Peta Keputusan Riset Machine Learning

Jawab setiap pertanyaan dengan ya atau tidak sampai tiba di kotak hasil. Hasil menunjukkan jenis tugas dan algoritma utamanya; rincian lengkap ada di halaman berikutnya.

Pertanyaan · YA · TIDAK Jawaban yang membawa ke kotak itu · Hasil (nomor H merujuk ke halaman rincian); baris kecil adalah algoritma utama · Baca dari kiri ke kanan, mulai dari kotak gelap.

Apakah data Anda memiliki label atau target yang ingin diprediksi?

- **YA** Apakah hanya sebagian kecil data yang berlabel, sedangkan sebagian besar belum?
  - **YA** Apakah Anda bisa meminta ahli melabeli data tambahan secara bertahap selama penelitian?
    - **YA** **H1 · Active learning**
      Uncertainty sampling (least confidence, margin, entropy), Query-by-Committee
    - **TIDAK** **H2 · Semi-supervised learning**
      Self-training/pseudo-labeling, Label Propagation, Label Spreading, co-training
  - **TIDAK** Apakah data utama berupa citra atau video?
    - **YA** Apakah datanya video (ada dimensi waktu antar-frame)?
      - **YA** **H3 · Klasifikasi video / pengenalan aksi**
        3D CNN (I3D, R(2+1)D), SlowFast, TSN/TSM
      - **TIDAK** Apakah Anda perlu mengetahui letak objek di dalam citra, bukan hanya satu label untuk seluruh citra?
        - **YA** Apakah batas objek dibutuhkan sampai tingkat piksel?
          - **YA** **H4 · Segmentasi citra**
            DeepLabv3+, Mask R-CNN (instance), U-Net++, nnU-Net (medis)
          - **TIDAK** **H5 · Deteksi objek**
            YOLO (v5/v8/v11), RetinaNet, Faster R-CNN dengan FPN
        - **TIDAK** **H6 · Klasifikasi citra**
          Transfer learning: ResNet, EfficientNet, DenseNet, MobileNet
    - **TIDAK** Apakah data utama berupa teks bahasa alami?
      - **YA** Apakah keluarannya berupa teks baru (terjemahan, ringkasan, atau jawaban)?
        - **YA** **H7 · Teks ke teks (terjemahan, ringkasan, tanya jawab)**
          Transformer encoder-decoder: T5, BART, mT5, mBART, IndoBART, IndoT5
        - **TIDAK** Apakah label diberikan untuk setiap kata/token?
          - **YA** **H8 · Pelabelan sekuens (NER, POS tagging)**
            BiLSTM-CRF dengan word/character embedding
          - **TIDAK** **H9 · Klasifikasi teks**
            fastText, TextCNN, BiLSTM dengan word embedding
      - **TIDAK** Apakah data utama berupa audio atau sinyal?
        - **YA** Apakah tujuannya mengubah ucapan menjadi teks?
          - **YA** **H10 · Pengenalan ucapan (speech to text)**
            wav2vec 2.0 atau HuBERT + CTC, Conformer
          - **TIDAK** **H11 · Klasifikasi audio / sinyal**
            CNN pada log-mel spectrogram, CRNN, 1D-CNN pada sinyal mentah
        - **TIDAK** Apakah data berbentuk graf/jaringan (simpul dan relasi)?
          - **YA** **H12 · Pembelajaran pada graf (node, link, atau graf)**
            GCN, GraphSAGE, GAT, GIN
          - **TIDAK** Apakah urutan waktu pada data penting (deret waktu)?
            - **YA** Apakah Anda ingin memprediksi nilai pada waktu mendatang?
              - **YA** **H13 · Peramalan deret waktu (forecasting)**
                XGBoost/LightGBM dengan fitur lag, LSTM, GRU, TCN
              - **TIDAK** **H14 · Klasifikasi deret waktu**
                ROCKET/MiniRocket, fitur statistik (tsfresh) + Random Forest, 1D-CNN, LSTM
            - **TIDAK** Apakah tujuannya merekomendasikan item kepada pengguna berdasarkan riwayat interaksi?
              - **YA** **H15 · Sistem rekomendasi**
                Matrix factorization (SVD, ALS), BPR, Factorization Machines, content-based (TF-IDF)
              - **TIDAK** Apakah target berupa kategori/kelas (bukan angka kontinu)?
                - **YA** Apakah satu data bisa memiliki beberapa label sekaligus?
                  - **YA** **H16 · Klasifikasi multi-label**
                    Classifier Chains, Label Powerset, ML-kNN, Random Forest multi-output, jaringan saraf dengan keluaran sigmoid
                  - **TIDAK** Apakah jumlah kelas lebih dari dua?
                    - **YA** **H17 · Klasifikasi multikelas (data tabular)**
                      Random Forest, SVM (one-vs-rest atau one-vs-one), XGBoost, LightGBM, CatBoost, MLP
                    - **TIDAK** **H18 · Klasifikasi biner (data tabular)**
                      Random Forest, SVM (kernel RBF), XGBoost, LightGBM, CatBoost, MLP
                - **TIDAK** Apakah target berupa waktu sampai suatu kejadian, dan sebagian data belum mengalami kejadian itu?
                  - **YA** **H19 · Analisis survival (waktu sampai kejadian)**
                    Random Survival Forest, Gradient Boosting Survival, Survival SVM, Accelerated Failure Time
                  - **TIDAK** Apakah tujuannya mengurutkan sekumpulan item menurut relevansi atau peringkat?
                    - **YA** **H20 · Learning to rank (pemeringkatan)**
                      RankNet, LambdaMART (LightGBM/XGBoost lambdarank), RankSVM
                    - **TIDAK** **H21 · Regresi (prediksi nilai kontinu)**
                      Random Forest, SVR, XGBoost, LightGBM, CatBoost, Gaussian Process Regression
- **TIDAK** Apakah sistem belajar dari interaksi coba-coba dengan lingkungan dan menerima reward?
  - **YA** Apakah setiap keputusan berdiri sendiri, tanpa memengaruhi keadaan berikutnya?
    - **YA** **H22 · Multi-armed bandit**
      UCB1, Thompson Sampling
    - **TIDAK** Apakah Anda hanya memiliki log interaksi lama dan tidak bisa mencoba langsung di lingkungan atau simulator?
      - **YA** **H23 · Offline reinforcement learning**
        CQL, IQL, BCQ, TD3+BC
      - **TIDAK** Apakah aksinya bernilai kontinu?
        - **YA** **H24 · Deep RL dengan aksi kontinu**
          PPO, SAC, TD3, DDPG
        - **TIDAK** Apakah jumlah keadaan kecil sehingga bisa ditulis dalam tabel?
          - **YA** **H25 · RL tabular**
            Q-Learning, SARSA, Expected SARSA, Monte Carlo control
          - **TIDAK** **H26 · Deep RL dengan aksi diskret**
            Double DQN, Dueling DQN, Rainbow, A2C, PPO
  - **TIDAK** Apakah tujuannya menghasilkan data baru (teks, citra, atau data sintetis)?
    - **YA** Apakah yang dihasilkan berupa teks?
      - **YA** **H27 · Generasi teks / model bahasa**
        Fine-tuning GPT-2/T5; LLM terbuka (Llama, Mistral, Qwen) dengan LoRA/QLoRA
      - **TIDAK** Apakah yang dihasilkan berupa citra?
        - **YA** **H28 · Generasi citra**
          DCGAN, StyleGAN, Pix2Pix, CycleGAN
        - **TIDAK** **H29 · Data tabular sintetis**
          CTGAN, TVAE
    - **TIDAK** Apakah tujuannya menemukan data yang menyimpang dari pola normal?
      - **YA** Apakah datanya deret waktu atau aliran data sensor/log?
        - **YA** **H30 · Deteksi anomali deret waktu**
          Isolation Forest pada fitur jendela, Matrix Profile, galat prediksi atau rekonstruksi LSTM/Autoencoder
        - **TIDAK** **H31 · Deteksi anomali (data tabular atau citra)**
          Isolation Forest, Local Outlier Factor, One-Class SVM, ECOD/COPOD, GMM
      - **TIDAK** Apakah tujuannya mengelompokkan data yang mirip ke dalam grup?
        - **YA** Apakah jumlah klaster sudah diketahui atau ditentukan?
          - **YA** Apakah klaster diperkirakan kompak (membulat) dan berukuran mirip?
            - **YA** **H32 · Clustering partisi (jumlah klaster diketahui)**
              Mini-Batch K-Means, Gaussian Mixture Model, K-Medoids, Fuzzy C-Means, K-Prototypes (data campuran)
            - **TIDAK** **H33 · Clustering untuk bentuk klaster tidak bulat**
              Spectral Clustering, Gaussian Mixture Model dengan kovarians penuh
          - **TIDAK** Apakah data mengandung banyak noise atau klaster berbentuk tidak beraturan?
            - **YA** **H34 · Clustering berbasis kepadatan**
              HDBSCAN, OPTICS, Mean Shift
            - **TIDAK** **H35 · Clustering dengan jumlah klaster belum diketahui**
              K-Means dengan metode elbow, silhouette, atau gap statistic; GMM dengan BIC/AIC; X-Means
        - **TIDAK** Apakah tujuannya mencari aturan 'jika A maka B' pada data transaksi?
          - **YA** **H36 · Aturan asosiasi / pola frekuen**
            FP-Growth, ECLAT
          - **TIDAK** Apakah tujuannya menemukan topik tersembunyi pada kumpulan dokumen?
            - **YA** **H37 · Pemodelan topik**
              Latent Dirichlet Allocation (LDA), NMF
            - **TIDAK** Apakah tujuannya mengurangi jumlah fitur atau memvisualkan data berdimensi tinggi?
              - **YA** Apakah tujuan utamanya visualisasi 2D/3D?
                - **YA** **H38 · Visualisasi data berdimensi tinggi**
                  t-SNE, UMAP
                - **TIDAK** **H39 · Reduksi dimensi dan seleksi fitur**
                  LDA (supervised), Kernel PCA, ICA, NMF; seleksi fitur: mutual information, chi-square, RFE, Lasso, Boruta
              - **TIDAK** Apakah tujuannya mempelajari representasi dari data tak berlabel untuk dipakai pada tugas lain?
                - **YA** **H40 · Self-supervised representation learning**
                  Citra: SimCLR, MoCo, BYOL, DINO, MAE. Teks: masked language modeling. Audio: wav2vec 2.0. Deret waktu: TS2Vec. Tabular: SCARF
                - **TIDAK** **H41 · Masalah belum terpetakan**
                  Rumuskan ulang: apa masukan, apa keluaran yang diinginkan, dan bagaimana keberhasilan diukur

