Pengenalan KLasifikasi

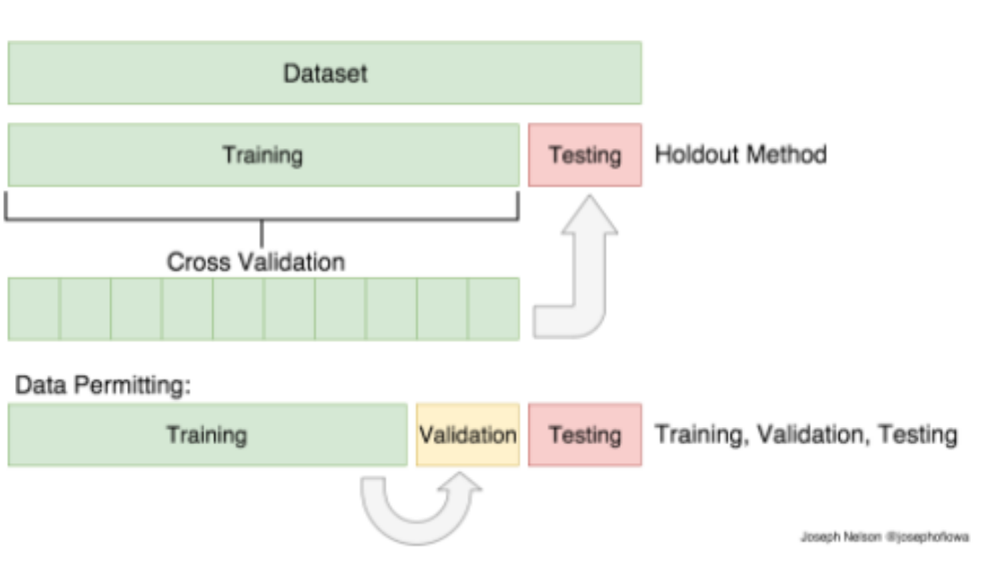
Apa itu Klasifikasi?

Klasifikasi dapat didefinisikan sebagai proses memprediksi kelas atau kategori dari nilai yang diamati atau titik data yang diberikan. Keluaran yang dikategorikan seperti “Yes” atau “No”, “Hitam” atau “Putih” dan “spam” atau “tidak spam”. Algoritma klasifikasi:

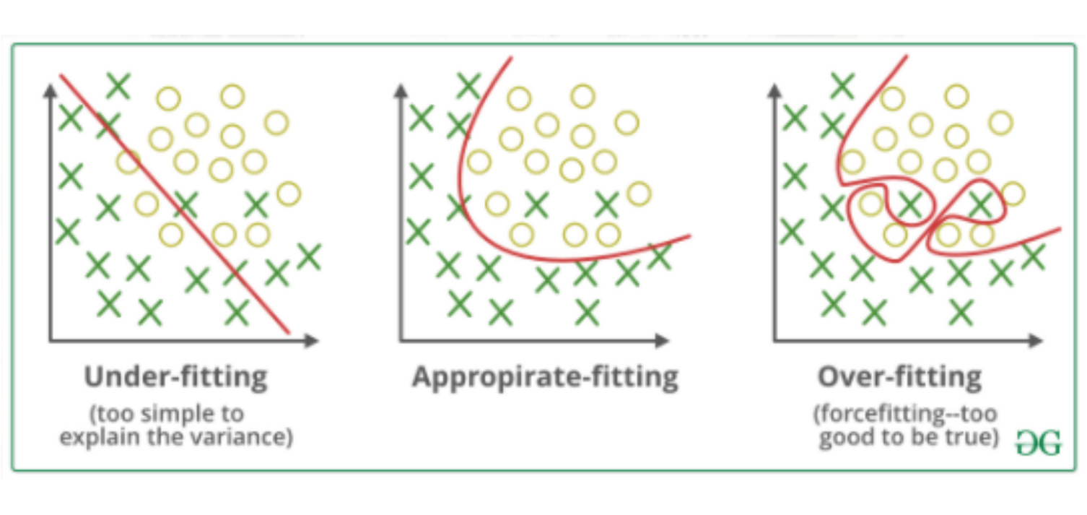
1. Regresi logistik
2. Support Vector Machine(SVM)
3. Decision Tree
4. Naïve Bayes
5. Random Forest

dll

Partition Data



General Problems in Machine Learning Model



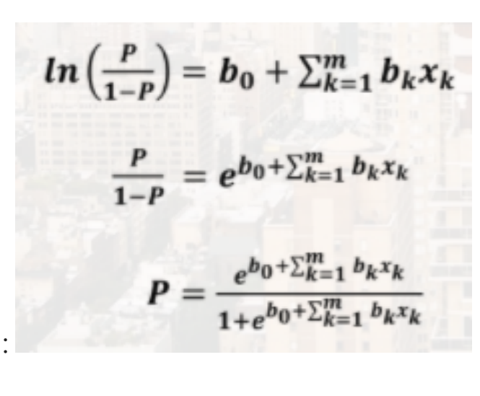
Akurasi vs Interpretasi

Regresi Logistik

Apa itu regresi logistik?

“Regresi logistik merupakan analisis regresi yang digunakan ketika variabel dependen berupa variabel kategorik.”

