Algoritma Machine Learning: K-Nearest Neighbor (KNN)

Kecerdasan Artifisial(CIF63310 / 2 sks)

Outline

- Deskripsi Klasifikasi
- Teknik Klasifikasi
- Instance-based Classifier
- K-Nearest Neighbor
- Ukuran Jarak

Definisi Klasifikasi

- Klasifikasi memerlukan data yang terdiri dari sekumpulan fitur, salah satu fiturnya harus berupa kelas/kategori data tersebut

Warna	Panjang (cm)	Berat (kg)	Jenis Makanan	Spesies
Coklat	180	175	Karnivora	Singa
Hijau	6	0,022	Herbivora	Kodok
……	……..	……..	…….	…….
Hitam Putih	250	360	Herbivora	Zebra

- Klasifikasi bertujuan memberi label data baru yang belum memiliki kelas

Warna	Panjang (cm)	Berat (kg)	Jenis Makanan	Spesies
Hitam	120	30	Omnivora	??????

Teknik Klasifikasi

- Terdapat beberapa teknik klasifikasi:
  - Instance based
  - Decision tree based
  - Rule-based
  - Neural Network
  - Naïve Bayes dan Bayesian Network
  - Support Vector Machine

Instance Based Classifier

- Instance based classifier bekerja dengan **membandingkan data uji dengan data latih** secara langsung
- Kelas data uji = kelas data latih yang **paling mirip**

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

DATASET IRIS

- 4 fitur, 3 kelas

Iris setosa

Iris virginica

Iris versicolor

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

Instance Based Classifier

Apakah kelas yang paling tepat untuk data berikut?

Sepal length	Sepal width	Petal length	Petal Width	Species
6.1	3	4.9	1.8	??????

Iris-virginica

Jenis Instance Based Classifier

- Terdapat dua jenis metode instance based classifier, yaitu **rote-learner** dan **nearest-neighbor**
- **Rote-learner** memberikan label pada data uji berdasarkan label pada data latih yang memiliki nilai **sama persis** dengan data uji.
- **Nearest-neighbor** memberi label pada data uji berdasarkan label pada data latih yang memiliki nilai **paling mirip** dengan data uji

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
4,3	**3**	**1,1**	**0,1**	**Iris-setosa**
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

Rote learner

Sepal length	Sepal width	Petal length	Petal Width	Species
4,3	3	1,1	0,1	Iris-setosa

Permasalahan Rote learner

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

Sepal length	Sepal width	Petal length	Petal Width	Species
4,5	3,1	1,5	0,1	??????

Rote learner gagal mengklasifikasikan data uji yang tidak sama persis dengan data latih

Nearest neighbor classifier

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	**3,1**	**1,5**	**0,2**	**Iris-setosa**
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

Sepal length	Sepal width	Petal length	Petal Width	Species
4,5	3,1	1,5	0,1	Iris-setosa

Permasalahan Nearest Neighbor

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	**3,1**	**1,5**	**0,2**	**Iris-virginica**
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