Rincian Hasil Peta Keputusan Machine Learning

Untuk setiap hasil: algoritma (baseline, utama, lanjutan), dataset pembanding, praproses, metrik, validasi, pustaka, jebakan umum, dan contoh judul.

Rincian setiap hasil

H1 · Active learning
Supervised learning dengan pelabelan bertahap

- **Baseline**: Random sampling (pembanding wajib)
- **Utama**: Uncertainty sampling (least confidence, margin, entropy), Query-by-Committee
- **Lanjutan**: BALD, Core-Set, BADGE, kombinasi dengan semi-supervised
- **Dataset**: Benchmark berlabel penuh (CIFAR-10, AG News, dan sejenisnya) dengan label disembunyikan untuk simulasi; data sendiri dengan ahli sebagai oracle
- **Praproses**: Sama seperti tugas supervised-nya; siapkan pool tak berlabel dan seed set kecil
- **Metrik**: Kurva belajar (performa terhadap jumlah label), area di bawah kurva belajar, penghematan label
- **Validasi**: Ulangi dengan beberapa seed awal; test set tetap dan tidak pernah dipilih untuk dilabeli
- **Pustaka**: modAL, scikit-activeml, Label Studio
- **Jebakan**: Tidak membandingkan dengan random sampling; sampling bias; biaya dan ketidakkonsistenan pelabel diabaikan
- **Contoh judul**: Pelabelan hemat untuk klasifikasi keluhan mahasiswa dengan uncertainty sampling

H2 · Semi-supervised learning
Semi-supervised learning

- **Baseline**: Model supervised yang dilatih hanya dengan data berlabel (batas bawah)
- **Utama**: Self-training/pseudo-labeling, Label Propagation, Label Spreading, co-training
- **Lanjutan**: FixMatch, MixMatch, Mean Teacher, FlexMatch, UDA
- **Dataset**: CIFAR-10 (250 atau 4.000 label), SVHN, STL-10; atau data sendiri dengan sebagian label disembunyikan
- **Praproses**: Sama seperti tugas supervised-nya; augmentasi lemah dan kuat untuk metode consistency
- **Metrik**: Metrik tugas (accuracy, F1) pada test set berlabel; kurva performa terhadap jumlah label
- **Validasi**: Test set berlabel yang bersih dan terpisah; ulangi dengan beberapa subset label acak
- **Pustaka**: scikit-learn (semi_supervised), USB (Unified SSL Benchmark), PyTorch
- **Jebakan**: Pseudo-label salah memperkuat dirinya sendiri (confirmation bias); distribusi data tak berlabel berbeda dari data berlabel
- **Contoh judul**: Klasifikasi citra medis dengan FixMatch saat hanya 5% data berlabel

H3 · Klasifikasi video / pengenalan aksi
Supervised deep learning

- **Baseline**: CNN per frame + rata-rata, atau CNN + LSTM
- **Utama**: 3D CNN (I3D, R(2+1)D), SlowFast, TSN/TSM
- **Lanjutan**: TimeSformer, Video Swin, VideoMAE
- **Dataset**: UCF101, HMDB51, Kinetics-400, Something-Something V2, NTU RGB+D
- **Praproses**: Sampling frame, resize, augmentasi temporal dan spasial; opsional ekstraksi pose/optical flow
- **Metrik**: Top-1 dan top-5 accuracy, F1 per kelas, mAP
- **Validasi**: Split resmi dataset; pembagian per subjek atau per video sumber
- **Pustaka**: PyTorchVideo, MMAction2, Hugging Face Transformers
- **Jebakan**: Klip dari video sumber yang sama berada di latih dan uji; biaya komputasi diremehkan
- **Contoh judul**: Pengenalan aktivitas mencurigakan saat ujian daring dengan VideoMAE

H4 · Segmentasi citra
Supervised deep learning

- **Baseline**: U-Net (semantik)
- **Utama**: DeepLabv3+, Mask R-CNN (instance), U-Net++, nnU-Net (medis)
- **Lanjutan**: SegFormer, Mask2Former, Segment Anything (SAM) dengan prompt atau fine-tuning
- **Dataset**: Cityscapes, ADE20K, Pascal VOC, MS COCO, BraTS (tumor otak), ISIC, DeepGlobe (citra satelit)
- **Praproses**: Anotasi mask piksel; augmentasi geometris yang diterapkan sama pada citra dan mask; patching untuk citra besar
- **Metrik**: mIoU, Dice coefficient, pixel accuracy, Hausdorff distance (medis)
- **Validasi**: K-fold per pasien atau per wilayah; test set terpisah
- **Pustaka**: segmentation_models.pytorch, MMSegmentation, MONAI, Detectron2
- **Jebakan**: Pixel accuracy tinggi karena latar belakang dominan; mask dan citra tidak selaras setelah augmentasi
- **Contoh judul**: Segmentasi lahan sawit dari citra satelit Sentinel-2 dengan SegFormer

H5 · Deteksi objek
Supervised deep learning

- **Baseline**: Faster R-CNN atau SSD pra-latih
- **Utama**: YOLO (v5/v8/v11), RetinaNet, Faster R-CNN dengan FPN
- **Lanjutan**: DETR, RT-DETR, Grounding DINO (open-vocabulary)
- **Dataset**: MS COCO, Pascal VOC, Open Images, KITTI, BDD100K, VisDrone
- **Praproses**: Anotasi bounding box (CVAT, Roboflow, LabelImg); augmentasi mosaic, skala, flip
- **Metrik**: mAP@0.5, mAP@[0.5:0.95], precision, recall, IoU, FPS/latensi
- **Validasi**: Train/validation/test tetap; pembagian per lokasi atau per video agar tidak bocor
- **Pustaka**: Ultralytics, Detectron2, MMDetection
- **Jebakan**: Frame berurutan dari video yang sama masuk ke latih dan uji; anotasi tidak konsisten; objek kecil terabaikan
- **Contoh judul**: Deteksi pelanggaran helm pengendara motor secara real-time dengan YOLO

