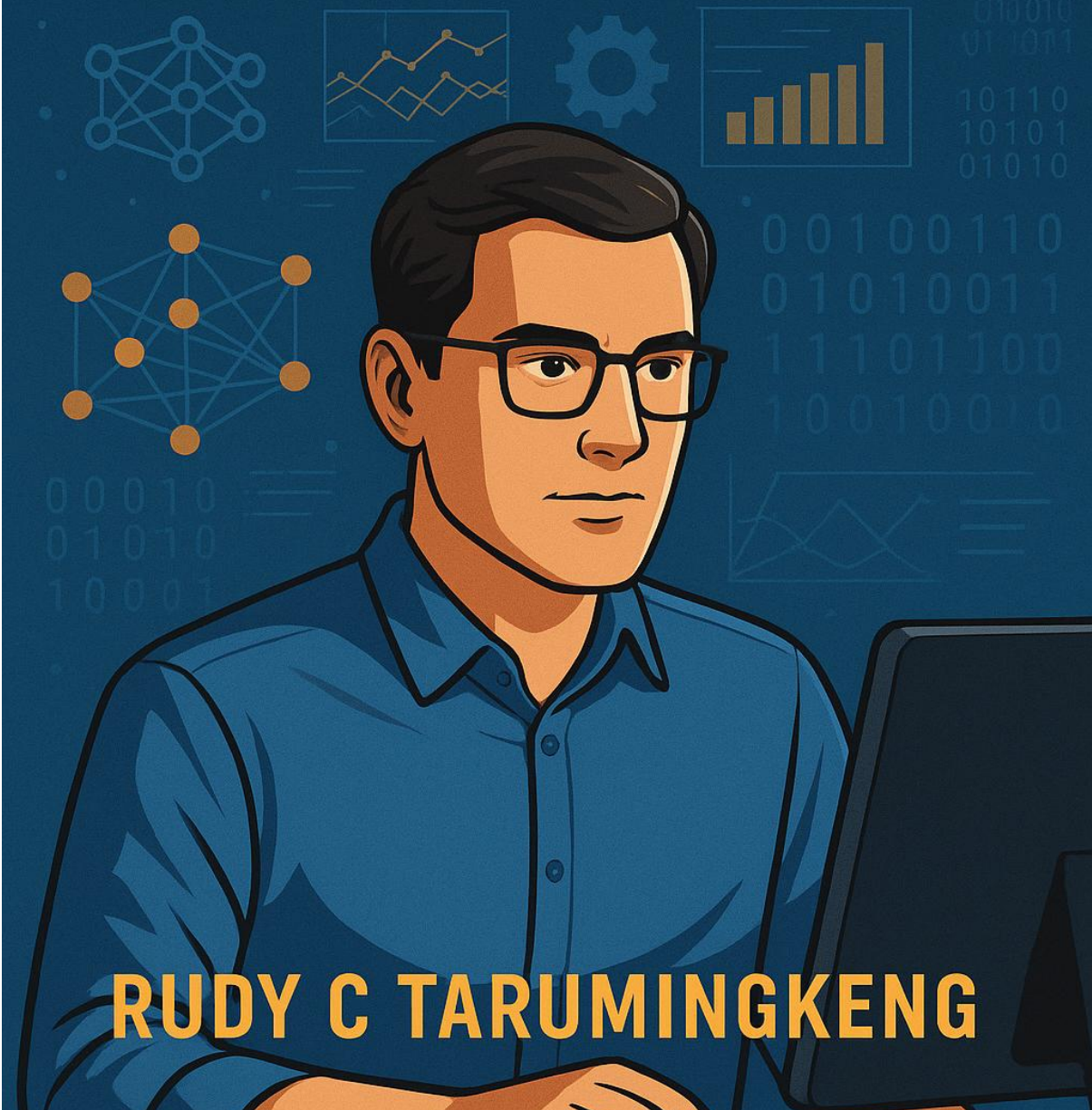


10 ALGORITMA MACHINE LEARNING YANG HARUS DIKETAHUI SEBAGAI DATA SCIENTIST



RUDY C TARUMINGKENG

*Rudy C Tarumingkeng: 10 Algoritma Machine Learning Yang Perlu
Diketahui Sebagai Data Scientist*

Prof Ir Rudy C Tarumingkeng, PhD

Professor of Management NUP: 9903252922

Rector, Cenderawasih State University, Papua (1978-1988)

Rector, Krida Wacana Christian University, Jakarta (1991-2000)

Chairman Board of Professors, IPB-University, Bogor (2005-2006)

Data Analyst and Chairman, Academic Senate, IBM-ASMI, Jakarta

© RUDYCT e-PRESS

rudymt75@gmail.com

Bogor, Indonesia

4 November 2025

10 ALGORITMA MACHINE LEARNING YANG PERLU DIKETAHUI SEBAGAI DATA SCIENTIST

RudyCT Academic Series 2025.

Pendahuluan

Dalam era data-driven economy, profesi *data scientist* menjadi salah satu pilar utama dalam pengambilan keputusan strategis berbasis analitik. Namun, di balik popularitas istilah *machine learning (ML)*, inti keahliannya terletak pada pemahaman terhadap algoritma—yakni fondasi matematis dan komputasional yang memungkinkan sistem belajar dari data.

