Pertemuan 9

Pembelajaran Mesin (Machine Learning)

A. Supervised vs Unsupervised Learning

Pembelajaran mesin (Machine Learning) secara umum dibagi menjadi dua pendekatan utama: **Supervised Learning** dan **Unsupervised Learning**. Keduanya digunakan untuk tujuan yang berbeda tergantung pada jenis data dan permasalahan yang ingin diselesaikan.

1. Supervised Learning

Supervised Learning adalah teknik pembelajaran di mana model dilatih menggunakan data yang memiliki label, yaitu setiap contoh data sudah disertai jawaban atau kategori yang benar. Melalui pasangan input–output tersebut, model belajar mengenali pola, hubungan, serta karakteristik data sehingga mampu membuat prediksi yang akurat terhadap data baru yang belum pernah dilihat sebelumnya. Pendekatan ini umum digunakan pada masalah klasifikasi dan regresi yang membutuhkan hasil prediksi terukur dan terverifikasi.

Artinya, setiap data input memiliki target output yang benar, sehingga model belajar memetakan input → output.

Contoh:

Jika diberikan data mahasiswa dengan label *"Lulus Tepat Waktu"* atau *"Tidak Tepat Waktu"*, maka model belajar mengenali pola untuk memprediksi label tersebut.

Membangun model prediktif yang mampu memprediksi nilai atau kelas dari data baru, yaitu merancang sistem yang dapat menggeneralisasi pola dari data latihan sehingga dapat memberikan hasil prediksi yang tepat ketika dihadapkan pada data yang belum pernah dipelajari sebelumnya. Tujuan ini mencakup kemampuan model untuk memahami hubungan input–output secara konsisten,

mengurangi kesalahan prediksi, serta memberikan estimasi yang andal dalam berbagai situasi nyata.

Contoh Algoritma

- k-Nearest Neighbor (k-NN)
- Decision Tree
- Support Vector Machine (SVM)
- Naïve Bayes
- Linear Regression / Logistic Regression

Contoh Kasus Nyata

- Memprediksi penyakit berdasarkan gejala.
- Menentukan kategori email: *spam* atau *non-spam*.
- Memprediksi IPK akhir mahasiswa dari nilai dan kehadiran.
- Deteksi kredit macet pada sistem keuangan.

Kelebihan

- Akurasi tinggi jika dataset cukup dan berkualitas.
- Hasil prediksi dapat dievaluasi dengan jelas karena tersedia label.

Kekurangan

- Membutuhkan dataset yang sudah dilabeli (mahal dan memakan waktu).
- Tidak optimal jika data banyak noise atau tidak terstruktur.

2. Unsupervised Learning

Unsupervised Learning adalah teknik pembelajaran di mana model dilatih menggunakan data tanpa label, sehingga tidak ada informasi jawaban atau kategori yang diberikan sebelumnya. Dalam kondisi ini, model berupaya memahami karakteristik alami dari data secara mandiri. Tujuannya adalah menemukan pola tersembunyi, struktur, atau pengelompokan dalam data, seperti kelompok objek yang mirip, hubungan antarvariabel, atau dimensi penting yang dapat merepresentasikan data secara lebih ringkas. Pendekatan ini sangat berguna

ketika kita ingin melakukan eksplorasi data atau memahami struktur data sebelum melakukan analisis lanjutan.

Contoh:

Mengelompokkan mahasiswa berdasarkan gaya belajar tanpa menentukan kategori sebelumnya.

Contoh Algoritma

- K-Means Clustering
- Hierarchical Clustering
- DBSCAN
- PCA (Principal Component Analysis) untuk reduksi dimensi
- Association Rule Mining (Apriori)

Contoh Kasus Nyata

- Segmentasi pelanggan berdasarkan perilaku belanja.
- Mengelompokkan artikel berita berdasarkan topik.
- Clustering citra dalam computer vision.
- Pengelompokan mahasiswa berdasarkan minat atau kemampuan logis-matematis.

Kelebihan

- Tidak membutuhkan label → lebih mudah dikumpulkan.
- Mampu menemukan pola yang tidak selalu terlihat secara manual.

Kekurangan

- Hasil sulit dievaluasi karena tidak ada label pembanding.
- Interpretasi clustering bisa subjektif.