H6 · Klasifikasi citra
Supervised deep learning

- **Baseline**: CNN sederhana, atau fitur HOG/LBP + SVM
- **Utama**: Transfer learning: ResNet, EfficientNet, DenseNet, MobileNet
- **Lanjutan**: Vision Transformer (ViT), Swin, ConvNeXt, model pra-latih self-supervised (DINOv2)
- **Dataset**: MNIST, Fashion-MNIST, CIFAR-10/100, ImageNet, PlantVillage, ISIC (kulit), ChestX-ray14
- **Praproses**: Resize, normalisasi, augmentasi (flip, rotasi, crop, color jitter, MixUp/CutMix)
- **Metrik**: Accuracy, top-5 accuracy, precision, recall, F1 per kelas, confusion matrix, ROC-AUC
- **Validasi**: Train/validation/test atau stratified k-fold; pembagian per pasien/objek bila satu sumber punya banyak citra
- **Pustaka**: PyTorch + torchvision/timm, TensorFlow/Keras, Albumentations, Grad-CAM
- **Jebakan**: Citra dari pasien atau objek yang sama tersebar di data latih dan uji; augmentasi diterapkan pada data uji; overfitting pada data kecil
- **Contoh judul**: Klasifikasi penyakit daun padi dengan EfficientNet dan Grad-CAM

H7 · Teks ke teks (terjemahan, ringkasan, tanya jawab)
Supervised learning / NLP generatif

- **Baseline**: Lead-3 atau TextRank (ringkasan), BM25 + ekstraksi (tanya jawab)
- **Utama**: Transformer encoder-decoder: T5, BART, mT5, mBART, IndoBART, IndoT5
- **Lanjutan**: LLM dengan LoRA/QLoRA, Retrieval-Augmented Generation (RAG)
- **Dataset**: WMT dan OPUS (terjemahan), CNN/DailyMail, XSum, IndoSum, Liputan6 (ringkasan), SQuAD, TyDi QA (tanya jawab)
- **Praproses**: Pembersihan teks, tokenisasi subword, pemotongan panjang masukan, pasangan sumber-target yang selaras
- **Metrik**: BLEU, chrF, ROUGE, METEOR, BERTScore, Exact Match dan F1 (tanya jawab), penilaian manusia
- **Validasi**: Split resmi benchmark; penilaian manusia pada sampel acak dengan lebih dari satu penilai
- **Pustaka**: Hugging Face Transformers, PEFT, SacreBLEU, evaluate, LangChain/LlamaIndex
- **Jebakan**: Hanya mengandalkan metrik tumpang-tindih kata; halusinasi fakta tidak terukur; data uji bocor ke data pra-latih LLM
- **Contoh judul**: Peringkasan otomatis skripsi berbahasa Indonesia dengan IndoT5

H8 · Pelabelan sekuens (NER, POS tagging)
Supervised learning / NLP

- **Baseline**: Conditional Random Field (CRF) dengan fitur buatan
- **Utama**: BiLSTM-CRF dengan word/character embedding
- **Lanjutan**: Fine-tuning BERT/IndoBERT/XLM-R untuk token classification; LLM dengan prompt terstruktur
- **Dataset**: CoNLL-2003, OntoNotes 5.0, WikiANN, Universal Dependencies, NERGrit dan POSP (IndoNLU)
- **Praproses**: Tokenisasi, skema label BIO/BIOES, penyelarasan label dengan subword
- **Metrik**: Precision, recall, F1 tingkat entitas (seqeval); accuracy untuk POS
- **Validasi**: Split per dokumen; k-fold bila data kecil
- **Pustaka**: sklearn-crfsuite, spaCy, Hugging Face Transformers, seqeval, Label Studio
- **Jebakan**: Menghitung F1 per token (seharusnya per entitas); kalimat dari dokumen yang sama di latih dan uji
- **Contoh judul**: Ekstraksi entitas pada dokumen peraturan akademik berbahasa Indonesia

H9 · Klasifikasi teks
Supervised learning / NLP

- **Baseline**: TF-IDF + Naive Bayes, Logistic Regression, atau SVM linear
- **Utama**: fastText, TextCNN, BiLSTM dengan word embedding
- **Lanjutan**: Fine-tuning BERT, IndoBERT, XLM-RoBERTa; LLM zero-shot/few-shot; SetFit untuk data sedikit
- **Dataset**: IMDb, SST-2, AG News, 20 Newsgroups, IndoNLU (SmSA, EmoT), IndoLEM
- **Praproses**: Case folding, pembersihan (URL, emoji, tanda baca), tokenisasi, stopword removal dan stemming (untuk model klasik), normalisasi kata tidak baku
- **Metrik**: Accuracy, precision, recall, macro-F1, confusion matrix
- **Validasi**: Stratified k-fold atau split resmi benchmark; laporkan rata-rata beberapa seed untuk model Transformer
- **Pustaka**: scikit-learn, Sastrawi, Hugging Face Transformers, fastText
- **Jebakan**: Vectorizer di-fit pada seluruh data sebelum split; duplikat teks di latih dan uji; stemming merusak masukan model Transformer
- **Contoh judul**: Analisis sentimen ulasan layanan akademik dengan IndoBERT

H10 · Pengenalan ucapan (speech to text)
Supervised deep learning

- **Baseline**: Whisper pra-latih tanpa pelatihan ulang (zero-shot)
- **Utama**: wav2vec 2.0 atau HuBERT + CTC, Conformer
- **Lanjutan**: Fine-tuning Whisper, penambahan language model, adaptasi domain/aksen
- **Dataset**: LibriSpeech, Mozilla Common Voice (tersedia bahasa Indonesia), FLEURS, TED-LIUM
- **Praproses**: Resampling 16 kHz, pemotongan senyap, normalisasi teks transkrip, augmentasi (SpecAugment, noise)
- **Metrik**: Word Error Rate (WER), Character Error Rate (CER), real-time factor
- **Validasi**: Pembagian per penutur; test set dengan kondisi akustik berbeda
- **Pustaka**: Hugging Face Transformers, SpeechBrain, ESPnet, torchaudio, jiwer
- **Jebakan**: Penutur yang sama di latih dan uji; normalisasi teks berbeda antara referensi dan hipotesis
- **Contoh judul**: Transkripsi otomatis kuliah berbahasa Indonesia dengan fine-tuning Whisper

H11 · Klasifikasi audio / sinyal
Supervised learning

- **Baseline**: Fitur MFCC atau statistik sinyal + SVM/Random Forest
- **Utama**: CNN pada log-mel spectrogram, CRNN, 1D-CNN pada sinyal mentah
- **Lanjutan**: Audio Spectrogram Transformer (AST), PANNs, embedding wav2vec 2.0/HuBERT
- **Dataset**: ESC-50, UrbanSound8K, AudioSet, Speech Commands, GTZAN; sinyal: PhysioNet MIT-BIH (EKG), CWRU Bearing, UCI HAR
- **Praproses**: Resampling, filtering, segmentasi jendela, ekstraksi spektrogram, normalisasi, augmentasi (time shift, noise)
- **Metrik**: Accuracy, macro-F1, confusion matrix, ROC-AUC; sensitivitas dan spesifisitas untuk data medis
- **Validasi**: Pembagian per penutur, subjek, atau rekaman (GroupKFold); fold resmi dataset
- **Pustaka**: librosa, torchaudio, SciPy, scikit-learn, PyTorch
- **Jebakan**: Jendela dari rekaman yang sama masuk ke latih dan uji sehingga akurasi semu sangat tinggi
- **Contoh judul**: Klasifikasi aritmia dari sinyal EKG dengan 1D-CNN

H12 · Pembelajaran pada graf (node, link, atau graf)
Supervised / semi-supervised pada graf

- **Baseline**: Fitur node + Logistic Regression/MLP, Label Propagation, DeepWalk atau node2vec
- **Utama**: GCN, GraphSAGE, GAT, GIN
- **Lanjutan**: Graph Transformer, GNN heterogen (R-GCN, HAN), GNN temporal (TGN)
- **Dataset**: Cora, CiteSeer, PubMed, Open Graph Benchmark (OGB), Reddit, PPI, MoleculeNet/QM9, Elliptic (transaksi Bitcoin)
- **Praproses**: Konstruksi graf (definisi node dan edge), fitur node, normalisasi adjacency
- **Metrik**: Node: accuracy, macro-F1. Link: ROC-AUC, Hits@k, MRR. Graf: accuracy atau MAE
- **Validasi**: Split resmi; untuk link prediction sembunyikan edge uji dari graf pelatihan; split temporal bila graf berkembang
- **Pustaka**: PyTorch Geometric, DGL, NetworkX
- **Jebakan**: Edge uji ikut terlihat saat pelatihan; over-smoothing pada GNN yang terlalu dalam
- **Contoh judul**: Prediksi kolaborasi peneliti dari jaringan ko-penulisan dengan GraphSAGE