Bagi seorang *data scientist*, mengenal berbagai algoritma ML bukan sekadar memahami rumus, tetapi juga memahami konteks penggunaannya: kapan digunakan, mengapa efektif, serta bagaimana menginterpretasi hasilnya. Algoritma yang tepat akan menghasilkan model prediksi yang akurat dan efisien; sebaliknya, pilihan algoritma yang salah dapat menimbulkan bias, *overfitting*, atau kesimpulan yang menyesatkan.

Tulisan ini akan mengulas **10 algoritma machine learning yang wajib dikuasai setiap *data scientist***—dilengkapi dengan penjelasan prinsip kerja, kelebihan, keterbatasan, dan contoh kasus aplikatif dalam dunia nyata.

1. Linear Regression: Fondasi Prediksi Kuantitatif

Konsep Dasar

Linear Regression adalah algoritma paling klasik dan fundamental dalam *supervised learning*, digunakan untuk memprediksi nilai kontinu berdasarkan hubungan linear antara variabel independen (X) dan dependen (Y). Modelnya sederhana:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Kelebihan

- Mudah diinterpretasikan.
- Komputasi cepat bahkan untuk dataset besar.
- Cocok untuk hubungan yang bersifat linear dan stabil.

Keterbatasan

- Tidak cocok untuk hubungan non-linear.
- Rentan terhadap *outlier*.
- Asumsi multikolinearitas dapat mengacaukan hasil.

Aplikasi Nyata

Dalam bidang manajemen bisnis, regresi linear digunakan untuk **memprediksi penjualan berdasarkan biaya iklan** atau **menganalisis hubungan antara harga dan permintaan produk**. Misalnya, perusahaan e-commerce menggunakan regresi linear untuk memperkirakan omset berdasarkan jumlah kunjungan situs dan pengeluaran iklan digital.

2. Logistic Regression: Prediksi Klasifikasi Biner

Konsep Dasar

Meski namanya "regression", algoritma ini digunakan untuk *classification problems*. Logistic Regression memetakan output menjadi probabilitas antara 0 dan 1 menggunakan fungsi sigmoid:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}}$$

Kelebihan

- Interpretasi probabilistik yang mudah.
- Efektif untuk klasifikasi biner sederhana.
- Cepat dan stabil secara numerik.

Keterbatasan

- Tidak cocok untuk hubungan kompleks atau non-linear.
- Kinerja menurun jika terdapat fitur interaksi yang kuat.

Aplikasi Nyata

Digunakan dalam **prediksi churn pelanggan**, **deteksi penipuan kartu kredit**, atau **diagnosis penyakit (ya/tidak)**. Misalnya, bank memanfaatkan Logistic Regression untuk mengklasifikasikan nasabah berisiko gagal bayar (1) atau tidak (0).

3. Decision Tree: Model Hierarkis Berbasis Keputusan

Konsep Dasar

Decision Tree bekerja dengan membagi data ke dalam subset berdasarkan *decision rules* dari atribut tertentu. Setiap percabangan (node) mewakili pertanyaan logis—misalnya "Apakah usia > 30

tahun?”—hingga mencapai *leaf node* yang menghasilkan keputusan akhir.

Kelebihan

- Mudah dipahami secara visual.
- Tidak memerlukan normalisasi data.
- Dapat menangani data kategorikal dan numerik.

Keterbatasan

- Rentan terhadap *overfitting*.
- Sensitif terhadap perubahan kecil pada data.

Aplikasi Nyata

Dalam dunia ritel, Decision Tree sering digunakan untuk **segmentasi pelanggan** berdasarkan perilaku pembelian. Misalnya, perusahaan supermarket membangun pohon keputusan untuk mengidentifikasi pelanggan yang berpotensi membeli produk premium.

4. Random Forest: Ensemble Learning untuk Akurasi Tinggi

Konsep Dasar

Random Forest adalah kumpulan (*ensemble*) dari banyak Decision Tree yang dilatih pada subset data dan fitur yang berbeda. Prediksi dihasilkan melalui voting mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi).

Kelebihan

- Mengurangi *overfitting* dibanding Decision Tree tunggal.
- Memberikan estimasi pentingnya fitur (*feature importance*).
- Efektif untuk dataset besar dan kompleks.

Keterbatasan

- Kurang interpretatif.
- Konsumsi memori dan waktu pelatihan tinggi.

Aplikasi Nyata

Dalam keuangan, Random Forest digunakan untuk **menilai kelayakan kredit** atau **mendeteksi anomali transaksi keuangan**. Di bidang medis, digunakan untuk **memprediksi risiko penyakit** berdasarkan kombinasi faktor biologis dan gaya hidup.

5. Support Vector Machine (SVM): Optimalisasi Margin Klasifikasi

Konsep Dasar

SVM bertujuan menemukan *hyperplane* terbaik yang memisahkan dua kelas dengan margin maksimum. Untuk data yang tidak linear, SVM menggunakan *kernel trick* (seperti RBF atau polynomial) untuk memproyeksikan data ke dimensi lebih tinggi.

Kelebihan

- Efektif untuk dataset kecil dan dimensi tinggi.
- Tahan terhadap *overfitting*.
- Dapat digunakan untuk klasifikasi dan regresi (SVR).

