HADLA
MEDIA INFORMASI

DASAR-DASAR MACHINE LEARNING:

Teori dan Aplikasi

Penulis :

Nada Arina Romli, M.I.Kom.
Nurhadi, S.Kom., M.Kom.
Ahmad Budi Trisnawan, S.T., M.Kom.
Eka Prasetya Adhy Sugara, S.T., M.Kom.
A. Taqwa Martadinata, M.Kom.
Imam Halim Mursyidin, S.Kom., M.Kom.
M. Rhifky Wayahdi, S.Kom., M.Kom.
Mifta Ardianti, S.T., M.Kom.
Joni Karman, M.Kom.
Muhammad Edya Rosadi, S.Kom., M.Kom.
Ahmad Khusaeri, M.Kom.
Novi Lestari, S.Kom.,M.Kom.
Budi Berlinton Sitorus, S.T, M.Sc.
Dr. Muhamad Akbar. S.T.,M.IT.
Muhammad Irvai, M.Kom.

EDITOR : NURHADI, S.KOM., M.KOM.

DASAR-DASAR MACHINE LEARNING:
Teori dan Aplikasi

Penulis

Nada Arina Romli
Nurhadi
Ahmad Budi Trisnawan
Eka Prasetya Adhy Sugara
A. Taqwa Martadinata
Imam Halim Mursyidin
M. Rhifky Wayahdi
Mifta Ardianti
Joni Karman
Muhammad Edya Rosadi
Ahmad Khusaeri
Novi Lestari
Budi Berlinton Sitorus
Muhamad Akbar
Muhammad Irvai

PENERBIT:

HADLA
MEDIA INFORMASI
Website: www.media.hadlacorp.com

UU No 28 tahun 2014 tentang Hak Cipta

Pasal 113

1) Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
2) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp. 500.000.000,00 (lima ratus juta rupiah).
3) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (l) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/ atau pidana denda paling banyak Rp1.000.000.000,00 (satu miliar rupiah).
4) Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp. 4.000.000.000,00 (empat miliar rupiah).

DASAR-DASAR MACHINE LEARNING: Teori dan Aplikasi

Tim Penulis:

Nada Arina Romli
Nurhadi
Ahmad Budi Trisnawan
Eka Prasetya Adhy Sugara
A. Taqwa Martadinata
Imam Halim Mursyidin
M. Rhifky Wayahdi
Mifta Ardianti
Joni Karman
Muhammad Edya Rosadi
Ahmad Khusaeri
Novi Lestari
Budi Berlinton Sitorus
Muhamad Akbar
Muhammad Irvai

Desain Cover:
Sulaiman
Tata Letak:
Sulaiman
Editor
Nurhadi

ISBN:

Cetakan Pertama:
januari, 2026

KATA PENGANTAR

Puji dan syukur ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya buku berjudul **"DASAR-DASAR MACHINE LEARNING: Teori dan Aplikasi"** ini dapat disusun dan diselesaikan dengan baik. Buku ini hadir sebagai respons atas pesatnya perkembangan teknologi informasi dan kebutuhan akan pemahaman yang kuat terhadap machine learning sebagai salah satu pilar utama dalam bidang kecerdasan buatan, data science, dan transformasi digital di berbagai sektor.

Machine learning telah menjadi teknologi kunci yang mendorong inovasi dalam beragam bidang, mulai dari industri, pendidikan, kesehatan, keuangan, transportasi, hingga pemerintahan. Kemampuannya dalam mempelajari pola dari data, membuat prediksi, serta mendukung pengambilan keputusan menjadikan machine learning sebagai kompetensi penting yang harus dikuasai oleh mahasiswa, akademisi, peneliti, maupun praktisi teknologi. Oleh karena itu, buku ini disusun untuk memberikan landasan konseptual dan teoretis yang kuat, sekaligus memperkenalkan penerapan machine learning secara praktis dan sistematis.

Buku ini membahas konsep-konsep dasar machine learning secara bertahap, dimulai dari pengenalan data dan proses pembelajaran mesin, jenis-jenis pembelajaran seperti supervised learning, unsupervised learning, dan semi-supervised learning, hingga pengenalan algoritma-algoritma populer yang banyak digunakan dalam berbagai studi kasus. Setiap pembahasan disusun dengan pendekatan yang terstruktur, disertai penjelasan yang mudah dipahami, ilustrasi konseptual, serta contoh aplikasi yang relevan dengan kebutuhan dunia nyata. Dengan demikian, pembaca diharapkan tidak hanya memahami teori, tetapi juga mampu mengaitkannya dengan permasalahan praktis yang dihadapi.

Selain aspek teoretis, buku ini juga menekankan pentingnya pemahaman terhadap proses pengolahan data, evaluasi model, serta tantangan dan etika dalam penerapan machine learning. Hal ini menjadi penting mengingat keberhasilan suatu model tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh kualitas data, pemilihan metode yang tepat, serta tanggung jawab dalam pemanfaatan teknologi. Melalui pendekatan ini, buku ini diharapkan dapat membentuk cara berpikir analitis, kritis, dan bertanggung jawab dalam mengembangkan solusi berbasis machine learning.

Buku **"Dasar-Dasar Machine Learning: Teori dan Aplikasi"** ini ditujukan sebagai buku ajar maupun referensi bagi mahasiswa jenjang sarjana, khususnya pada bidang ilmu komputer, sistem informasi, teknik informatika, dan disiplin ilmu terkait. Selain itu, buku ini juga dapat dimanfaatkan oleh dosen, peneliti, serta praktisi yang ingin memperkuat pemahaman konseptual sekaligus memperluas wawasan mengenai penerapan machine learning di berbagai bidang. Penyusunan materi diselaraskan dengan kebutuhan pembelajaran akademik dan perkembangan teknologi terkini, sehingga diharapkan relevan dan aplikatif.

Penulis menyadari sepenuhnya bahwa buku ini tidak akan terwujud tanpa dukungan dan kontribusi dari berbagai pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada rekan sejawat, kolega akademik, serta semua pihak yang telah memberikan masukan, saran, dan dorongan selama proses penulisan. Ucapan terima kasih juga penulis sampaikan kepada para pembaca yang telah memilih buku ini sebagai bagian dari proses belajar dan pengembangan diri.

Akhir kata, penulis berharap buku ini dapat memberikan manfaat yang luas, menjadi sumber pembelajaran yang bermakna, serta berkontribusi dalam pengembangan ilmu pengetahuan dan teknologi, khususnya di bidang machine learning. Penulis menyadari bahwa buku ini masih memiliki keterbatasan, oleh karena itu kritik dan saran yang konstruktif sangat diharapkan sebagai bahan perbaikan dan penyempurnaan pada edisi-edisi berikutnya.

Semoga buku ini dapat menjadi langkah awal yang kokoh bagi pembaca dalam memahami dan menguasai dasar-dasar machine learning, serta mengaplikasikannya secara bijak dan bertanggung jawab.

.