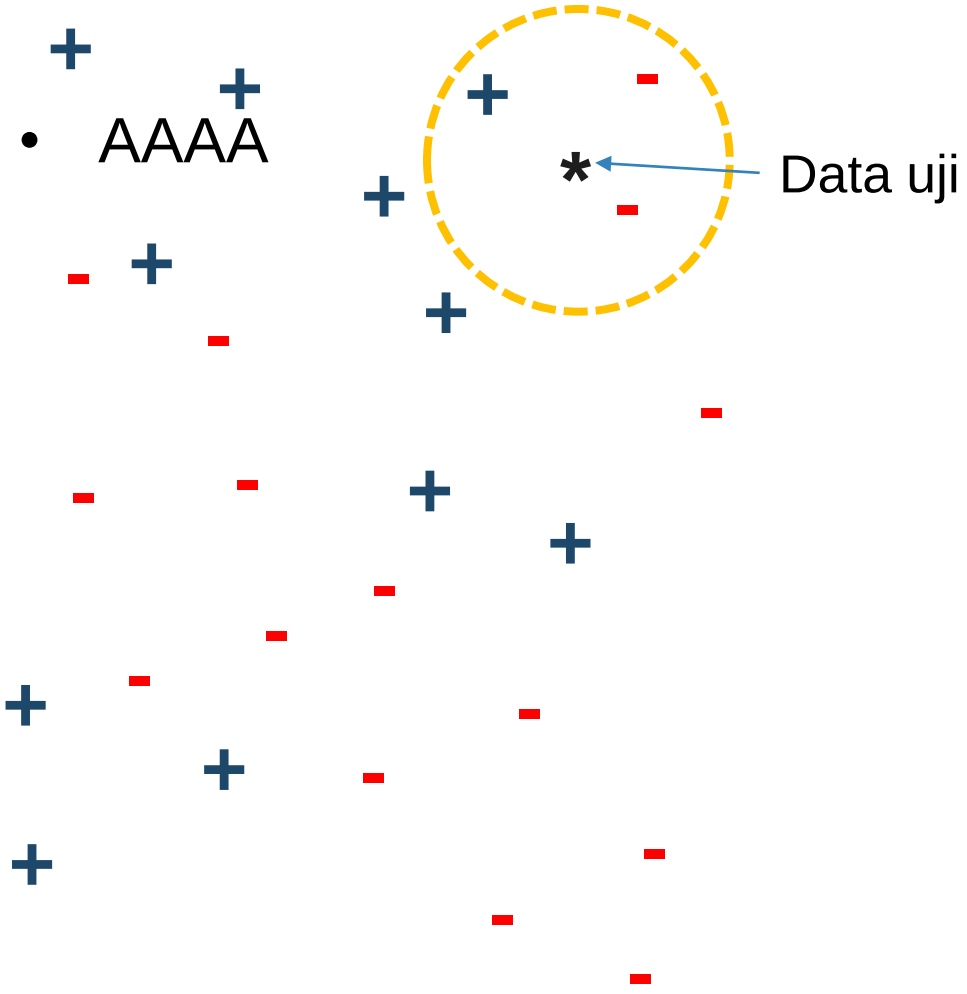
Sepal length	Sepal width	Petal length	Petal Width	Species
4,5	3,1	1,5	0,1	Iris-virginica

Bagaimana jika terdapat kesalahan label pada data latih?

