PETA UTAMA

Panduan Memilih Metode Machine Learning

Machine learning adalah cara membuat komputer belajar dari data. Panduan ini membantu Anda memilih metode yang cocok hanya dengan menjawab ya atau tidak.

1. Jawab pertanyaan besar di bawah, lalu ikuti kolom **Ya** atau **Tidak**.
2. Baca pertanyaan dari atas. Berhenti di pertanyaan pertama yang jawabannya **Ya**, lalu buka halaman jalurnya.
3. Di halaman jalur, jawab ya atau tidak sampai tiba di **kartu hasil**. Kartu itu berisi metode dan data yang disarankan.

Apakah data Anda sudah punya kunci jawaban?

Artinya setiap data sudah diberi jawaban yang benar. Contoh: foto daun sudah diberi nama penyakitnya, data mahasiswa sudah ditandai lulus tepat waktu atau tidak. Istilah resminya: label.

YA — Data sudah punya kunci jawaban

1 Apakah hanya sedikit data yang punya kunci jawaban?
Misalnya 500 dari 50.000 data, karena memberi jawaban itu mahal.
→ **Jalur A · Kunci jawaban hanya sedikit** halaman 2

Tidak ↓

2 Apakah datanya gambar atau video?
Foto, hasil rontgen, citra satelit, rekaman CCTV.
→ **Jalur B · Gambar dan video** halaman 2

Tidak ↓

3 Apakah datanya tulisan?
Ulasan, berita, dokumen, percakapan.
→ **Jalur C · Tulisan** halaman 3

Tidak ↓

4 Apakah datanya suara atau sinyal?
Rekaman suara, detak jantung, getaran mesin.
→ **Jalur D · Suara dan sinyal** halaman 3

Tidak ↓

5 Apakah datanya jaringan hubungan?
Pertemanan, artikel yang saling mengutip, transaksi antar-rekening.
→ **Jalur E · Jaringan hubungan** halaman 3

Tidak ↓

6 Apakah datanya berurutan menurut waktu?
Data harian, bulanan, atau bacaan sensor dari waktu ke waktu.
→ **Jalur F · Data berurutan menurut waktu** halaman 4

Tidak ↓

7 Apakah Anda ingin menyarankan pilihan kepada pengguna?
Seperti saran film, buku, atau mata kuliah.
→ **Jalur G · Menyarankan pilihan** halaman 4

Tidak ↓

8 Apakah yang ingin ditebak berupa kategori?
Kategori: lulus atau tidak, jenis penyakit. Bukan kategori: harga, IPK.
→ **Jalur H · Tabel: menebak kategori** halaman 4

Tidak ↓

✓ **Kalau semua jawabannya Tidak**
Data Anda berbentuk tabel dan yang ditebak berupa angka, waktu, atau urutan.
→ **Jalur I · Tabel: menebak angka, waktu, atau urutan** halaman 5

TIDAK — Data belum punya kunci jawaban

1 Apakah komputer belajar dengan mencoba, lalu diberi hadiah atau hukuman?
Robot, permainan, pengaturan lampu lalu lintas.
→ **Jalur J · Belajar dari coba-coba (bagian 1)** halaman 5

Tidak ↓

2 Apakah Anda ingin komputer membuat data baru?
Tulisan, gambar, atau data tiruan.
→ **Jalur L · Membuat data baru** halaman 6

Tidak ↓

3 Apakah Anda ingin menemukan data yang janggal?
Penipuan, serangan jaringan, kerusakan mesin.
→ **Jalur M · Menemukan data yang janggal** halaman 7

Tidak ↓

4 Apakah Anda ingin mengelompokkan data yang mirip?
Membagi pelanggan atau mahasiswa menjadi beberapa kelompok.
→ **Jalur N · Mengelompokkan data yang mirip** halaman 7

Tidak ↓

✓ **Kalau semua jawabannya Tidak**
Anda ingin menemukan pola lain atau meringkas data.
→ **Jalur O · Menemukan pola** halaman 8

Jalur A, Jalur B

Jalur A — Kunci jawaban hanya sedikit

Sebagian besar data belum diberi jawaban. Ada dua jalan keluar.

Apakah ada ahli yang bisa memberi jawaban tambahan sedikit demi sedikit?
Misalnya dokter atau dosen yang bersedia menjawab data yang dipilih komputer.

YA — Meminta bantuan ahli secara hemat

Istilah resmi: Active learning

Komputer memilih sendiri data mana yang paling perlu dijawab oleh ahli, sehingga pekerjaan ahli jauh lebih sedikit.

Mulai dari	Random sampling (pembanding wajib)
Lalu coba	Uncertainty sampling (least confidence, margin, entropy), Query-by-Committee
Data latihan	Benchmark berlabel penuh (CIFAR-10, AG News, dan sejenisnya) dengan label disembunyikan untuk simulasi; data sendiri dengan ahli sebagai oracle
Ukur dengan	Kurva belajar (performa terhadap jumlah label), area di bawah kurva belajar, penghematan label

TIDAK — Belajar dari sedikit jawaban

Istilah resmi: Semi-supervised learning

Komputer belajar dari sedikit data berjawaban, lalu memanfaatkan banyak data tanpa jawaban untuk memperkuat dirinya.