Januari 2026,

Hormat kami,

Tim penulis

PENGANTAR EDITOR

Bismillahirrahmanirrahim,

Alhamdulillah kami panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas segala rahmat, taufik, dan hidayah-Nya sehingga buku ini yang berjudul **"Dasar-Dasar Machine Learning: Teori dan Aplikasi"** dapat diterbitkan dan hadir di tengah-tengah pembaca. Buku ini disusun sebagai respons akademik terhadap perkembangan pesat teknologi informasi dan meningkatnya kebutuhan akan pemahaman yang mendasar, sistematis, dan aplikatif mengenai machine learning. Dalam konteks tersebut, kehadiran buku ini menjadi sangat relevan sebagai sumber pembelajaran dan rujukan ilmiah.

Sebagai editor, kami menilai bahwa machine learning tidak lagi dapat dipandang sebagai bidang yang bersifat khusus atau terbatas pada kalangan tertentu, melainkan telah menjadi kompetensi lintas disiplin yang berperan penting dalam berbagai sektor. Pemanfaatan machine learning dalam analisis data, sistem rekomendasi, pengenalan pola, otomasi cerdas, dan pengambilan keputusan telah mendorong perubahan signifikan dalam cara manusia bekerja dan berinteraksi dengan teknologi. Oleh karena itu, diperlukan buku dasar yang tidak hanya menjelaskan konsep dan teori, tetapi juga mampu mengaitkannya dengan konteks aplikasi nyata secara proporsional. Buku ini hadir untuk menjawab kebutuhan tersebut.

Proses penyuntingan buku ini dilakukan dengan tujuan menjaga keseimbangan antara ketepatan ilmiah dan keterbacaan materi. Editor berupaya memastikan bahwa setiap bab tersusun secara runtut dan memiliki keterkaitan logis antarbab, dimulai dari pengenalan konsep dasar machine learning, jenis-jenis pembelajaran mesin, hingga pembahasan algoritma dan contoh penerapannya. Istilah-istilah teknis disesuaikan dengan kaidah keilmuan yang berlaku, sekaligus disajikan dengan bahasa yang komunikatif agar mudah dipahami oleh pembaca pemula tanpa mengurangi kedalaman substansi.

Buku **"Dasar-Dasar Machine Learning: Teori dan Aplikasi"** memiliki kekuatan pada pendekatan penyajiannya yang mengintegrasikan aspek teoretis dengan sudut pandang praktis. Pembaca diajak untuk memahami tidak hanya bagaimana suatu algoritma bekerja, tetapi juga mengapa algoritma tersebut digunakan, bagaimana data diproses, serta bagaimana hasil model dievaluasi. Selain itu, buku ini juga menyinggung

berbagai tantangan dalam implementasi machine learning, seperti kualitas data, bias model, interpretabilitas, serta isu etika dan tanggung jawab dalam pemanfaatan teknologi. Pembahasan ini penting untuk membentuk pola pikir kritis dan reflektif bagi pembaca.

Dari sudut pandang editor, buku ini sangat tepat digunakan sebagai buku ajar pada jenjang pendidikan tinggi, khususnya program studi yang berkaitan dengan ilmu komputer, teknik informatika, sistem informasi, dan data science. Struktur materi yang sistematis memungkinkan dosen dan mahasiswa menggunakannya sebagai pegangan utama dalam proses pembelajaran. Di sisi lain, buku ini juga relevan bagi pembaca umum dan praktisi pemula yang ingin membangun pemahaman konseptual yang kuat sebelum melangkah ke tahap implementasi yang lebih lanjut.

Editor menyadari bahwa penyusunan dan penyuntingan sebuah buku akademik merupakan proses yang berkelanjutan. Oleh karena itu, kami sangat terbuka terhadap kritik, saran, dan masukan dari para pembaca demi penyempurnaan isi dan penyajian buku ini pada edisi berikutnya. Ucapan apresiasi yang setinggi-tingginya kami sampaikan kepada penulis atas kerja keras, dedikasi, serta keterbukaannya dalam menerima masukan selama proses penyuntingan, sehingga buku ini dapat disajikan dengan kualitas yang optimal.

Akhir kata, kami berharap buku **"Dasar-Dasar Machine Learning: Teori dan Aplikasi"** dapat memberikan kontribusi nyata dalam pengembangan literasi dan pemahaman machine learning, serta menjadi referensi yang bermanfaat bagi pembaca dalam memahami dan menerapkan konsep-konsep dasar machine learning secara tepat, kritis, dan bertanggung jawab.

Januari 2026

Hormat saya

Nurhadi
Editor

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1 PENDAHULUAN MACHINE LEARNING

Nada Arina Romli, M.I.Kom.

A. DEFINISI MACHINE LEARNING

Machine learning adalah subbidang kecerdasan buatan (AI) yang berfokus pada algoritma yang dapat "mempelajari" pola data pelatihan dan, selanjutnya, membuat kesimpulan yang akurat tentang data baru. Kemampuan pengenalan pola ini memungkinkan model machine learning untuk membuat keputusan atau prediksi tanpa instruksi eksplisit yang dikodekan secara langsung.

Machine learning telah mendominasi bidang AI: ia menyediakan tulang punggung sebagian besar sistem AI modern, mulai dari model peramalan hingga kendaraan otonom hingga model bahasa besar (LLM) dan alat AI generatif lainnya.

Premis utama machine learning (ML) adalah bahwa jika Anda mengoptimalkan kinerja model pada kumpulan data tugas yang cukup menyerupai masalah dunia nyata yang akan digunakan—melalui proses yang disebut pelatihan model—model tersebut dapat membuat prediksi yang akurat pada data baru yang dilihatnya dalam kasus penggunaan akhirnya.

Pelatihan itu sendiri hanyalah sarana untuk mencapai tujuan: generalisasi, penerjemahan kinerja yang kuat pada data pelatihan ke hasil yang bermanfaat dalam skenario dunia nyata, adalah tujuan mendasar dari machine learning. Pada intinya, model terlatih menerapkan pola yang dipelajarinya dari data pelatihan untuk menyimpulkan output yang benar untuk tugas dunia nyata: penerapan model AI oleh karena itu disebut inferensi AI.

Pembelajaran mendalam (deep learning), sub-bidang machine learning yang didorong oleh jaringan saraf tiruan yang besar—atau lebih tepatnya, "dalam"—telah muncul selama beberapa dekade terakhir sebagai arsitektur model AI mutakhir di hampir setiap domain di mana AI digunakan. Berbeda dengan algoritma yang didefinisikan secara eksplisit dari machine learning tradisional, pembelajaran mendalam

bergantung pada "jaringan" operasi matematika terdistribusi yang memberikan kemampuan yang tak tertandingi untuk mempelajari nuansa rumit dari data yang sangat kompleks. Karena pembelajaran mendalam membutuhkan sejumlah besar data dan sumber daya komputasi, kemunculannya bertepatan dengan meningkatnya pentingnya "big data" dan unit pemrosesan grafis (GPU).

Disiplin ilmu machine learning sangat terkait erat dengan ilmu data. Dalam arti tertentu, machine learning dapat dipahami sebagai kumpulan algoritma dan teknik untuk mengotomatiskan analisis data dan (yang lebih penting) menerapkan pembelajaran dari analisis tersebut untuk eksekusi tugas-tugas yang relevan secara otonom..

