Evaluasi Model & Data Leakage

Fokus:

Accuracy • Precision/Recall/F1 • ROC–AUC • MSE/MAE • Data Leakage

Tujuan:

- Paham arti & kapan pakai tiap metrik
- Bisa membaca hasil model dengan benar
- Tahu dan menghindari **data leakage**

Kenapa metrik itu penting?

Model **bagus** itu tergantung **tujuan**.

- Kelas **imbang** (50–50) → Accuracy sering cukup.
- Kelas **timpang** (1–99) → Accuracy bisa menipu, pakai **Precision/Recall/F1**.
- Perlu lihat kualitas **skor probabilitas** → **ROC–AUC**.
- Tugas prediksi **angka** → **MSE/MAE**.

> Pilih metrik = pilih “cara menilai” yang sesuai bisnis/tujuan.

Accuracy (Klasifikasi)

Definisi: proporsi prediksi yang benar.

$$\text{Accuracy} = \frac{\text{benar}}{\text{total}}$$

Kapan cocok: kelas cukup **seimbang**, cost salah kira relatif mirip.

Waspada: pada kelas **timpang**, model bisa “terlihat bagus” padahal tidak berguna (tebak mayoritas terus).

Precision, Recall, F1 (Klasifikasi)

Precision (ketepatan): dari yang diprediksi **positif**, berapa yang benar?

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

Recall (sensitivitas): dari semua **positif nyata**, berapa yang tertangkap?

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

F1-score: rata-rata harmonik Precision & Recall (seimbang keduanya).

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R}$$

Kapan dipakai: data **timpang**, perlu kontrol salah tangkap (FP) vs lolos (FN).

ROC–AUC (Klasifikasi Berbasis Skor/Probabilitas)

ROC curve: plot **TPR(=Recall)** vs **FPR** saat **ambang (threshold)** diubah-ubah.
AUC: luas area di bawah kurva (0.5 = acak, 1.0 = sempurna).

Makna praktis: seberapa baik model **mengurutkan** positif > negatif terlepas dari threshold.
Kapan pakai: saat kita punya **skor/probabilitas** dan ingin menilai kualitas ranking model.

> Untuk data **sangat timpang**, pertimbangkan juga **PR–AUC** (Precision–Recall AUC).

MSE & MAE (Regresi)

MAE (Mean Absolute Error): rata-rata **selisih absolut**. Mudah dipahami, **robust** terhadap outlier.

$$\mathrm{MAE} = \frac{1}{n}\sum |y - \hat{y}|$$

MSE (Mean Squared Error): rata-rata **selisih kuadrat**. Menghukum error besar lebih keras.

$$\mathrm{MSE} = \frac{1}{n}\sum (y - \hat{y})^2$$

Kapan pakai:

- **MAE** saat ingin ukuran “rata-rata meleset berapa unit”.
- **MSE/RMSE** saat ingin penalti ekstra untuk error besar.

Data Leakage (Kebocoran Data)

Definisi: informasi dari **validation/test** “bocor” ke **training**, bikin skor latihan **fantastis** tapi gagal di dunia nyata.

Contoh umum:

- **Scaling/encoding** di-fit pada **seluruh data** sebelum split.
- Menggunakan fitur **masa depan** (tgl pengiriman saat memprediksi “akan terlambat?”).
- Menggabung data pengguna di train dan test tanpa anonimisasi yang benar (ID jadi sinyal).

Aturan emas:

- **Split dulu**, **fit** preprocessing **hanya di train**, lalu **apply** ke val/test (pakai **Pipeline/ColumnTransformer**).

Contoh Sederhana — Klasifikasi (angka kecil)

Misal 10 kasus, label 1 = positif, 0 = negatif:

- **y_true:** 1 0 1 0 0 1 0 0 1 0
- **y_pred:** 1 0 0 0 0 1 0 0 1 0

Hitung:

- Benar = 9/10? **Tidak** → cek satu-satu: salah di posisi ke-3 saja → **Benar=9**, **Accuracy=0.9**
- **TP** = 3 (pos 1,6,9), **FP** = 0, **FN** = 1 (pos 3), **TN** = 6
- **Precision** = 3/(3+0)=1.00
- **Recall** = 3/(3+1)=0.75
- **F1** ≈ 0.86

> Akurasi tinggi. Precision sempurna, tapi Recall masih bisa ditingkatkan (ada 1 positif yang miss).

Contoh Sederhana — ROC–AUC (intuisi)

Skor probabilitas model (untuk 6 kasus positif/negatif):

y_true	skor_model
1	0.90
0	0.80
1	0.70
0	0.60
1	0.55
0	0.10

Model cenderung memberi skor **lebih tinggi** ke yang positif → **AUC tinggi**.
Jika skor positif dan negatif **tumpang tindih** parah, **AUC turun** mendekati 0.5.

Contoh Sederhana — Regresi

Nilai asli vs prediksi:

- **y_true**: 10, 12, 13, 9
- **y_pred**: 11, 11, 15, 8

Error: 1, 1, 2, 1 →

- **MAE** = (|1|+|1|+|2|+|1|)/4 = **1.25**
- **MSE** = (1^2+1^2+2^2+1^2)/4 = (1+1+4+1)/4 = **1.75**

> MAE bilang “rata-rata meleset 1.25 unit”, MSE memberi bobot lebih untuk error 2 (kuadrat=4).

Contoh Leakage — yang SALAH

- Ambil semua data → **fit StandardScaler** → baru **split** train/test → **latih**.
  Akibatnya, info **mean/std** dari test sudah “terlihat” saat scaling → skor palsu bagus.

Solusi benar:

- **Split dulu**, lalu di dalam **Pipeline**, scaler **di-fit di train** dan **apply** ke test.

Ringkas Pilih Metrik

- **Accuracy**: kelas **imbang**, mudah.
- **Precision/Recall/F1**: kelas **timpang** atau cost FP/FN berbeda.
- **ROC–AUC**: kualitas **ranking probabilitas**, threshold-agnostic.
- **MAE/MSE**: tugas **regresi**; MAE = mudah dipahami, MSE = penalti error besar.
- **Selalu** hindari **leakage** dengan **Pipeline**.

Mini-Cheat Sheet (Do/Don’t)

Do

- Split dulu, pakai **Pipeline/ColumnTransformer**.
- Laporkan metrik yang sesuai konteks.
- Tampilkan contoh salah (confusion matrix atau contoh residu besar).

Don’t

- Fit scaler/encoder di seluruh data.
- Hanya lihat accuracy saat data timpang.
- Pakai fitur “masa depan”.

Mini Task

1. Buat vektor kecil **y_true, y_pred, y_proba** (data random).
2. Hitung Accuracy, P/R/F1, ROC–AUC.
3. Buat contoh **regresi** (4–6 angka), hitung MAE/MSE.
4. Tulis 3 baris penjelasan **kapan metrikmu cocok**.
5. Jelaskan contoh **leakage** yang mungkin di datamu & cara mencegahnya.