Mulai dari	Model supervised yang dilatih hanya dengan data berlabel (batas bawah)
Lalu coba	Self-training/pseudo-labeling, Label Propagation, Label Spreading, co-training
Data latihan	CIFAR-10 (250 atau 4.000 label), SVHN, STL-10; atau data sendiri dengan sebagian label disembunyikan
Ukur dengan	Metrik tugas (accuracy, F1) pada test set berlabel; kurva performa terhadap jumlah label

Jalur B — Gambar dan video

Tentukan dulu apa yang ingin diketahui dari gambar.

Apakah datanya video?
Gambar bergerak, misalnya rekaman CCTV.

YA — Mengenali isi video

Istilah resmi: Klasifikasi video / pengenalan aksi

Menentukan apa yang terjadi dalam video, misalnya orang berjalan, jatuh, atau menyontek.

Mulai dari	CNN per frame + rata-rata, atau CNN + LSTM
Lalu coba	3D CNN (I3D, R(2+1)D), SlowFast, TSN/TSM
Data latihan	UCF101, HMDB51, Kinetics-400, Something-Something V2, NTU RGB+D
Ukur dengan	Top-1 dan top-5 accuracy, F1 per kelas, mAP

TIDAK — Apakah Anda perlu tahu di mana letak bendanya?

Bukan hanya 'ada motor', tetapi 'motornya di sebelah sini'.

Apakah perlu sampai tepi benda yang persis?
Misalnya untuk menghitung luas tumor atau luas lahan.

YA — Mewarnai benda sampai tepinya

Istilah resmi: Segmentasi citra

Menandai setiap titik gambar milik benda apa, misalnya untuk mengukur luas tumor atau lahan sawah.

Mulai dari	U-Net (semantik)
Lalu coba	DeepLabv3+, Mask R-CNN (instance), U-Net++, nnU-Net (medis)
Data latihan	Cityscapes, ADE20K, Pascal VOC, MS COCO, BraTS (tumor otak), ISIC, DeepGlobe (citra satelit)
Ukur dengan	mIoU, Dice coefficient, pixel accuracy, Hausdorff distance (medis)

TIDAK — Menemukan dan mengotaki benda

Istilah resmi: Deteksi objek

Menunjukkan letak setiap benda dengan kotak, misalnya semua motor di jalan.

Mulai dari	Faster R-CNN atau SSD pra-latih
Lalu coba	YOLO (v5/v8/v11), RetinaNet, Faster R-CNN dengan FPN
Data latihan	MS COCO, Pascal VOC, Open Images, KITTI, BDD100K, VisDrone
Ukur dengan	mAP@0.5, mAP@[0.5:0.95], precision, recall, IoU, FPS/latensi

TIDAK — Memberi nama pada gambar

Istilah resmi: Klasifikasi citra

Menentukan sebuah gambar termasuk jenis apa, misalnya daun sehat atau daun sakit.

Mulai dari	CNN sederhana, atau fitur HOG/LBP + SVM
Lalu coba	Transfer learning: ResNet, EfficientNet, DenseNet, MobileNet
Data latihan	MNIST, Fashion-MNIST, CIFAR-10/100, ImageNet, PlantVillage, ISIC (kulit), ChestX-ray14
Ukur dengan	Accuracy, top-5 accuracy, precision, recall, F1 per kelas, confusion matrix, ROC-AUC

Jalur C, Jalur D, Jalur E

Jalur C — Tulisan

Tentukan dulu bentuk hasil yang Anda inginkan.

Apakah hasilnya berupa tulisan baru?
Terjemahan, ringkasan, atau jawaban atas pertanyaan.

YA

Mengubah tulisan menjadi tulisan lain

Istilah resmi: Teks ke teks (terjemahan, ringkasan, tanya jawab)

Menerjemahkan, meringkas, atau menjawab pertanyaan.

Mulai dari	Lead-3 atau TextRank (ringkasan), BM25 + ekstraksi (tanya jawab)
Lalu coba	Transformer encoder-decoder: T5, BART, mT5, mBART, IndoBART, IndoT5
Data latihan	WMT dan OPUS (terjemahan), CNN/DailyMail, XSum, IndoSum, Liputan6 (ringkasan), SQuAD, TyDi QA (tanya jawab)
Ukur dengan	BLEU, chrF, ROUGE, METEOR, BERTScore, Exact Match dan F1 (tanya jawab), penilaian manusia

TIDAK

Apakah setiap kata perlu diberi tanda?
Misalnya menandai kata mana yang nama orang dan nama tempat.

YA

Menandai kata-kata penting

Istilah resmi: Pelabelan sekuens (NER, POS tagging)

Menandai kata mana yang merupakan nama orang, tempat, atau lembaga di dalam kalimat.

Mulai dari	Conditional Random Field (CRF) dengan fitur buatan
Lalu coba	BiLSTM-CRF dengan word/character embedding
Data latihan	CoNLL-2003, OntoNotes 5.0, WikiANN, Universal Dependencies, NERGrit dan POSP (IndoNLU)
Ukur dengan	Precision, recall, F1 tingkat entitas (seqeval); accuracy untuk POS

TIDAK

Memilah tulisan ke dalam kategori

Istilah resmi: Klasifikasi teks

Menentukan sebuah tulisan termasuk jenis apa, misalnya ulasan positif atau negatif.

Mulai dari	TF-IDF + Naive Bayes, Logistic Regression, atau SVM linear
Lalu coba	fastText, TextCNN, BiLSTM dengan word embedding
Data latihan	IMDb, SST-2, AG News, 20 Newsgroups, IndoNLU (SmSA, EmoT), IndoLEM
Ukur dengan	Accuracy, precision, recall, macro-F1, confusion matrix

Jalur D — Suara dan sinyal

Rekaman suara atau sinyal dari alat ukur.