3. Perbandingan Supervised vs Unsupervised Learning

Aspek	Supervised Learning	Unsupervised Learning
Label Data	Ada (data berlabel)	Tidak ada (data mentah)
Tujuan	Prediksi output baru	Menemukan pola & struktur data
Jenis Masalah	Klasifikasi, regresi	Clustering, asosiasi, deteksi outlier

Contoh Algoritma	k-NN, SVM, Decision Tree	K-Means, PCA, DBSCAN
Evaluasi Model	Akurasi, precision, recall, F1-score	Silhouette Score, Davies–Bouldin Index
Kelebihan	Prediksi akurat & terukur	Menemukan pola tersembunyi
Kekurangan	Perlu label → mahal	Sulit interpretasi & evaluasi

B. Algoritma populer: k-NN, Decision Tree, SVM, Clustering

Pada machine learning, terdapat sejumlah algoritma penting yang sering digunakan untuk berbagai tugas klasifikasi, prediksi, maupun pengelompokan data. Berikut penjelasan konsep, kelebihan, kelemahan, dan rumus dasar dari masing-masing algoritma.

1. k-Nearest Neighbor (k-NN)

k-NN adalah algoritma berbasis kemiripan (similarity-based), di mana kelas dari data baru ditentukan berdasarkan k tetangga terdekat.

a. Prinsip Kerja

- Tentukan nilai *k* (jumlah tetangga).
- Hitung jarak antara data baru dengan seluruh data pada training set.
- Pilih *k* data yang jaraknya paling dekat.
- Voting kelas mayoritas → hasil prediksi.

b. Rumus Jarak

Rumus yang paling umum digunakan adalah Euclidean Distance:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Jika ingin menggunakan Manhattan Distance:

$$d(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$

c. Kelebihan

- Mudah diimplementasikan.
- Tidak membutuhkan training intensif.

d. Kekurangan

- Lambat untuk dataset besar.
- Sensitif terhadap skala data (perlu normalisasi).

2. Decision Tree

Decision Tree adalah model yang menggunakan struktur pohon keputusan untuk memetakan input → output berdasarkan aturan pemisahan (*splitting rules*).

a. Prinsip Kerja

- Dataset dibagi berdasarkan fitur yang memberikan pemisahan terbaik.
- Proses berulang hingga mencapai kondisi berhenti.

b. Rumus Pemilihan Split (Information Gain)

Menggunakan Entropy:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Kemudian:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

c. Rumus Gini Index (Alternatif Entropy)

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2$$

d. Kelebihan

- Mudah dipahami (interpretable).
- Cocok untuk data kategorikal maupun numerik.

e. Kekurangan

- Cenderung overfitting.
- Perubahan kecil pada data bisa mengubah struktur pohon.

3. Support Vector Machine (SVM)

SVM adalah algoritma yang mencari hyperplane terbaik untuk memisahkan kelas dalam ruang fitur.

a. Prinsip Kerja

- Cari garis/pesawat pemisah dengan margin terbesar antara dua kelas.
- Menggunakan kernel untuk data non-linear.

b. Rumus Hyperplane

Jika data linear separable:

$$f(x) = w \cdot x + b$$

Tujuan SVM adalah memaksimalkan margin:

$$Margin = \frac{2}{\|w\|}$$

c. Fungsi Kernel (Contoh RBF Kernel)

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

d. Kelebihan

- Akurasi tinggi, terutama untuk data high-dimensional.
- Mampu menangani data non-linear dengan kernel.

e. Kekurangan

- Tidak cocok untuk dataset sangat besar.
- Parameter tuning (C, gamma) cukup kompleks.

4. Clustering (Unsupervised Learning)

Clustering bertujuan mengelompokkan data berdasarkan kemiripan tanpa label. Algoritma paling umum adalah K-Means Clustering.

a. Prinsip Kerja

- Tentukan jumlah cluster (k).
- Inisialisasi centroid secara acak.
- Hitung jarak setiap data ke centroid terdekat.
- Update centroid menggunakan rata-rata anggota cluster.
- Ulangi hingga konvergen.

b. Rumus Update Centroid

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

c. Fungsi Objektif (Meminimalkan SSE)

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

d. Kelebihan

- Cepat dan sederhana.
- Cocok untuk data besar.

e. Kekurangan

- Sensitif terhadap nilai k dan outlier.
- Hanya efektif untuk cluster berbentuk bulat (spherical).