H13 · Peramalan deret waktu (forecasting)
Supervised learning pada deret waktu

- **Baseline**: Naive, seasonal naive, ARIMA/SARIMA, Exponential Smoothing (ETS), Prophet
- **Utama**: XGBoost/LightGBM dengan fitur lag, LSTM, GRU, TCN
- **Lanjutan**: N-BEATS, N-HiTS, Temporal Fusion Transformer, PatchTST, DLinear, model fondasi (Chronos, TimesFM)
- **Dataset**: M4, M5, ETT, Electricity, Traffic, Weather; data lokal dari BPS, BMKG, Bank Indonesia, Yahoo Finance
- **Praproses**: Penanganan nilai hilang, uji stasioneritas (ADF), differencing, dekomposisi musiman, fitur lag dan kalender, scaling yang di-fit pada data latih
- **Metrik**: MAE, RMSE, sMAPE, MASE; untuk probabilistik: CRPS dan pinball loss
- **Validasi**: Rolling-origin atau time series split; data tidak boleh diacak; uji Diebold-Mariano antar-model
- **Pustaka**: statsmodels, Prophet, Darts, sktime, NeuralForecast, GluonTS
- **Jebakan**: Mengacak data sebelum split; memakai informasi masa depan sebagai fitur; tidak dibandingkan dengan naive forecast
- **Contoh judul**: Peramalan jumlah pendaftar mahasiswa baru dengan Temporal Fusion Transformer

H14 · Klasifikasi deret waktu
Supervised learning pada deret waktu

- **Baseline**: 1-NN dengan Dynamic Time Warping
- **Utama**: ROCKET/MiniRocket, fitur statistik (tsfresh) + Random Forest, 1D-CNN, LSTM
- **Lanjutan**: InceptionTime, HIVE-COTE 2.0, Transformer deret waktu
- **Dataset**: UCR/UEA Time Series Archive, UCI HAR, WISDM, PhysioNet
- **Praproses**: Segmentasi jendela, normalisasi z per deret, resampling, filtering
- **Metrik**: Accuracy, macro-F1, confusion matrix
- **Validasi**: Split resmi arsip UCR; pembagian per subjek (leave-one-subject-out) untuk data sensor manusia
- **Pustaka**: sktime, aeon, tsai, tslearn
- **Jebakan**: Jendela bertumpang-tindih dari subjek yang sama di latih dan uji
- **Contoh judul**: Pengenalan aktivitas manusia dari sensor ponsel dengan MiniRocket

H15 · Sistem rekomendasi
Supervised / collaborative filtering

- **Baseline**: Most-popular, item-based KNN, user-based KNN
- **Utama**: Matrix factorization (SVD, ALS), BPR, Factorization Machines, content-based (TF-IDF)
- **Lanjutan**: Neural Collaborative Filtering, LightGCN, SASRec/BERT4Rec (sekuensial), two-tower
- **Dataset**: MovieLens (100K/1M/20M), Amazon Reviews, Netflix Prize, Book-Crossing, Yelp, Last.fm
- **Praproses**: Matriks interaksi user-item; pilih umpan balik eksplisit (rating) atau implisit (klik); filter user/item dengan interaksi minimal
- **Metrik**: Recall@k, NDCG@k, Hit Rate@k, MAP; RMSE/MAE untuk prediksi rating; coverage dan diversity
- **Validasi**: Split temporal atau leave-one-out per pengguna; negative sampling yang dilaporkan dengan jelas
- **Pustaka**: Surprise, implicit, LightFM, RecBole, PyTorch
- **Jebakan**: Split acak membocorkan interaksi masa depan; bias popularitas; cold start pengguna atau item baru
- **Contoh judul**: Rekomendasi mata kuliah pilihan dengan LightGCN dari riwayat pengambilan mata kuliah

H16 · Klasifikasi multi-label
Supervised learning

- **Baseline**: Binary Relevance dengan Logistic Regression
- **Utama**: Classifier Chains, Label Powerset, ML-kNN, Random Forest multi-output, jaringan saraf dengan keluaran sigmoid
- **Lanjutan**: Transformer dengan kepala sigmoid, model yang memodelkan korelasi label (label attention, graph label)
- **Dataset**: Yeast, Emotions, Scene (repositori Mulan), Reuters-21578, label objek MS-COCO, Toxic Comment (Jigsaw)
- **Praproses**: Sama seperti klasifikasi biasa; label dibuat menjadi vektor biner (multi-hot)
- **Metrik**: Hamming loss, subset accuracy, micro-F1, macro-F1, Jaccard score, ranking loss
- **Validasi**: Iterative stratification (k-fold khusus multi-label); ambang keputusan ditala per label pada data validasi
- **Pustaka**: scikit-learn (MultiOutputClassifier, ClassifierChain), scikit-multilearn, PyTorch
- **Jebakan**: Memakai softmax (seharusnya sigmoid per label); mengabaikan label langka; subset accuracy saja terlalu keras
- **Contoh judul**: Klasifikasi multi-label topik artikel jurnal berbahasa Indonesia

H17 · Klasifikasi multikelas (data tabular)
Supervised learning

- **Baseline**: Multinomial Logistic Regression, Naive Bayes, KNN, Decision Tree
- **Utama**: Random Forest, SVM (one-vs-rest atau one-vs-one), XGBoost, LightGBM, CatBoost, MLP
- **Lanjutan**: Stacking ensemble, TabNet, FT-Transformer, klasifikasi hierarkis bila kelas bertingkat
- **Dataset**: Iris, Wine, Covertype, Dry Bean, Letter Recognition, Predict Students' Dropout and Academic Success (UCI)
- **Praproses**: Imputasi, encoding kategori, scaling, seleksi fitur; periksa jumlah sampel per kelas
- **Metrik**: Accuracy, macro-F1, weighted-F1, balanced accuracy, Cohen's kappa, confusion matrix per kelas
- **Validasi**: Stratified k-fold CV; uji Friedman + Nemenyi bila membandingkan banyak model
- **Pustaka**: scikit-learn, XGBoost, LightGBM, CatBoost, Optuna, SHAP
- **Jebakan**: Hanya melaporkan accuracy padahal kelas kecil gagal dikenali; kelas dengan sampel terlalu sedikit untuk dibagi ke semua fold
- **Contoh judul**: Klasifikasi tingkat risiko putus studi mahasiswa menjadi tiga kelas dengan LightGBM

H18 · Klasifikasi biner (data tabular)
Supervised learning

- **Baseline**: Logistic Regression, Naive Bayes, KNN, Decision Tree
- **Utama**: Random Forest, SVM (kernel RBF), XGBoost, LightGBM, CatBoost, MLP
- **Lanjutan**: Stacking ensemble, TabNet, FT-Transformer, TabPFN (data kecil), AutoML (AutoGluon)
- **Dataset**: UCI Adult, Breast Cancer Wisconsin, Pima Indians Diabetes, Bank Marketing, Telco Customer Churn, Credit Card Fraud (ULB)
- **Praproses**: Imputasi nilai hilang, encoding kategori (one-hot atau target encoding), scaling untuk model berbasis jarak dan SVM, seleksi fitur
- **Metrik**: Accuracy, precision, recall, F1, ROC-AUC, PR-AUC, MCC, confusion matrix
- **Validasi**: Stratified 10-fold cross validation; nested CV bila ada tuning; uji McNemar atau Wilcoxon antar-model
- **Pustaka**: scikit-learn, XGBoost, LightGBM, CatBoost, imbalanced-learn, Optuna, SHAP
- **Jebakan**: Data leakage (scaling atau SMOTE dilakukan sebelum split); accuracy menyesatkan pada kelas timpang; data uji sama dengan data latih
- **Contoh judul**: Prediksi kelulusan tepat waktu mahasiswa dengan gradient boosting dan penjelasan SHAP

H19 · Analisis survival (waktu sampai kejadian)
Supervised learning dengan data tersensor