Apakah Anda ingin mengubah ucapan menjadi tulisan?
Misalnya menuliskan isi rekaman kuliah.

YA

Mengubah ucapan menjadi tulisan

Istilah resmi: Pengenalan ucapan (speech to text)

Menuliskan apa yang diucapkan di dalam rekaman.

Mulai dari	Whisper pra-latih tanpa pelatihan ulang (zero-shot)
Lalu coba	wav2vec 2.0 atau HuBERT + CTC, Conformer
Data latihan	LibriSpeech, Mozilla Common Voice (tersedia bahasa Indonesia), FLEURS, TED-LIUM
Ukur dengan	Word Error Rate (WER), Character Error Rate (CER), real-time factor

TIDAK

Mengenali jenis suara atau sinyal

Istilah resmi: Klasifikasi audio / sinyal

Menentukan sebuah rekaman termasuk jenis apa, misalnya detak jantung normal atau tidak.

Mulai dari	Fitur MFCC atau statistik sinyal + SVM/Random Forest
Lalu coba	CNN pada log-mel spectrogram, CRNN, 1D-CNN pada sinyal mentah
Data latihan	ESC-50, UrbanSound8K, AudioSet, Speech Commands, GTZAN; sinyal: PhysioNet MIT-BIH (EKG), CWRU Bearing, UCI HAR
Ukur dengan	Accuracy, macro-F1, confusion matrix, ROC-AUC; sensitivitas dan spesifisitas untuk data medis

Jalur E — Jaringan hubungan

Data berupa titik-titik yang saling terhubung.

Belajar dari jaringan hubungan

Istilah resmi: Pembelajaran pada graf (node, link, atau graf)

Menebak sifat suatu titik atau hubungan baru di dalam jaringan, misalnya peneliti mana yang mungkin bekerja sama.

Mulai dari	Fitur node + Logistic Regression/MLP, Label Propagation, DeepWalk atau node2vec
Lalu coba	GCN, GraphSAGE, GAT, GIN
Data latihan	Cora, CiteSeer, PubMed, Open Graph Benchmark (OGB), Reddit, PPI, MoleculeNet/QM9, Elliptic (transaksi Bitcoin)
Ukur dengan	Node: accuracy, macro-F1. Link: ROC-AUC, Hits@k, MRR. Graf: accuracy atau MAE

Jalur F, Jalur G, Jalur H

Jalur F — Data berurutan menurut waktu

Urutan waktu pada data tidak boleh diacak.

Apakah Anda ingin menebak nilai di masa depan?
Misalnya jumlah pendaftar tahun depan.

YA

Meramal angka masa depan

Istilah resmi: Peramalan deret waktu (forecasting)

Menebak nilai berikutnya dari data masa lalu, misalnya jumlah pendaftar tahun depan.

Mulai dari	Naive, seasonal naive, ARIMA/SARIMA, Exponential Smoothing (ETS), Prophet
Lalu coba	XGBoost/LightGBM dengan fitur lag, LSTM, GRU, TCN
Data latihan	M4, M5, ETT, Electricity, Traffic, Weather; data lokal dari BPS, BMKG, Bank Indonesia, Yahoo Finance
Ukur dengan	MAE, RMSE, sMAPE, MASE; untuk probabilistik: CRPS dan pinball loss

TIDAK

Mengenali pola pada data berurutan

Istilah resmi: Klasifikasi deret waktu

Menentukan sepotong data sensor termasuk kegiatan apa, misalnya berjalan atau berlari.

Mulai dari	1-NN dengan Dynamic Time Warping
Lalu coba	ROCKET/MiniRocket, fitur statistik (tsfresh) + Random Forest, 1D-CNN, LSTM
Data latihan	UCR/UEA Time Series Archive, UCI HAR, WISDM, PhysioNet
Ukur dengan	Accuracy, macro-F1, confusion matrix

Jalur G — Menyarankan pilihan

Data berupa riwayat siapa memilih apa.

Menyarankan pilihan

Istilah resmi: Sistem rekomendasi

Menyarankan barang, film, atau mata kuliah yang mungkin disukai seseorang berdasarkan riwayatnya.

Mulai dari	Most-popular, item-based KNN, user-based KNN
Lalu coba	Matrix factorization (SVD, ALS), BPR, Factorization Machines, content-based (TF-IDF)
Data latihan	MovieLens (100K/1M/20M), Amazon Reviews, Netflix Prize, Book-Crossing, Yelp, Last.fm
Ukur dengan	Recall@k, NDCG@k, Hit Rate@k, MAP; RMSE/MAE untuk prediksi rating; coverage dan diversity

Jalur H — Tabel: menebak kategori

Data berbentuk baris dan kolom, dan jawabannya berupa kategori.

Apakah satu data boleh punya beberapa kategori sekaligus?
Misalnya satu artikel bertopik 'AI' dan juga 'kesehatan'.

YA

Memberi beberapa kategori sekaligus

Istilah resmi: Klasifikasi multi-label

Satu data boleh masuk lebih dari satu kategori.