Rangkuman Singkat

Algoritma	Tipe	Rumus Utama	Kegunaan
k-NN	Supervised	Euclidean Distance	Klasifikasi/prediksi berdasarkan tetangga
Decision Tree	Supervised	Entropy, Gini, Information Gain	Pohon keputusan, interpretasi mudah
SVM	Supervised	Hyperplane & Margin	Pemisahan kelas linear & non-linear
K-Means	Unsupervised	Update centroid, SSE	Pengelompokan data

C. Evaluasi model: akurasi, precision, recall

Evaluasi model machine learning sangat penting untuk mengetahui kualitas prediksi yang dihasilkan. Tiga metrik paling dasar yang digunakan pada masalah klasifikasi adalah Akurasi, Precision, dan Recall. Ketiganya dihitung berdasarkan *Confusion Matrix*, yaitu tabel yang menggambarkan prediksi model terhadap kondisi aktual.

1. Confusion Matrix

Confusion Matrix terdiri dari empat komponen:

	Prediksi Positif	Prediksi Negatif
Aktual Positif	True Positive (TP)	False Negative (FN)
Aktual Negatif	False Positive (FP)	True Negative (TN)

Penjelasan singkat:

- True Positive (TP): Model memprediksi positif dan benar.
- False Positive (FP): Model memprediksi positif tetapi salah.
- False Negative (FN): Model memprediksi negatif tetapi salah.
- True Negative (TN): Model memprediksi negatif dan benar.

2. Akurasi (Accuracy)

a. Pengertian

Akurasi mengukur seberapa banyak prediksi model yang benar dibandingkan seluruh data. Cocok digunakan jika jumlah kelas seimbang.

b. Rumus Akurasi

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

c. Contoh Singkat

Jika model menghasilkan:

TP = 40, TN = 50, FP = 5, FN = 5

$$Accuracy = \frac{40 + 50}{40 + 50 + 5 + 5} = \frac{90}{100} = 0.90 = 90\%$$

3. Precision

a. Pengertian

Precision mengukur tingkat ketepatan prediksi positif model. Tinggi jika FP (prediksi positif yang salah) rendah.

Penting untuk kasus seperti:

- deteksi spam
- deteksi penyakit langka (menghindari false alarm)

b. Rumus Precision

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

c. Contoh Singkat

TP = 40, FP = 10

$$Precision = \frac{40}{40 + 10} = \frac{40}{50} = 0.80 = 80\%$$

4. Recall

a. Pengertian

Recall mengukur seberapa baik model dapat menangkap seluruh kasus positif. Tinggi jika FN (positif yang terlewat) rendah.

Penting untuk kasus:

- diagnosis kanker
- deteksi kecurangan
- sistem keamanan

b. Rumus Recall

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

c. Contoh Singkat

TP = 40, FN = 20

$$Recall = \frac{40}{40 + 20} = \frac{40}{60} = 0.67 = 67\%$$

5. Hubungan Precision dan Recall

Precision dan Recall sering mengalami trade-off:

- Precision tinggi → model lebih ketat dalam memprediksi positif.
- Recall tinggi → model menangkap sebanyak mungkin data positif.

Untuk menyeimbangkan keduanya, digunakan metrik tambahan: F1-Score.

Rumus F1-Score

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

Tabel Perbandingan Metrik

Metrik	Fokus	Cocok Untuk	Sensitif Terhadap
Akurasi	Prediksi keseluruhan	Dataset seimbang	Ketidakseimbangan kelas
Precision	Ketepatan prediksi positif	Menghindari FP	False Positive
Recall	Kelengkapan prediksi positif	Menghindari FN	False Negative

Contoh Ilustrasi Kasus

Kasus: Model deteksi pneumonia

- Jika Precision rendah, banyak orang sehat diklasifikasikan sebagai pneumonia → panik, biaya tinggi.
- Jika Recall rendah, pasien pneumonia bisa tidak terdeteksi → berbahaya.