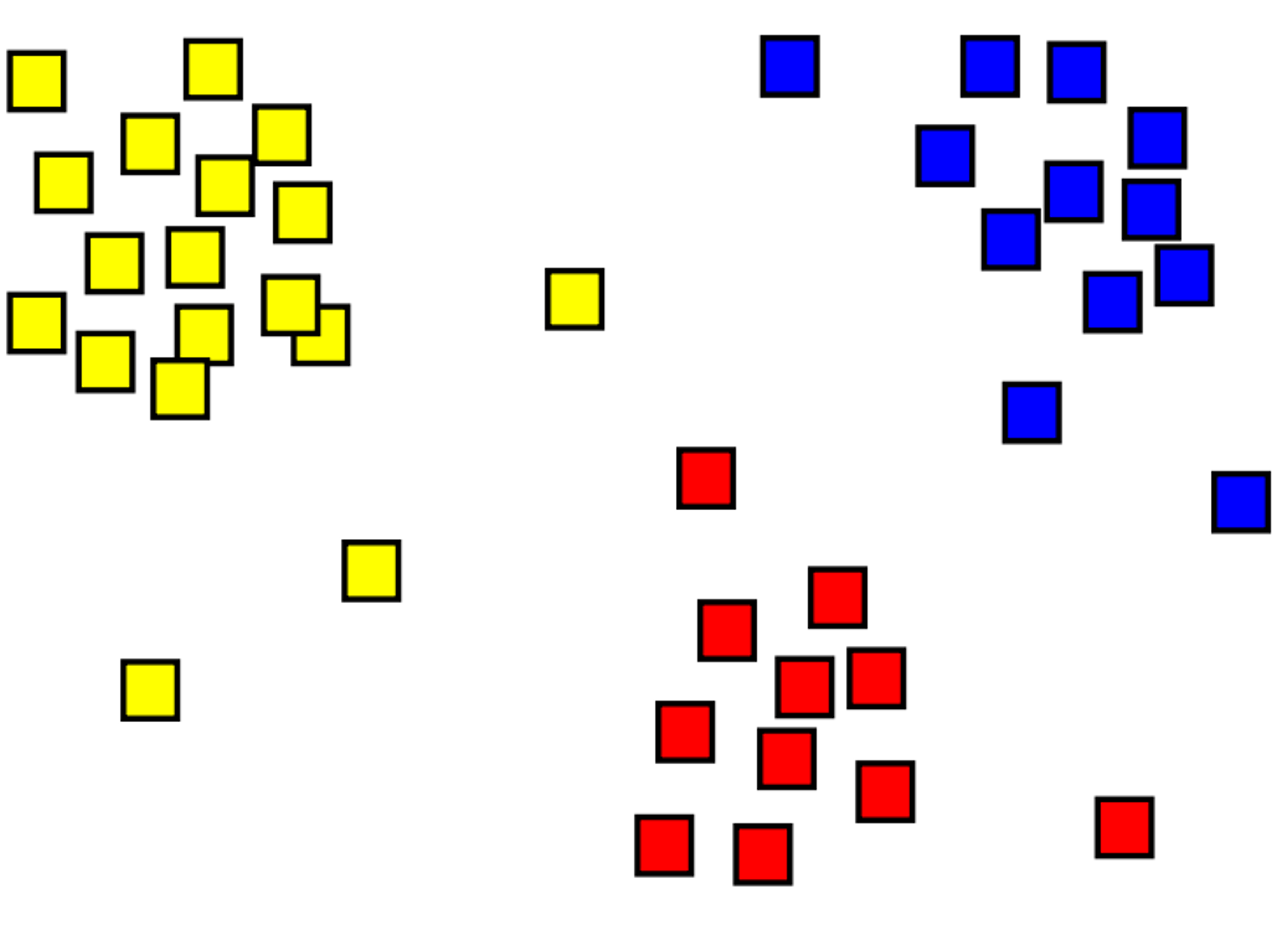
**03 - Entrenamiento y evaluación de modelos de agrupación en clústeres**

# Introducción

Completado100 XP

* 2 minutos

La agrupación en clústeres es el proceso de agrupar objetos con objetos similares. Por ejemplo, en la imagen siguiente tenemos una colección de coordenadas 2D que se han agrupado en tres categorías: superior izquierda (amarillo), inferior (rojo) y superior derecha (azul).



Una diferencia importante entre los modelos de agrupación en clústeres y clasificación es que la agrupación en clústeres es un método "no supervisado", donde el "entrenamiento" se realiza sin etiquetas. En su lugar, los modelos identifican ejemplos que tienen una colección similar de características. En la imagen anterior, los ejemplos que se encuentran en una ubicación similar se agrupan.