Mulai dari	Binary Relevance dengan Logistic Regression
Lalu coba	Classifier Chains, Label Powerset, ML-kNN, Random Forest multi-output, jaringan saraf dengan keluaran sigmoid
Data latihan	Yeast, Emotions, Scene (repositori Mulan), Reuters-21578, label objek MS-COCO, Toxic Comment (Jigsaw)
Ukur dengan	Hamming loss, subset accuracy, micro-F1, macro-F1, Jaccard score, ranking loss

TIDAK

Apakah pilihannya lebih dari dua?
Dua pilihan: ya atau tidak. Lebih dari dua: rendah, sedang, tinggi.

YA

Memilih satu dari banyak kategori

Istilah resmi: Klasifikasi multikelas (data tabular)

Menebak satu kategori dari tiga pilihan atau lebih, misalnya risiko rendah, sedang, atau tinggi.

Mulai dari	Multinomial Logistic Regression, Naive Bayes, KNN, Decision Tree
Lalu coba	Random Forest, SVM (one-vs-rest atau one-vs-one), XGBoost, LightGBM, CatBoost, MLP
Data latihan	Iris, Wine, Covertype, Dry Bean, Letter Recognition, Predict Students' Dropout and Academic Success (UCI)
Ukur dengan	Accuracy, macro-F1, weighted-F1, balanced accuracy, Cohen's kappa, confusion matrix per kelas

TIDAK

Menebak ya atau tidak

Istilah resmi: Klasifikasi biner (data tabular)

Menebak satu dari dua kemungkinan, misalnya lulus tepat waktu atau tidak.

Mulai dari	Logistic Regression, Naive Bayes, KNN, Decision Tree
Lalu coba	Random Forest, SVM (kernel RBF), XGBoost, LightGBM, CatBoost, MLP
Data latihan	UCI Adult, Breast Cancer Wisconsin, Pima Indians Diabetes, Bank Marketing, Telco Customer Churn, Credit Card Fraud (ULB)
Ukur dengan	Accuracy, precision, recall, F1, ROC-AUC, PR-AUC, MCC, confusion matrix

Jalur I, Jalur J

Jalur I — Tabel: menebak angka, waktu, atau urutan

Data berbentuk baris dan kolom, dan jawabannya bukan kategori.

Apakah yang ditebak adalah lamanya waktu sampai sesuatu terjadi?
Misalnya lama studi sampai lulus, padahal sebagian mahasiswa masih kuliah.

YA → **Menebak berapa lama sampai sesuatu terjadi**
Istilah resmi: Analisis survival (waktu sampai kejadian)

Memperkirakan lama waktu sampai suatu kejadian, termasuk bila sebagian data belum mengalaminya.

Mulai dari	Kaplan-Meier (deskriptif), Cox Proportional Hazards
Lalu coba	Random Survival Forest, Gradient Boosting Survival, Survival SVM, Accelerated Failure Time
Data latihan	METABRIC, SUPPORT, GBSG2, Rossi recidivism, Veterans' Lung Cancer
Ukur dengan	Concordance index (C-index), integrated Brier score, time-dependent AUC

TIDAK → **Apakah Anda ingin mengurutkan dari yang terbaik?**
Misalnya urutan hasil pencarian atau peringkat kinerja.

YA → **Mengurutkan dari yang terbaik**
Istilah resmi: Learning to rank (pemeringkatan)

Menyusun daftar menurut peringkat, misalnya hasil pencarian yang paling cocok berada di atas.

Mulai dari	BM25 atau skor heuristik; regresi pointwise
Lalu coba	RankNet, LambdaMART (LightGBM/XGBoost lambdarank), RankSVM
Data latihan	MSLR-WEB10K/30K, Yahoo! Learning to Rank Challenge, Istella LETOR, MS MARCO
Ukur dengan	NDCG@k, MAP, MRR, Precision@k, Kendall tau

TIDAK → **Menebak angka**
Istilah resmi: Regresi (prediksi nilai kontinu)

Menebak nilai berupa angka, misalnya harga rumah atau IPK.

Mulai dari	Linear Regression, Ridge, Lasso, ElasticNet, KNN Regressor
Lalu coba	Random Forest, SVR, XGBoost, LightGBM, CatBoost, Gaussian Process Regression
Data latihan	California Housing, Ames Housing, Bike Sharing, Concrete Compressive Strength, Wine Quality, Student Performance (UCI)
Ukur dengan	MAE, RMSE, R^2, MAPE; plot residual dan plot prediksi vs aktual

Jalur J — Belajar dari coba-coba (bagian 1)

Komputer mencoba suatu tindakan, lalu belajar dari hadiah atau hukuman.

Apakah setiap pilihan berdiri sendiri, tidak memengaruhi langkah berikutnya?
Misalnya memilih iklan mana yang ditampilkan untuk tiap pengunjung.

YA → **Memilih yang terbaik sambil mencoba**
Istilah resmi: Multi-armed bandit

Mencoba beberapa pilihan, lalu semakin sering memilih yang hasilnya paling baik.

Mulai dari	Epsilon-greedy
Lalu coba	UCB1, Thompson Sampling
Data latihan	Open Bandit Dataset, Yahoo! R6 (rekomendasi berita), simulasi dari MovieLens
Ukur dengan	Cumulative regret, reward rata-rata, off-policy evaluation (IPS, doubly robust)

TIDAK → **Apakah Anda hanya punya catatan lama dan tidak bisa mencoba langsung?**
Misalnya catatan keputusan petugas, tanpa simulator.

YA → **Belajar dari catatan keputusan lama**
Istilah resmi: Offline reinforcement learning

Belajar strategi dari catatan keputusan masa lalu, tanpa boleh mencoba langsung.

