1. **Melatih dan mengevaluasi model pengklusteran**

.Pengklusteran adalah sejenis pembelajaran mesin yang digunakan untuk mengelompokkan item serupa ke dalam kluster.

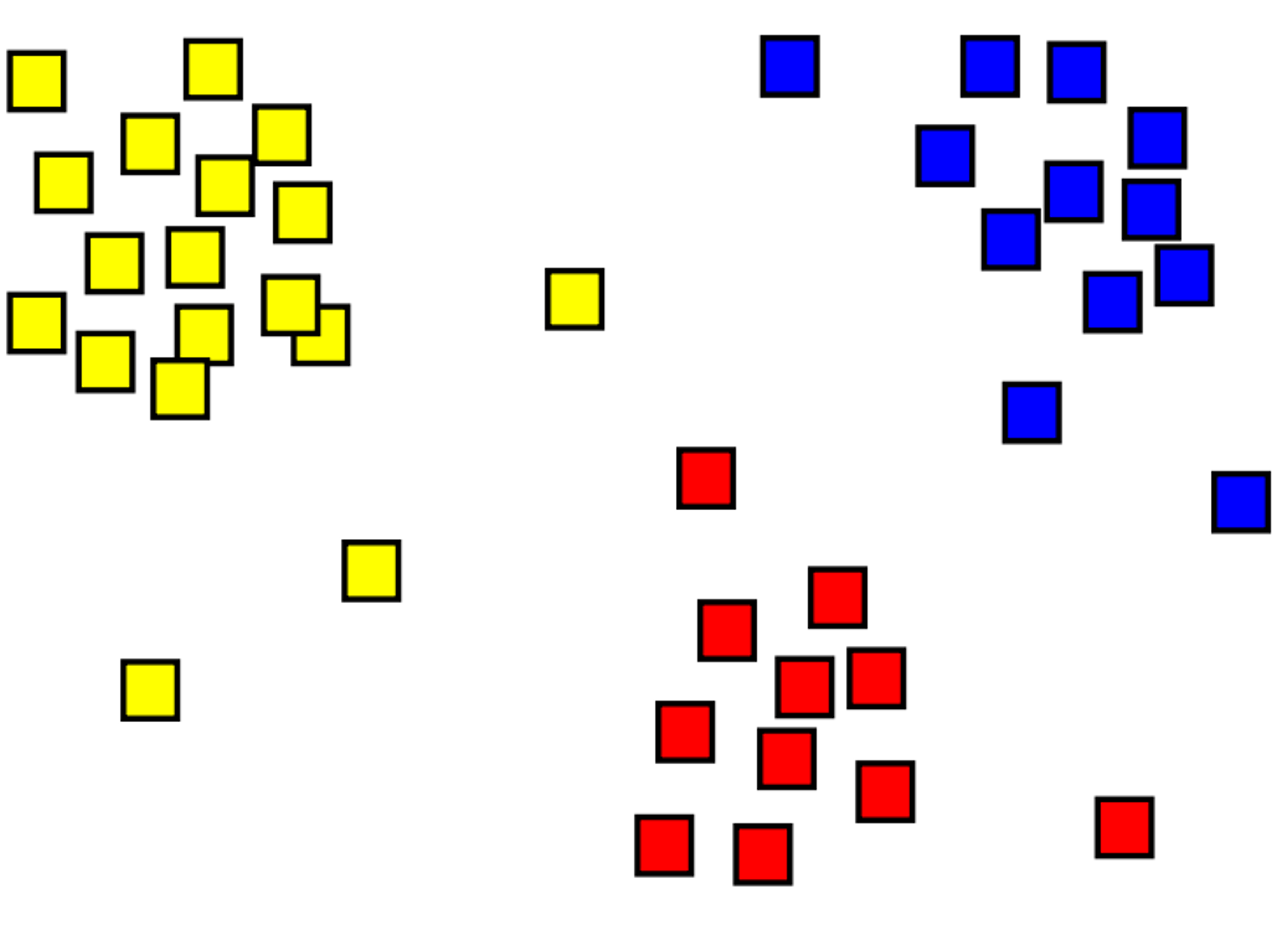
**Tujuan pembelajaran**

Dalam modul ini, Anda akan mempelajari:

* Kapan harus menggunakan pengklusteran
* Cara melatih dan mengevaluasi model pengklusteran menggunakan kerangka kerja scikit-learn

1. **Pengantar**

Pengklusteran adalah proses pengelompokan objek dengan objek yang serupa. Misalnya, pada gambar di bawah ini kami memiliki kumpulan koordinat 2D yang telah dikelompokkan ke dalam tiga kategori - kiri atas (kuning), bawah (merah), dan kanan atas (biru).