La agrupación en clústeres es común y útil para explorar nuevos datos en los que aún no se conocen patrones entre puntos de datos, como categorías de alto nivel. Se usa en muchos campos que necesitan etiquetar automáticamente datos complejos, incluido el análisis de redes sociales, la conectividad cerebral o el filtrado de correo no deseado, entre otros.

## Siguiente unidad: ¿Qué es la agrupación en clústeres?

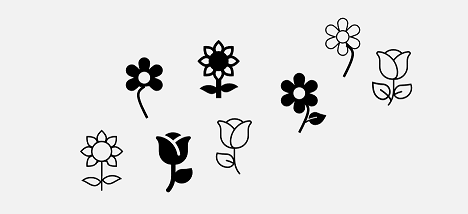
# ¿Qué es la agrupación en clústeres?

Completado100 XP

* 5 minutos

La agrupación en clústeres es un tipo de aprendizaje automático sin supervisión en el que las observaciones se agrupan en clústeres en función de las similitudes en sus valores de datos o características. Este tipo de aprendizaje automático se considera no supervisado porque no usa valores de etiqueta conocidos previamente para entrenar un modelo. En un modelo de agrupación en clústeres, la etiqueta es el clúster al que se asigna la observación, únicamente en función en sus características.

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



Puede ser útil agrupar estas flores en clústeres en función de las similitudes entre sus características.

Hay muchas maneras de hacerlo. Por ejemplo, si la mayoría de las flores tienen el mismo número de hojas, podrían agruparse en aquellas con muchos o pocos pétalos. Por otra parte, si el número de pétalos y hojas varía considerablemente, puede haber un patrón que detectar, como que las que tienen muchas hojas también tienen muchos pétalos. El objetivo del algoritmo de agrupación en clústeres es encontrar la manera óptima de dividir el conjunto de datos en grupos. El significado de "óptima" depende del algoritmo utilizado y del conjunto de datos que se proporciona.

Aunque este ejemplo de la flor puede ser sencillo para un humano con solo unas pocas muestras, a medida que el conjunto de datos crece a miles de muestras o a más de dos características, los algoritmos de agrupación en clústeres se vuelven muy útiles para diseccionar rápidamente un conjunto de datos en grupos.

## Siguiente unidad: Ejercicio: Entrenamiento y evaluación de un modelo de agrupación en clústeres

# Evaluación de diferentes tipos de agrupación en clústeres

Completado100 XP

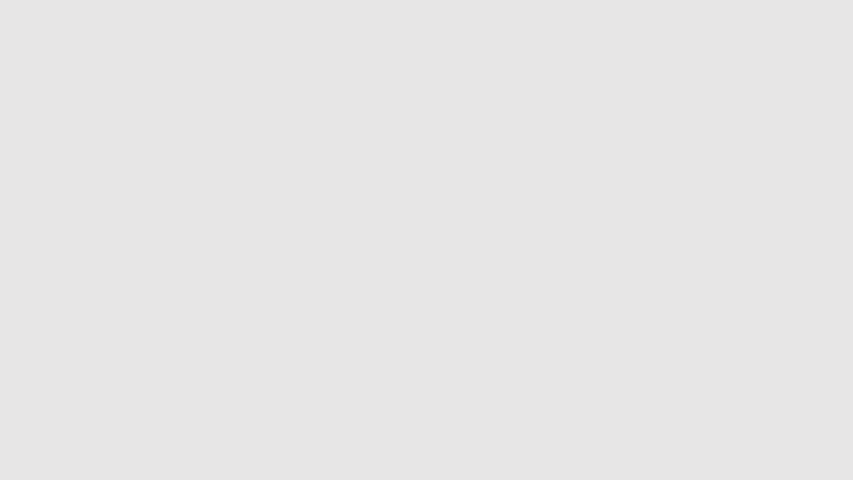
* 5 minutos

## Entrenamiento de un modelo de agrupación en clústeres

Hay varios algoritmos que puede usar para la agrupación en clústeres. Uno de los algoritmos más usados es la agrupación en clústeres k-means que, en su forma más sencilla, comprende los pasos siguientes:

1. Los valores de las características se vectorizan para definir coordenadas de n dimensiones (donde n es el número de características). En el ejemplo de las flores, tenemos dos características (el número de pétalos y el número de hojas), por lo que el vector de características tiene dos coordenadas que se pueden usar para trazar conceptualmente los puntos de datos en un espacio de dos dimensiones.
2. Decida cuántos clústeres quiere usar para agrupar las flores y llame a este valor ***k***. Por ejemplo, para crear tres clústeres, usaría un valor k de 3. Después, se representan los puntos k en coordenadas aleatorias. En última instancia, estos puntos serán los puntos centrales de cada clúster, por lo que se hace referencia a ellos como centroides.
3. Cada punto de datos (en este caso, cada flor) se asigna a su centroide más cercano.
4. Cada centroide se mueve al centro de los puntos de datos asignados en función de la distancia media entre los puntos.
5. Después de mover el centroide, los puntos de datos pueden pasar a estar más cerca de otro centroide, por lo que se reasignan a los clústeres en función del nuevo centroide más cercano.
6. Los pasos de movimiento de centroides y reasignación de clústeres se repiten hasta que los clústeres se estabilizan o se alcanza un número máximo predeterminado de iteraciones.

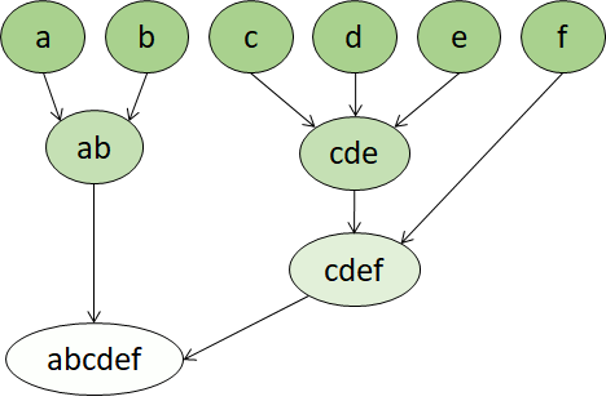
En la siguiente animación se ilustra este proceso:



## Agrupación en clústeres jerárquica

La agrupación en clústeres jerárquica es otro tipo de algoritmo de agrupación en clústeres en el que los propios clústeres pertenecen a un grupo más grande, que pertenecen a grupos incluso más grandes, y así sucesivamente. El resultado es que los puntos de datos pueden agruparse con diferentes grados de precisión: con un gran número de grupos muy pequeños y precisos, o con un pequeño número de grupos más grandes.

Por ejemplo, si aplicamos la agrupación a los significados de las palabras, podemos obtener un grupo que contenga adjetivos específicos de las emociones ("enfadado", "feliz", etc.), que a su vez pertenezca a un grupo que contenga todos los adjetivos relacionados con el ser humano ("feliz", "guapo", "joven"), y este a su vez pertenezca a un grupo aún mayor que contenga todos los adjetivos ("feliz", "verde", "guapo", "duro", etc.).



La agrupación en clústeres jerárquicas es útil no solo para dividir los datos en grupos, sino también para comprender las relaciones entre estos grupos. Una ventaja importante de la agrupación en clústeres jerárquica es que no requiere que el número de clústeres se defina de antemano y, a veces, puede proporcionar resultados más interpretables que los enfoques no jerárquicos. El principal inconveniente es que estos enfoques pueden tardar mucho más tiempo en calcularse que los enfoques más sencillos y, a veces, no son adecuados para grandes conjuntos de datos.

## Siguiente unidad: Ejercicio: Entrenamiento y evaluación de modelos de agrupación en clústeres avanzados

# Resumen

Completado100 XP

* 1 minuto

En este módulo, ha aprendido a usar la agrupación en clústeres para crear modelos de aprendizaje automático no supervisados que agrupan observaciones de datos en clústeres. Después, ha usado el marco **scikit-learn** en Python para entrenar un modelo de agrupación en clústeres.

Aunque scikit-learn es un marco popular para escribir código con el fin de entrenar modelos de agrupación en clústeres, también puede crear soluciones de aprendizaje automático para la agrupación en clústeres con las herramientas gráficas de Microsoft Azure Machine Learning. Para obtener más información sobre el desarrollo sin código de modelos de agrupación en clústeres mediante Azure Machine Learning, consulte el módulo [Creación de un modelo de agrupación en clústeres con el diseñador de Azure Machine Learning](https://learn.microsoft.com/es-es/learn/modules/create-clustering-model-azure-machine-learning-designer/).

## Desafío: Agrupación de datos sin etiquetar

Ahora que ha visto cómo crear un modelo de agrupación en clústeres, ¿por qué no prueba con uno? Encontrará un desafío de agrupación en clústeres en el cuaderno [04 - Clustering Challenge.ipynb](https://github.com/MicrosoftDocs/ml-basics/blob/master/challenges/04%20-%20Clustering%20Challenge.ipynb).