- **Baseline**: Kaplan-Meier (deskriptif), Cox Proportional Hazards
- **Utama**: Random Survival Forest, Gradient Boosting Survival, Survival SVM, Accelerated Failure Time
- **Lanjutan**: DeepSurv, DeepHit, model competing risks
- **Dataset**: METABRIC, SUPPORT, GBSG2, Rossi recidivism, Veterans' Lung Cancer
- **Praproses**: Definisikan waktu mulai, kejadian, dan penyensoran dengan jelas; imputasi; encoding
- **Metrik**: Concordance index (C-index), integrated Brier score, time-dependent AUC
- **Validasi**: K-fold CV terstratifikasi menurut status kejadian; uji asumsi proportional hazards (Schoenfeld)
- **Pustaka**: lifelines, scikit-survival, pycox
- **Jebakan**: Membuang data tersensor atau memperlakukannya sebagai regresi biasa; fitur yang baru diketahui setelah waktu mulai
- **Contoh judul**: Pemodelan waktu sampai mahasiswa putus studi dengan Random Survival Forest

H20 · Learning to rank (pemeringkatan)
Supervised learning

- **Baseline**: BM25 atau skor heuristik; regresi pointwise
- **Utama**: RankNet, LambdaMART (LightGBM/XGBoost lambdarank), RankSVM
- **Lanjutan**: Neural ranker: cross-encoder BERT, ColBERT, listwise loss (ListNet, ApproxNDCG)
- **Dataset**: MSLR-WEB10K/30K, Yahoo! Learning to Rank Challenge, Istella LETOR, MS MARCO
- **Praproses**: Fitur per pasangan kueri-item; data dikelompokkan per kueri
- **Metrik**: NDCG@k, MAP, MRR, Precision@k, Kendall tau
- **Validasi**: Pembagian data per kueri (bukan per baris); k-fold berbasis grup
- **Pustaka**: LightGBM, XGBoost, TensorFlow Ranking, Sentence-Transformers, Pyserini
- **Jebakan**: Baris dari kueri yang sama masuk ke data latih dan uji sekaligus; bias posisi pada data klik
- **Contoh judul**: Pemeringkatan kinerja dosen yang adil dan dapat dijelaskan dengan LambdaMART

H21 · Regresi (prediksi nilai kontinu)
Supervised learning

- **Baseline**: Linear Regression, Ridge, Lasso, ElasticNet, KNN Regressor
- **Utama**: Random Forest, SVR, XGBoost, LightGBM, CatBoost, Gaussian Process Regression
- **Lanjutan**: Stacking, MLP atau TabNet, quantile regression untuk interval prediksi
- **Dataset**: California Housing, Ames Housing, Bike Sharing, Concrete Compressive Strength, Wine Quality, Student Performance (UCI)
- **Praproses**: Imputasi, penanganan outlier, transformasi target (log) bila sangat miring, encoding, scaling
- **Metrik**: MAE, RMSE, R^2, MAPE; plot residual dan plot prediksi vs aktual
- **Validasi**: K-fold CV (berulang bila data kecil); uji Wilcoxon atas galat antar-model
- **Pustaka**: scikit-learn, XGBoost, LightGBM, statsmodels, Optuna, SHAP
- **Jebakan**: MAPE meledak saat nilai aktual mendekati nol; R^2 tinggi karena leakage; ekstrapolasi di luar rentang data latih
- **Contoh judul**: Prediksi IPK akhir mahasiswa dari data dua semester pertama

H22 · Multi-armed bandit
Reinforcement learning (tanpa transisi keadaan)

- **Baseline**: Epsilon-greedy
- **Utama**: UCB1, Thompson Sampling
- **Lanjutan**: Contextual bandit: LinUCB, LinTS, neural bandit
- **Dataset**: Open Bandit Dataset, Yahoo! R6 (rekomendasi berita), simulasi dari MovieLens
- **Praproses**: Definisikan lengan (aksi), konteks, dan reward; catat probabilitas aksi pada log
- **Metrik**: Cumulative regret, reward rata-rata, off-policy evaluation (IPS, doubly robust)
- **Validasi**: Simulasi berulang dengan banyak seed; evaluasi off-policy pada data log
- **Pustaka**: Vowpal Wabbit, Open Bandit Pipeline, MABWiser
- **Jebakan**: Mengevaluasi kebijakan baru pada log lama tanpa koreksi bias; reward tertunda
- **Contoh judul**: Pemilihan materi latihan adaptif untuk mahasiswa dengan contextual bandit

H23 · Offline reinforcement learning
Reinforcement learning dari data log

- **Baseline**: Behavior Cloning
- **Utama**: CQL, IQL, BCQ, TD3+BC
- **Lanjutan**: Decision Transformer, model-based offline RL (MOPO, COMBO)
- **Dataset**: D4RL, Minari, RL Unplugged; log operasional sendiri
- **Praproses**: Susun log menjadi tupel (keadaan, aksi, reward, keadaan berikutnya); normalisasi keadaan
- **Metrik**: Normalized return, off-policy evaluation (Fitted Q Evaluation, importance sampling)
- **Validasi**: Evaluasi di simulator bila ada; beberapa seed; laporkan kualitas data log (acak, medium, expert)
- **Pustaka**: d3rlpy, CORL, Minari
- **Jebakan**: Nilai aksi di luar distribusi data ditaksir terlalu tinggi; kebijakan diuji langsung di dunia nyata tanpa pengaman
- **Contoh judul**: Optimasi penjadwalan dari log historis dengan Conservative Q-Learning

H24 · Deep RL dengan aksi kontinu
Reinforcement learning

- **Baseline**: Kebijakan acak, pengendali PID, atau REINFORCE
- **Utama**: PPO, SAC, TD3, DDPG
- **Lanjutan**: Model-based RL (Dreamer, MBPO), sim-to-real dengan domain randomization
- **Dataset**: Lingkungan: MuJoCo (Gymnasium), DeepMind Control Suite, PyBullet, Isaac Lab
- **Praproses**: Rancang ruang keadaan, aksi, dan fungsi reward; normalisasi observasi dan reward
- **Metrik**: Average return, sample efficiency, tingkat keberhasilan; interquartile mean dengan selang kepercayaan
- **Validasi**: Minimal 5 sampai 10 seed; evaluasi kebijakan deterministik pada episode terpisah
- **Pustaka**: Stable-Baselines3, CleanRL, RLlib, rliable
- **Jebakan**: Satu seed saja; reward shaping menghasilkan perilaku yang tidak diinginkan; hasil di simulator tidak berpindah ke dunia nyata
- **Contoh judul**: Kendali lengan robot dengan Soft Actor-Critic

H25 · RL tabular
Reinforcement learning

- **Baseline**: Kebijakan acak
- **Utama**: Q-Learning, SARSA, Expected SARSA, Monte Carlo control
- **Lanjutan**: Dyna-Q; value iteration atau policy iteration bila model lingkungan diketahui
- **Dataset**: Lingkungan: FrozenLake, Taxi, CliffWalking (Gymnasium), GridWorld buatan sendiri
- **Praproses**: Diskretisasi keadaan dan aksi; rancang fungsi reward
- **Metrik**: Reward kumulatif per episode, jumlah langkah menuju tujuan, kurva konvergensi
- **Validasi**: Rata-rata banyak run; bandingkan jadwal epsilon dan learning rate
- **Pustaka**: Gymnasium, NumPy
- **Jebakan**: Eksplorasi kurang; ruang keadaan ternyata terlalu besar untuk tabel
- **Contoh judul**: Penentuan rute robot gudang pada grid dengan Q-Learning

H26 · Deep RL dengan aksi diskret
Reinforcement learning

- **Baseline**: DQN
- **Utama**: Double DQN, Dueling DQN, Rainbow, A2C, PPO
- **Lanjutan**: MuZero, IMPALA, multi-agent (QMIX, MAPPO)
- **Dataset**: Lingkungan: CartPole, LunarLander, Atari (ALE), MiniGrid, PettingZoo (multi-agent)
- **Praproses**: Rancang observasi, aksi, reward; frame stacking dan normalisasi
- **Metrik**: Average return, human-normalized score (Atari), sample efficiency
- **Validasi**: Minimal 5 seed; evaluasi terpisah dari pelatihan; laporkan selang kepercayaan
- **Pustaka**: Stable-Baselines3, CleanRL, RLlib, PettingZoo
- **Jebakan**: Pelatihan tidak stabil dan sensitif hyperparameter; satu seed; lingkungan uji sama persis dengan lingkungan latih
- **Contoh judul**: Pengaturan lampu lalu lintas adaptif dengan PPO pada simulator SUMO

