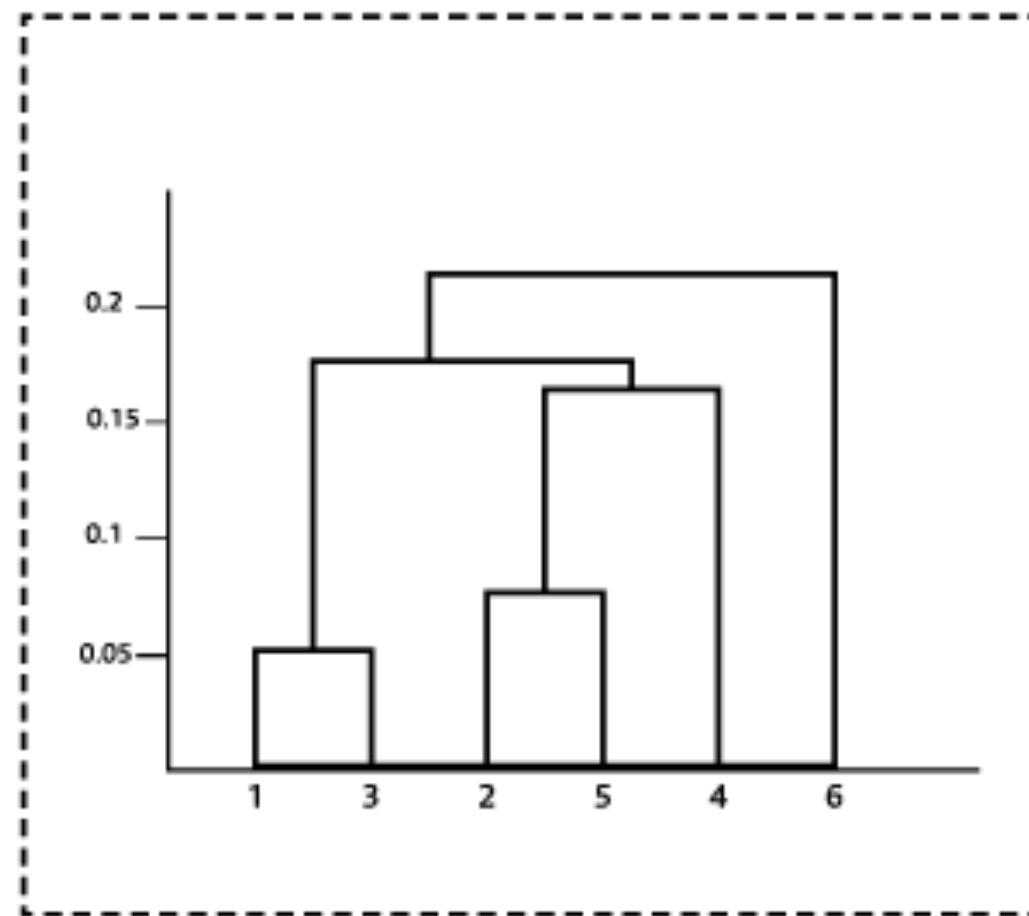
<b>setosa</b>	0	0	50	0
<b>versicolor</b>	26	0	0	24
<b>virginica</b>	1	32	0	17

