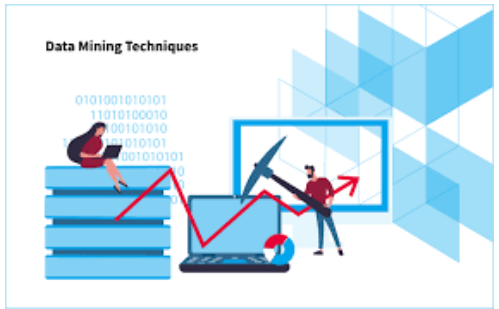
DATA MINING

04 - Algoritma K-NN

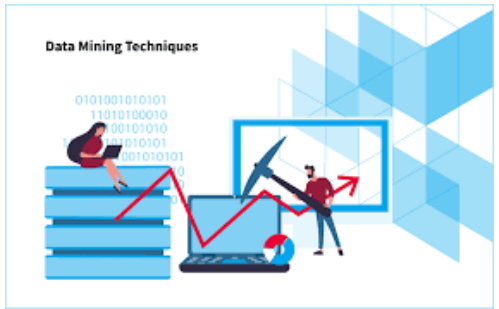
Oleh: Leny Tritanto N., M.Kom.



Algoritma Supervised dan UnsupervisedLearning:

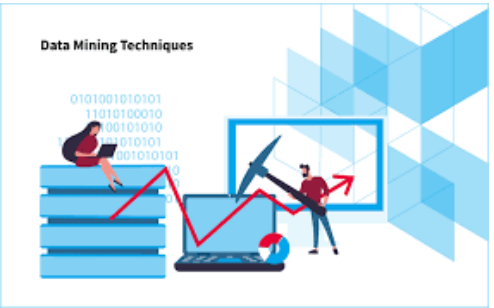
Supervised Learning	Unsupervised Learning
▪ Linear Regression	▪ K-Means
✓ Decision Tree and Random Forest	▪ Hierarchical Clustering
✓ Naive Bayes Classifier	▪ DBSCAN
→ Nearest Neighbour Classifier (KNN)	▪ Association Rule
▪ Artificial Neural Network	▪ Apriori Algorithm
▪ Support Vector Machine (SVM)	

UNBIN



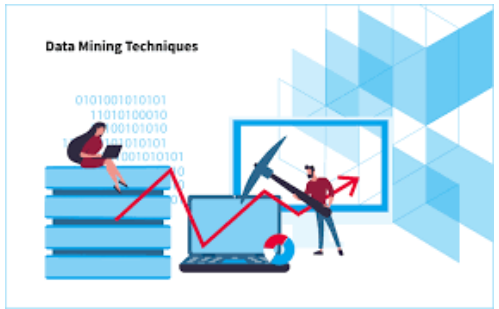
Metode Klasifikasi Nearest Neighbor (K-NN)?

UNBIN



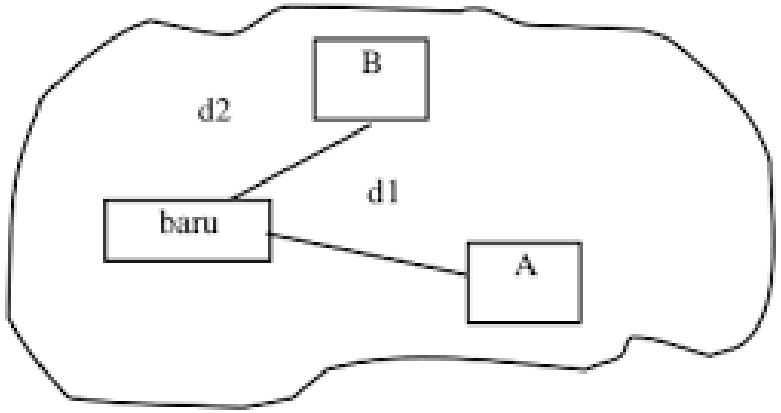
APA ITU K-NEAREST NEIGHBORS?