Persamaan regresi logistik :



$$ln\left(\frac{P}{1-P}\right) = b_0 + \sum_{k=1}^{m} b_k x_k$$

$$\frac{P}{1-P} = e^{b_0 + \sum_{k=1}^{m} b_k x_k}$$

$$P = \frac{e^{b_0 + \sum_{k=1}^{m} b_k x_k}}{1+e^{b_0 + \sum_{k=1}^{m} b_k x_k}}$$

```
Hasil regresi logistik 0 dan 1. Jika hasil prediksi > 0,5 maka akan dibulatkan menjadi 1. Jika hasil
prediksi < 0,5 maka akan dibulatkan menjadi 0.
```

Jenis regresi logistik

1. Biner atau binomial

Merupakan jenis klasifikasi yang mana variabel dependen hanya memiliki dua kemungkinan label yakni 1 atau 0. Misalnya sukses atau gagal, iya atau tidak, menang atau kalah, dll.

Menggunakan distribusi binomial yang mana outputnya nominal dan berjumlah 2, misal yes dan no.

1. Multinomial

Merupakan jenis klasifikasi yang mana variabel dependen dapat terdiri dari 3 atau lebih kemungkinan label yang tak berurutan dan tidak memiliki signifikansi kuantitatif. Misalnya “jenis A” atau “jenis B” atau “jenis C”.

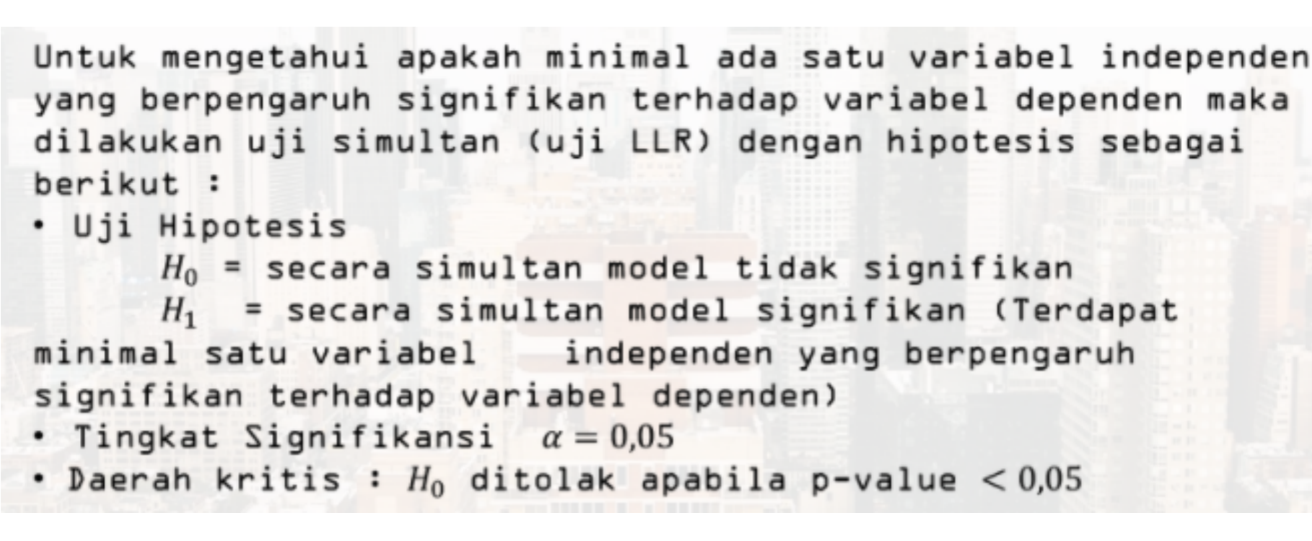
Menggunakan distribusi binomial yang mana outputnya nominal dan berjumlah > 2.

1. Ordinal

Merupakan jenis klasifikasi yang mana variabel dependen dapat terdiri dari 3 atau lebih kemungkinan label yang berurutan dan memiliki signifikansi kuantitatif. Misalnya “buruk”, “baik”, “sangat baik”, atau “luar biasa” dan setiap kategori dapat mempunyai nilai seperti 0, 1, 2, 3.

Outputnya berupa Kategori Ordinal yang mana terdapat tingkatkan.

Uji Simultan (Uji LLR)