Keterbatasan

- Sulit diinterpretasikan.
- Pemilihan kernel sangat berpengaruh pada hasil.

Aplikasi Nyata

SVM banyak digunakan untuk **klasifikasi teks dan spam detection**, **pengenalan wajah**, dan **bioinformatika**. Sebagai contoh, SVM digunakan untuk membedakan email "spam" dan "non-spam" berdasarkan pola kata.

6. Naïve Bayes: Probabilitas dan Teorema Bayes

Konsep Dasar

Naïve Bayes mengasumsikan bahwa fitur bersifat independen satu sama lain, dan menggunakan Teorema Bayes untuk menghitung probabilitas suatu kelas berdasarkan fitur yang diamati.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Kelebihan

- Cepat dan efisien untuk dataset besar.
- Cocok untuk data teks (seperti klasifikasi dokumen).
- Tidak memerlukan banyak data untuk pelatihan.

Keterbatasan

- Asumsi independensi sering tidak realistis.
- Sensitif terhadap data dengan korelasi tinggi antar fitur.

Aplikasi Nyata

Digunakan secara luas dalam **analisis sentimen**, **klasifikasi berita**, dan **spam filtering**. Contoh: sistem email Gmail menggunakan Naïve Bayes untuk menandai pesan "spam" berdasarkan probabilitas kata-kata tertentu.

7. K-Nearest Neighbors (KNN): Pembelajaran Berdasarkan Kedekatan

Konsep Dasar

KNN mengklasifikasikan data baru berdasarkan mayoritas kelas dari K tetangga terdekat dalam ruang fitur. Jarak yang umum digunakan antara lain *Euclidean distance*.

Kelebihan

- Implementasi sederhana.
- Tidak memerlukan pelatihan eksplisit (lazy learner).
- Efektif untuk data dengan pola lokal yang kuat.

Keterbatasan

- Lambat untuk dataset besar.
- Sensitif terhadap *outlier* dan pemilihan nilai K .

Aplikasi Nyata

KNN sering digunakan dalam **sistem rekomendasi**, **pengenalan pola**, dan **deteksi anomali**. Misalnya, platform streaming menggunakan KNN untuk merekomendasikan film berdasarkan kesamaan preferensi pengguna lain.

8. K-Means Clustering: Pembelajaran Tak Terawasi untuk Segmentasi

Konsep Dasar

Berbeda dari algoritma sebelumnya, K-Means adalah *unsupervised learning algorithm* yang mengelompokkan data ke dalam K kluster berdasarkan kedekatan terhadap pusat (*centroid*). Tujuannya adalah meminimalkan *intra-cluster variance*.

Kelebihan

- Efisien dan mudah diimplementasikan.
- Baik untuk segmentasi data dengan struktur yang jelas.

Keterbatasan

- Harus menentukan jumlah kluster (K) sebelumnya.
- Tidak cocok untuk data dengan bentuk non-sferis.
- Rentan terhadap *outlier*.

Aplikasi Nyata

K-Means digunakan dalam **segmentasi pasar**, **analisis perilaku pelanggan**, dan **pengelompokan dokumen**. Misalnya, perusahaan telekomunikasi dapat mengelompokkan pelanggan berdasarkan frekuensi dan durasi panggilan.

9. Gradient Boosting dan XGBoost: Model Berbasis Ensemble Adaptif

Konsep Dasar

Gradient Boosting membangun model secara bertahap dengan menggabungkan banyak *weak learners* (biasanya Decision Trees kecil) untuk memperbaiki kesalahan model sebelumnya. XGBoost adalah implementasi modern yang lebih cepat dan efisien.

Kelebihan

- Akurasi sangat tinggi dalam kompetisi *machine learning*.
- Dapat menangani data tidak seimbang.
- Mendukung paralelisasi dan regularisasi.

Keterbatasan

- Kompleks dan sulit diinterpretasikan.

- Membutuhkan *tuning* parameter yang hati-hati.

Aplikasi Nyata

Banyak digunakan dalam kompetisi *Kaggle*, keuangan (penilaian risiko), dan pemasaran (prediksi churn). Misalnya, bank menggunakan XGBoost untuk memperkirakan kemungkinan nasabah mengajukan pinjaman baru.

10. Neural Networks dan Deep Learning: Model Inspirasi Otak Manusia

Konsep Dasar

Neural Networks meniru cara kerja neuron biologis melalui lapisan (*layers*) yang saling terhubung. Setiap neuron melakukan transformasi matematis atas masukan dan mengirimkan sinyal ke lapisan berikutnya. *Deep Learning* adalah jaringan dengan banyak lapisan (*deep architecture*).

Kelebihan

- Mampu menangani data kompleks seperti gambar, suara, dan teks.
- Fleksibel dan kuat dalam pembelajaran non-linear.
- Dapat melakukan *feature extraction* secara otomatis.

Keterbatasan

- Membutuhkan data dan komputasi besar.
- Kurang transparan (black box problem).
- Sulit ditafsirkan secara intuitif.

Aplikasi Nyata

Neural Networks digunakan pada **pengenalan wajah (face recognition)**, **penerjemahan otomatis (Google Translate)**, **mobil otonom**, dan **analisis citra medis**.