- Termasuk ke dalam pembelajaran supervised learning atau ada kelas output
- Menggunakan distance function atau similarity metric
- Bisa dipakai untuk permasalahan klasifikasi dan regresi
- Digunakan untuk mengklasifikasi atau memprediksi suatu kelas tertentu berdasarkan kelas mayoritas ketetanggaan terdekat.

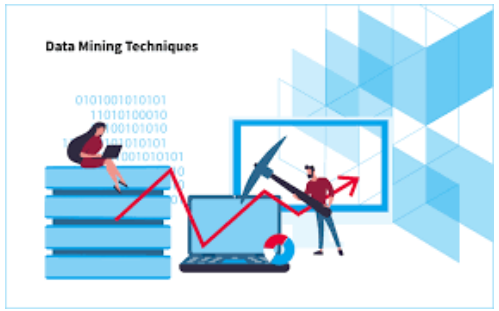


BAGAIMANA PENERAPAN K-NN?

- Pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama
- Berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada
- Misalkan: akan dicari solusi terhadap seorang pasien baru dengan menggunakan solusi dari pasien lama
- Untuk mencari kasus pasien mana yang akan digunakan, maka **dihitung kedekatan** kasus pasien baru dengan semua kasus pasien lama
- Kasus pasien lama dengan **kedekatan terbesarlah** yang akan diambil solusinya untuk digunakan pada kasus pasien baru



UNBIN

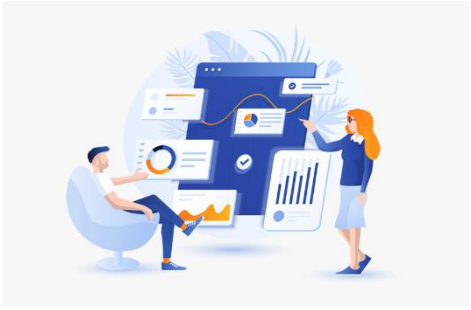


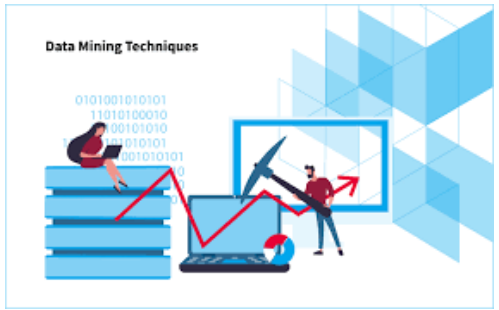
Contoh dataset yang dapat digunakan pada KNN

Input variabel

Output/Class

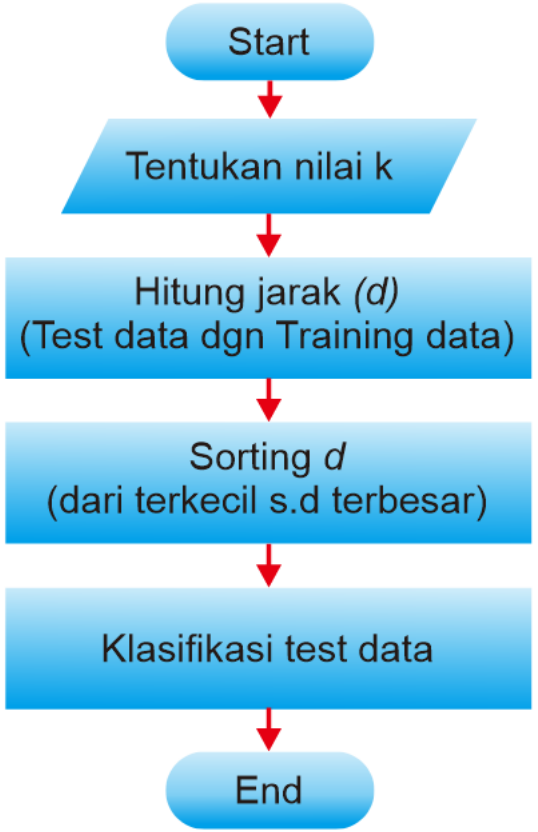
Cuaca	Angin	Keputusan Main
Cerah	Kencang	Tidak
Mendung	Lemah	Ya
Hujan	Lemah	Ya
Cerah	Kencang	Ya
Cerah	Kencang	Ya
Mendung	Kencang	Ya
Hujan	Kencang	Tidak
Hujan	Lemah	Tidak
Cerah	Lemah	Ya
Hujan	Kencang	Ya
Cerah	Kencang	Tidak
Mendung	Lemah	Ya
Mendung	Lemah	Ya
Hujan	Kencang	Tidak