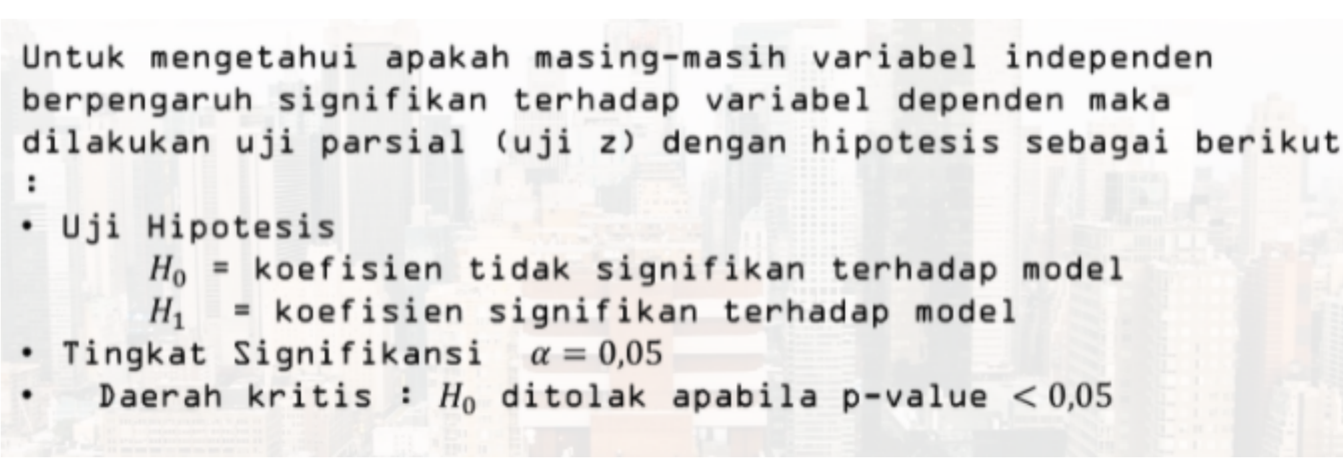


Untuk mengetahui apakah minimal ada satu variabel independen yang berpengaruh signifikan terhadap variabel dependen maka dilakukan uji simultan (uji LLR) dengan hipotesis sebagai berikut :

- Uji Hipotesis
  H_0 = secara simultan model tidak signifikan
  H_1 = secara simultan model signifikan (Terdapat minimal satu variabel independen yang berpengaruh signifikan terhadap variabel dependen)
- Tingkat Signifikansi $\alpha = 0{,}05$
- Daerah kritis : H_0 ditolak apabila p-value $< 0{,}05$

Jika di regresi linear menggunakan Uji F sedangkan di Uji regresi logistik menggunakan Uji LLR.

Uji Parsial (Uji z)



Untuk mengetahui apakah masing-masih variabel independen berpengaruh signifikan terhadap variabel dependen maka dilakukan uji parsial (uji z) dengan hipotesis sebagai berikut :

- Uji Hipotesis
  H_0 = koefisien tidak signifikan terhadap model
  H_1 = koefisien signifikan terhadap model
- Tingkat Signifikansi $\alpha = 0{,}05$
- Daerah kritis : H_0 ditolak apabila p-value $< 0{,}05$

Jika di regresi linear menggunakan Uji t sedangkan di Uji regresi logistik menggunakan Uji z.

Asumsi regresi logistik

Dikarenakan regresi logistik merupakan metode parametrik, maka terdapat asumsi yang harus dipenuhi. Yaitu:

- Tidak ada multikolinearitas dalam model yang berarti setiap variabel independen tidak saling berkorelasi/berhubungan satu sama lain.

Non Multikolinearitas

Variabel independen dikatakan tidak saling berhubungan atau berkorelasi satu sama lain apabila memiliki nilai Variance Inflation Factor (VIF) kurang dari 10 dan korelasi kurang dari +-0.8.

Kelebihan dan kekurangan regresi logistik

Kelebihan :