H27 · Generasi teks / model bahasa
Generative AI

- **Baseline**: Model n-gram atau template
- **Utama**: Fine-tuning GPT-2/T5; LLM terbuka (Llama, Mistral, Qwen) dengan LoRA/QLoRA
- **Lanjutan**: Retrieval-Augmented Generation (RAG), agen dengan tool, instruction tuning
- **Dataset**: WikiText-103, OpenWebText, dataset instruksi (Alpaca, Dolly), korpus domain sendiri
- **Praproses**: Pembersihan korpus, format instruksi-jawaban, chunking dokumen untuk RAG
- **Metrik**: Perplexity, BLEU/ROUGE, BERTScore, faithfulness dan answer relevancy (RAGAS), penilaian manusia
- **Validasi**: Set uji yang tidak pernah dilihat; penilaian buta oleh beberapa penilai; uji kesepakatan antar-penilai
- **Pustaka**: Hugging Face Transformers, PEFT, LangChain, LlamaIndex, Ollama, vLLM
- **Jebakan**: Halusinasi; data uji sudah ada di data pra-latih; evaluasi hanya dengan metrik otomatis
- **Contoh judul**: Chatbot layanan akademik berbasis RAG atas dokumen peraturan kampus

H28 · Generasi citra
Generative AI

- **Baseline**: Variational Autoencoder (VAE)
- **Utama**: DCGAN, StyleGAN, Pix2Pix, CycleGAN
- **Lanjutan**: DDPM, Stable Diffusion dengan LoRA, DreamBooth, atau ControlNet
- **Dataset**: MNIST, CIFAR-10, CelebA, FFHQ, LSUN, LAION (teks ke citra)
- **Praproses**: Resize dan crop seragam, normalisasi ke [-1, 1], pasangan teks-citra bila bersyarat
- **Metrik**: FID, KID, Inception Score, precision/recall generatif, CLIP score
- **Validasi**: FID dihitung terhadap data nyata yang tidak dipakai melatih; jumlah sampel memadai (umumnya puluhan ribu)
- **Pustaka**: PyTorch, Diffusers, torch-fidelity
- **Jebakan**: Mode collapse pada GAN; FID bias pada sampel sedikit; model menghafal data latih; isu hak cipta
- **Contoh judul**: Augmentasi citra medis langka dengan diffusion model

H29 · Data tabular sintetis
Generative AI

- **Baseline**: SMOTE atau Gaussian Copula
- **Utama**: CTGAN, TVAE
- **Lanjutan**: TabDDPM, TabSyn, model berbasis LLM (GReaT)
- **Dataset**: Adult, Covertype, Credit Default; data sendiri yang sensitif
- **Praproses**: Tandai tipe kolom (kategori, numerik, tanggal), tangani nilai hilang
- **Metrik**: Fidelity (uji KS, selisih korelasi), utility (train on synthetic, test on real), privasi (distance to closest record, membership inference)
- **Validasi**: Bandingkan model yang dilatih pada data sintetis terhadap model pada data asli, diuji pada data asli yang terpisah
- **Pustaka**: SDV, Synthcity, SDMetrics
- **Jebakan**: Data sintetis menyalin baris asli (bocor privasi); relasi antar-kolom rusak
- **Contoh judul**: Pembangkitan data akademik sintetis yang menjaga privasi dengan CTGAN

H30 · Deteksi anomali deret waktu
Unsupervised learning pada deret waktu

- **Baseline**: Rolling z-score, residual dekomposisi STL
- **Utama**: Isolation Forest pada fitur jendela, Matrix Profile, galat prediksi atau rekonstruksi LSTM/Autoencoder
- **Lanjutan**: USAD, OmniAnomaly, Anomaly Transformer, TranAD
- **Dataset**: NAB, Yahoo S5, SMD, SMAP/MSL, SWaT/WADI, UCR Anomaly Archive
- **Praproses**: Resampling, imputasi, normalisasi yang di-fit pada data latih, pembentukan jendela geser
- **Metrik**: PR-AUC, F1 berbasis rentang, VUS, keterlambatan deteksi; bila memakai point-adjusted F1, laporkan juga versi tanpa penyesuaian
- **Validasi**: Split kronologis; latih pada periode normal; ambang dari data validasi
- **Pustaka**: PyOD, stumpy, Merlion, TSB-AD
- **Jebakan**: Point adjustment membuat skor tampak tinggi; benchmark dengan anomali yang terlalu mudah; split acak
- **Contoh judul**: Deteksi anomali trafik jaringan kampus dengan Anomaly Transformer

H31 · Deteksi anomali (data tabular atau citra)
Unsupervised / one-class learning

- **Baseline**: Z-score, IQR, galat rekonstruksi PCA
- **Utama**: Isolation Forest, Local Outlier Factor, One-Class SVM, ECOD/COPOD, GMM
- **Lanjutan**: Autoencoder, VAE, Deep SVDD, DAGMM; PatchCore untuk citra
- **Dataset**: ODDS, ADBench, NSL-KDD, UNSW-NB15, CIC-IDS2017, N-BaIoT, IoT-23, Credit Card Fraud; citra: MVTec AD
- **Praproses**: Scaling, encoding, seleksi fitur; pisahkan data normal untuk pelatihan bila skema one-class
- **Metrik**: ROC-AUC, PR-AUC, F1 pada ambang terpilih, precision@k, false positive rate
- **Validasi**: Latih pada data normal atau tanpa label; label hanya dipakai untuk evaluasi; ambang ditentukan pada data validasi
- **Pustaka**: PyOD, scikit-learn, PyTorch, Anomalib (citra)
- **Jebakan**: Ambang ditala pada data uji; asumsi proporsi kontaminasi keliru; accuracy menyesatkan karena anomali sangat jarang
- **Contoh judul**: Deteksi botnet IoT tanpa label dengan autoencoder pada N-BaIoT

H32 · Clustering partisi (jumlah klaster diketahui)
Unsupervised learning

- **Baseline**: K-Means (inisialisasi k-means++)
- **Utama**: Mini-Batch K-Means, Gaussian Mixture Model, K-Medoids, Fuzzy C-Means, K-Prototypes (data campuran)
- **Lanjutan**: Deep Embedded Clustering (DEC), clustering pada embedding hasil autoencoder
- **Dataset**: Iris, Wine, Mall Customers, Online Retail (segmentasi RFM), MNIST
- **Praproses**: Scaling wajib; penanganan outlier; reduksi dimensi (PCA) bila fitur banyak
- **Metrik**: Silhouette, Davies-Bouldin, Calinski-Harabasz; ARI dan NMI bila ada label pembanding
- **Validasi**: Ulangi dengan banyak inisialisasi; uji stabilitas dengan bootstrap; interpretasi klaster oleh ahli domain
- **Pustaka**: scikit-learn, scikit-learn-extra, kmodes, scikit-fuzzy
- **Jebakan**: Lupa scaling; memaksa data kategori ke K-Means; klaster tidak bermakna meski silhouette tinggi
- **Contoh judul**: Segmentasi mahasiswa berdasarkan pola belajar di LMS dengan K-Means

H33 · Clustering untuk bentuk klaster tidak bulat
Unsupervised learning

- **Baseline**: Agglomerative Clustering (Ward atau average linkage)
- **Utama**: Spectral Clustering, Gaussian Mixture Model dengan kovarians penuh
- **Lanjutan**: Subspace clustering, deep clustering
- **Dataset**: Data sintetis (moons, circles) untuk validasi metode; data jaringan atau citra
- **Praproses**: Scaling; pemilihan ukuran kemiripan (kernel RBF, k-nearest neighbor graph)
- **Metrik**: Silhouette, Davies-Bouldin; ARI dan NMI bila ada label pembanding
- **Validasi**: Uji stabilitas terhadap parameter kemiripan; bandingkan dengan K-Means
- **Pustaka**: scikit-learn, SciPy
- **Jebakan**: Spectral clustering mahal untuk data besar; hasil sangat bergantung pada parameter graf kemiripan
- **Contoh judul**: Pengelompokan wilayah berdasarkan pola mobilitas dengan Spectral Clustering