Flowchart Algoritma K-NN

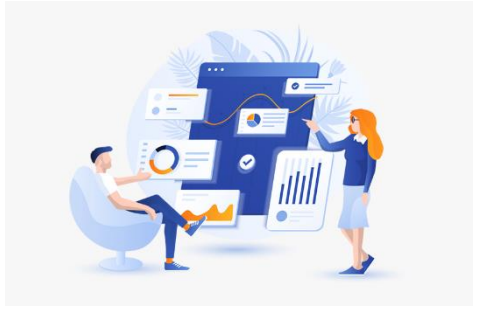
- K pada K-NN terkait dengan jumlah tetangga terdekat yang dipilih
- Nilai ini harus ditentukan di awal
- Umumnya K yang dipilih berjumlah ganjil untuk menghindari munculnya jumlah jarak yang sama
- Umumnya nilai K akan bertambah sebanding dengan jumlah dataset
- Nilai K yang baik dapat dipilih dengan optimasi parameter menggunakan *cross validation*

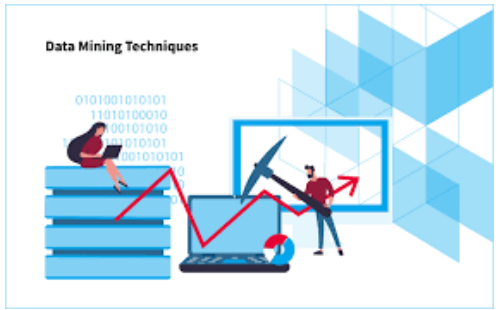


Training data

Test data

Cuaca	Angin	Keputusan Main
Cerah	Kencang	Tidak
Mendung	Lemah	Ya
Hujan	Lemah	Ya
Cerah	Kencang	Ya
Cerah	Kencang	Ya
Mendung	Kencang	Ya
Hujan	Kencang	Tidak
Hujan	Lemah	Tidak
Cerah	Lemah	Ya
Hujan	Kencang	Ya
Cerah	Kencang	Tidak
Mendung	Lemah	Ya
Mendung	Lemah	Ya
Hujan	Kencang	Tidak
Hujan	Lemah	?





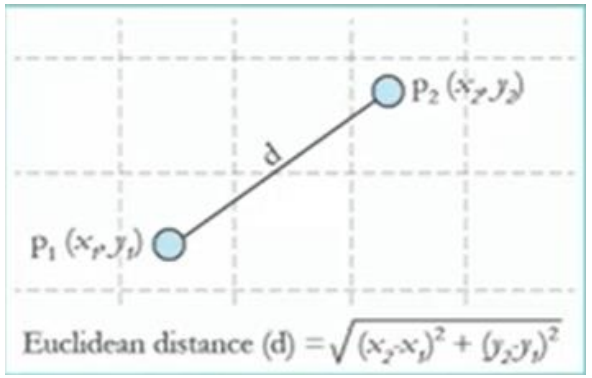
RUMUS MENGHITUNG JARAK

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

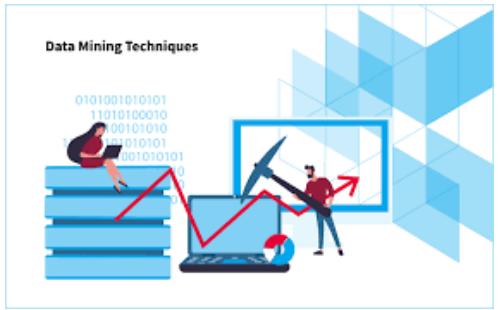
Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$



Rumus jarak yang digunakan:

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

UNBIN



LANJUT PADA LATIHAN STUDI KASUS MENGGUNAKAN MICROSOFT EXCEL