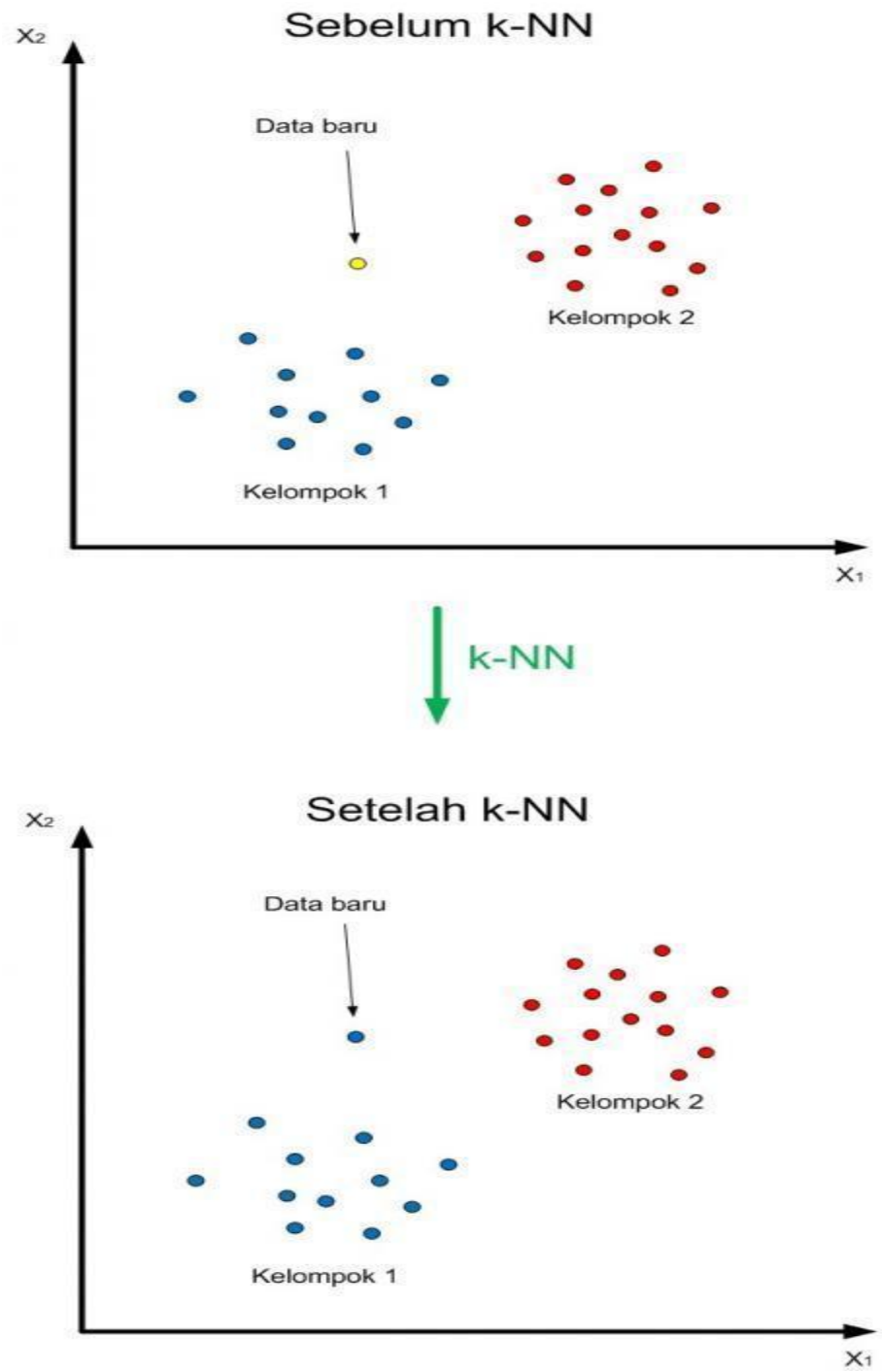
Machine Learning: K-nearest Neighbors

Kali ini kita akan belajar bersama tentang teknik klasifikasi yang lain yaitu K-nearest neighbors (k-NN). Jika diartikan ke dalam bahasa Indonesia, artinya adalah tetangga terdekat sebanyak K buah.. Perlu diperhatikan bahwa K yang dimaksud berbeda dengan K-Means Clustering yang merupakan salah satu teknik *clustering.*

Untuk bisa memahami konsep k-NN secara mudah, mari kita lihat ilustrasi di bawah ini:



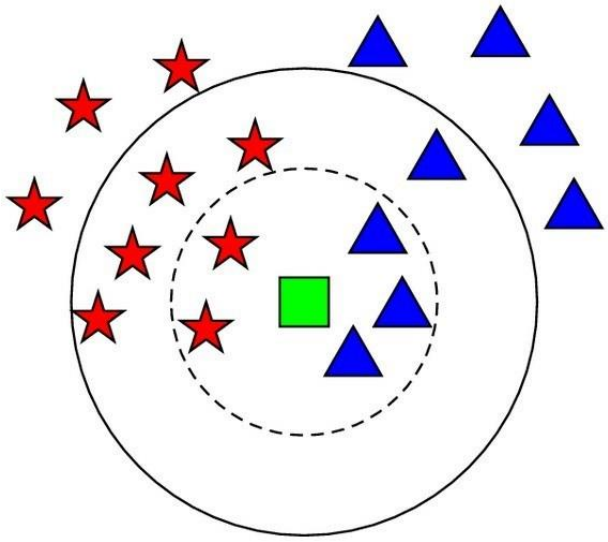
Ilustrasi k-NN sebelum dan sesudah.

Pada gambar di atas dapat dilihat bahwa kita sudah memiliki data yang dibagi ke dalam dua kelompok, misal keputusan pelanggan untuk beli/tidak. Kemudian kita mendapatkan 1 data tambahan. Pertanyaannya, ia masuk ke kelompok biru atau merah? Melalui k-NN ternyata ia masuk ke dalam kelompok biru. Bagaimana caranya?

Berikut adalah langkah-langkah k-NN:

1. Tentukan jumlah kelompok neighbors (K) nya. Umumnya adalah 5.
2. Ambil data K terdekat (K neighbors) dari data terbaru (umumnya 5 buah K) berdasarkan jarak euclidean antar keduanya.
3. Dari K-neighbors ini, hitung berapa banyak data poin yang masuk di masing-masing kategori.
4. Masukkan data baru ini ke dalam kelompok yang memiliki jumlah K-neighbors terbanyak.

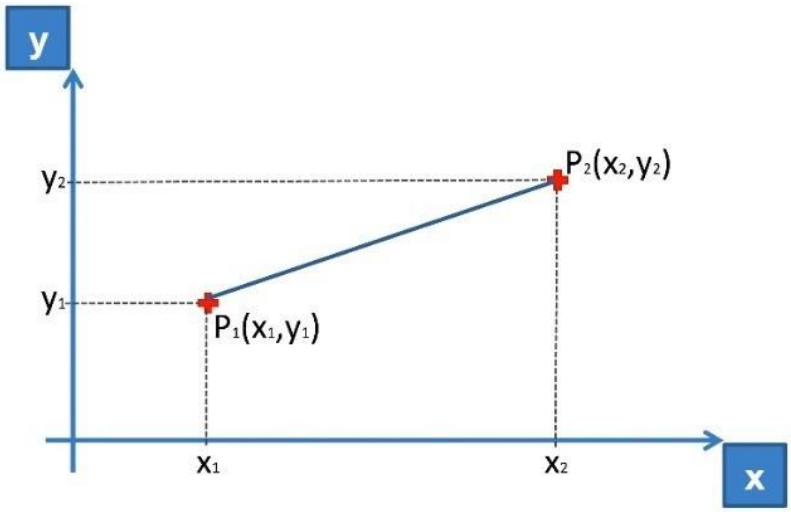
Bisa dilihat pada ilustrasi di bawah ini:



Ilustrasi metode k-NN. Lingkaran awal adalah k-NN dengan K=5, sementara lingkaran luar adalah k-NN dengan K=10.