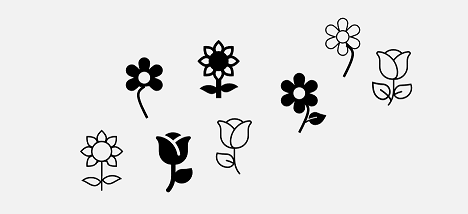
Perbedaan utama antara model pengelompokan dan klasifikasi adalah bahwa pengelompokan adalah metode 'tanpa pengawasan', tempat 'pelatihan' dilakukan tanpa label. Sebagai gantinya, model mengidentifikasi contoh yang memiliki koleksi fitur serupa. Pada gambar di atas, contoh yang berada di lokasi yang sama dikelompokkan bersama.

Pengklusteran bersifat umum dan berguna untuk menjelajahi data baru tempat pola antara titik data, seperti kategori tingkat tinggi, belum diketahui. Pengklusteran ini digunakan di banyak bidang yang perlu secara otomatis memberi label pada data kompleks, termasuk analisis jaringan sosial, konektivitas otak, pemfilteran spam, dan sebagainya.

1. **Apa itu pengklusteran?**

Pengklusteran adalah bentuk pembelajaran mesin tanpa pengawasan di mana pengamatan dikelompokkan ke dalam kelompok berdasarkan kesamaan dalam nilai datanya, atau fitur. Pembelajaran mesin semacam ini dianggap tidak diawasi karena tidak menggunakan nilai label yang diketahui sebelumnya untuk melatih model; dalam model pengklusteran, label adalah kluster yang menjadi tujuan pengamatan, berdasarkan fitur-fitur model semata.

Misalnya, seorang ahli botani mengamati sampel bunga dan mencatat jumlah kelopak dan daun pada setiap bunga.



Hal ini mungkin berguna untuk mengelompokkan bunga-bunga ini ke dalam kluster berdasarkan kesamaan di antara fitur-fiturnya.

Ada banyak cara yang bisa dilakukan. Misalnya, jika sebagian besar bunga memiliki jumlah daun yang sama, bunga itu dapat dikelompokkan menjadi bunga yang memiliki banyak kelopak vs sedikit kelopak. Sebagai alternatif, jika jumlah kelopak dan daun sangat bervariasi, mungkin ada pola yang dapat ditemukan, seperti yang memiliki banyak daun dan juga memiliki banyak kelopak. Tujuan dari algoritma pengklusteran adalah untuk menemukan cara yang optimal untuk membagi himpunan data menjadi beberapa kelompok. Apa arti dari ‘optimal’ bergantung pada algoritme yang digunakan dan himpunan data yang disediakan.

Meskipun contoh bunga ini mungkin sederhana untuk dicapai manusia dengan hanya beberapa sampel, karena himpunan data tumbuh menjadi ribuan sampel atau lebih dari dua fitur, algoritme pengklusteran menjadi sangat berguna untuk membedah himpunan data dengan cepat ke dalam kelompok.

1. **Latihan - Melatih dan mengevaluasi model pengklusteran**

**Clustering introduction:** D:\Data Scientist Indonesia - Cloud Skills Challenge from Microsoft\4 - Melatih dan mengevaluasi model pengklusteran\01-clustering-introduction.ipynb