Perbandingan dan Analisis Strategis

Algoritma	Tipe	Kelebihan Utama	Keterbatasan	Aplikasi
Linear Regression	Supervised	Interpretatif dan cepat	Tidak cocok untuk data non-linear	Prediksi harga
Logistic Regression	Supervised	Probabilistik	Terbatas pada dua kelas	Deteksi churn
Decision Tree	Supervised	Mudah dipahami	Overfitting	Segmentasi pelanggan
Random Forest	Ensemble	Akurat	Berat komputasi	Kredit scoring
SVM	Supervised	Margin optimal	Sulit diinterpretasi	Deteksi spam
Naïve Bayes	Supervised	Cepat	Asumsi independen	Klasifikasi teks
KNN	Supervised	Sederhana	Tidak efisien untuk data besar	Rekomendasi

Algoritma	Tipe	Kelebihan Utama	Keterbatasan	Aplikasi
K-Means	Unsupervised	Efisien	Harus tentukan K	Segmentasi
XGBoost	Ensemble	Akurasi tinggi	Kompleks	Risiko keuangan
Neural Networks	Deep Learning	Fleksibel	Black box	Citra & suara

Konteks Penggunaan dalam Dunia Data Science

Seorang *data scientist* yang handal tidak hanya memilih algoritma terbaik, tetapi juga memahami konteks bisnis dan karakteristik data. Misalnya:

- **Model interpretable** (seperti Linear atau Logistic Regression) cocok untuk laporan manajerial.
- **Model non-linear** (seperti Random Forest dan XGBoost) digunakan untuk kompetisi prediksi dengan data besar.
- **Model deep learning** digunakan untuk pemrosesan visual, suara, dan teks alami.

Pendekatan holistik melibatkan:

1. Eksplorasi data (*EDA*).
 2. Pemilihan algoritma berdasarkan tujuan (klasifikasi, regresi, atau clustering).
 3. Validasi hasil dengan metrik seperti *accuracy*, *F1-score*, *RMSE*, atau *AUC-ROC*.
-

Refleksi dan Diskusi

Kemajuan *machine learning* tidak hanya soal kecepatan komputasi, tetapi juga pemahaman etika dan tanggung jawab sosial dalam penggunaan data. Misalnya, bias algoritmik dalam perekrutan karyawan atau pemberian kredit dapat menyebabkan diskriminasi jika data pelatihan tidak representatif.

Oleh karena itu, *data scientist modern* harus memadukan tiga kompetensi utama:

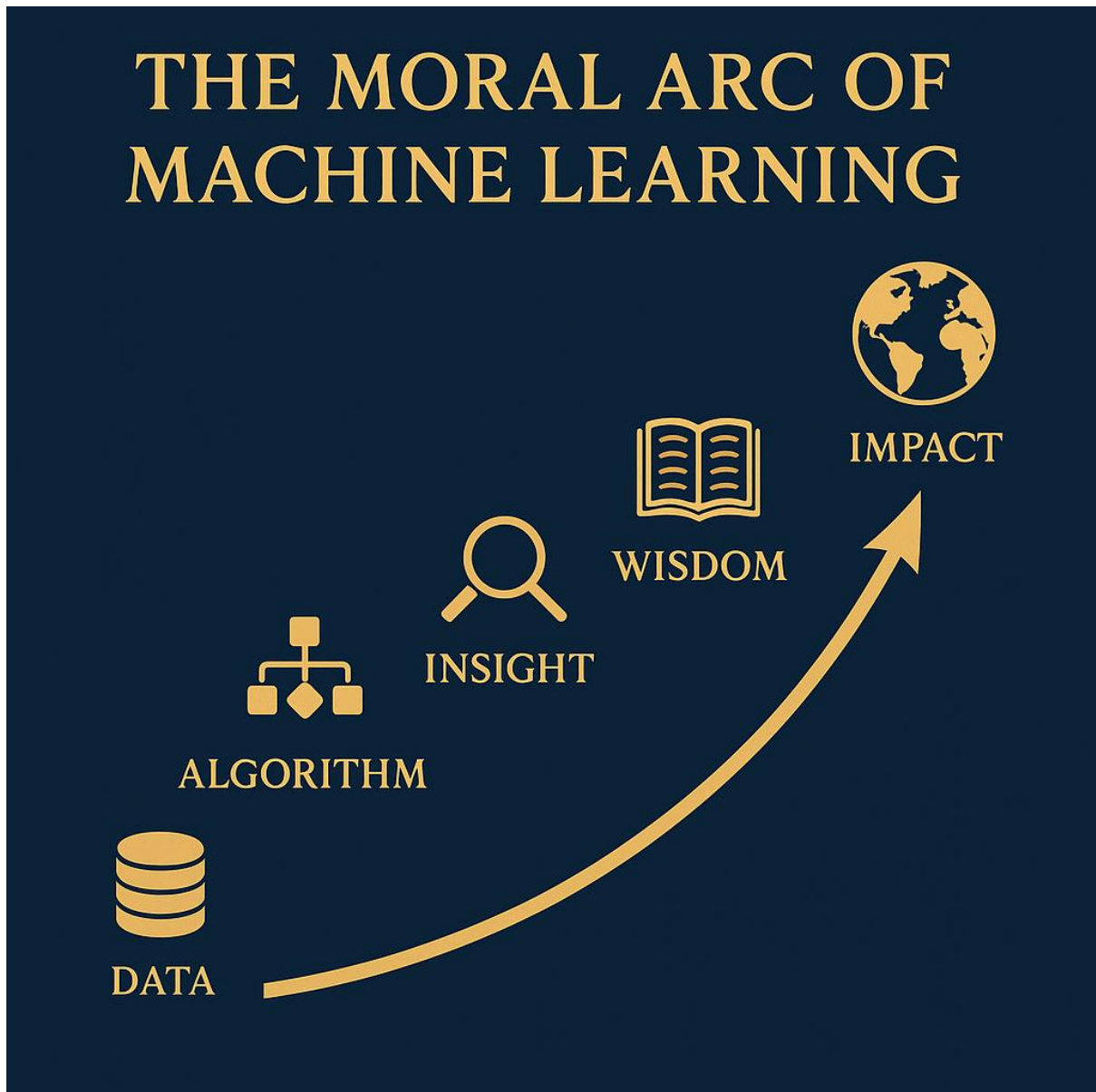
1. **Analitik dan statistik** — memahami logika di balik model.
 2. **Pemrograman dan komputasi** — mengimplementasikan model secara efisien.
 3. **Etika dan konteks sosial** — memastikan penggunaan AI untuk kebaikan bersama (*AI for Good*).
-