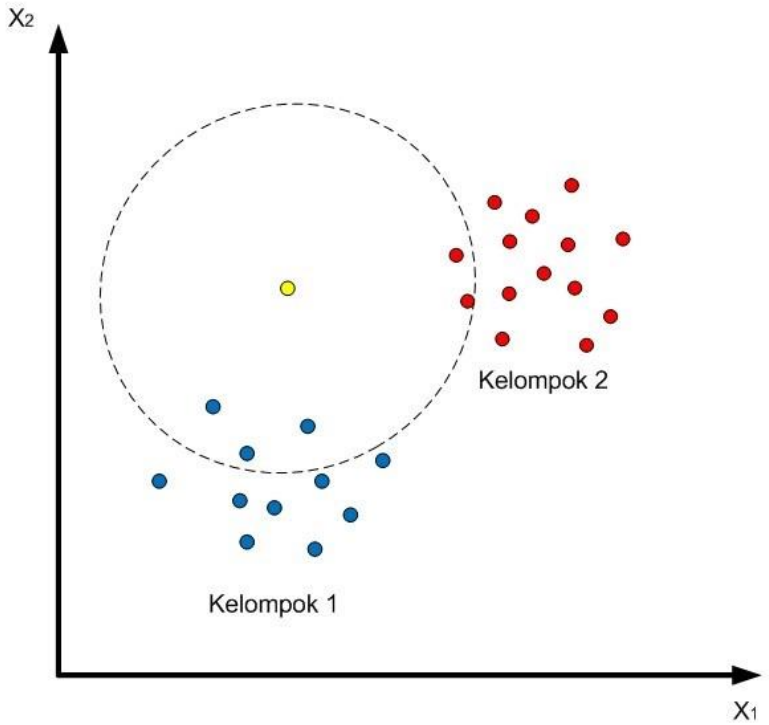
Melalui gambar di atas, kita bisa tahu bahwa penentuan jumlah K, berpengaruh terhadap pengambilan keputusan. Jika K=5 maka ia masuk kelompok biru, dan jika K=10, ia masuk kelompok merah.

Perlu diingat, untuk mencari K terdekat yang dilihat adalah jarak euclideannya. Seperti ini formulanya:



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$ Ilustrasi euclidean distance.

Dengan mengambil 5 data poin terdekat melalui jarak euclideannya untuk kasus kita di atas tadi, maka ilustrasinya akan tampak sebagai berikut:



Ilustrasi k-NN dengan mencari 5 titik dengan jarak euclidean terdekat.

Melalui 5 data ini dapat dilihat bahwa jumlah biru lebih banyak dari jumlah data merah, sehingga data baru ini masuk ke dalam kelompok biru.

Algoritma ***k-Nearest Neighbor*** adalah algoritma *supervised learning* dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori **k**-tetangga terdekat.

Tujuan dari algoritma ini adalah untuk mengklasifikasikan obyek baru berdasarkan atribut dan *sample-sample* dari *training data*.

Algoritma *k-Nearest Neighbor* menggunakan *Neighborhood Classification* sebagai nilai prediksi dari nilai *instance* yang baru.

Contoh Kasus

Misalnya ada sebuah rumah yang **berada tepat di tengah perbatasan** antara Kota Bandung dan Kabupaten Bandung, sehingga pemerintah kesulitan untuk menentukan **apakah rumah tersebut termasuk kedalam wilayah Kota Bandung atau Kabupaten Bandung**.

Kita bisa menentukannya dengan menggunakan **Algoritma k-NN**, yaitu dengan melibatkan jarak antara rumah tersebut dengan rumah-rumah yang ada disekitarnya (tetangganya).

Pertama, kita harus menentukan jumlah tetangga yg akan kita perhitungkan (k), misalnya kita tentukan **3 tetangga terdekat** (**k = 3**).

Kedua, hitung jarak setiap tetangga terhadap rumah tersebut, lalu urutkan hasilnya berdasarkan jarak, mulai dari yang terkecil ke yang terbesar.

Ketiga, ambil 3 (k) tetangga yg paling dekat, lalu kita lihat masing-masing dari tetangga tersebut apakah termasuk kedalam wilayah Kota atau Kabupaten. Ada 2 kemungkinan:

- Bila dari 3 tetangga tersebut terdapat ada 2 rumah yg termasuk kedalam wilayah Kota Bandung, maka rumah tersebut termasuk kedalam wilayah Kota Bandung.
- Sebaliknya, bila dari 3 tetangga tersebut terdapat 2 rumah yg termasuk kedalam wilayah Kabupaten Bandung, maka rumah tersebut termasuk kedalam wilayah Kabupaten Bandung.

***Dalam menentukan nilai k**, bila **jumlah klasifikasi kita genap** maka sebaiknya kita gunakan **nilai k ganjil**, dan begitu pula sebaliknya bila **jumlah klasifikasi kita ganjil** maka sebaiknya gunakan **nilai k genap**, karena jika tidak begitu, ada kemungkinan kita **tidak akan mendapatkan jawaban**.*

Pembahasan Lebih Detil

Pada kasus diatas, kita menghitung jarak suatu rumah terhadap tetangga-tetangganya, itu berarti kita harus mengetahui posisi dari setiap rumah. Kita bisa menggunakan *latitude* dan *longitude* (atau garis lintang dan garis bujur) sebagai posisi.