**This modul answer this qustion:** how do you know how many clusters to separate your data into?

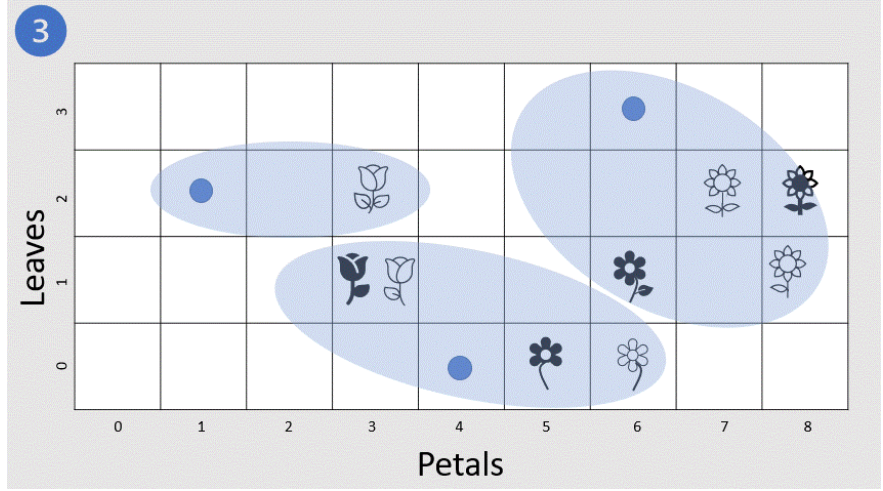
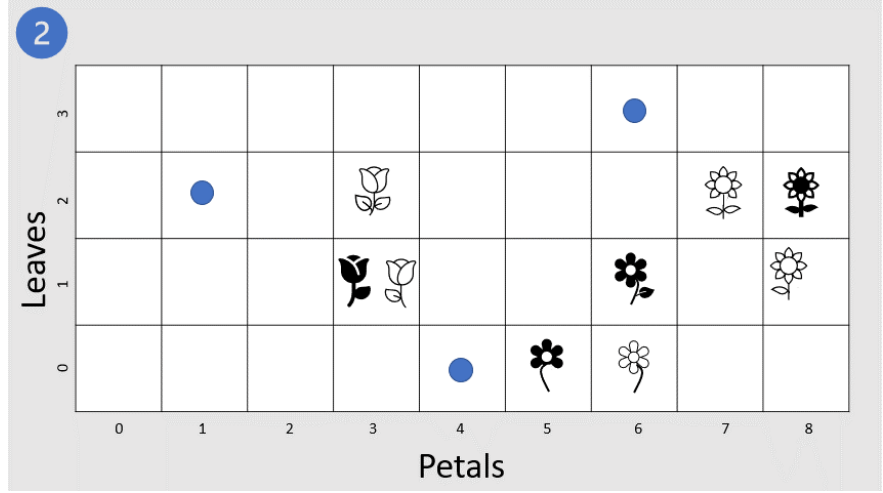
1. **Mengevaluasi berbagai jenis pengklusteran**

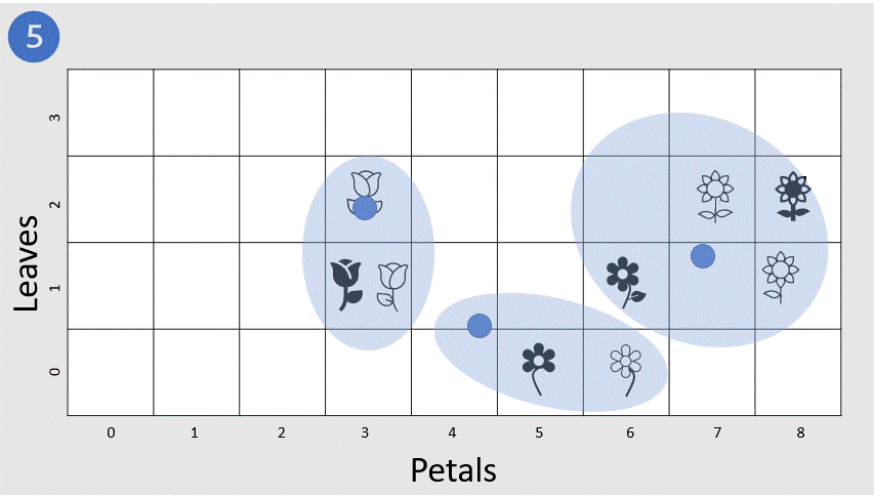
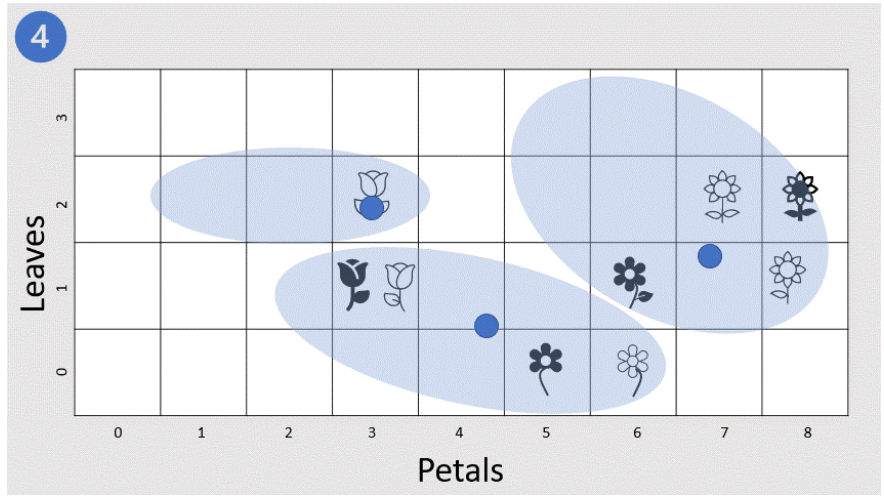
**Melatih model pengklusteran**