Kesimpulan

Menguasai sepuluh algoritma ini merupakan langkah awal menuju profesionalisme sebagai *data scientist*. Masing-masing memiliki keunggulan dan konteks penggunaan spesifik:

- Mulai dari **Linear Regression** yang sederhana namun kuat,
- hingga **Neural Networks** yang mendasari revolusi *deep learning*.

Namun, lebih dari sekadar alat teknis, seorang *data scientist sejati* adalah **arsitek pengetahuan**—yang mampu mengubah data menjadi kebijakan, prediksi menjadi keputusan, dan algoritma menjadi nilai kemanusiaan.



ELABORASI DAN DISKUSI LANJUTAN

Berikut bagian **elaborasi dan diskusi lanjutan** untuk makalah berjudul  *"10 Algoritma Machine Learning yang Harus Diketahui sebagai Data Scientist"* yang menekankan konteks filosofis, profesional, dan aplikatif bagi dunia *data science* modern.

1. Machine Learning sebagai Evolusi Epistemologis

Dalam sejarah perkembangan ilmu pengetahuan, *machine learning* bukan hanya produk kemajuan komputasi, tetapi merupakan **evolusi cara manusia memahami realitas melalui data**. Jika pada masa lalu manusia menggunakan hukum Newton untuk menjelaskan dunia fisik, kini kita menggunakan *algoritma* untuk menjelaskan dunia sosial, ekonomi, dan biologis yang kompleks.

Proses pembelajaran mesin merepresentasikan bentuk baru dari epistemologi modern: dari **observasi empiris (data)** menuju **abstraksi komputasional (model)**, dan akhirnya menghasilkan **pengetahuan prediktif (insight)**.

Dengan demikian, machine learning dapat dilihat sebagai *the second enlightenment* — pencerahan baru yang menggabungkan logika matematis dengan kemampuan eksplorasi tak terbatas dari data digital.

2. Paradigma “From Data to Wisdom” dalam Konteks Moral

Infografik “*The Moral Arc of Machine Learning — From Data → Algorithm → Insight → Wisdom → Impact*” menggambarkan transformasi pengetahuan menjadi kebijaksanaan.

- **Data:** Fakta mentah yang tak bermakna tanpa konteks.
- **Algorithm:** Mekanisme rasional yang mengorganisasi data menjadi pola.
- **Insight:** Pemahaman yang muncul dari interpretasi hasil model.
- **Wisdom:** Kemampuan untuk menggunakan insight secara etis dan berorientasi manusia.
- **Impact:** Penerapan nyata yang membawa manfaat sosial, ekonomi, dan moral.

Dalam konteks ini, *machine learning* bukan hanya soal akurasi prediksi, melainkan juga **tanggung jawab dalam penerapan hasilnya**. Sebuah model yang mampu memprediksi perilaku konsumen, misalnya, juga memiliki potensi untuk memanipulasi keputusan manusia. Oleh karena itu, moralitas dalam ML menjadi penting — agar hasil analisis tidak sekadar “benar secara statistik”, tetapi juga “baik secara etis”.

3. Integrasi Machine Learning dalam Keputusan Manajerial

Seorang *data scientist* modern harus memahami bahwa algoritma tidak berdiri sendiri. Ia adalah bagian dari **rantai pengambilan keputusan organisasi**. Tiga level integrasi penting adalah:

1. **Strategic Level** — penggunaan model prediktif untuk *forecasting* pasar, risiko investasi, dan perencanaan jangka panjang.
2. **Tactical Level** — pemanfaatan analisis data untuk *customer segmentation*, *inventory optimization*, atau *dynamic pricing*.
3. **Operational Level** — penerapan model secara otomatis melalui *AI-driven systems*, misalnya dalam *chatbot*, *fraud detection*, atau *real-time analytics*.