1. Penerapan yang mudah dan efektif.

2. Tidak diperlukan tuning hyperparameter

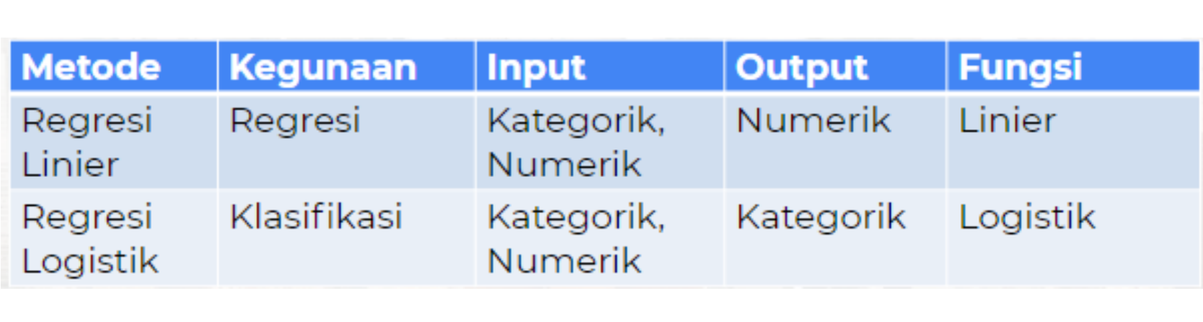
Kekurangan

1. Performa yang buruk pada data non linear

2. Performa yang buruk dengan fitur yang tidak relevan dan berkorelasi tinggi

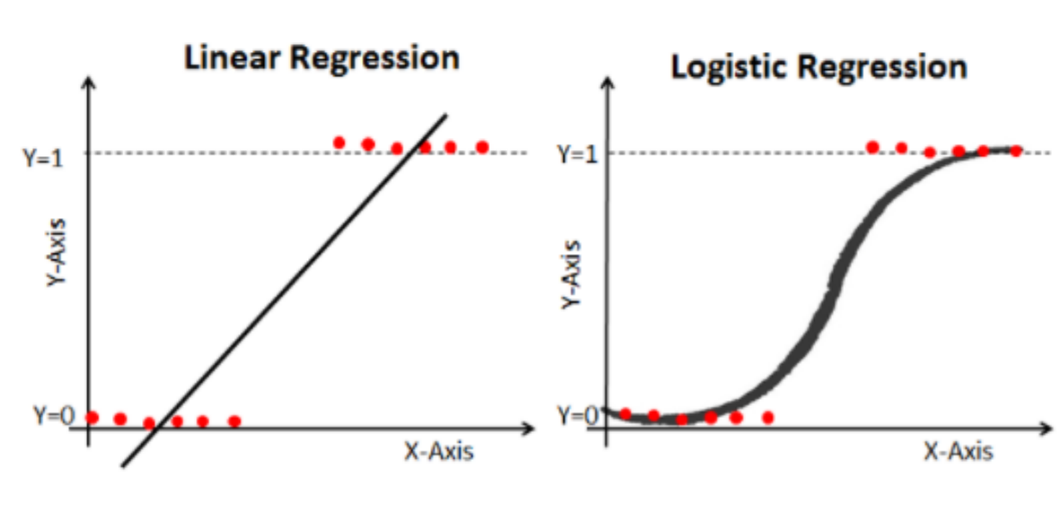
3. Algoritma yang tidak terlalu kuat dan mudah diungguli oleh algoritma lainnya.

Perbedaan regresi logistik dengan regresi linear



Metode	Kegunaan	Input	Output	Fungsi
Regresi Linier	Regresi	Kategorik, Numerik	Numerik	Linier
Regresi Logistik	Klasifikasi	Kategorik, Numerik	Kategorik	Logistik

Perbedaan regresi logistik dengan regresi linear



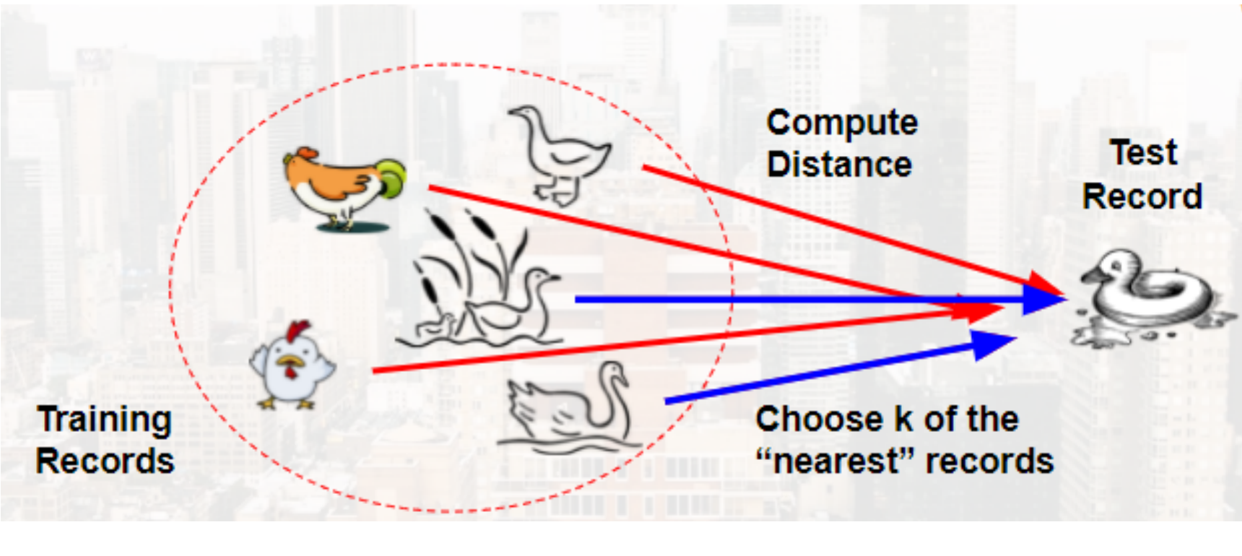
K-Nearest Neighbours

Jumlah tetangga terdekat. Jika K = 5 berarti 5 tetangga terdekat.

Apa itu K-Nearest Neighbours (KNN)

Algoritma K-nearest neighbours (KNN) adalah jenis supervised ML yang dapat digunakan baik untuk klasifikasi maupun regresi. KNN adalah algoritma lazy classifier karena tidak memiliki fase pelatihan khusus dan menggunakan semua data untuk pelatihan saat klasifikasi. KNN juga merupakan algoritma pembelajaran non-parametrik.

Ide Dasar KNN



Ide dasar menggunakan tetangga terdekat. Misal kita memiliki data training yang berada di dalam lingkaran merah. Misal data input kita merupakan ciri-ciri hewan, sedangkan outputnya berapa kategori, contohnya ayam dan bebek.

Kenapa disebut lazy classifier, karena metode ini mau bekerja ketika terdapat data baru yang ingin di prediksi. Jadi pada saat proses training, metode ini tidak bekerja sebelum terdapat data baru yang ingin di prediksi.

Cara agar metode ini mau running adalah pada data baru (Test Record), metode ini akan mencari jarak terdekat antara data baru, dengan data training. Lalu kita tentukan, kita ingin berapa jumlah tentangga terdekat, misalnya k = 3 atau 3 tetangga terdekat. Dari ketiga tetangga tersebut, kita lakukan voting jumlah terbanyak. Berdasarkan gambar diketahui bahwa paling banyak bebek, maka output yang muncul dari data baru tersebut diklasifikasikan sebagai bebek.