Mulai dari	Behavior Cloning
Lalu coba	CQL, IQL, BCQ, TD3+BC
Data latihan	D4RL, Minari, RL Unplugged; log operasional sendiri
Ukur dengan	Normalized return, off-policy evaluation (Fitted Q Evaluation, importance sampling)

TIDAK → Lanjut ke Jalur K · Belajar dari coba-coba (bagian 2) (halaman 6)

Jalur K, Jalur L

Jalur K — Belajar dari coba-coba (bagian 2)

Komputer bisa mencoba langsung di simulator atau lingkungan nyata.

Apakah yang diatur berupa nilai yang bisa naik-turun secara halus?
Halus: sudut kemudi, kecepatan. Tidak halus: kiri, kanan, maju.

YA

Belajar mengendalikan gerakan halus

Istilah resmi: Deep RL dengan aksi kontinu

Belajar mengatur nilai yang bisa naik-turun sedikit demi sedikit, misalnya gerak lengan robot.

Mulai dari	Kebijakan acak, pengendali PID, atau REINFORCE
Lalu coba	PPO, SAC, TD3, DDPG
Data latihan	Lingkungan: MuJoCo (Gymnasium), DeepMind Control Suite, PyBullet, Isaac Lab
Ukur dengan	Average return, sample efficiency, tingkat keberhasilan; interquartile mean dengan selang kepercayaan

TIDAK

Apakah semua keadaan bisa didaftar dalam tabel kecil?
Misalnya posisi pada papan 5×5.

YA

Belajar strategi di dunia kecil

Istilah resmi: RL tabular

Belajar langkah terbaik pada masalah sederhana yang semua keadaannya bisa didaftar, misalnya labirin kecil.

Mulai dari	Kebijakan acak
Lalu coba	Q-Learning, SARSA, Expected SARSA, Monte Carlo control
Data latihan	Lingkungan: FrozenLake, Taxi, CliffWalking (Gymnasium), GridWorld buatan sendiri
Ukur dengan	Reward kumulatif per episode, jumlah langkah menuju tujuan, kurva konvergensi

TIDAK

Belajar strategi di dunia rumit

Istilah resmi: Deep RL dengan aksi diskret

Belajar memilih langkah terbaik pada masalah rumit, misalnya permainan atau lampu lalu lintas.

Mulai dari	DQN
Lalu coba	Double DQN, Dueling DQN, Rainbow, A2C, PPO
Data latihan	Lingkungan: CartPole, LunarLander, Atari (ALE), MiniGrid, PettingZoo (multi-agent)
Ukur dengan	Average return, human-normalized score (Atari), sample efficiency

Jalur L — Membuat data baru

Komputer menghasilkan sesuatu yang belum ada.

Apakah yang dibuat berupa tulisan?
Jawaban chatbot, draf dokumen.

YA

Membuat tulisan baru

Istilah resmi: Generasi teks / model bahasa

Komputer menulis jawaban, ringkasan, atau draf, misalnya chatbot layanan kampus.

Mulai dari	Model n-gram atau template
Lalu coba	Fine-tuning GPT-2/T5; LLM terbuka (Llama, Mistral, Qwen) dengan LoRA/QLoRA
Data latihan	WikiText-103, OpenWebText, dataset instruksi (Alpaca, Dolly), korpus domain sendiri
Ukur dengan	Perplexity, BLEU/ROUGE, BERTScore, faithfulness dan answer relevancy (RAGAS), penilaian manusia

TIDAK

Apakah yang dibuat berupa gambar?
Gambar medis tambahan, ilustrasi.

YA

Membuat gambar baru

Istilah resmi: Generasi citra

Komputer membuat gambar yang belum pernah ada, misalnya menambah contoh gambar penyakit langka.

Mulai dari	Variational Autoencoder (VAE)
Lalu coba	DCGAN, StyleGAN, Pix2Pix, CycleGAN
Data latihan	MNIST, CIFAR-10, CelebA, FFHQ, LSUN, LAION (teks ke citra)
Ukur dengan	FID, KID, Inception Score, precision/recall generatif, CLIP score

TIDAK

Membuat data tiruan

Istilah resmi: Data tabular sintetis

Membuat tabel data buatan yang mirip aslinya, agar data asli yang rahasia tidak perlu dibagikan.

Mulai dari	SMOTE atau Gaussian Copula
Lalu coba	CTGAN, TVAE
Data latihan	Adult, Covertype, Credit Default; data sendiri yang sensitif
Ukur dengan	Fidelity (uji KS, selisih korelasi), utility (train on synthetic, test on real), privasi (distance to closest record, membership inference)

Jalur M, Jalur N

Jalur M — Menemukan data yang janggal

Mencari data yang berbeda dari kebanyakan.

Apakah datanya berurutan menurut waktu?
Bacaan sensor atau catatan sistem dari waktu ke waktu.

YA — Mencari kejadian janggal dari waktu ke waktu

Istilah resmi: Deteksi anomali deret waktu

Menemukan saat ketika sensor atau sistem berperilaku tidak seperti biasanya.