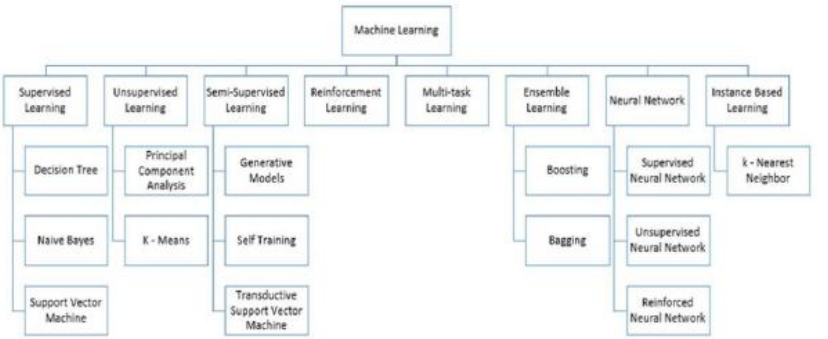
Arthur Samuel mendefinisikan pembelajaran mesin sebagai cabang penelitian yang memungkinkan komputer untuk belajar tanpa pemrograman eksplisit. Program catur Arthur Samuel telah membuatnya terkenal. Untuk melatih mesin agar dapat menangani data secara lebih efektif, pembelajaran mesin (ML) digunakan. Terkadang kita tidak mampu menguraikan atau mengekstrak informasi dari data setelah melihatnya. Sekarang kita menggunakan pembelajaran mesin. Kebutuhan akan pembelajaran mesin hanya akan terus meningkat karena banyaknya dataset yang tersedia saat ini. Pembelajaran mesin digunakan oleh beberapa sektor untuk mengambil data yang relevan.

Tujuan pembelajaran mesin adalah untuk belajar dari data. Membuat robot belajar tanpa pemrograman eksplisit telah menjadi subjek penelitian yang luas.

Topik ini mencakup kumpulan data yang sangat besar, dan banyak matematikawan dan programmer telah menggunakan berbagai cara untuk menemukan jawabannya.

Untuk mengatasi masalah berbasis data, pembelajaran mesin menggunakan berbagai metode. Ilmuwan data sering menekankan bahwa tidak ada satu algoritma pun yang cocok untuk setiap situasi. Jenis masalah yang ingin Anda selesaikan, jumlah variabel, jenis model terbaik, dan faktor-faktor lainnya semuanya memengaruhi jenis metode yang digunakan.

Berikut sekilas tentang beberapa algoritma yang umum digunakan dalam machine learning (ML).



Gambar 1. 1 Set Analisis dalam Machine Learning

B. ALGORITMA DALAM MACHINE LEARNING

Semua metode machine learning dapat dikategorikan ke dalam tiga paradigma pembelajaran yang berbeda: Tergantung pada jenis data pelatihan yang digunakan dan sifat tujuan pelatihan, pembelajaran dapat berupa pembelajaran terawasi, pembelajaran tak terawasi, atau pembelajaran penguatan.

Pembelajaran terawasi melatih model untuk memprediksi keluaran yang "benar" untuk masukan tertentu. Ini berlaku untuk tugas-tugas yang membutuhkan tingkat akurasi tertentu relatif terhadap "kebenaran dasar" eksternal, seperti klasifikasi atau regresi.

Pembelajaran tak terawasi melatih model untuk membedakan pola intrinsik, ketergantungan, dan korelasi dalam data. Tidak seperti dalam pembelajaran terawasi, tugas pembelajaran tak terawasi tidak melibatkan kebenaran dasar eksternal apa pun yang menjadi acuan perbandingan keluarannya.

Pembelajaran penguatan (reinforcement learning/RL) melatih model untuk mengevaluasi lingkungannya dan mengambil tindakan yang akan menghasilkan imbalan terbesar. Skenario RL tidak melibatkan keberadaan kebenaran dasar tunggal, tetapi melibatkan keberadaan tindakan "baik" dan "buruk" (atau netral).

Proses pelatihan ujung-ke-ujung untuk suatu model dapat, dan sering kali, melibatkan pendekatan hibrida yang memanfaatkan lebih dari

satu paradigma pembelajaran ini. Misalnya, pembelajaran tanpa pengawasan sering digunakan untuk memproses data terlebih dahulu untuk digunakan dalam pembelajaran terawasi atau pembelajaran penguatan. Model bahasa besar (LLM) biasanya menjalani pelatihan awal (pra-pelatihan) dan penyempurnaan melalui varian pembelajaran terawasi, diikuti oleh penyempurnaan lebih lanjut melalui teknik RL seperti pembelajaran penguatan dari umpan balik manusia (RLHF).

Dalam praktik yang serupa tetapi berbeda, berbagai metode pembelajaran ensemble menggabungkan keluaran dari beberapa algoritma.

1. Supervised Learning

Semua metode machine learning dapat dikategorikan ke dalam tiga paradigma pembelajaran yang berbeda: Pembelajaran terawasi, pembelajaran tak terawasi, atau pembelajaran penguatan, tergantung pada jenis data pelatihan yang digunakan dan sifat tujuan pelatihan.

Pembelajaran terawasi melatih model untuk memprediksi keluaran yang "benar" untuk masukan tertentu. Ini berlaku untuk tugas-tugas yang membutuhkan tingkat akurasi tertentu relatif terhadap "kebenaran dasar" eksternal, seperti klasifikasi atau regresi.

Pembelajaran tak terawasi melatih model untuk membedakan pola intrinsik, ketergantungan, dan korelasi dalam data. Tidak seperti dalam pembelajaran terawasi, tugas pembelajaran tak terawasi tidak melibatkan kebenaran dasar eksternal apa pun yang menjadi acuan perbandingan keluarannya.

Pembelajaran penguatan (reinforcement learning/RL) melatih model untuk mengevaluasi lingkungannya dan mengambil tindakan yang akan menghasilkan imbalan terbesar. Skenario RL tidak melibatkan keberadaan kebenaran dasar tunggal, tetapi melibatkan keberadaan tindakan "baik" dan "buruk" (atau netral).

Proses pelatihan ujung-ke-ujung untuk suatu model dapat, dan sering kali, melibatkan pendekatan hibrida yang memanfaatkan lebih dari satu paradigma pembelajaran ini. Misalnya, pembelajaran tanpa pengawasan sering digunakan untuk memproses data terlebih dahulu untuk digunakan dalam pembelajaran terawasi atau pembelajaran penguatan. Model bahasa besar (LLM) biasanya menjalani pelatihan awal (pra-pelatihan) dan penyempurnaan melalui varian pembelajaran

terawasi, diikuti oleh penyempurnaan lebih lanjut melalui teknik RL seperti pembelajaran penguatan dari umpan balik manusia (RLHF).

Dalam praktik yang serupa tetapi berbeda, berbagai metode pembelajaran ensemble menggabungkan keluaran dari beberapa algoritma.