Permasalahan Nearest Neighbor

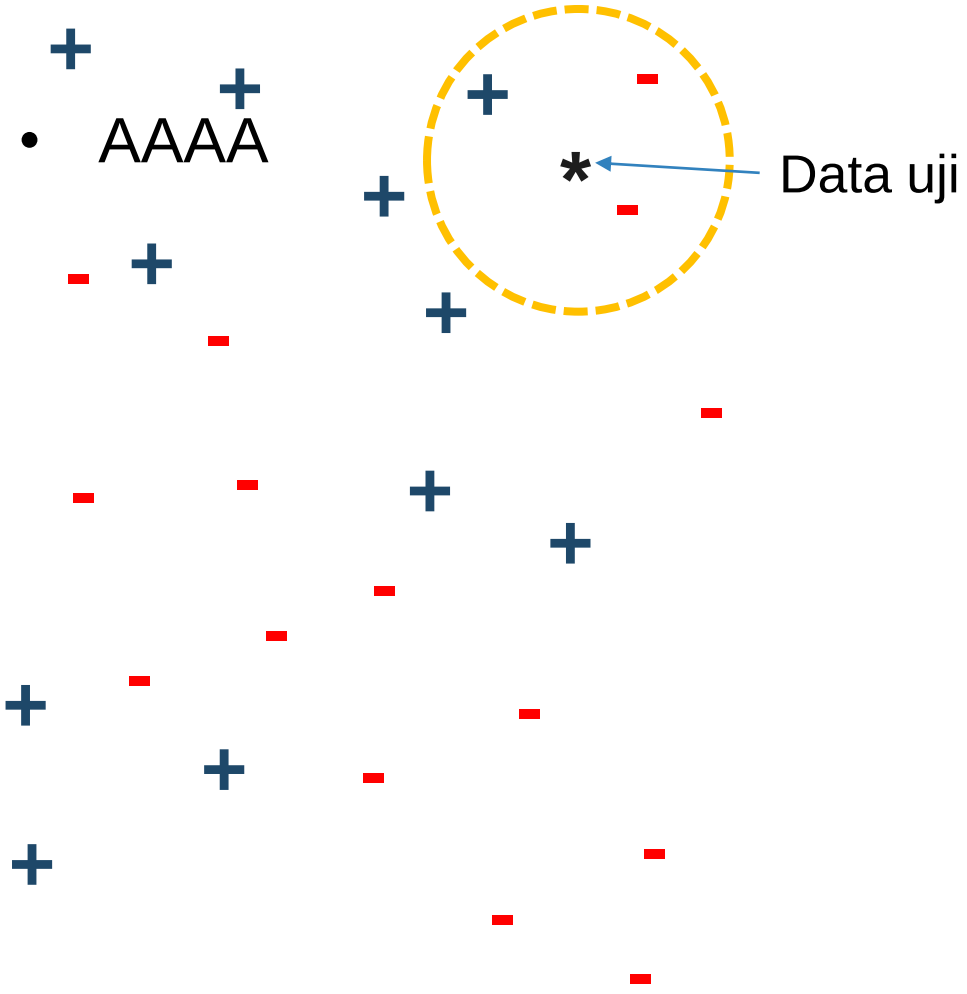
- Nearest neighbor classifier sangat sensitif terhadap noise
- Kesalahan pemberian label pada data latih menyebabkan kesalahan klasifikasi pada data uji
- **IDE** : jangan berpatokan pada satu data termirip, tapi gunakan **beberapa** data yang **paling mirip**

K-Nearest Neighbor (KNN)



- KNN memerlukan tiga komponen :
  - Kumpulan data latih
  - Ukuran jarak untuk menghitung jarak antar data
  - Nilai k : berapa banyak tetangga yang akan diambil

K-Nearest Neighbor (KNN)



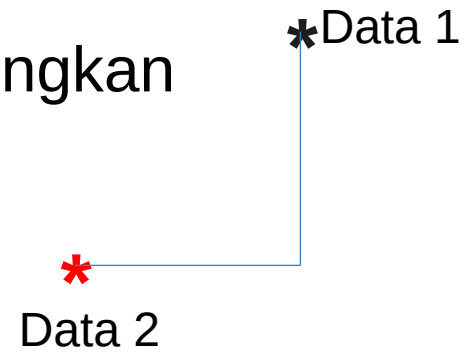
- Cara kerja KNN
  - Hitung jarak data uji ke **setiap** data latih.
  - Ambil **k** data latih yang paling dekat (memiliki jarak terkecil).
  - Tentukan kelas data uji menggunakan **mayoritas** kelas data latih.

Ukuran jarak

- Ukuran jarak digunakan untuk mengetahui **kemiripan** antara dua data
- Jarak besar = tidak mirip, jarak kecil = mirip
- Metode perhitungan jarak yang sering digunakan:
  - Manhattan/ City block Distance
  - Euclidean Distance
  - Minkowski Distance

Manhattan distance

- Manhattan distance menghitung jarak dua vektor data berdasarkan panjang total dari proyeksi garis menghubungkan kedua vektor pada masing-masing axis
- Perhitungan menggunakan jumlah selisih absolut dari kedua vektor



$$d = \sum_{i}^{n} |x_i - y_i|$$

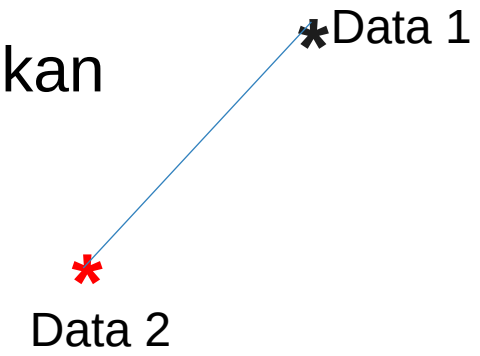
Manhattan Distance

Sepal length	Sepal width	Petal length	Petal Width
5,1	3,5	1,4	0,2
6,4	2,7	5,3	1,9