Mulai dari	Rolling z-score, residual dekomposisi STL
Lalu coba	Isolation Forest pada fitur jendela, Matrix Profile, galat prediksi atau rekonstruksi LSTM/Autoencoder
Data latihan	NAB, Yahoo S5, SMD, SMAP/MSL, SWaT/WADI, UCR Anomaly Archive
Ukur dengan	PR-AUC, F1 berbasis rentang, VUS, keterlambatan deteksi; bila memakai point-adjusted F1, laporkan juga versi tanpa penyesuaian

TIDAK — Mencari data yang janggal

Istilah resmi: Deteksi anomali (data tabular atau citra)

Menemukan data yang berbeda dari kebanyakan, misalnya transaksi curang atau serangan jaringan.

Mulai dari	Z-score, IQR, galat rekonstruksi PCA
Lalu coba	Isolation Forest, Local Outlier Factor, One-Class SVM, ECOD/COPOD, GMM
Data latihan	ODDS, ADBench, NSL-KDD, UNSW-NB15, CIC-IDS2017, N-BaIoT, IoT-23, Credit Card Fraud; citra: MVTec AD
Ukur dengan	ROC-AUC, PR-AUC, F1 pada ambang terpilih, precision@k, false positive rate

Jalur N — Mengelompokkan data yang mirip

Komputer membagi data menjadi kelompok tanpa diberi tahu jawabannya.

Apakah jumlah kelompok sudah Anda tentukan?
Misalnya harus menjadi 4 kelompok.

YA — Apakah kelompoknya kira-kira bulat dan sama besar?
Kalau ragu, jawab Ya.

YA — Membagi menjadi sejumlah kelompok

Istilah resmi: Clustering partisi (jumlah klaster diketahui)

Membagi data menjadi kelompok sebanyak yang Anda tentukan; anggota tiap kelompok saling mirip.

Mulai dari	K-Means (inisialisasi k-means++)
Lalu coba	Mini-Batch K-Means, Gaussian Mixture Model, K-Medoids, Fuzzy C-Means, K-Prototypes (data campuran)
Data latihan	Iris, Wine, Mall Customers, Online Retail (segmentasi RFM), MNIST
Ukur dengan	Silhouette, Davies-Bouldin, Calinski-Harabasz; ARI dan NMI bila ada label pembanding

TIDAK — Mengelompokkan bentuk yang tidak beraturan

Istilah resmi: Clustering untuk bentuk klaster tidak bulat

Membagi data menjadi kelompok meskipun bentuk kelompoknya memanjang atau melengkung.

Mulai dari	Agglomerative Clustering (Ward atau average linkage)
Lalu coba	Spectral Clustering, Gaussian Mixture Model dengan kovarians penuh
Data latihan	Data sintetis (moons, circles) untuk validasi metode; data jaringan atau citra
Ukur dengan	Silhouette, Davies-Bouldin; ARI dan NMI bila ada label pembanding

TIDAK — Apakah ada banyak data terpencil atau kelompok berbentuk aneh?
Misalnya titik-titik lokasi di peta.

YA — Mengelompokkan berdasarkan keramaian

Istilah resmi: Clustering berbasis kepadatan

Menemukan kelompok di tempat data berkumpul padat dan mengabaikan data yang terpencil.

Mulai dari	DBSCAN
Lalu coba	HDBSCAN, OPTICS, Mean Shift
Data latihan	Data spasial atau GPS (titik gempa, lokasi kecelakaan), data sintetis (moons), data dengan banyak outlier
Ukur dengan	DBCV, silhouette (tanpa titik noise), proporsi noise; ARI dan NMI bila ada label

TIDAK — Mengelompokkan tanpa tahu jumlah kelompok

Istilah resmi: Clustering dengan jumlah klaster belum diketahui

Mencari sendiri berapa kelompok yang paling masuk akal, lalu membagi data ke dalamnya.

Mulai dari	Agglomerative Clustering dengan dendrogram
Lalu coba	K-Means dengan metode elbow, silhouette, atau gap statistic; GMM dengan BIC/AIC; X-Means
Data latihan	Iris, Wine, Wholesale Customers, Country Data, data survei sendiri
Ukur dengan	Silhouette, Davies-Bouldin, Calinski-Harabasz, cophenetic correlation

Jalur O, Jalur P

Jalur O — Menemukan pola

Mencari hubungan atau tema yang tersembunyi di dalam data.

Apakah Anda mencari hal yang sering muncul bersama?
Misalnya barang yang sering dibeli bersamaan.

YA

Mencari hal yang sering muncul bersama

Istilah resmi: Aturan asosiasi / pola frekuen

Menemukan aturan seperti 'yang membeli roti biasanya juga membeli selai'.

Mulai dari	Apriori
Lalu coba	FP-Growth, ECLAT
Data latihan	Groceries, Instacart Market Basket, Online Retail (UCI), repositori FIMI
Ukur dengan	Support, confidence, lift, conviction, leverage

TIDAK

Apakah Anda mencari tema dalam kumpulan tulisan?
Misalnya topik utama dari ribuan abstrak.

YA

Menemukan tema dalam kumpulan tulisan

Istilah resmi: Pemodelan topik

Menemukan topik-topik utama dari ribuan dokumen tanpa membacanya satu per satu.

Mulai dari	Latent Semantic Analysis (LSA)
Lalu coba	Latent Dirichlet Allocation (LDA), NMF
Data latihan	20 Newsgroups, Reuters-21578, abstrak arXiv, berita atau cuitan berbahasa Indonesia
Ukur dengan	Topic coherence (C_v, NPMI), topic diversity, perplexity, penilaian manusia

TIDAK

Lanjut ke Jalur P · Meringkas data dan lainnya (halaman 8)

Jalur P — Meringkas data dan lainnya

Data terlalu rumit dan perlu disederhanakan.