Algoritma pembelajaran terawasi melatih model untuk tugas-tugas yang membutuhkan akurasi, seperti klasifikasi atau regresi. Machine learning terawasi mendukung model pembelajaran mendalam (deep learning) canggih dan berbagai model ML tradisional yang masih banyak digunakan di berbagai industri.

Model regresi memprediksi nilai kontinu, seperti harga, durasi, suhu, atau ukuran. Contoh algoritma regresi tradisional meliputi regresi linier, regresi polinomial, dan model ruang keadaan (state space).

Model klasifikasi memprediksi nilai diskrit, seperti kategori (atau kelas) tempat suatu titik data berada, keputusan biner, atau tindakan spesifik yang harus diambil. Contoh algoritma klasifikasi tradisional meliputi mesin vektor pendukung (SVM), Naïve Bayes, dan regresi logistik.

Banyak algoritma ML terawasi dapat digunakan untuk kedua tugas tersebut. Misalnya, output dari algoritma regresi dapat digunakan untuk memprediksi klasifikasi.

Agar dapat diukur dan dioptimalkan akurasinya, output model harus dibandingkan dengan kebenaran dasar (ground truth): output ideal atau "benar" untuk setiap input yang diberikan. Dalam pembelajaran terawasi konvensional, kebenaran dasar tersebut disediakan oleh data berlabel. Model deteksi spam email dilatih pada kumpulan data email yang masing-masing telah diberi label sebagai SPAM atau BUKAN SPAM. Model segmentasi gambar dilatih pada gambar di mana setiap piksel individual telah dianotasi oleh klasifikasinya. Tujuan pembelajaran terawasi adalah untuk menyesuaikan parameter model hingga outputnya secara konsisten sesuai dengan kebenaran dasar yang diberikan oleh label tersebut.

Hal penting dalam pembelajaran terawasi adalah penggunaan fungsi kerugian yang mengukur perbedaan ("kerugian") antara output model dan kebenaran dasar di seluruh kumpulan input pelatihan. Tujuan pembelajaran terawasi didefinisikan secara matematis sebagai meminimalkan output dari fungsi kerugian. Setelah kerugian dihitung, berbagai algoritma optimasi—yang sebagian besar melibatkan

BAB 2 Hubungan Machine Learning dengan Artificial Intelligence dan Data Science

Nurhadi, S.Kom., M.Kom.

A. KONSEP DASAR ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, DAN DATA SCIENCE

Transformasi digital yang terjadi secara global telah mengubah cara manusia bekerja, berkomunikasi, dan mengambil keputusan. Di balik perubahan tersebut, terdapat peran penting teknologi berbasis data dan kecerdasan buatan. Artificial Intelligence (AI), Machine Learning (ML), dan Data Science (DS) merupakan tiga bidang utama yang menjadi fondasi dalam pengembangan sistem cerdas modern. Ketiganya tidak hanya berkembang pesat secara teknologi, tetapi juga memberikan dampak signifikan dalam berbagai sektor, seperti pendidikan, kesehatan, industri, keuangan, dan pemerintahan.

Dalam konteks akademik, pemahaman terhadap konsep dasar AI, ML, dan Data Science menjadi kompetensi esensial bagi mahasiswa bidang Ilmu Komputer, Sistem Informasi, dan bidang terkait. Meskipun sering digunakan secara bersamaan, ketiga istilah tersebut memiliki fokus, ruang lingkup, dan tujuan yang berbeda. Oleh karena itu, diperlukan pembahasan konseptual yang komprehensif agar tidak terjadi kesalahpahaman dalam penerapannya.

1. Konsep Dasar Artificial Intelligence

Artificial Intelligence atau kecerdasan buatan merupakan cabang ilmu komputer yang berfokus pada pengembangan sistem yang mampu meniru atau mensimulasikan kecerdasan manusia. Menurut Russell dan Norvig, AI adalah studi tentang agen cerdas (intelligent agents), yaitu sistem yang mampu mengamati lingkungan dan mengambil tindakan yang memaksimalkan peluang keberhasilan dalam mencapai tujuan tertentu.

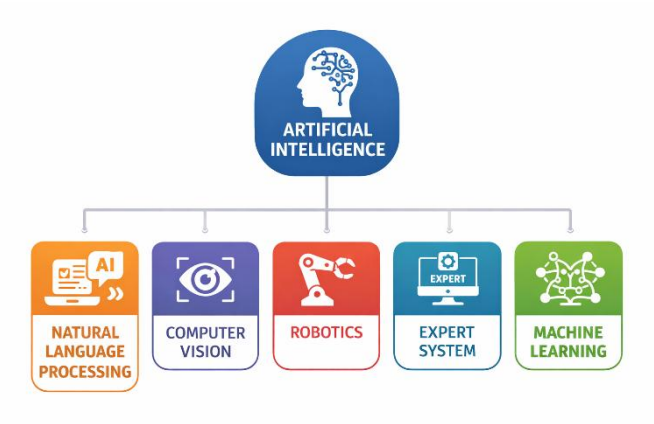
Secara historis, AI berkembang dari pendekatan berbasis aturan (rule-based systems) hingga pendekatan berbasis pembelajaran dan data. Pada tahap awal, AI banyak mengandalkan logika simbolik dan

representasi pengetahuan, seperti sistem pakar (expert systems) yang menggunakan aturan "if–then". Namun, pendekatan ini memiliki keterbatasan dalam menangani ketidakpastian dan kompleksitas dunia nyata.

Ruang lingkup Artificial Intelligence meliputi berbagai bidang, antara lain:

- **Knowledge Representation and Reasoning**, untuk menyimpan dan memanipulasi pengetahuan
- **Search and Optimization**, seperti algoritma pencarian heuristik
- **Natural Language Processing (NLP)**, untuk memahami dan menghasilkan bahasa manusia
- **Computer Vision**, untuk memahami informasi visual
- **Robotics**, untuk mengintegrasikan kecerdasan dengan sistem fisik

Tujuan utama AI adalah menciptakan sistem yang mampu berpikir dan bertindak secara rasional, baik dalam konteks yang terdefinisi dengan baik maupun dalam lingkungan yang dinamis dan kompleks.



Gambar 2. 1 Ruang Lingkup Artificial Intelligence

2. Konsep Dasar Machine Learning

Machine Learning merupakan subbidang dari Artificial Intelligence yang berfokus pada pengembangan algoritma yang memungkinkan sistem belajar dari data dan pengalaman. Berbeda dengan pemrograman tradisional, pada Machine Learning aturan atau pola tidak ditentukan secara eksplisit oleh programmer, melainkan dipelajari secara otomatis dari data.

Tom Mitchell mendefinisikan Machine Learning sebagai berikut: *"A computer program is said to learn from experience E with respect to some task T and performance measure P, if its performance at task T improves with experience E."* Definisi ini menegaskan bahwa inti dari ML adalah peningkatan kinerja melalui pembelajaran berbasis data.

Secara umum, Machine Learning diklasifikasikan menjadi tiga pendekatan utama:

1. **Supervised Learning**
   Menggunakan data berlabel untuk melatih model, contohnya klasifikasi dan regresi.