Algoritma KNN

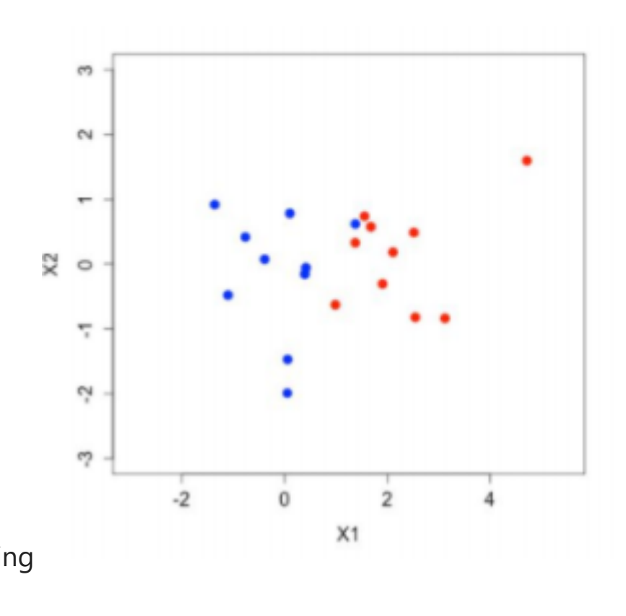
langkah-langkah algoritma:

1. Langkah 1 - Untuk mengimplementasikan algoritma apa pun, kami membutuhkan kumpulan data. Jadi langkah pertama KNN, kita harus memuat data pelatihan serta data pengujian.
2. Langkah 2 - Selanjutnya, kita perlu memilih nilai K yaitu titik data terdekat. K bisa berupa bilangan bulat apa pun.
3. Langkah 3 - Untuk setiap poin dalam data pengujian lakukan hal berikut -

> 3.1 - Menghitung jarak antara data pengujian dan setiap baris data pelatihan dengan bantuan salah satu metode yaitu: jarak Euclidean, Manhattan atau Hamming. Metode yang paling umum digunakan untuk menghitung jarak adalah Euclidean.
>
> 3.2 - Sekarang, berdasarkan nilai jarak, urutkan jarak dari yang terkecil ke terbesar.
>
> 3.3 - Selanjutnya, pilih K observasi terdekat (Jarak terdekat).
>
> 3.4 – Prediksi data pengujian dengan melakukan voting dari k observasi terdekat.