Apakah Anda ingin meringkas data yang kolomnya sangat banyak?
Misalnya ratusan kolom ingin dijadikan belasan.

YA

Apakah tujuannya untuk digambar dan dilihat?
Kalau tujuannya mempercepat model, jawab Tidak.

YA

Menggambar data rumit menjadi peta titik

Istilah resmi: Visualisasi data berdimensi tinggi

Meringkas data berkolom banyak menjadi gambar dua dimensi agar polanya terlihat oleh mata.

Mulai dari	PCA dua komponen
Lalu coba	t-SNE, UMAP
Data latihan	MNIST, Fashion-MNIST, COIL-20, data ekspresi gen, embedding hasil model
Ukur dengan	Trustworthiness, continuity, kNN preservation

TIDAK

Meringkas kolom data

Istilah resmi: Reduksi dimensi dan seleksi fitur

Mengurangi jumlah kolom tanpa banyak kehilangan informasi, agar model lebih cepat dan sederhana.

Mulai dari	PCA, variance threshold, penyaringan korelasi
Lalu coba	LDA (supervised), Kernel PCA, ICA, NMF; seleksi fitur: mutual information, chi-square, RFE, Lasso, Boruta
Data latihan	Madelon, Arcene, data ekspresi gen (TCGA), MNIST
Ukur dengan	Explained variance, galat rekonstruksi, performa dan waktu latih model hilir

TIDAK

Apakah Anda ingin komputer berlatih sendiri dari data mentah untuk dipakai pada tugas lain?
Misalnya jutaan data tanpa jawaban, lalu dipakai untuk tugas dengan sedikit jawaban.

YA

Berlatih sendiri dari data mentah

Istilah resmi: Self-supervised representation learning

Komputer membuat soal latihan sendiri dari data tanpa jawaban agar paham polanya, lalu dipakai untuk tugas lain.

Mulai dari	Model supervised dari nol, atau model pra-latih umum (ImageNet) sebagai pembanding
Lalu coba	Citra: SimCLR, MoCo, BYOL, DINO, MAE. Teks: masked language modeling. Audio: wav2vec 2.0. Deret waktu: TS2Vec. Tabular: SCARF
Data latihan	ImageNet, CIFAR-10/100, STL-10; korpus teks atau data domain tak berlabel berukuran besar
Ukur dengan	Linear probing, fine-tuning dengan sedikit label, akurasi kNN pada embedding

TIDAK

Belum ketemu: rumuskan ulang

Istilah resmi: Masalah belum terpetakan

Tulis dulu dalam satu kalimat: apa masukannya dan apa hasil yang diinginkan. Mungkin masalah Anda tidak memerlukan machine learning.

Mulai dari	Mulai dari exploratory data analysis (statistik deskriptif, visualisasi distribusi dan korelasi)
Lalu coba	Rumuskan ulang: apa masukan, apa keluaran yang diinginkan, dan bagaimana keberhasilan diukur
Data latihan	Belum dapat ditentukan sebelum masalah dirumuskan
Ukur dengan	Belum dapat ditentukan

Sesuaikan dengan keadaan Anda

Periksa daftar ini satu per satu. Kalau jawabannya Ya, tambahkan sarannya ke rancangan penelitian Anda.

Kalau jawabannya Ya...

Data Anda sedikit?

Kurang dari sekitar 1.000 baris.

- Pakai model sederhana, atau model yang sudah dilatih orang lain (transfer learning).
- Perbanyak data dengan variasi buatan (augmentasi).
- Uji berulang kali (cross validation) agar hasilnya bukan kebetulan.

Satu kategori jauh lebih banyak dari yang lain?

Misalnya 95% normal dan 5% penipuan.

- Jangan hanya melihat akurasi; pakai F1 atau PR-AUC.
- Beri bobot lebih pada kategori yang sedikit, atau seimbangkan dengan SMOTE pada data latih saja.

Keputusan model harus bisa dijelaskan?

Keputusan akademik, medis, atau keuangan.

- Pilih model yang mudah dibaca: pohon keputusan atau regresi logistik.
- Untuk model rumit, tambahkan alat penjelas: SHAP atau LIME.

Data tersebar dan tidak boleh dikumpulkan?

Data pasien antar-rumah sakit, data perangkat IoT.

- Pakai federated learning: yang berkeliling adalah modelnya, bukan datanya.
- Mulai dari FedAvg; pustaka yang mudah: Flower.

Data datang terus-menerus dan polanya bisa berubah?

Transaksi, lalu lintas jaringan.

- Pakai model yang belajar sedikit demi sedikit (online learning), pustaka River.
- Pasang pendeteksi perubahan pola (drift), misalnya ADWIN.

Model harus jalan di ponsel atau alat kecil?

Atau harus menjawab sangat cepat.

- Pilih model kecil, lalu kecilkan lagi dengan quantization dan pruning.
- Laporkan juga ukuran dan kecepatan model, bukan hanya ketepatannya.

Kolom data sangat banyak?

Ratusan atau ribuan kolom.

- Pilih kolom yang penting saja (seleksi fitur) atau ringkas dengan PCA.
- Lakukan pemilihan kolom hanya pada data latih.

Banyak isian kosong atau nilai aneh?

Formulir tidak lengkap, sensor gagal baca.

- Isi yang kosong dengan nilai tengah (median), atau pakai KNN imputer.
- Pakai model berbasis pohon (XGBoost, LightGBM) yang tahan terhadap data kotor.