2. **Unsupervised Learning**
   Menganalisis data tanpa label untuk menemukan pola tersembunyi, seperti clustering dan association rules.

3. **Reinforcement Learning**
   Sistem belajar melalui interaksi dengan lingkungan menggunakan mekanisme reward dan punishment.

Algoritma Machine Learning banyak digunakan dalam berbagai aplikasi, seperti prediksi penjualan, deteksi penipuan, pengenalan wajah, serta sistem rekomendasi pada platform digital.

Gambar 2. 2 Alur Kerja Machine Learning

3. Konsep Dasar Data Science

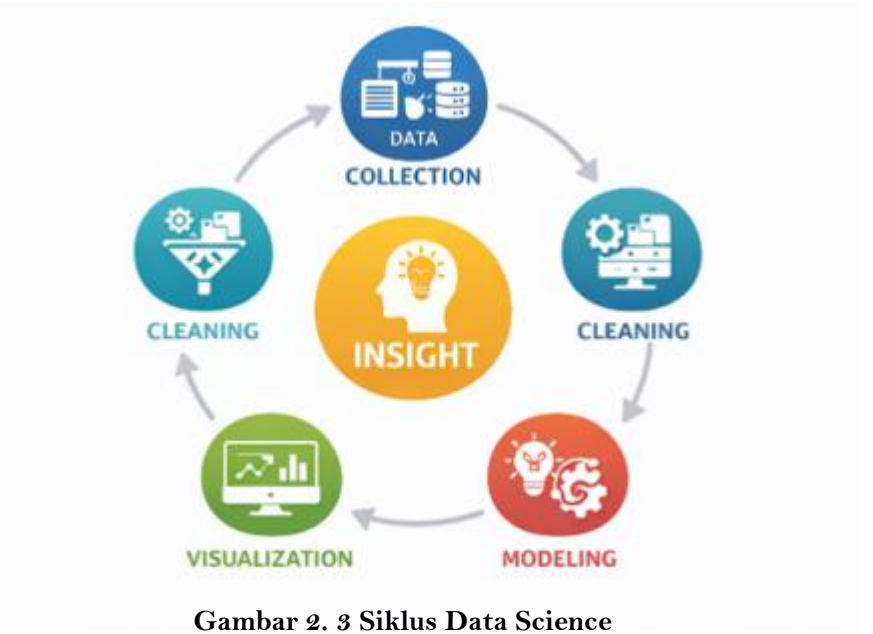
Data Science merupakan disiplin ilmu interdisipliner yang menggabungkan statistika, matematika, pemrograman, dan pengetahuan domain untuk mengekstraksi informasi bernilai dari data. Fokus utama Data Science adalah menghasilkan wawasan (insight) yang dapat digunakan sebagai dasar pengambilan keputusan strategis.

Data Science tidak hanya berkaitan dengan analisis data, tetapi juga mencakup keseluruhan siklus pengelolaan data, mulai dari pengumpulan hingga visualisasi hasil. Seorang data scientist dituntut untuk memahami karakteristik data, memilih metode analisis yang tepat, serta mampu mengkomunikasikan hasil analisis secara efektif.

Tahapan utama dalam Data Science meliputi:

- **Data Collection**, pengumpulan data dari berbagai sumber
- **Data Cleaning**, mengatasi data yang tidak lengkap atau tidak konsisten
- **Exploratory Data Analysis**, memahami pola dan distribusi data
- **Modeling**, menggunakan teknik statistik atau Machine Learning
- **Visualization and Interpretation**, menyajikan hasil analisis secara informatif

Data Science berperan sebagai jembatan antara data mentah dan pengetahuan yang dapat dimanfaatkan oleh organisasi.



Gambar 2. 3 Siklus Data Science

B. MACHINE LEARNING SEBAGAI INTI PENGEMBANGAN ARTIFICIAL INTELLIGENCE

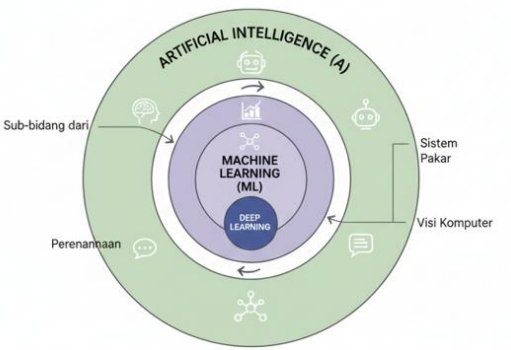
Artificial Intelligence (AI) merupakan bidang ilmu komputer yang bertujuan untuk mengembangkan sistem yang mampu menampilkan perilaku cerdas seperti manusia, termasuk kemampuan berpikir, belajar, dan mengambil keputusan. Dalam perkembangannya, AI mengalami perubahan pendekatan yang signifikan. Pada awalnya, sistem AI dibangun menggunakan pendekatan berbasis aturan (rule-based systems) yang sangat bergantung pada pengetahuan eksplisit dari pakar.

Seiring meningkatnya kompleksitas permasalahan dan ketersediaan data dalam jumlah besar, pendekatan berbasis aturan menjadi kurang efektif. Kondisi ini mendorong munculnya Machine Learning (ML) sebagai pendekatan utama dalam pengembangan AI modern. Machine Learning memungkinkan sistem AI belajar langsung dari data dan meningkatkan kinerjanya secara otomatis, sehingga menjadikannya inti dari pengembangan Artificial Intelligence saat ini.

Dalam konteks AI, Machine Learning memberikan kemampuan adaptif yang tidak dimiliki oleh sistem berbasis aturan. AI berbasis

Machine Learning mampu memperbaiki performa seiring bertambahnya data dan pengalaman, sehingga lebih fleksibel dalam menghadapi lingkungan yang dinamis.

Pendekatan utama dalam Machine Learning meliputi supervised learning, unsupervised learning, dan reinforcement learning. Ketiga pendekatan ini menjadi fondasi bagi berbagai aplikasi AI modern.



Gambar 2. 4 Posisi Machine Learning dalam Artificial Intelligence

Machine Learning berperan sebagai mekanisme pembelajaran utama dalam Artificial Intelligence. Tanpa Machine Learning, sistem AI hanya mampu menjalankan instruksi yang telah ditentukan sebelumnya dan sulit beradaptasi terhadap perubahan.

Peran utama Machine Learning dalam pengembangan AI meliputi:

1. Memungkinkan sistem belajar dari data historis
2. Meningkatkan akurasi prediksi dan keputusan
3. Mengurangi ketergantungan pada aturan manual
4. Mendukung otomatisasi dan adaptasi sistem

Dengan peran tersebut, Machine Learning menjadi elemen kunci yang menghubungkan data dengan kecerdasan sistem AI.

Dalam pengembangan sistem AI, Machine Learning diterapkan melalui alur kerja yang sistematis. Proses ini dimulai dari pengumpulan data hingga implementasi model ke dalam sistem AI.