Sinergi antara manusia dan algoritma menjadi inti dari **decision intelligence** — paradigma baru yang menempatkan *machine learning* sebagai mitra kognitif dalam pengambilan keputusan.

4. Dilema Etika dan Tantangan Keadilan Algoritmik

Seiring meningkatnya kekuatan prediktif algoritma, muncul pula tantangan besar terkait **bias, transparansi, dan akuntabilitas**. Kasus-kasus seperti sistem *credit scoring* yang mendiskriminasi kelompok minoritas atau algoritma rekrutmen yang bias terhadap gender menunjukkan bahwa **data bukanlah netral**.

Bias dapat muncul dari:

- Data historis yang tidak representatif.
- Pemilihan fitur yang mencerminkan stereotip sosial.
- Interpretasi hasil yang mengabaikan konteks manusia.

Untuk itu, diperlukan pendekatan “Fair AI” dan “Explainable AI (XAI)” agar algoritma dapat diaudit, dijelaskan, dan dipertanggungjawabkan. Prinsip keadilan algoritmik menegaskan bahwa *accuracy is not enough* — keadilan dan kemanusiaan juga harus diukur.

5. Dimensi Sosial dan Humanistik dalam Machine Learning

Machine learning bukan sekadar perangkat teknis, tetapi juga **alat transformasi sosial**. Aplikasinya mencakup:

- **Kesehatan publik** (prediksi penyakit menular, personalisasi pengobatan).
- **Pendidikan** (sistem pembelajaran adaptif berbasis perilaku siswa).
- **Lingkungan hidup** (pemantauan hutan, prediksi polusi, efisiensi energi).
- **Pemerintahan digital** (analisis kebijakan publik berbasis data).

Namun, ketika teknologi mengambil peran dalam menentukan nasib manusia — misalnya dalam *AI-driven hiring* atau *automated justice systems* — muncul kebutuhan untuk menegaskan prinsip *human-in-the-loop*, yakni memastikan **keputusan akhir tetap berada di tangan manusia**.

6. Keahlian Masa Depan bagi Data Scientist

Menjadi *data scientist* di abad ke-21 bukan lagi hanya tentang menguasai Python, TensorFlow, atau R, tetapi tentang mengembangkan

mindset multidisipliner.

Kompetensi masa depan mencakup:

Kompetensi	Deskripsi	Implikasi
Technical Literacy	Pemahaman mendalam tentang algoritma dan statistik	Membangun model yang valid dan efisien
Ethical Literacy	Kesadaran terhadap dampak sosial dan bias data	Menghindari diskriminasi dan manipulasi
Business Literacy	Kemampuan menerjemahkan hasil analisis menjadi strategi bisnis	Mengubah insight menjadi keputusan
Cultural Literacy	Sensitivitas terhadap konteks sosial dan budaya	Menyesuaikan model dengan nilai lokal
Reflective Thinking	Kemampuan menimbang nilai dan makna dari hasil teknologi	Membangun keilmuan yang humanistik

Integrasi kelima literasi ini menciptakan *data scientist holistik* — bukan hanya ahli analisis, tetapi juga **arsitek kebijakan berbasis data**.

7. Machine Learning dan Era AI Generatif

Perkembangan terbaru menunjukkan pergeseran menuju **AI generatif (Generative AI)** yang mampu menciptakan teks, gambar, bahkan kode. Model seperti GPT, DALL-E, dan Gemini adalah hasil dari fondasi algoritmik yang sama: *supervised learning*, *deep learning*, dan *reinforcement learning*.

Perbedaannya, kini model tidak hanya *mempelajari pola*, tetapi juga *menghasilkan kreativitas baru*. Ini menandai **era co-creation** —

kolaborasi antara manusia dan mesin dalam mencipta pengetahuan, seni, maupun strategi bisnis.

Namun, muncul pula pertanyaan etis:

“Jika mesin bisa mencipta, di manakah letak otoritas intelektual manusia?”

Jawabannya bukan dalam kompetisi, tetapi **ko-evolusi** — manusia dan mesin berkembang bersama dalam ekosistem kognitif baru. Manusia tetap pemegang moral, sementara mesin menjadi katalis intelektual.

8. Dimensi Spiritualitas dalam Sains Data

Dalam refleksi yang lebih mendalam, *machine learning* mengajarkan kita **kerendahan hati epistemologis**. Semakin besar data yang kita miliki, semakin kita sadar akan keterbatasan pengetahuan kita.

Seperti ditegaskan filsuf ilmu Karl Popper, “*All knowledge is provisional.*”

Model yang hari ini akurat, bisa saja besok usang.

Kebenaran dalam *data science* bukanlah dogma, melainkan proses korektif berkelanjutan.

Dari sini, lahirlah kesadaran spiritual bahwa di balik jutaan data dan algoritma, terdapat panggilan moral untuk **menggunakan pengetahuan demi kemaslahatan manusia dan ciptaan**.

9. Machine Learning untuk Indonesia: Konteks Pembangunan

Dalam konteks nasional, penerapan *machine learning* dapat memperkuat:

- **Ketahanan pangan**, melalui prediksi cuaca dan hasil panen.
- **Kesehatan masyarakat**, melalui analisis epidemiologi berbasis data.
- **Pendidikan adaptif**, dengan *AI tutor* bagi siswa di daerah terpencil.