Perlu tahu seberapa yakin model?

Agar model boleh berkata 'saya ragu'.

- Pakai conformal prediction (pustaka MAPIE) untuk mendapat rentang jawaban.
- Periksa apakah angka peluang dari model bisa dipercaya (kalibrasi).

Keputusan menyangkut nasib orang?

Beasiswa, penilaian kinerja, penerimaan.

- Periksa apakah model adil untuk tiap kelompok (jenis kelamin, usia, daerah).
- Pustaka: Fairlearn, AIF360.

Ingin tahu sebab-akibat, bukan sekadar menebak?

Apakah beasiswa membuat mahasiswa lulus lebih cepat?

- Model penebak tidak cukup; pakai metode sebab-akibat (propensity score, difference-in-differences).
- Pustaka: DoWhy, EconML.

Komputer Anda tanpa GPU?

Hanya laptop biasa.

- Utamakan metode klasik dan gradient boosting; laptop sudah cukup.
- Untuk gambar dan tulisan, pakai Google Colab atau Kaggle yang gratis.

Kebiasaan baik untuk semua penelitian

1. Tulis tujuan dalam satu kalimat: apa masukannya, apa hasilnya, dan diukur dengan apa.
2. Sisihkan data uji sejak awal dan jangan disentuh sampai percobaan terakhir.
3. Selalu mulai dari metode paling sederhana sebagai pembanding (baris 'Mulai dari' pada kartu).
4. Bandingkan paling sedikit tiga metode dengan ukuran yang sama.
5. Ulangi percobaan beberapa kali dan laporkan rata-ratanya, bukan hasil terbaik saja.
6. Catat semua pengaturan supaya orang lain bisa mengulang penelitian Anda.
7. Cari data pembanding di UCI Machine Learning Repository, Kaggle, OpenML, Hugging Face, atau data.go.id.
8. Periksa izin dan kerahasiaan data sebelum mulai.

Kamus istilah

Istilah yang muncul di kartu hasil, dijelaskan dengan bahasa sehari-hari.

Data
Kumpulan contoh yang dipelajari komputer. Satu baris tabel, satu foto, atau satu kalimat adalah satu data.

Label (kunci jawaban)
Jawaban benar yang menempel pada setiap data. Contoh: 'lulus tepat waktu'.

Fitur
Kolom atau ciri yang dipakai untuk menebak. Contoh: IPK, kehadiran, jumlah SKS.

Model
Hasil belajar komputer: semacam rumus yang bisa menebak jawaban untuk data baru.

Algoritma
Cara belajarnya. Contoh: Decision Tree, K-Means.

Melatih (training)
Proses komputer mempelajari data sampai menghasilkan model.

Data latih dan data uji
Data latih untuk belajar; data uji disimpan untuk ujian. Seperti soal latihan dan soal ujian yang harus berbeda.

Supervised learning
Belajar dari data yang punya kunci jawaban.

Unsupervised learning
Mencari pola sendiri pada data tanpa kunci jawaban.

Reinforcement learning
Belajar dari coba-coba dengan hadiah dan hukuman.

Klasifikasi
Menebak kategori. Contoh: spam atau bukan.

Regresi
Menebak angka. Contoh: harga rumah.

Clustering
Mengelompokkan data yang mirip.

Baseline
Metode paling sederhana yang dijadikan pembanding. Metode baru harus mengalahkannya.

Dataset pembanding (benchmark)
Data umum yang dipakai banyak peneliti, supaya hasil bisa dibandingkan.

Akurasi
Persentase tebakan yang benar dari semua tebakan.

Precision
Dari semua yang ditebak 'ya', berapa yang memang 'ya'.

Recall
Dari semua yang sebenarnya 'ya', berapa yang berhasil ditemukan.

F1
Gabungan precision dan recall dalam satu angka.

Confusion matrix
Tabel yang menunjukkan tebakan mana yang benar dan mana yang tertukar.

ROC-AUC dan PR-AUC
Angka 0 sampai 1 yang menilai kemampuan model memisahkan dua kategori. Semakin dekat ke 1 semakin baik.

MAE dan RMSE
Rata-rata selisih antara tebakan dan nilai sebenarnya. Semakin kecil semakin baik.

Cross validation
Menguji model berkali-kali dengan pembagian data yang berbeda, agar hasilnya bukan kebetulan.

Overfitting
Model menghafal data latih tetapi gagal pada data baru. Seperti hafal soal latihan tanpa paham materinya.

Data leakage
Informasi dari data uji bocor ke proses belajar, sehingga hasil tampak bagus padahal palsu.

Praproses
Membersihkan dan menyiapkan data sebelum dipelajari.

Hyperparameter
Pengaturan yang dipilih peneliti sebelum melatih, misalnya jumlah pohon atau nilai K.

Ensemble
Menggabungkan banyak model agar tebakannya lebih kuat. Contoh: Random Forest.

Deep learning
Machine learning dengan jaringan saraf berlapis banyak; kuat untuk gambar, tulisan, dan suara.

Transfer learning
Memakai model yang sudah dilatih orang lain, lalu menyesuaikannya dengan data sendiri.

Peta Keluarga Machine Learning

Setiap warna adalah satu keluarga. Kotak berwarna menjelaskan keluarganya dengan bahasa sederhana; semakin jauh dari tengah, semakin spesifik sampai ke nama metode.