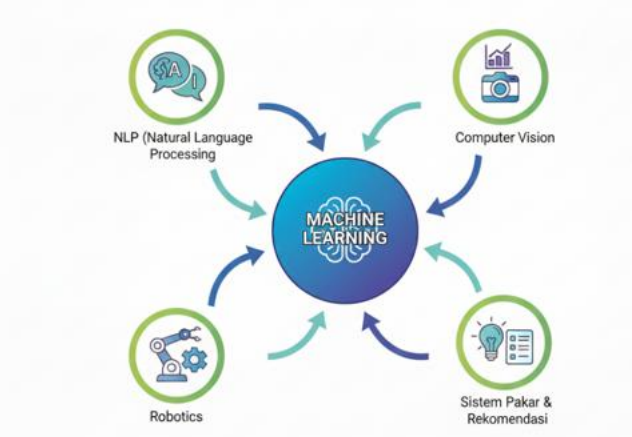
Tahapan umum alur kerja Machine Learning meliputi:

- Pengumpulan dan seleksi data
- Pra-pemrosesan dan pembersihan data
- Pelatihan model Machine Learning
- Evaluasi dan validasi model
- Implementasi model ke dalam sistem AI

Setiap tahapan memiliki peran penting dalam menentukan kualitas kecerdasan sistem AI yang dihasilkan.

Machine Learning menjadi fondasi bagi berbagai subbidang Artificial Intelligence lainnya. Dalam Computer Vision, Machine Learning digunakan untuk mengenali objek dan pola visual. Dalam Natural Language Processing, Machine Learning memungkinkan sistem memahami dan menghasilkan bahasa alami. Pada bidang Robotics, Machine Learning membantu robot belajar dari interaksi dengan lingkungan.

Hubungan ini menunjukkan bahwa Machine Learning tidak berdiri sendiri, melainkan menjadi penggerak utama kecerdasan pada berbagai aplikasi AI.



Gambar 2. 5 Machine Learning Sebagai Penggerak Sub Bidang AI

Perkembangan Artificial Intelligence modern tidak dapat dilepaskan dari kemajuan Machine Learning, khususnya Deep Learning. Deep Learning memungkinkan sistem AI memproses data kompleks seperti citra, suara, dan teks dengan tingkat akurasi yang tinggi.

Kemajuan ini didukung oleh meningkatnya ketersediaan data besar (big data), daya komputasi yang lebih kuat, serta pengembangan algoritma pembelajaran yang semakin efisien. Akibatnya, AI modern menjadi lebih cerdas, adaptif, dan mampu diterapkan pada berbagai permasalahan nyata.

Meskipun menjadi inti pengembangan AI, penggunaan Machine Learning juga menghadapi berbagai tantangan, antara lain:

- Ketergantungan pada kualitas dan jumlah data
- Risiko bias dan ketidakadilan dalam model
- Kurangnya transparansi dan interpretabilitas
- Kebutuhan sumber daya komputasi yang tinggi

Tantangan ini menuntut pendekatan yang hati-hati dan etis dalam pengembangan AI berbasis Machine Learning.

C. PERAN MACHINE LEARNING DALAM PROSES DATA SCIENCE

Perkembangan teknologi informasi telah mendorong meningkatnya ketersediaan data dalam berbagai bentuk dan skala. Data yang dihasilkan dari aktivitas manusia, sistem digital, dan perangkat cerdas memerlukan pendekatan khusus agar dapat diolah menjadi informasi dan pengetahuan yang bernilai. Dalam konteks inilah Data Science hadir sebagai disiplin ilmu yang mengintegrasikan statistika, pemrograman, analisis data, dan pengetahuan domain untuk menghasilkan insight berbasis data.

Salah satu komponen paling penting dalam Data Science adalah Machine Learning (ML). Machine Learning memungkinkan sistem mempelajari pola dari data secara otomatis dan menghasilkan model yang dapat digunakan untuk prediksi, klasifikasi, maupun pengambilan keputusan. Tanpa Machine Learning, proses Data Science akan terbatas pada analisis deskriptif dan eksploratif semata. Oleh karena itu, Machine Learning memiliki peran strategis dalam keseluruhan proses Data Science.

Proses Data Science umumnya digambarkan sebagai suatu siklus yang berkesinambungan. Tahapan dalam proses ini tidak bersifat linear, melainkan saling berkaitan dan dapat diulang sesuai kebutuhan. Secara umum, proses Data Science meliputi beberapa tahapan utama, yaitu pengumpulan data, pembersihan data, analisis eksploratif, pemodelan data, visualisasi, serta interpretasi hasil.

Pada setiap tahapan tersebut, Data Science bertujuan untuk memastikan bahwa data yang digunakan relevan, berkualitas, dan dapat mendukung tujuan analisis. Machine Learning berperan dominan pada tahap pemodelan, namun kontribusinya juga terasa pada tahap eksplorasi dan evaluasi data. Hal ini menunjukkan bahwa Machine Learning merupakan bagian integral dari keseluruhan proses Data Science.

Gambar 2. 6 Siklus Proses Data Science

Machine Learning menempati posisi strategis dalam Data Science sebagai metode utama untuk membangun model berbasis data. Fungsi utama Machine Learning dalam Data Science adalah mengubah data mentah menjadi model yang mampu merepresentasikan pola dan hubungan antar variabel.

Dalam praktik Data Science, Machine Learning digunakan untuk berbagai tujuan, seperti prediksi nilai di masa depan, pengelompokan data, deteksi anomali, dan pengenalan pola kompleks. Pendekatan ini

BAB 3 KONSEP DATA DALAM MACHINE LEARNING

Ahmad Budi Trisnawan, S.T., M.Kom.

A. PERAN DATA DALAM *MACHINE LEARNING*

Data memegang peranan fundamental dalam *Machine Learning* karena seluruh proses pembelajaran model bergantung pada kualitas, kuantitas, dan relevansi data yang digunakan (Muzakir et al., 2024). Tanpa data yang memadai, algoritma ML tidak dapat menemukan pola atau hubungan bermakna yang diperlukan untuk membuat prediksi yang akurat(Bintoro et al., 2024). Dalam konteks ini, data dapat dianggap sebagai "bahan mentah" yang akan diolah oleh algoritma untuk menghasilkan model yang dapat melakukan generalisasi terhadap kasus-kasus baru di dunia nyata. Semakin baik kualitas bahan mentah tersebut, semakin besar peluang model untuk memberikan hasil yang andal dan stabil.

Gambar 3. 1 Data dalam Machine Learning

Selain menjadi dasar bagi proses belajar, data juga berfungsi sebagai tolok ukur untuk mengevaluasi performa model. Model *Machine Learning* tidak cukup hanya diuji pada data yang digunakan untuk pelatihan, melainkan harus dievaluasi menggunakan data baru yang belum pernah dilihat sebelumnya (Permana et al., 2023). Hal ini penting untuk memastikan bahwa model tidak hanya menghafal pola yang ada, tetapi benar-benar memahami hubungan mendasar yang dapat berlaku pada data baru. Oleh karena itu, pemisahan data menjadi *train*, *validation*, dan *test set* menjadi langkah penting dalam *pipeline Machine Learning*.