- **Smart governance**, dengan transparansi penganggaran publik berbasis data.

Namun, tantangan terbesarnya adalah **kesenjangan digital (digital divide)** — di mana sebagian besar masyarakat belum memiliki literasi data. Oleh karena itu, selain membangun infrastruktur AI, Indonesia perlu membangun **ekosistem etika dan pendidikan data** untuk melahirkan generasi *data-literate citizens*.

10. Menuju Data Science Humanis dan Berkeadilan

Tujuan akhir dari pembelajaran mesin seharusnya bukan sekadar efisiensi atau laba, melainkan **pemberdayaan manusia**.

Dalam pandangan *human-centered AI*, teknologi harus memperkuat nilai-nilai Pancasila:

- *Kemanusiaan yang adil dan beradab* — melalui AI yang inklusif.
- *Keadilan sosial bagi seluruh rakyat Indonesia* — melalui akses data yang merata.
- *Persatuan Indonesia* — melalui kolaborasi lintas bidang dan daerah.

Dengan demikian, moral arc dari machine learning bukan hanya dari *Data → Impact*, tetapi juga dari *Code → Compassion*.

Ketika teknologi dan etika berjalan beriringan, maka ilmu data menjadi bukan sekadar alat prediksi, tetapi **jalan kebijaksanaan bagi kemanusiaan**.

Penutup: From Intelligence to Integrity

Perjalanan *machine learning* menggambarkan dinamika manusia dalam mencari makna di tengah kompleksitas dunia digital.

Data memberikan kita *intelligence*, tetapi hanya nilai dan moral yang dapat memberikan *integrity*.

Seorang *data scientist sejati* bukan hanya pencipta model, melainkan juga penjaga nilai—yang menempatkan teknologi di bawah kendali kebijaksanaan.

Dengan begitu, ilmu data tidak berhenti pada kecerdasan buatan (*artificial intelligence*), melainkan menuntun pada **kecerdasan bijak** (**augmented wisdom**)—sebuah sinergi antara algoritma dan hati nurani.



Berikut bagian **Glosarium dan Referensi** untuk makalah berjudul:

■ "10 Algoritma Machine Learning yang Harus Diketahui sebagai Data Scientist"



GLOSARIUM

A

- **Accuracy** — Ukuran performa model klasifikasi yang menunjukkan proporsi prediksi yang benar dibandingkan total prediksi.
- **Algorithm** — Sekumpulan instruksi logis untuk memproses data dan menghasilkan keputusan atau prediksi.
- **Artificial Intelligence (AI)** — Cabang ilmu komputer yang berfokus pada pembuatan sistem yang dapat berpikir, belajar, dan beradaptasi seperti manusia.

B

- **Bias (Model Bias)** — Kecenderungan algoritma untuk menghasilkan kesalahan sistematis karena asumsi yang salah terhadap data.
- **Boosting** — Teknik *ensemble learning* yang menggabungkan beberapa model lemah menjadi satu model yang kuat secara iteratif.

C

- **Clustering** — Metode *unsupervised learning* untuk mengelompokkan data berdasarkan kemiripan karakteristik.
- **Classification** — Proses pengelompokan data ke dalam kategori atau label yang telah ditentukan.

- **Cross-validation** — Teknik untuk mengukur kemampuan generalisasi model dengan membagi data menjadi beberapa subset pelatihan dan pengujian.

D

- **Data Cleaning** — Proses pembersihan data dari kesalahan, duplikasi, dan nilai hilang sebelum analisis dilakukan.
- **Decision Tree** — Algoritma berbasis pohon keputusan yang membagi data berdasarkan kondisi tertentu untuk mencapai hasil klasifikasi atau regresi.
- **Deep Learning** — Cabang dari *machine learning* yang menggunakan jaringan saraf (neural network) dengan banyak lapisan untuk mengenali pola kompleks.

E

- **Ensemble Learning** — Teknik penggabungan beberapa model (misalnya *Random Forest* atau *Gradient Boosting*) untuk meningkatkan akurasi.
- **Ethical AI** — Prinsip etika dalam pengembangan dan penggunaan kecerdasan buatan agar adil, transparan, dan bertanggung jawab.

F

- **Feature Engineering** — Proses menciptakan, memilih, dan mentransformasi variabel agar lebih relevan untuk algoritma ML.
- **Feature Importance** — Indikator kontribusi setiap variabel dalam menentukan hasil model prediktif.

G

- **Gradient Boosting** — Teknik ensemble yang membangun model bertahap, di mana setiap model baru memperbaiki kesalahan dari model sebelumnya.

H

- **Hyperparameter Tuning** — Proses menemukan kombinasi parameter optimal untuk meningkatkan performa model.

I

- **Insight** — Pengetahuan baru yang dihasilkan dari analisis data yang dapat dijadikan dasar pengambilan keputusan.
- **Interpretability** — Kemampuan untuk menjelaskan bagaimana dan mengapa suatu model menghasilkan prediksi tertentu.