K vs Homogeneity Score



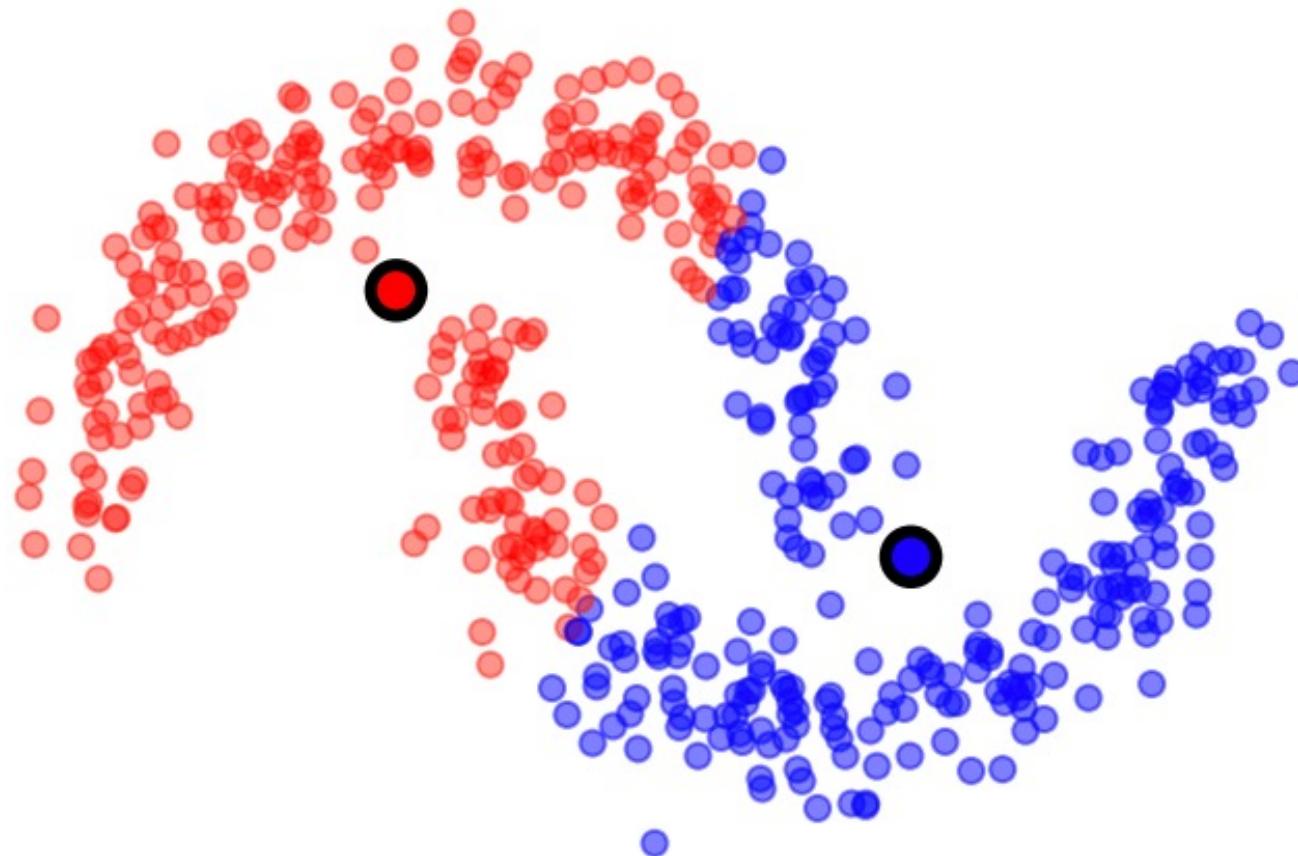
5

## Agrupamiento jerárquico



Agrupamiento jerárquicos

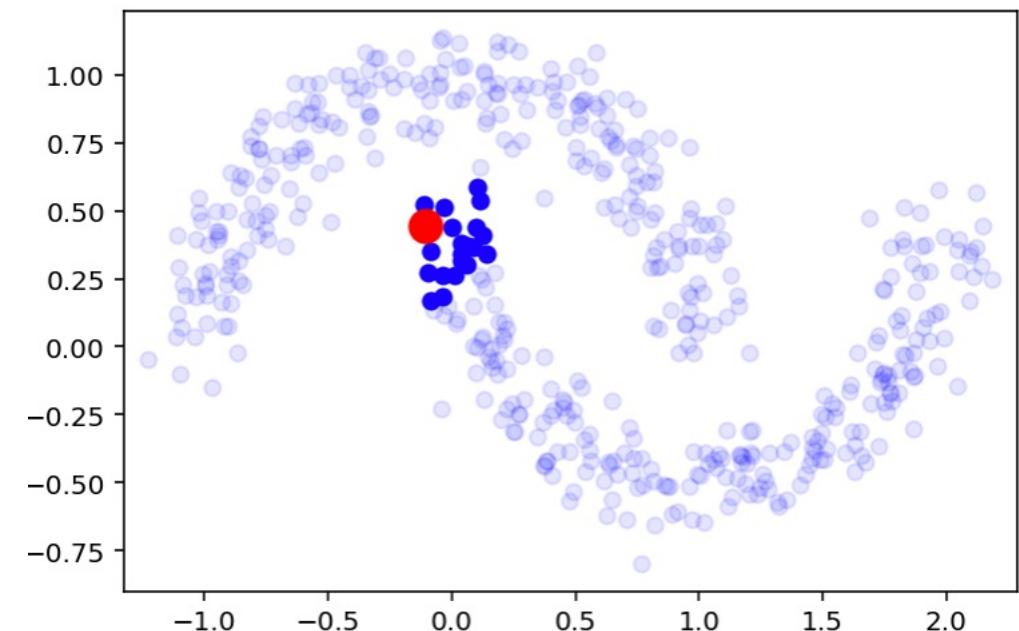
K-means sobre grupos no globulares



## Agrupamiento jerárquicos

## Agrupamiento jerárquico aglomerativo

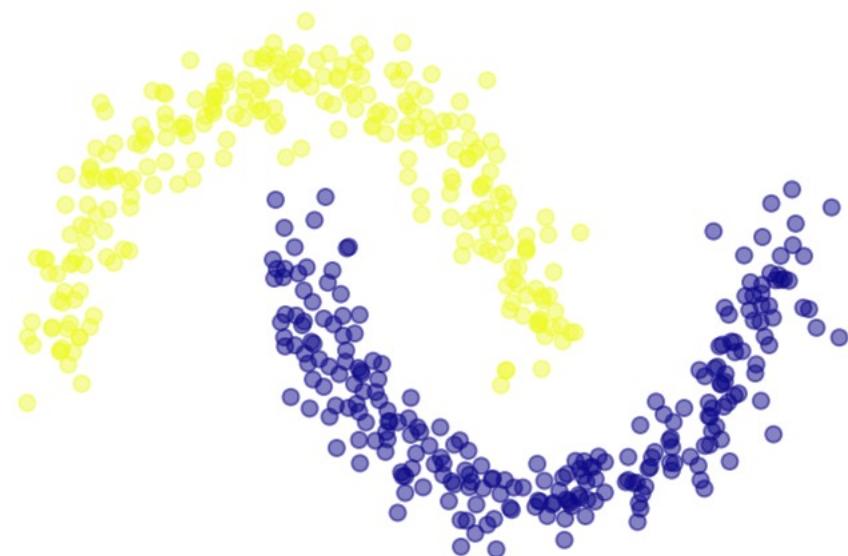
- Los métodos jerárquicos construyen grupos anidados
- El agrupamiento jerárquico aglomerativo empieza con grupos con un solo elemento,
- y los fusiona de acuerdo a las relaciones de vecindad
- En scikit learn esta relación se especifica con un grafo de conectividad
- La alternativa típica es usar el grafo de los k vecinos más cercanos



## Agrupamiento jerárquicos

## Agrupamiento jerárquico aglomerativo en scikit learn

```
1 from sklearn.neighbors import kneighbors_graph
2
3 knn_graph = kneighbors_graph(X, 4, include_self=False)
4 ac = AgglomerativeClustering(connectivity=knn_graph, linkage="average")
```





## Referencias

Manning, C., Raghavan, D. & Schütze, H. (2008). Introduction to Information Retrieval [Introducción a la recuperación de información]. Cambridge, UK: Cambridge University Press. ISBN: 978-0-521-86571-5

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis [Siluetas: una ayuda gráfica para la interpretación y validación el análisis de conglomerados].

<https://wis.kuleuven.be/stat/robust/papers/publications-1987/rousseeuw-silhouettes-jcam-scienceopenarchiv.pdf>

Shapiro, L. & Stockman, G. (2001). Computer Vision [Visión por computador]. New Jersey: Prentice-Hall. ISBN 0-13-030796-3

Tan, P., Steinbach, M., Karpatne, A. & Kumar, V. (2005). Introduction to Data Mining (2da edición) [Introducción de minería de datos].