Data juga berperan dalam mengendalikan tingkat bias dan variansi pada model. *Dataset* yang tidak seimbang, misalnya ketika jumlah sampel

pada satu kelas jauh lebih banyak dibanding kelas lainnya, dapat menyebabkan model condong memberi prediksi pada kelas dominan. Pola ini dikenal sebagai *class imbalance* dan dapat berdampak serius pada aplikasi dunia nyata, terutama dalam deteksi anomali, diagnosis medis, atau sistem keamanan (Elwirehardja et al., 2023). Dengan demikian, memastikan data representatif dan bebas dari bias adalah bagian penting dari proses desain sistem *Machine Learning* yang adil dan dapat dipercaya.

Selain itu, jenis data yang tersedia sering kali menentukan algoritma yang paling tepat digunakan. Data berformat numerik atau tabular biasanya cocok untuk algoritma *supervised learning* klasik, seperti *decision tree*, *random forest*, atau regresi linier. Sebaliknya, data, seperti gambar, teks, atau audio memerlukan metode dan arsitektur yang lebih kompleks, seperti *Convolutional Neural Networks* (CNN), *Recurrent Neural Networks* (RNN), *Transformer*, atau model-model *embedding* lainnya (Ichsan, 2025). Dengan kata lain, karakteristik data tidak hanya mempengaruhi proses pembelajaran, tetapi juga menentukan arah pengembangan model secara keseluruhan.

Terakhir, data juga berperan dalam proses penyempurnaan model setelah *deployment*. Dalam banyak aplikasi modern, sistem *Machine Learning* memerlukan data baru secara terus-menerus untuk meningkatkan akurasi seiring waktu. Mekanisme, seperti *online learning*, *model retraining*, dan *continuous monitoring* hanya dapat berjalan efektif, apabila tersedia data berkualitas dari proses operasional (Wijoyo et al., 2024). Dalam konteks ini, data tidak lagi dianggap sebagai aset statis, melainkan sebagai entitas dinamis yang terus memperkaya kemampuan model dan menjaga kualitas layanannya dalam jangka panjang.

B. JENIS-JENIS DATA DALAM *MACHINE LEARNING*

Data dalam *Machine Learning* memiliki beragam bentuk dan karakteristik. Memahami jenis data ini sangat penting karena setiap tipe memerlukan teknik representasi, preprocessing, dan algoritma yang berbeda (Aliyah et al., 2025). Kesalahan dalam mengenali jenis data dapat berdampak pada pemilihan model yang tidak tepat, hasil prediksi yang keliru, serta performa sistem yang buruk.

Jenis Data Berdasarkan Struktur

Jenis data berdasarkan struktur dalam konteks *machine learning* maupun pengolahan data secara umum.

a. Data Terstruktur *(Structured Data)* adalah data yang tersimpan dalam format tabel yang memiliki baris dan kolom dengan tipe data yang jelas. Formatnya mirip, seperti *spreadsheet* atau tabel *database* relasional.
b. Data Tidak Terstruktur *(Unstructured Data)* adalah data yang tidak memiliki format atau struktur baku. Karena tidak terorganisasi, data jenis ini membutuhkan teknik *preprocessing* dan representasi khusus sebelum dapat digunakan dalam ML.
c. Data Semi-Terstruktur *(Semi-Structured Data)* adalah data yang tidak tersimpan dalam format tabular, tetapi masih memiliki elemen struktural, seperti *tag* atau hierarki.

Jenis Data Berdasarkan Format atau Isi

Jenis data berdasarkan format atau isi dalam konteks komputasi maupun machine learning.

a. Data Numerik *(Numerical Data)* adalah data berbentuk angka dan dapat dilakukan operasi matematika.
b. Data Kategorik *(Categorical Data)* adalah data berupa kategori atau label tertentu.
c. Data Teks *(Text Data)* adalah salah satu data paling umum pada era digital, terutama dalam aplikasi NLP.
d. Data Gambar *(Image Data)* adalah data visual berupa kumpulan *pixel*.
e. Data Audio *(Audio Data)* adalah data berupa sinyal suara yang biasanya berupa gelombang *(waveform)*.
f. Data Video merupakan kombinasi dari gambar bergerak + audio.
g. Data Waktu *(Time-Series Data)* adalah data yang dicatat berurutan berdasarkan waktu.
h. Data Spasial atau Geospasial *(Spatial/Geospatial Data)* adalah data yang terkait lokasi atau posisi pada permukaan bumi.
i. Data Graf *(Graph Data)* adalah data dalam bentuk *node* (entitas) dan *edge* (hubungan).
j. Data Biner *(Binary Data)* adalah data dalam format 0 dan 1, sering digunakan sebagai representasi internal komputer.

Jenis Data Berdasarkan Label

Jenis data berdasarkan label dalam machine learning merujuk pada pengelompokan data berdasarkan ada atau tidaknya label (target/*output)* yang menyertai setiap sampel data (Bintoro et al.,

2024). Klasifikasi ini sangat penting karena menentukan jenis algoritma dan pendekatan pembelajaran yang akan digunakan.

a. Data Berlabel *(Labeled Data)* adalah setiap sampel sudah memiliki label yang benar.
b. Data Tidak Berlabel *(Unlabeled Data)* adalah data tanpa label, sehingga tidak diketahui kategori atau nilai targetnya.
c. *Weakly-Labeled* atau *Noisy Label* adalah label yang diberikan oleh sistem otomatis atau *crowd labeling*, sehingga tidak sepenuhnya akurat.
d. *Partially-Labeled Data* adalah sebagian data memiliki label, sebagian lain tidak.

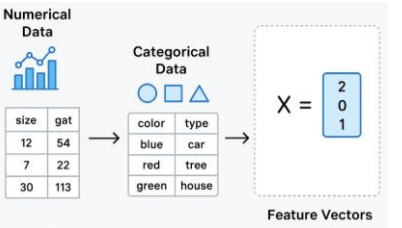
Jenis Data Berdasarkan Waktu

Jenis data berdasarkan waktu adalah pengelompokan data berdasarkan bagaimana data tersebut tercipta, diperbarui, atau berubah seiring berjalannya waktu. dalam konteks *machine learning*, analisis data, maupun sistem informasi, pengelompokan ini penting karena memengaruhi bagaimana data diproses, disimpan, dan dianalisis (Permana et al., 2023).

a. Data Statis *(Static Data)* adalah data yang tidak berubah atau jarang sekali berubah setelah dikumpulkan.
b. Data Dinamis *(Dynamic Data)* adalah data yang berubah secara berkala atau bahkan *real-time*.

C. REPRESENTASI DATA DALAM MACHINE LEARNING

Representasi data adalah cara bagaimana data diubah atau dikodekan menjadi bentuk yang dapat dipahami dan diproses oleh algoritma *Machine Learning* (Muhammad Hermawan et al., 2024). Hampir seluruh model ML bekerja dengan nilai numerik, sehingga data dalam bentuk kategori, teks, gambar, maupun audio harus dikonversi ke dalam format numerik terlebih dahulu (Elwirehardja et al., 2023). Representasi yang tepat akan memengaruhi kemampuan model dalam mengenali pola, sedangkan representasi yang buruk dapat menyebabkan model gagal belajar, *overfitting*, atau kehilangan informasi penting (Mitra Novitri Waruwu et al., 2024). Untuk itu, memilih representasi data yang sesuai dengan jenis data menjadi langkah fundamental dalam *pipeline Machine Learning*.