- $d = |5{,}1 - 6{,}4| + |3{,}5 - 2{,}7| + |1{,}4 - 5{,}3| + |0{,}2 - 1{,}9| = 7{,}7$

Euclidean distance

- Euclidean distance menghitung jarak dua vektor data berdasarkan panjang dari garis lurus yang menghubungkan kedua vektor
- Perhitungan menggunakan akar dari jumlah kuadrat selisih kedua vektor



$$d = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$$

Euclidean Distance

Sepal length	Sepal width	Petal length	Petal Width
5,1	3,5	1,4	0,2
6,4	2,7	5,3	1,9

- $d = \sqrt{(5{,}1 - 6{,}4)^2 + (3{,}5 - 2{,}7)^2 + (1{,}4 - 5{,}3)^2 + (0{,}2 - 1{,}9)^2} = 4{,}52$

Minkowski Distance

- Minkowski distance merupakan bentuk umum dari fungsi jarak dua vektor

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- Jika p = 1, nilainya sama dengan cityblock distance
- Jika p =2, nilainya sama dengan Euclidean distance

Algoritma KNN

- Berdasarkan sekelompok data latih x_i dan sebuah data uji y, lakukan langkah sebagai berikut:
  1. Tentukan nilai *k*
  2. Hitung jarak dari y ke semua x_i
  3. Pilihlah *k* buah x_i yang memiliki jarak terkecil ke y
  4. Lakukan majority voting untuk menentukan kelas dari y, berdasarkan data dari langkah 3

Algoritma KNN

Sepal length	Sepal width	Petal Length	Petal width	Species
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor
6,6	3	4,4	1,4	Iris-versicolor
6,8	2,8	4,8	1,4	Iris-versicolor
6,7	3	5	1,7	Iris-versicolor
6	2,9	4,5	1,5	Iris-versicolor
6,7	2,5	5,8	1,8	Iris-virginica
6,4	2,7	5,3	1,9	Iris-virginica
7,7	3	6,1	2,3	Iris-virginica
6,3	3,4	5,6	2,4	Iris-virginica
6,4	3,1	5,5	1,8	Iris-virginica
6	3	4,8	1,8	Iris-virginica

Sepal length	Sepal width	Petal length	Petal Width	Species
6.1	3	4.9	1.8	??????

Langkah 1 : Tentukan nilai *k*

$$k = 3$$

Algoritma KNN

Sepal length	Sepal width	Petal Length	Petal width	Species	Jarak Euclidean
5,1	3,5	1,4	0,2	Iris-setosa	4,007
4,9	3	1,4	0,2	Iris-setosa	4,031
4,7	3,2	1,3	0,2	Iris-setosa	4,186
4,6	3,1	1,5	0,2	Iris-setosa	4,047
4,3	3	1,1	0,1	Iris-setosa	4,535
5,8	4	1,2	0,2	Iris-setosa	4,164
7	3,2	4,7	1,4	Iris-versicolor	1,025
5,5	2,3	4	1,3	Iris-versicolor	1,382
6,6	3	4,4	1,4	Iris-versicolor	0,812
6,8	2,8	4,8	1,4	Iris-versicolor	0,837
6,7	3	5	1,7	Iris-versicolor	0,616
6	2,9	4,5	1,5	Iris-versicolor	0,520
6,7	2,5	5,8	1,8	Iris-virginica	1,192
6,4	2,7	5,3	1,9	Iris-virginica	0,592
7,7	3	6,1	2,3	Iris-virginica	2,062
6,3	3,4	5,6	2,4	Iris-virginica	1,025
6,4	3,1	5,5	1,8	Iris-virginica	0,678
6	3	4,8	1,8	Iris-virginica	0,141

Sepal length	Sepal width	Petal length	Petal Width	Species
6.1	3	4.9	1.8	??????