[https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7\\_basic\\_cluster\\_analysis.pdf](https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf)

Zhao Y. & Karypis G. (2011) Document Clustering. In: Sammut C [Agrupación de documentos. En: Sammut C] Webb G.I. (eds). Encyclopedia of Machine Learning. Springer, Boston, MA  
[https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8\\_231](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_231)



## Recursos adicionales

### Aprendizaje Computacional

Alpaydin, E. (2010). Introduction to Machine Learning. [Introducción al aprendizaje de máquinas]  
[https://kkpatel7.files.wordpress.com/2015/04/alpaydin\\_machinelearning\\_2010.pdf](https://kkpatel7.files.wordpress.com/2015/04/alpaydin_machinelearning_2010.pdf)

### Bibliografía

Mayo, M. (Mayo de 2018). Marcos para abordar el proceso de aprendizaje automático. Kdnuggets.  
<https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html>

Mayo, M. ( s.f.). El proceso de ciencia de datos, redescubierto. Kdnuggets. <https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html>

Google developers. (s.f.). Introducción a la estructura de problemas de aprendizaje automático. <https://developers.google.com/machine-learning/problem-framing>



## Derechos de imágenes

Tan, P., Steinbach, M., Karpatne, A. y Kumar V. (2005). Ejemplo de agrupamiento. [Gráfica]. Adaptada de Introduction to Data Mining, Addison-Wesley [https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7\\_basic\\_cluster\\_analysis.pdf](https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf)

Subpng. (s.f.). Segmentación de Mercado. [Vector]. <https://www.subpng.com/png-jxj3q3/>

IBM Knowledge Center. (s.f.). Agrupamiento de documentos. [Gráfica]. Adaptada de [https://www.ibm.com/support/knowledgecenter/en/SSBRAM\\_8.8.0/com.ibm.classify.workbench.doc/c\\_WBG\\_Taxonomy\\_Proposer.html](https://www.ibm.com/support/knowledgecenter/en/SSBRAM_8.8.0/com.ibm.classify.workbench.doc/c_WBG_Taxonomy_Proposer.html)

Tan, P., Steinbach, M., Karpatne, A. y Kumar V. (2005). Agrupamiento Particional. [Gráfica]. Adaptada de Introduction to Data Mining, Addison-Wesley [https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7\\_basic\\_cluster\\_analysis.pdf](https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf)

Tan, P., Steinbach, M., Karpatne, A. y Kumar V. (2005). Agrupamiento Jerárquico. [Gráfica]. Adaptada de Introduction to Data Mining, Addison-Wesley [https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7\\_basic\\_cluster\\_analysis.pdf](https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf)

UC Business Analytics R Programming Guide (s.f.). Plot coeficiente de silueta y mejor valor de K. [Gráfica]. Adaptada de [https://ucr.github.io/kmeans\\_clustering](https://ucr.github.io/kmeans_clustering)



## Derechos de imágenes

Aprendizaje computacional <https://pixabay.com/vectors/machine-learning-information-brain-5433370/>  
[https://www.freepik.es/vector-gratis/ordenador-portatil-pantalla-blanca-teclado\\_7222477.htm#page=1&query=computador&position=1](https://www.freepik.es/vector-gratis/ordenador-portatil-pantalla-blanca-teclado_7222477.htm#page=1&query=computador&position=1)

<https://www.freepik.es/vectores/negocios> Vector de Negocios creado por ibrandify



## Créditos

*Facultad de*  
**INGENIERÍA**

**Autores**

Fabio Augusto González Osorio, PhD

**Asistente docente**

Miguel Ángel Ortiz Marín

**Diseño instruccional**

Claudia Patricia Rodríguez Sánchez

**Diseño gráfico**

Clara Valeria Suárez Caballero

Milton R. Pachón Pinzón

**Diagramadora PPT**

Daniela Duque

**Fecha**  
2021-I