H34 · Clustering berbasis kepadatan
Unsupervised learning

- **Baseline**: DBSCAN
- **Utama**: HDBSCAN, OPTICS, Mean Shift
- **Lanjutan**: Clustering kepadatan pada embedding UMAP; ST-DBSCAN untuk data ruang-waktu
- **Dataset**: Data spasial atau GPS (titik gempa, lokasi kecelakaan), data sintetis (moons), data dengan banyak outlier
- **Praproses**: Scaling; pemilihan metrik jarak (haversine untuk koordinat); k-distance plot untuk memilih eps
- **Metrik**: DBCV, silhouette (tanpa titik noise), proporsi noise; ARI dan NMI bila ada label
- **Validasi**: Analisis sensitivitas eps dan min_samples
- **Pustaka**: scikit-learn, hdbscan
- **Jebakan**: Kepadatan klaster yang berbeda-beda menggagalkan DBSCAN (pakai HDBSCAN); dimensi tinggi membuat jarak tidak bermakna
- **Contoh judul**: Pemetaan titik rawan kecelakaan dengan HDBSCAN pada data koordinat

H35 · Clustering dengan jumlah klaster belum diketahui
Unsupervised learning

- **Baseline**: Agglomerative Clustering dengan dendrogram
- **Utama**: K-Means dengan metode elbow, silhouette, atau gap statistic; GMM dengan BIC/AIC; X-Means
- **Lanjutan**: HDBSCAN, Affinity Propagation, Bayesian GMM (Dirichlet process)
- **Dataset**: Iris, Wine, Wholesale Customers, Country Data, data survei sendiri
- **Praproses**: Scaling; penanganan outlier; pemilihan linkage dan metrik jarak
- **Metrik**: Silhouette, Davies-Bouldin, Calinski-Harabasz, cophenetic correlation
- **Validasi**: Bandingkan beberapa nilai k; uji stabilitas; validasi makna klaster dengan ahli domain
- **Pustaka**: scikit-learn, SciPy (dendrogram), yellowbrick
- **Jebakan**: Elbow tidak jelas; agglomerative lambat untuk data besar; jumlah klaster dipilih agar cocok dengan harapan
- **Contoh judul**: Pengelompokan program studi berdasarkan indikator mutu dengan clustering hierarkis

H36 · Aturan asosiasi / pola frekuen
Unsupervised learning (pattern mining)

- **Baseline**: Apriori
- **Utama**: FP-Growth, ECLAT
- **Lanjutan**: Sequential pattern mining (PrefixSpan), high-utility itemset mining
- **Dataset**: Groceries, Instacart Market Basket, Online Retail (UCI), repositori FIMI
- **Praproses**: Ubah data menjadi format transaksi (one-hot per item); diskretisasi atribut numerik
- **Metrik**: Support, confidence, lift, conviction, leverage
- **Validasi**: Tidak ada train/test baku; uji kestabilan aturan pada periode atau subset lain
- **Pustaka**: mlxtend, SPMF, arules (R)
- **Jebakan**: Ambang support terlalu rendah menghasilkan ribuan aturan; confidence tinggi tetapi lift mendekati 1; asosiasi dianggap sebab-akibat
- **Contoh judul**: Pola pengambilan mata kuliah yang berkaitan dengan keterlambatan lulus dengan FP-Growth

H37 · Pemodelan topik
Unsupervised learning / NLP

- **Baseline**: Latent Semantic Analysis (LSA)
- **Utama**: Latent Dirichlet Allocation (LDA), NMF
- **Lanjutan**: BERTopic, Top2Vec, Contextualized Topic Model, model topik dinamis
- **Dataset**: 20 Newsgroups, Reuters-21578, abstrak arXiv, berita atau cuitan berbahasa Indonesia
- **Praproses**: Pembersihan, tokenisasi, stopword removal, stemming/lemmatisasi (untuk LDA), bigram; untuk BERTopic cukup pembersihan ringan
- **Metrik**: Topic coherence (C_v, NPMI), topic diversity, perplexity, penilaian manusia
- **Validasi**: Bandingkan beberapa jumlah topik; uji kestabilan antar-seed; interpretasi topik oleh ahli
- **Pustaka**: Gensim, BERTopic, OCTIS, Sastrawi, pyLDAvis
- **Jebakan**: Jumlah topik dipilih sembarang; topik diberi nama berlebihan; dokumen sangat pendek menyulitkan LDA
- **Contoh judul**: Pemetaan tren topik riset informatika di jurnal terindeks SINTA dengan BERTopic

H38 · Visualisasi data berdimensi tinggi
Unsupervised learning (reduksi dimensi)

- **Baseline**: PCA dua komponen
- **Utama**: t-SNE, UMAP
- **Lanjutan**: PaCMAP, TriMap, PHATE
- **Dataset**: MNIST, Fashion-MNIST, COIL-20, data ekspresi gen, embedding hasil model
- **Praproses**: Scaling; PCA ke sekitar 50 dimensi sebelum t-SNE untuk data sangat besar
- **Metrik**: Trustworthiness, continuity, kNN preservation
- **Validasi**: Coba beberapa nilai perplexity atau n_neighbors dan beberapa seed
- **Pustaka**: scikit-learn, umap-learn, openTSNE
- **Jebakan**: Jarak antar-klaster dan ukuran klaster pada t-SNE tidak bermakna; satu konfigurasi dianggap kebenaran
- **Contoh judul**: Visualisasi embedding dokumen skripsi dengan UMAP untuk melihat rumpun topik

H39 · Reduksi dimensi dan seleksi fitur
Unsupervised / praproses

- **Baseline**: PCA, variance threshold, penyaringan korelasi
- **Utama**: LDA (supervised), Kernel PCA, ICA, NMF; seleksi fitur: mutual information, chi-square, RFE, Lasso, Boruta
- **Lanjutan**: Autoencoder, seleksi fitur metaheuristik (Genetic Algorithm, PSO)
- **Dataset**: Madelon, Arcene, data ekspresi gen (TCGA), MNIST
- **Praproses**: Scaling sebelum PCA; tangani nilai hilang
- **Metrik**: Explained variance, galat rekonstruksi, performa dan waktu latih model hilir
- **Validasi**: Reduksi di-fit hanya pada data latih di dalam tiap fold (gunakan Pipeline)
- **Pustaka**: scikit-learn, Boruta, PyTorch
- **Jebakan**: Seleksi fitur dilakukan pada seluruh data sebelum CV (leakage klasik); komponen PCA sulit diinterpretasi
- **Contoh judul**: Seleksi fitur trafik jaringan untuk deteksi intrusi yang ringan

H40 · Self-supervised representation learning
Self-supervised learning

- **Baseline**: Model supervised dari nol, atau model pra-latih umum (ImageNet) sebagai pembanding
- **Utama**: Citra: SimCLR, MoCo, BYOL, DINO, MAE. Teks: masked language modeling. Audio: wav2vec 2.0. Deret waktu: TS2Vec. Tabular: SCARF
- **Lanjutan**: Pra-pelatihan lanjutan pada data domain, model fondasi, pembelajaran multimodal (CLIP)
- **Dataset**: ImageNet, CIFAR-10/100, STL-10; korpus teks atau data domain tak berlabel berukuran besar
- **Praproses**: Rancang augmentasi atau tugas pretext yang sesuai jenis data
- **Metrik**: Linear probing, fine-tuning dengan sedikit label, akurasi kNN pada embedding
- **Validasi**: Evaluasi pada tugas hilir dengan test set berlabel; variasikan proporsi label (1%, 10%, 100%)
- **Pustaka**: PyTorch Lightning, lightly, solo-learn, Hugging Face Transformers
- **Jebakan**: Representation collapse; butuh komputasi besar; augmentasi yang merusak informasi penting
- **Contoh judul**: Pra-pelatihan kontrastif pada trafik IoT tak berlabel untuk deteksi botnet

H41 · Masalah belum terpetakan
Perlu perumusan ulang

- **Baseline**: Mulai dari exploratory data analysis (statistik deskriptif, visualisasi distribusi dan korelasi)
- **Utama**: Rumuskan ulang: apa masukan, apa keluaran yang diinginkan, dan bagaimana keberhasilan diukur
- **Lanjutan**: Bila tujuannya optimasi (penjadwalan, rute): pertimbangkan metaheuristik atau pemrograman matematis, bukan machine learning
- **Dataset**: Belum dapat ditentukan sebelum masalah dirumuskan
- **Praproses**: Kumpulkan dan pahami data lebih dahulu
- **Metrik**: Belum dapat ditentukan
- **Validasi**: Belum dapat ditentukan
- **Pustaka**: pandas, matplotlib, seaborn, ydata-profiling
- **Jebakan**: Memaksakan machine learning pada masalah yang cukup diselesaikan dengan aturan, statistik, atau optimasi
- **Contoh judul**: Ulangi pandu setelah keluaran penelitian dirumuskan dalam satu kalimat

Pertanyaan kondisi khusus (ya/tidak) setelah hasil ditemukan

Data sedikit
Apakah data Anda sedikit? Jika ya:

- Pilih model sederhana atau model pra-latih (transfer learning); hindari melatih deep learning dari nol.
- Gunakan augmentasi data, dan few-shot learning bila sampel per kelas sangat sedikit.
- Validasi dengan repeated stratified k-fold atau leave-one-out; laporkan selang kepercayaan.
- Untuk tabular kecil, coba TabPFN atau model linear dengan regularisasi.

