

Machine Learning

Mohammad Febryan Khamim

1 Pengantar Machine Learning

Machine Learning adalah teknik yang memungkinkan sistem komputer untuk belajar dan mengambil keputusan pada tugas tertentu tanpa harus diprogram secara eksplisit.

Konsep Sederhana : *Trial and Error* → **Machine learning** memungkinkan komputer untuk meningkatkan performanya untuk menuntaskan tugas tertentu secara otomatis.

1.1 Karakteristik Utama Machine Learning

- **Belajar dari data** : Model menemukan pola tersembunyi
- **Perbaikan otomatis** : Kinerja meningkat seiring pertambahan data
- **Generalisasi** : Tidak mengenali pola dalam data tertentu saja

1.2 Kapan Menggunakan Machine Learning?

- Ketika tugas atau permasalahan mengandung pola tertentu yang terlalu kompleks untuk diselesaikan dengan *rules*
- Terdapat data yang tak diketahui hubungan antarfiturnya
- Permasalahan mengandung probabilitas dan ketidakpastian
- Tugas memerlukan adaptasi atau pengembangan seiring pertambahan waktu.

1.3 Jenis-Jenis Machine Learning

Terdapat beberapa jenis Machine Learning, di antaranya sebagai berikut.

- **Unsupervised Learning** : Data tak berlabel
- **Supervised Learning** : Data berlabel
- **Reinforcement Learning** : Mempelajari data dengan *trial n error* untuk optimalkan *reward*

1.4 Proses Machine Learning

Input data → Algoritma ML → Pelatihan Model → Feedback loop → Proses iterasi → Evaluasi dan generalisasi

2 Supervised Learning

Dalam *supervised learning*, kita melatih model menggunakan data yang telah memiliki kelas atau label. Setelah proses *training*, model diuji dengan *data tes* untuk mengetahui seberapa baik performa prediksi yang dihasilkan.

2.1 Metrik Evaluasi dan Robustness

- **Robustness:** Untuk mengukur seberapa *robust* sebuah model, dapat digunakan *Mean Square Error* (MSE).
- **Tipe Tugas:**
 - **Klasifikasi:** Memprediksi label atau kategori data.
 - **Regresi:** Memprediksi nilai kontinyu atau jumlah, contohnya prediksi harga rumah berdasarkan fitur-fitur tertentu.

2.2 Pembagian Data

Terdapat 3 jenis data utama dalam menyelesaikan tugas *Machine Learning*:

1. **Training Data:** Digunakan untuk melatih parameter model.
2. **Validation Data:** Digunakan untuk menentukan *hyperparameter* yang perlu disuaikan (*adjusted*) serta memilih model terbaik.
3. **Test Data:** Digunakan untuk mengevaluasi performa akhir dan menentukan metrik evaluasi.

3 Metrik Evaluasi Klasifikasi

Evaluasi performa klasifikasi menggunakan beberapa metrik utama: Akurasi, *Recall*, *Precision*, dan *F1-Score*.

- **Akurasi:** Jumlah total jawaban benar dibagi dengan banyaknya jumlah prediksi. Metrik ini baik digunakan pada data yang seimbang (*balanced*), tetapi kurang baik untuk *imbalanced data*.
- **Recall:** Menunjukkan kemampuan model menemukan semua kasus relevan pada data.

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

- **Precision:** Kemampuan model klasifikasi untuk mengidentifikasi titik data yang relevan secara tepat.

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

- **F1-Score:** Metrik campuran optimal (*optimal blend*) antara *precision* dan *recall*.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.1 Confusion Matrix

Tabel berikut merepresentasikan *Confusion Matrix* untuk evaluasi prediksi:

	Pred. Positif	Pred. Negatif
Cond. Positif	TP (True Positive)	FN (False Negative)
Cond. Negatif	FP (False Positive)	TN (True Negative)

4 Tugas-Tugas pada Machine Learning

4.1 Supervised Learning

1. **Klasifikasi:** Decision Tree, SVM, Random Forest.
2. **Regresi:** Linear Regression, Logistic Regression.
3. **Time-series:** LSTM, ARIMA.

4.2 Unsupervised Learning

1. **Klasterisasi:** K-means, Hierarchical, Gaussian Mixture.
2. **Reduksi Dimensi:** PCA, t-SNE, Autoencoder.
3. **Association Rule:** Apriori.
4. **Anomaly Detection:** GMM.

5 Konsep Lanjutan Machine Learning

5.1 Bias-Variance Tradeoff

Ketidakseimbangan antara bias dan varians menentukan kemampuan generalisasi model:

- **Underfitting:** Terjadi saat model terlalu sederhana (*High Bias*).
- **Overfitting:** Terjadi saat model terlalu kompleks dan menghafal *noise* pada data (*High Variance*).

5.2 Teknik Regularisasi

Digunakan untuk mencegah *overfitting* dengan menambahkan penalti pada fungsi *loss*:

- **Lasso (L1):** $\text{Loss} + \lambda \sum |w|$ (Dapat melakukan seleksi fitur).
- **Ridge (L2):** $\text{Loss} + \lambda \sum w^2$ (Mengecilkan bobot secara merata).

5.3 Cross-Validation

Teknik untuk mengevaluasi performa model dengan membagi data menjadi k lipatan (k -fold). Model dilatih k kali, menggunakan lipatan yang berbeda sebagai data validasi setiap kalinya untuk memastikan hasil evaluasi yang lebih objektif.