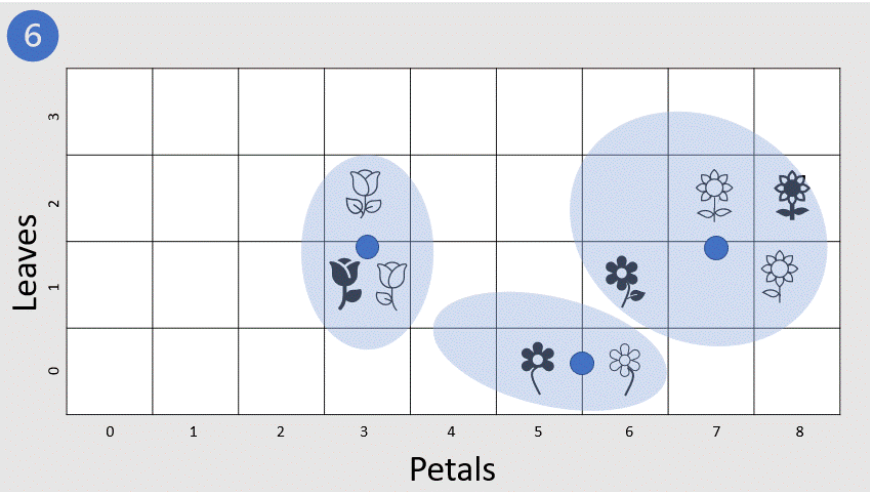
Ada beberapa algoritma yang dapat Anda gunakan untuk pengklusteran. Salah satu algoritme yang paling umum digunakan adalah pengklusteran K-Means yang, dalam bentuknya yang paling sederhana, terdiri dari langkah-langkah berikut:

1. Nilai fitur divektorkan untuk menentukan koordinat n-dimensi (di mana n adalah jumlah fitur). Dalam contoh bunga, kita memiliki dua fitur (jumlah kelopak dan jumlah daun), sehingga vektor fitur memiliki dua koordinat yang dapat kita gunakan untuk secara konseptual memplot poin data dalam ruang dua dimensi.
2. Anda memutuskan berapa banyak kelompok yang ingin Anda gunakan untuk mengelompokkan bunga, dan menyebut nilai ini k. Misalnya, untuk membuat tiga kluster, Anda akan menggunakan nilai k dari 3. Kemudian titik k diplot pada koordinat acak. Titik-titik ini pada akhirnya akan menjadi titik pusat untuk setiap kluster, sehingga disebut sebagai centroid.
3. Setiap poin data (dalam kasus ini bunga) ditetapkan ke pusat terdekatnya.
4. Setiap pusat dipindahkan ke pusat poin data yang ditetapkan berdasarkan jarak rata-rata antara titik.
5. Setelah memindahkan pusat, poin data sekarang lebih dekat ke pusat yang berbeda, sehingga poin data dipindahkan ke kluster berdasarkan pusat terdekat yang baru.
6. Pergerakan pusat dan langkah alokasi ulang kluster diulang hingga kluster menjadi stabil atau jumlah perulangan maksimum yang telah ditentukan tercapai.