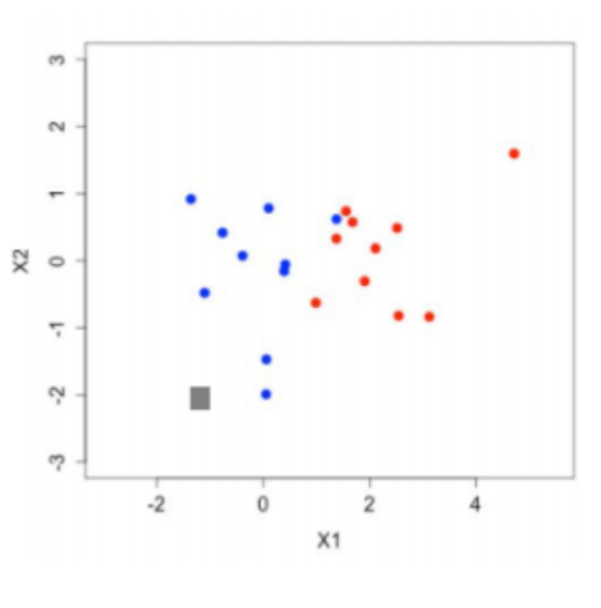
1. Langkah 4 - Selesai

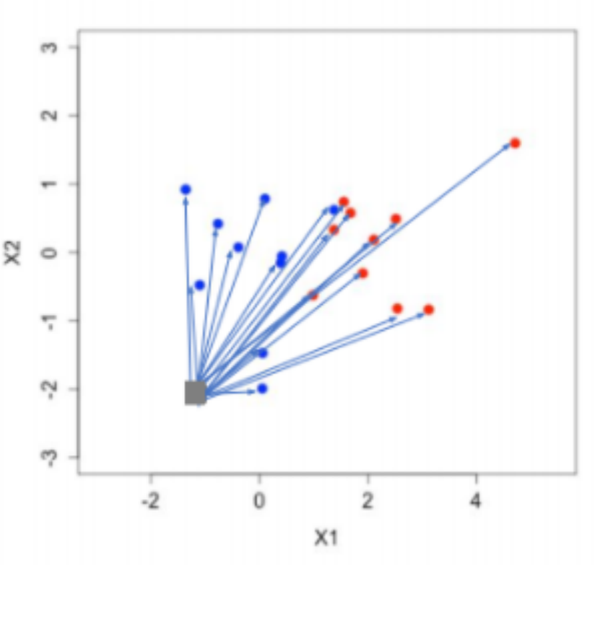
Algoritma KNN



1. Kumpulkan dataset training

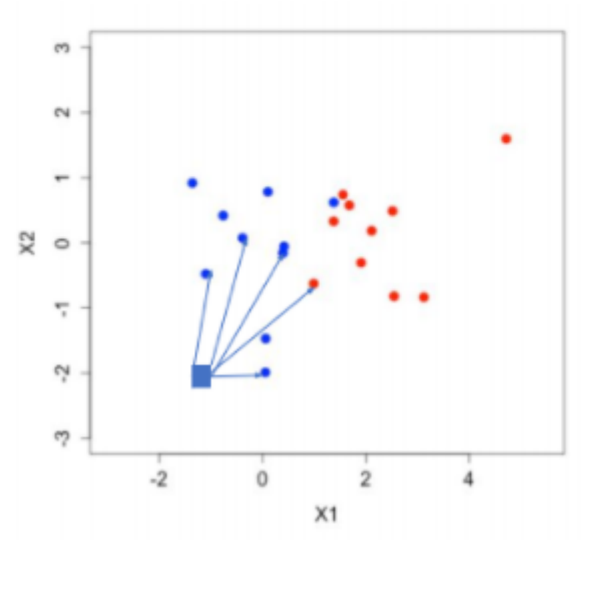
1. Masukkan data observasi baru, misalkan (-1,-2)





1. Bandingkan (Ukur jarak) dari setiap observasi

1. Misal K=6, maka cari 6 observasi yang terdekat. Dan assign sebagai kelas yang voting mayoritas.

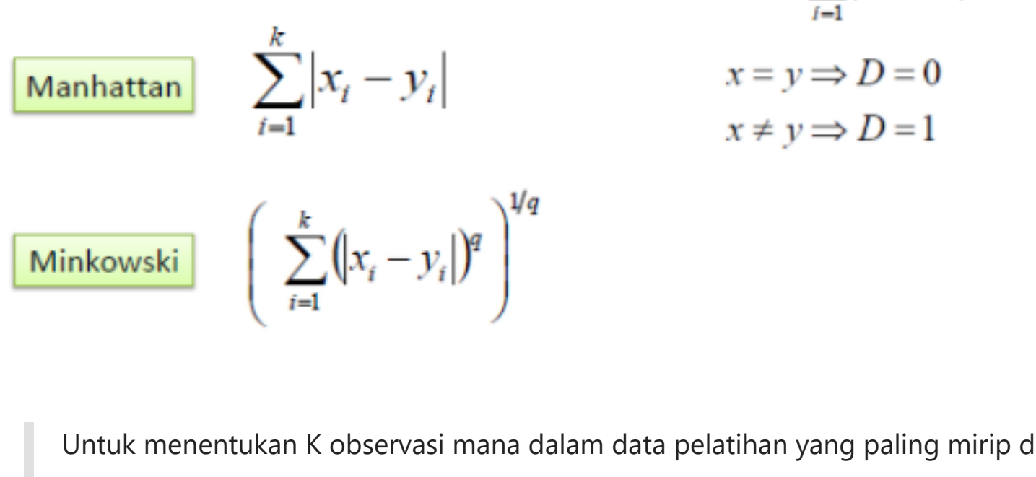


Maka data baru akan diprediksi masuk ke kelas Biru.

Jarak

Jika kita memiliki variabel numerik, maka cara untuk menghitung jarak antar data baru dengan tetangga terdekat, menggunakan 3 cara yaitu: Euclidean, Manhattan, Minkowski.

Jika kita memiliki variabel kategorik, maka cara untuk menghitung jarak antar data baru dengan tetangga terdekat, menggunakan cara Hamming Distance.



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

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

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

Hamming Distance

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

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

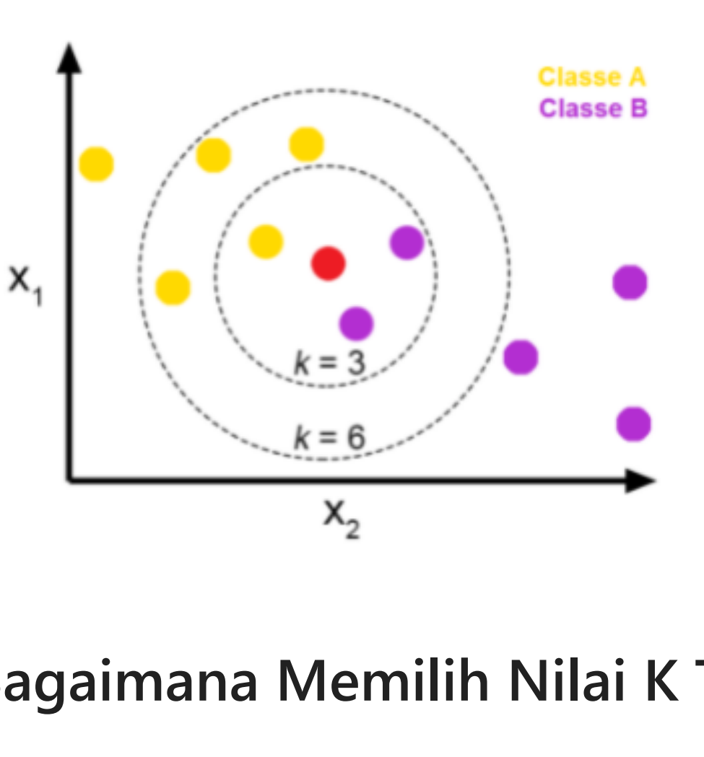
> Untuk menentukan K observasi mana dalam data pelatihan yang paling mirip dengan input baru, pengukur jarak digunakan. Untuk variabel input bernilai numerik, ukuran jarak paling populer adalah jarak Euclidean.
>
> Meskipun jarak Euclidean berguna dalam dimensi rendah, jarak tersebut tidak berfungsi dengan baik dalam dimensi tinggi dan untuk variabel kategori.