Untuk mempermudah pemahaman, saya akan coba menggunakan data yang nilainya sederhana. Data yang akan digunakan adalah sebagai berikut:

Rumah	Lat	Long	Lokasi
A	11	26	Kota
B	15	29	Kota
C	19	28	Kota
D	18	30	Kota
E	16	26	Kota
F	23	25	Kabupaten
G	25	22	Kabupaten
H	21	24	Kabupaten
I	23	25	Kabupaten
J	29	24	Kabupaten
X	19	25	?

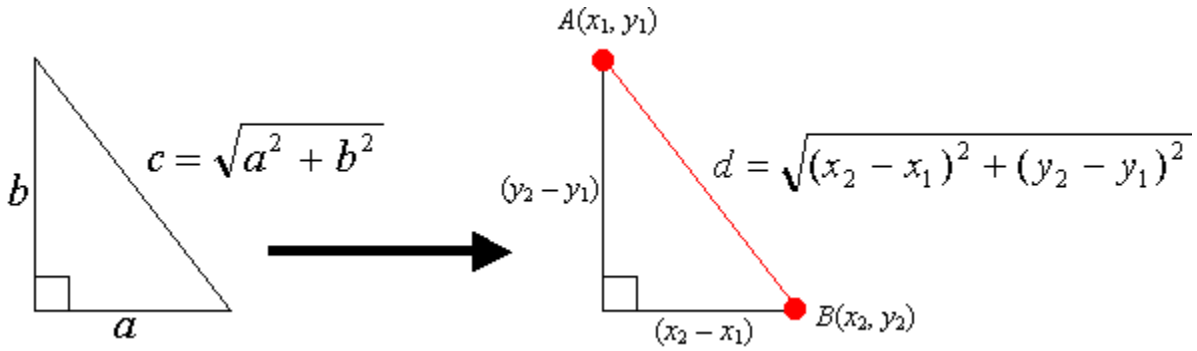
Dari *data* diatas, kita mendapatkan beberapa informasi, diantaranya:

- ***Independent Variables***, yaitu variable yang nilainya **tidak dipengaruhi** oleh variable lain. Pada contoh *data* diatas, yang termasuk *independent variable* adalah ***Lat***, dan ***Long***.
- ***Dependent Variables***, yaitu *variable* yang nilainya **dipengaruhi** oleh *variable* lain. Pada contoh *data* diatas, yang termasuk *dependent variable* adalah **Lokasi**.
- **Rumah A-E** adalah rumah yang masuk ke dalam wilayah **Kota**.
- **Rumah F-J** adalah rumah yang masuk ke dalam wilayah **Kabupaten**.

- **Rumah X** adalah rumah yang akan kita prediksi menggunakan algoritma kNN apakah termasuk ke dalam wilayah Kota atau Kabupaten.

Didalam dunia ***Machine Learning***, ***Independent Variables*** sering disebut juga sebagai ***Features***.

Selanjutnya kita hitung jarak antara rumah X terhadap rumah A-G dengan menggunakan rumus ***pythagoras***:



Pythagoras Formula. Source: Devon Maths Tuition

Diketahui, dimana **x** adalah ***Lat***, **y** adalah ***Long***, sedangkan **(x1, y1)** adalah *lat* dan *long* dari **rumah X**, dan **(x2, y2)** adalah *lat* dan *long* dari **masing-masing tetangganya**.

Setelah dihitung, selanjutnya adalah **urutkan jarak tersebut dari yang paling kecil ke yang paling besar**, hasilnya adalah sebagai berikut:

Rumah	Lat	Long	Jarak Terhadap Rumah X
H	21	24	2.24
C	19	28	3.00
E	16	26	3.16
F	23	25	4.00
I	23	25	4.00
D	18	30	5.10
B	15	29	5.66
G	25	22	6.71
A	11	26	8.06
J	29	24	10.05

Dapat dilihat dari hasil perhitungan diatas, bahwa ternyata 3 tetangga terdekat dari rumah X adalah:

- **Rumah H** (Kabupaten) yang memiliki jarak **2.24**,
- **Rumah C** (Kota) yang memiliki jarak **3**, dan
- **Rumah E** (Kota) yang memiliki jarak **3.16**.

Dari ke-3 tetangga terdekat, terdapat **2 rumah** yang termasuk kedalam wilayah **Kota** dan **1 rumah** yang masuk ke dalam wilayah **Kabupaten**. Sehingga dapat disimpulkan, bahwa **Rumah X adalah rumah yang termasuk kedalam wilayah Kota Bandung**.