Animasi berikut menunjukkan proses ini:



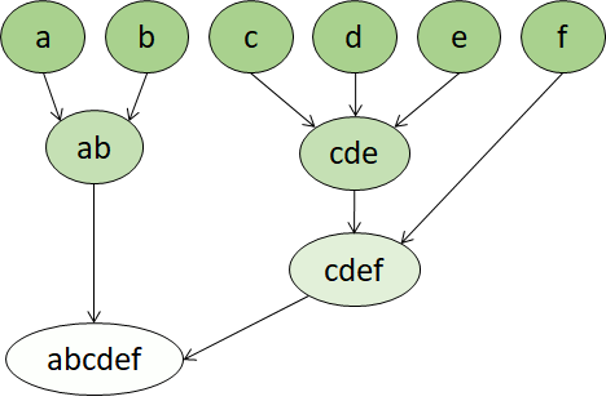




**Pengklusteran Hirarkis**

Pengklusteran hierarkis adalah jenis lain dari algoritma pengklusteran di mana kluster itu sendiri termasuk dalam kelompok yang lebih besar, yang termasuk dalam kelompok yang lebih besar, dan seterusnya. Hasilnya adalah bahwa poin data dapat berupa kluster dalam derajat presisi yang berbeda: dengan sejumlah besar grup yang sangat kecil dan presisi, atau sejumlah kecil grup yang lebih besar.

Misalnya, jika kita menerapkan pengklusteran pada makna kata, kita mungkin mendapatkan kelompok yang berisi kata sifat khusus untuk emosi (‘marah’, ‘bahagia’, dan sebagainya), yang dengan sendirinya termasuk dalam kelompok yang berisi semua kata sifat yang berhubungan dengan manusia ( ‘bahagia’, ‘tampan’, ‘muda’), dan ini termasuk dalam kelompok yang lebih tinggi yang berisi semua kata sifat (‘bahagia’, ‘hijau’, ‘tampan’, ‘keras’ dll.).



Pengklusteran hierarkis berguna untuk tidak hanya memecah data menjadi kelompok-kelompok, tetapi juga memahami hubungan antara kelompok-kelompok ini. Keuntungan utama dari pengklusteran hierarkis adalah tidak memerlukan jumlah kluster yang harus ditentukan sebelumnya, dan terkadang dapat memberikan hasil yang lebih dapat diinterpretasikan daripada pendekatan non-hierarkis. Kelemahan utama adalah bahwa pendekatan ini dapat memakan waktu lebih lama untuk dihitung daripada pendekatan yang lebih sederhana dan terkadang tidak cocok untuk himpunan data yang besar.

1. **Latihan - Melatih dan mengevaluasi model pengklusteran tingkat lanjut**

**clustering automatically:** D:\Data Scientist Indonesia - Cloud Skills Challenge from Microsoft\4 - Melatih dan mengevaluasi model pengklusteran\02-kmeans-and-hierachial.ipynb

1. **Ringkasan**

Dalam modul ini, Anda mempelajari bagaimana pengklusteran dapat digunakan untuk membuat model pembelajaran mesin tanpa pengawasan yang mengelompokkan pengamatan data ke dalam kluster. Anda kemudian menggunakan kerangka kerja scikit-learn di Python untuk melatih model pengklusteran.

Meskipun scikit-learn adalah kerangka kerja populer untuk menulis kode untuk melatih model pengklusteran, Anda juga dapat membuat solusi pembelajaran mesin untuk pengklusteran menggunakan alat grafis di Azure Machine Learning. Anda dapat mempelajari selengkapnya tentang pengembangan model pengklusteran tanpa kode menggunakan Azure Machine Learning di modul Membuat model pengklusteran dengan perancang Azure Machine Learning.

**Tantangan: Kluster Data Tanpa Label**

Sekarang setelah Anda melihat cara membuat model pengklusteran, mengapa tidak mencobanya sendiri? Anda akan menemukan tantangan pengklusteran di buku catatan 04 - Clustering Challenge.ipynb!

<https://github.com/MicrosoftDocs/ml-basics/blob/master/challenges/04%20-%20Clustering%20Challenge.ipynb>