Langkah 2 : Hitung jarak data uji ke data latih

$$k = 3$$

Algoritma KNN

Sepal length	Sepal width	Petal Length	Petal width	Species	Jarak Euclidean
5,1	3,5	1,4	0,2	Iris-setosa	4,007
4,9	3	1,4	0,2	Iris-setosa	4,031
4,7	3,2	1,3	0,2	Iris-setosa	4,186
4,6	3,1	1,5	0,2	Iris-setosa	4,047
4,3	3	1,1	0,1	Iris-setosa	4,535
5,8	4	1,2	0,2	Iris-setosa	4,164
7	3,2	4,7	1,4	Iris-versicolor	1,025
5,5	2,3	4	1,3	Iris-versicolor	1,382
6,6	3	4,4	1,4	Iris-versicolor	0,812
6,8	2,8	4,8	1,4	Iris-versicolor	0,837
6,7	3	5	1,7	Iris-versicolor	0,616
6	2,9	4,5	1,5	Iris-versicolor	0,520
6,7	2,5	5,8	1,8	Iris-virginica	1,192
6,4	2,7	5,3	1,9	Iris-virginica	0,592
7,7	3	6,1	2,3	Iris-virginica	2,062
6,3	3,4	5,6	2,4	Iris-virginica	1,025
6,4	3,1	5,5	1,8	Iris-virginica	0,678
6	3	4,8	1,8	Iris-virginica	0,141

Sepal length	Sepal width	Petal length	Petal Width	Species
6.1	3	4.9	1.8	??????

Langkah 3 : Pilih *k* data dengan jarak terkecil

$$k = 3$$

Permasalahan KNN

Sepal length	Sepal width	Petal Length	Petal width	Species	Jarak Euclidean
5,1	3,5	1,4	0,2	Iris-setosa	4,007
4,9	3	1,4	0,2	Iris-setosa	4,031
4,7	3,2	1,3	0,2	Iris-setosa	4,186
4,6	3,1	1,5	0,2	Iris-setosa	4,047
4,3	3	1,1	0,1	Iris-setosa	4,535
5,8	4	1,2	0,2	Iris-setosa	4,164
7	3,2	4,7	1,4	Iris-versicolor	1,025
5,5	2,3	4	1,3	Iris-versicolor	1,382
6,6	3	4,4	1,4	Iris-versicolor	0,812
6,8	2,8	4,8	1,4	Iris-versicolor	0,837
6,7	3	5	1,7	Iris-versicolor	0,616
6	2,9	4,5	1,5	Iris-versicolor	0,520
6,7	2,5	5,8	1,8	Iris-virginica	1,192
6,4	2,7	5,3	1,9	Iris-virginica	0,592
7,7	3	6,1	2,3	Iris-virginica	2,062
6,3	3,4	5,6	2,4	Iris-virginica	1,025
6,4	3,1	5,5	1,8	Iris-virginica	0,678
6	3	4,8	1,8	Iris-virginica	0,141

Sepal length	Sepal width	Petal length	Petal Width	Species
6.1	3	4.9	1.8	??????

Jika k=7, kelas apa yang sesuai untuk data uji?

Permasalahan KNN

- Berapa nilai *k* yang tepat?
  - Jika *k* terlalu kecil, sensitive terhadap *noise/outlier* (ingat metode nearest-neighbor!!!)
  - Jika *k* terlalu besar, penentuan kelas dapat dipengaruhi oleh data dari kelas yang berbeda.
- Kompleksitas tinggi
  - Perhitungan jarak dilakukan terhadap **setiap** data latih.
  - Semakin banyak data latih, semakin lama prosesnya
- KNN merupakan metode *lazy learner*
  - Proses **belajar** (perhitungan jarak)selalu dilakukan untuk setiap data uji.
  - Sebenarnya, **tidak ada proses training** dalam KNN

DISKUSI

TERIMA KASIH