Pemilihan K

Nilai K merupakan Hyperparameter, penentuan model terbaik bergantung pada besaran nilai K. Biasanya dipilih angka ganjil agar dapat menghasilkan output yang jelas pada saat voting. Biasanya K nya 3 atau 5.

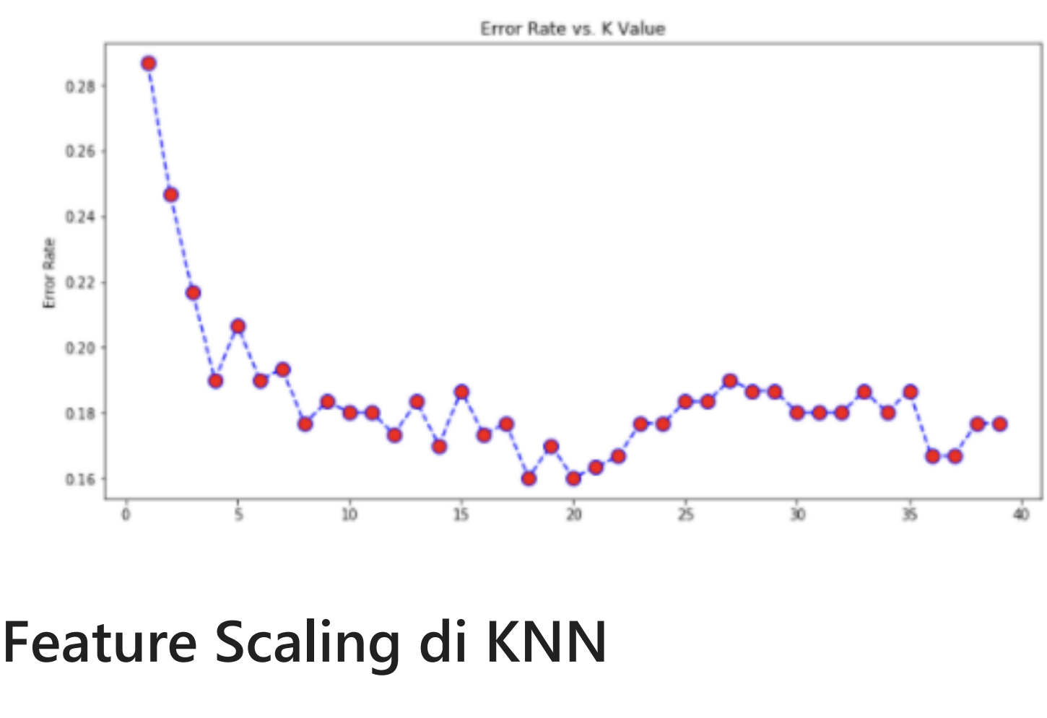
Memilih nilai k:

> Jika k terlalu kecil, sensitif terhadap observasi noise dan subjektif
>
> Jika k terlalu besar, Neighbour dapat menyertakan observasi dari kelas lain



Bagaimana Memilih Nilai K Terbaik?

Menggunakan Tunning Hyperparameter. Jadi masing-masing nilai K akan dicoba satu per satu, dan dipilihan yang mana nilai K memiliki nilai error terkecil.



Feature Scaling di KNN

KNN sensitif terhadap skala data, maka harus dilakukan Feature Scaling.

Karena kuadrat dalam Persamaan, jarak Euclidean lebih sensitif terhadap pencilan. Selain itu, sebagian besar pengukuran jarak sensitif terhadap skala fitur. Data dengan fitur yang memiliki skala berbeda akan membiaskan ukuran jarak karena prediktor dengan nilai terbesar akan berkontribusi paling besar pada jarak antara dua sampel.

Pro

1. Sederhana dan intuitif
2. Ini bagus untuk kumpulan data non linier yang dapat dipisahkan
3. Hanya terdapat beberapa hyperparameter (K dan jarak)

Kontra

1. Biaya prediksi tinggi (lebih buruk untuk kumpulan data besar)
2. Tidak bagus dengan data dimensi tinggi
3. Variabel input kategorik tidak berfungsi dengan baik
4. Waktu prediksi lambat

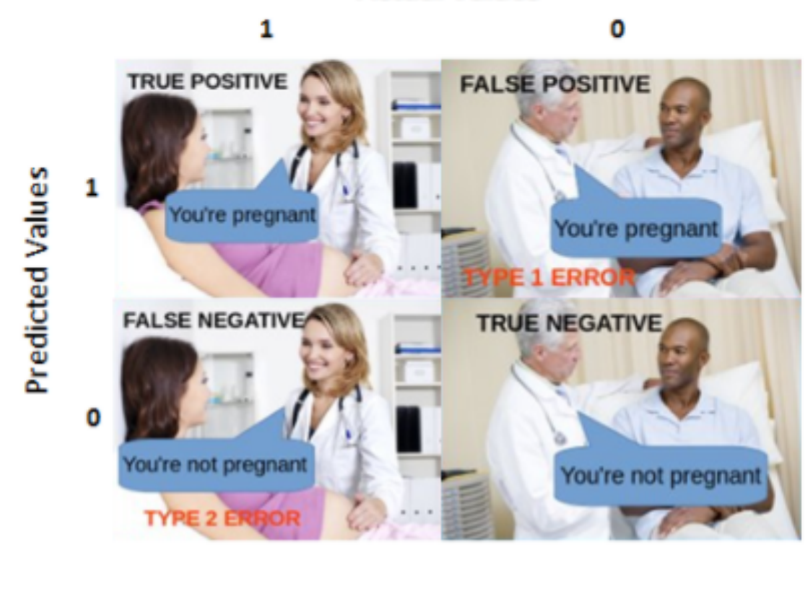
Performa dan Evaluasi Model

Kenapa Evaluasi Model Penting?