K

- **K-Means** — Algoritma *unsupervised* untuk mengelompokkan data ke dalam K klaster berdasarkan jarak terhadap centroid.
- **K-Nearest Neighbors (KNN)** — Algoritma *supervised* yang mengklasifikasikan data berdasarkan kedekatannya dengan K tetangga terdekat.
- **Kernel Trick** — Metode dalam SVM untuk mengubah data non-linear menjadi linear di ruang berdimensi lebih tinggi.

L

- **Linear Regression** — Algoritma statistik untuk memodelkan hubungan linear antara variabel input dan output.
- **Logistic Regression** — Model prediktif untuk klasifikasi biner yang menghasilkan probabilitas keluaran.

M

- **Machine Learning (ML)** — Proses pembelajaran otomatis dari data untuk membuat prediksi atau keputusan tanpa pemrograman eksplisit.

- **Model Overfitting** — Kondisi di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga gagal melakukan generalisasi terhadap data baru.

N

- **Naïve Bayes** — Algoritma probabilistik berdasarkan Teorema Bayes yang mengasumsikan independensi antar fitur.
- **Neural Network** — Struktur komputasional yang meniru sistem saraf manusia dengan neuron-neuron buatan.

O

- **Outlier** — Nilai ekstrem dalam data yang berbeda jauh dari observasi lain, dapat memengaruhi hasil model.

P

- **Precision & Recall** — Metrik evaluasi model klasifikasi yang menilai ketepatan dan kelengkapan hasil prediksi.
- **Predictive Analytics** — Cabang analitik data yang memanfaatkan model statistik untuk memperkirakan kejadian masa depan.

R

- **Random Forest** — Algoritma ensemble berbasis banyak *decision tree* yang hasilnya diambil melalui voting atau rata-rata.
- **Regularization** — Teknik untuk mencegah *overfitting* dengan menambahkan penalti terhadap kompleksitas model.

S

- **Support Vector Machine (SVM)** — Algoritma klasifikasi yang mencari garis pemisah (*hyperplane*) dengan margin maksimum antar kelas.
- **Supervised Learning** — Jenis ML yang belajar dari data berlabel untuk membuat prediksi.

T

- **Training Set** — Subset data yang digunakan untuk melatih model.
- **Test Set** — Subset data yang digunakan untuk menguji kemampuan generalisasi model.

U

- **Unsupervised Learning** — Jenis ML yang menemukan pola tanpa label atau target tertentu (misalnya clustering).

W

- **Wisdom** — Tahap tertinggi dalam hierarki pengetahuan di mana insight digunakan secara bijaksana untuk tujuan kemanusiaan.
-



DAFTAR REFERENSI

Buku dan Jurnal Akademik

1. Alpaydin, Ethem. *Introduction to Machine Learning*. MIT Press, 2021.
2. Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2016.
3. Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly, 2023.
4. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2017.
5. James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2023.
6. Murphy, Kevin P. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

*Rudy C Tarumíngkeng: 10 Algoritma Machine Learning Yang Perlu
Diketahui Sebagai Data Scientist*

7. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2017.
8. Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, 2015.
9. Russell, Stuart & Norvig, Peter. *Artificial Intelligence: A Modern Approach*. Pearson, 2021.
10. Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997 (klasik, fundamental).

Artikel dan Sumber Daring

1. Google AI Blog. *Responsible AI Practices and Ethical Frameworks*. (2023).
2. IBM Research. *Explainable Artificial Intelligence (XAI) and Trust in Data Systems*. (2022).
3. Kaggle Documentation. *Machine Learning Algorithm Benchmarks and Leaderboards*.
4. McKinsey Global Institute. *The State of AI in 2024: Scaling Value and Managing Risk*. (2024).
5. OpenAI Research Papers. *Generative Models and Reinforcement Learning from Human Feedback*. (2023–2024).
6. European Commission. *Ethics Guidelines for Trustworthy AI*. (2021).
7. World Economic Forum. *Data for Humanity: Principles for an Ethical Data Economy*. (2022).

Referensi Kontekstual Indonesia

1. Bappenas RI. *Strategi Nasional Kecerdasan Artifisial 2020–2045*.
2. Kemenkominfo RI. *Pedoman Etika AI dan Literasi Data Nasional*. (2023).

3. Telkom University. *AI and Data Science Roadmap for Indonesia 2045*.
4. Tarumingkeng, Rudy C. *Digital Ethics and Leadership 5.0*. RudyCT Academic Series, 2025.
5. Tarumingkeng, Rudy C. *Manajemen Inovasi di Era AI: Mengubah Ide Menjadi Solusi Nyata*. (2025).

✦ Refleksi Penutup

Dalam pandangan filosofis, penguasaan *machine learning* bukanlah tujuan akhir, melainkan **langkah menuju kesadaran intelektual dan moral** dalam dunia digital.

Ketika *data scientist* memahami bahwa setiap baris kode memengaruhi kehidupan manusia, maka teknologi berubah dari alat menjadi amanah.

Sebagaimana tertulis di dalam *kompas data scientist*.

"From Intelligence → Integrity → Impact → Inclusion"

Artinya, kecerdasan sejati tidak hanya menciptakan model yang cerdas, tetapi juga dunia yang lebih adil, inklusif, dan berkeadaban.

ChatGPT 5. Writer's account. Accessed 4 November 2025

<https://chatgpt.com/c/6909654c-03fc-8322-b33a-c15c284974b4>