Kelas tidak seimbang
Apakah jumlah data antarkelas sangat timpang? Jika ya:

- Gunakan class weight, focal loss, atau resampling (SMOTE, ADASYN, undersampling) hanya pada data latih.
- Ganti accuracy dengan PR-AUC, F1, MCC, atau balanced accuracy.
- Gunakan stratified split dan tala ambang keputusan pada data validasi.
- Bila kelas minoritas sangat jarang, rumuskan sebagai deteksi anomali.

Butuh interpretabilitas (XAI)
Apakah keputusan model harus dapat dijelaskan kepada manusia? Jika ya:

- Utamakan model glass-box: Logistic Regression, Decision Tree, rule-based, Explainable Boosting Machine.
- Untuk model kompleks tambahkan SHAP, LIME, permutation importance, partial dependence plot.
- Untuk citra gunakan Grad-CAM; untuk teks gunakan atribusi token.
- Evaluasi penjelasan itu sendiri (fidelity, stabilitas) dan validasi dengan ahli domain.

Data terdistribusi dan privat (federated learning)
Apakah data tersebar di banyak pihak atau perangkat dan tidak boleh dikumpulkan di satu tempat? Jika ya:

- Gunakan Federated Learning: FedAvg sebagai baseline, FedProx atau SCAFFOLD untuk data non-IID.
- Simulasikan partisi non-IID (distribusi Dirichlet) dan laporkan pengaruhnya.
- Tambahkan differential privacy atau secure aggregation bila privasi menjadi klaim penelitian.
- Laporkan jumlah ronde komunikasi dan biaya komunikasi, selain akurasi.
- Pustaka: Flower, FedML, TensorFlow Federated.

Data streaming dan concept drift
Apakah data datang terus-menerus dan polanya dapat berubah seiring waktu? Jika ya:

- Gunakan online/incremental learning: Hoeffding Tree, Adaptive Random Forest, SGD dengan partial_fit.
- Pasang detektor drift: ADWIN, DDM, Page-Hinkley.
- Evaluasi dengan skema prequential (test-then-train), bukan k-fold.
- Pustaka: River, scikit-multiflow.

Perangkat terbatas atau real-time
Apakah model harus berjalan real-time atau di perangkat dengan sumber daya terbatas? Jika ya:

- Pilih arsitektur ringan: MobileNet, EfficientNet-Lite, DistilBERT, model pohon kecil.
- Terapkan quantization, pruning, dan knowledge distillation.
- Laporkan ukuran model, latensi, penggunaan memori, dan konsumsi energi bersama akurasi.
- Pustaka: TensorFlow Lite, ONNX Runtime, TinyML (Edge Impulse).

Dimensi tinggi
Apakah jumlah fitur sangat banyak dibandingkan jumlah data? Jika ya:

- Lakukan seleksi fitur: filter (mutual information, chi-square), wrapper (RFE), embedded (Lasso, importance pohon).
- Pertimbangkan PCA atau autoencoder, dan model dengan regularisasi kuat.
- Seleksi fitur harus berada di dalam tiap fold cross validation.
- Laporkan jumlah fitur terpilih dan kestabilannya antar-fold.

Data kotor
Apakah data mengandung banyak nilai hilang, noise, atau outlier? Jika ya:

- Imputasi: median/modus, KNN imputer, MICE; tambahkan indikator nilai hilang.
- Gunakan RobustScaler dan model yang tahan outlier (berbasis pohon, Huber loss).
- Gradient boosting (XGBoost, LightGBM, CatBoost) menangani nilai hilang secara bawaan.
- Laporkan proporsi nilai hilang dan bandingkan beberapa strategi imputasi.

Kuantifikasi ketidakpastian
Apakah Anda membutuhkan ukuran ketidakpastian dari setiap prediksi? Jika ya:

- Gunakan conformal prediction untuk interval atau himpunan prediksi dengan jaminan cakupan.
- Alternatif: quantile regression, Gaussian Process, MC Dropout, deep ensemble.
- Periksa kalibrasi probabilitas (reliability diagram, ECE, Brier score) dan kalibrasi ulang (Platt, isotonic, temperature scaling).
- Pustaka: MAPIE, scikit-learn CalibratedClassifierCV.

Keadilan (fairness)
Apakah keputusan model berdampak langsung pada orang dan menyangkut atribut sensitif? Jika ya:

- Ukur demographic parity, equalized odds, dan disparate impact per kelompok.
- Mitigasi: reweighing (pra), adversarial debiasing (saat latih), penyesuaian ambang (pasca).
- Laporkan trade-off antara akurasi dan keadilan.
- Pustaka: Fairlearn, AIF360.

Inferensi kausal
Apakah tujuan Anda memahami sebab-akibat, bukan sekadar memprediksi? Jika ya:

- Model prediktif tidak cukup; gunakan kerangka kausal: DAG, propensity score matching, difference-in-differences, instrumental variable.
- Untuk efek perlakuan heterogen: causal forest, meta-learner (T-, S-, X-learner), double machine learning.
- Nyatakan asumsi (tidak ada confounder tersembunyi) dan lakukan analisis sensitivitas.
- Pustaka: DoWhy, EconML, CausalML.

Komputasi terbatas
Apakah komputasi Anda terbatas (tanpa GPU yang memadai)? Jika ya:

- Utamakan machine learning klasik dan gradient boosting, yang cukup dengan CPU.
- Untuk citra/teks: pakai model pra-latih sebagai pengekstrak fitur beku, lalu latih classifier ringan.
- Manfaatkan Google Colab atau Kaggle Notebooks untuk GPU gratis.
- Kecilkan ruang pencarian hyperparameter (Optuna dengan pruning) dan gunakan subset data untuk eksperimen awal.

Kerangka umum untuk semua riset

1. Tulis pertanyaan riset dan hipotesis dalam satu kalimat: masukan, keluaran, dan ukuran keberhasilan.
2. Cari dataset pembanding di UCI Machine Learning Repository, Kaggle, OpenML, Hugging Face Datasets, Papers with Code, PhysioNet, IEEE DataPort, Zenodo, Google Dataset Search, atau Satu Data Indonesia (data.go.id) dan BPS.
3. Pisahkan data uji sebelum praproses apa pun; semua transformasi di-fit hanya pada data latih.
4. Selalu sertakan baseline sederhana; kebaruan diukur terhadap baseline dan metode terbaik yang sudah ada.
5. Tala hyperparameter pada data validasi; data uji dipakai satu kali di akhir.
6. Lakukan ablation study: lepaskan tiap komponen usulan untuk menunjukkan kontribusinya.
7. Uji signifikansi: McNemar atau Wilcoxon untuk dua model; Friedman + Nemenyi untuk banyak model pada banyak dataset.
8. Laporkan rata-rata ± simpangan baku dari beberapa seed atau fold, bukan satu angka terbaik.
9. Jaga reprodusibilitas: catat seed, versi pustaka, konfigurasi; publikasikan kode dan data bila diizinkan.
10. Periksa izin, privasi, dan etika penggunaan data sejak awal.

Rekomendasi adalah titik awal. Periksa literatur terbaru pada topik Anda sebelum menetapkan metode.

Taksonomi Machine Learning

Peta pikiran dari paradigma belajar sampai algoritma, termasuk deep learning, ensemble, paradigma khusus, dan alur kerja riset.

Setiap warna adalah satu cabang utama. Semakin jauh dari pusat, semakin spesifik (kategori, lalu algoritma atau teknik).