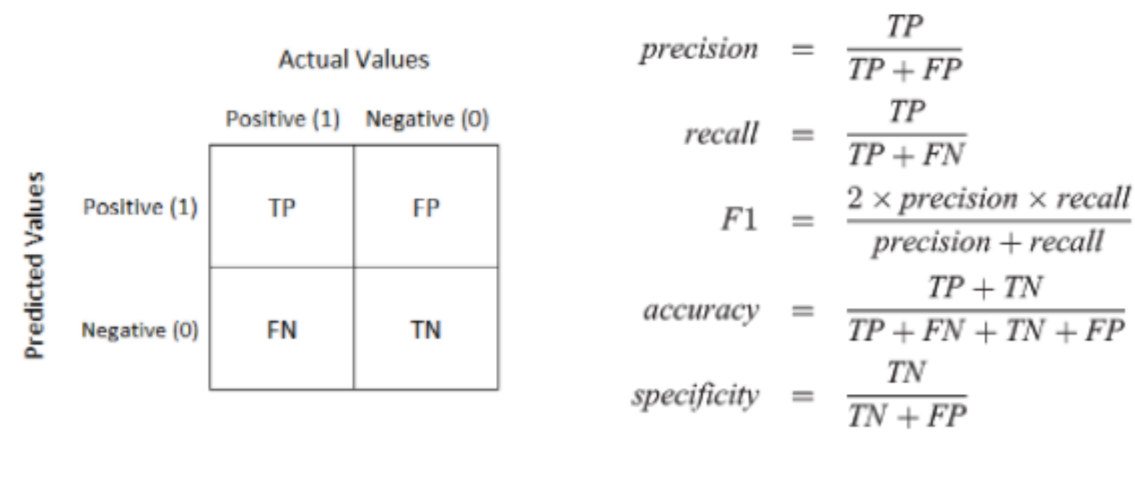
1. Sebagai bagian dari setiap proyek ilmu data
2. Untuk menjelaskan bagaimana model yang baik bekerja dan apakah hasilnya dapat dipercaya atau tidak
3. Untuk membantu memutuskan algoritma mana yang harus digunakan

Confusion Matrix

Di Klasifikasi, evaluasi model menggunakan Confusion Matrix.



Kriteria Evaluasi



Actual Values

Predicted Values

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP}$$

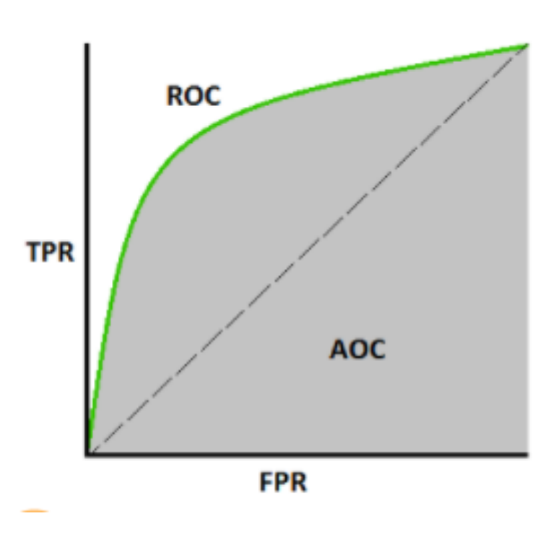
$$specificity = \frac{TN}{TN+FP}$$

Recall: Untuk melihat berapa persen label positif atau yes. Specifity: sama seperti Recall, tetapi untuk label negatif atau no. Accuracy: Untuk melihat berapa persen tingkat ke akuratan. Misalnya memiliki data benar 80, jumlah data 100. Maka tingkat akurasinya 80/100 atau 80%. Precision: Hampir sama seperti recall. Misal dari 100 data, kita memprediksi 60 data positif, 40 data negatif. Dari 60 data positif kita mengetahui bahwa 40 data benar positif sedangkan 20 data salah prediksi atau seharusnya negatif.

ROC/AUC

Outputnya berupa plot. AUC singkatan dari Area Under Curve. Untuk menentukan bagus atau tidaknya model. AUC menghitung besaran luas. Misalnya luas AUC hanya berada di garis titik titik, maka AUCnya hanya 0.5

AUC dibangun menggunakan Recall (sumbu y) dan 1 - Specificity (sumbu x)



TPR= Recall= Sensitivity FPR=1-Spesificity

AUC thumb rules:

1. .90-1 = excellent (A)
2. .80-.90 = good (B)
3. .70-.80 = fair (C)
4. .60-.70 = poor (D)
5. .50-.60 = fail (F)