Gambar 3. 2 Representasi Data dalam Machine Learning

1. Representasi Data Numerik

Data numerik adalah jenis data yang secara langsung dapat diproses oleh sebagian besar algoritma ML tanpa perlu encoding tambahan.

a. Normalisasi *(Normalization)* menyesuaikan skala data agar berada dalam rentang tertentu, biasanya 0–1. Teknik umum dari Min-Max Scalling

$$x' = \frac{x - xmin}{xmax - xmin}$$

Digunakan ketika data memiliki rentang nilai berbeda-beda.

b. Standarisasi *(Standardization)* mengubah data agar memiliki mean = 0 dan standar deviasi = 1. Biasanya digunakan untuk algoritma, seperti SVM, *Logistic Regression*, dan *Neural Networks*.

2. Representasi Data Kategorik

Data kategorik tidak dapat digunakan langsung oleh model karena berupa teks atau label. Oleh sebab itu, perlu teknik *encoding*, sebagai berikut:

a. *One-Hot Encoding* mengubah kategori menjadi vektor biner.
b. *Label Encoding* mengubah kategori menjadi angka ordinal.
c. Target *Encoding* mengganti kategori dengan nilai rata-rata target, dan digunakan pada dataset besar dengan banyak kategori unik.
d. *Binary Encoding / Hashing Encoding* berguna untuk *high-cardinality features* (ribuan kategori).

3. Representasi Data Teks

Teks adalah jenis data tidak terstruktur yang memerlukan pemrosesan khusus. Representasinya telah berevolusi dari teknik sederhana hingga model *deep learning* modern.

a. Representasi Tradisional adalah cara menggambarkan, menyimpan, atau mengorganisasi data dalam bentuk yang sederhana, langsung, dan biasanya belum melalui proses transformasi kompleks.
b. *Word Embedding* adalah representasi kata dalam bentuk vektor berdimensi rendah yang mencerminkan makna.
c. *Contextual Embedding* (Modern NLP) adalah menghasilkan representasi kata berdasarkan konteks kalimat.

4. Representasi Data Gambar

Data gambar merupakan matriks angka *(array)* yang mewakili intensitas *pixel.*

a. *Pixel Matrix* dengan setiap gambar direpresentasikan sebagai Grayscale *(H x W x 1H)* dan RGB *(H x W x 3).*
b. *Feature Extraction* memiliki dua pendekatan, yaitu *Manual Feature Extraction* (klasik), seperti SIFT, HOG, dan SURF yang digunakan sebelum era *deep learning*, serta *Automatic Feature Extraction (Deep Learning)*, dimana CNN secara otomatis mengekstrak fitur hierarkis, seperti *edge*, *texture*, *shape*, dan *object parts.* Hasil ekstraksi gambar dapat direpresentasikan sebagai *embedding* berdimensi tinggi.

5. Representasi Data Audio

Audio direpresentasikan sebagai sinyal kontinu dan harus diubah ke bentuk yang lebih informatif.

a. *Waveform* adalah representasi dasar berupa amplitudo vs waktu.
b. *Spectrogram* adalah transformasi dari sinyal suara ke domain frekuensi.
c. *Mel-Spectrogram* adalah spektrogram yang menggunakan skala mel, dan lebih sesuai persepsi manusia.
d. MFCC *(Mel-Frequency Cepstral Coefficients)* adalah representasi paling populer untuk *speech recognition* dan *speaker identification.*

BAB 4 REPRESENTASI DAN TRANSFORMASI DATA

Eka Prasetya Adhy Sugara, S.T., M.Kom.

A. KONSEP REPRESENTASI DATA

Representasi data merupakan konsep fundamental dalam *Machine Learning* karena menjadi jembatan antara fenomena dunia nyata dan model matematis yang digunakan oleh algoritma pembelajaran mesin. Algoritma *Machine Learning* pada dasarnya tidak memiliki pemahaman semantik terhadap objek, peristiwa, maupun konteks dunia nyata. Algoritma hanya mampu memproses data dalam bentuk simbolik atau numerik. Oleh karena itu, bagaimana data direpresentasikan akan sangat menentukan informasi apa yang dapat dipelajari oleh model dan sejauh mana model tersebut mampu melakukan generalisasi terhadap data baru.

Secara umum, representasi data dapat didefinisikan sebagai proses pemetaan objek atau fenomena ke dalam struktur data formal yang dapat diproses secara komputasional. Dalam *Machine Learning*, struktur ini umumnya berupa vektor numerik atau matriks. Pemilihan representasi data yang tepat merupakan bagian integral dari desain sistem pembelajaran mesin, karena representasi yang buruk dapat menyebabkan model gagal mempelajari pola yang sebenarnya ada di dalam data (Wicaksana and Rachman 2018).

Dalam praktik *Machine Learning*, setiap objek data direpresentasikan sebagai *feature vector*. Fitur adalah atribut atau karakteristik terukur yang dianggap relevan untuk menggambarkan objek tersebut. Sebagai contoh, pada kasus prediksi kelulusan mahasiswa, fitur dapat berupa IPK, jumlah SKS, lama studi, dan status kehadiran. Pemilihan fitur ini tidak bersifat netral, melainkan mencerminkan asumsi perancang sistem terhadap faktor-faktor yang memengaruhi target prediksi. Fitur yang baik harus informatif, stabil, dan memiliki hubungan yang jelas dengan variabel target (Kuhn and Johnson 2019).

Representasi data juga sangat berkaitan dengan konsep *feature engineering*, yaitu proses sistematis untuk merancang, memilih, dan

memodifikasi fitur agar informasi yang terkandung dalam data dapat dimanfaatkan secara optimal oleh algoritma. Dalam banyak aplikasi nyata, peningkatan performa model lebih sering diperoleh melalui perbaikan representasi data dibandingkan dengan penggantian algoritma ke model yang lebih kompleks (Géron 2019). Hal ini menunjukkan bahwa representasi data memiliki peran yang lebih strategis daripada sekadar tahap awal preprocessing.

Secara matematis, dataset dalam *Machine Learning* biasanya direpresentasikan dalam bentuk matriks $X \in \mathbb{R}^{nxd}$, dengan n menyatakan jumlah sampel dan d menyatakan jumlah fitur. Setiap baris matriks merepresentasikan satu objek data, sedangkan setiap kolom merepresentasikan satu fitur. Representasi ini memungkinkan penerapan operasi aljabar linear yang menjadi dasar banyak algoritma *Machine Learning*, seperti regresi linear, *principal component analysis*, *support vector machine*, dan *neural network*. Formulasi matematis semacam ini memberikan kerangka kerja yang seragam untuk menganalisis berbagai algoritma pembelajaran mesin (Zhou 2021).

Selain aspek matematis, representasi data juga harus mempertimbangkan sifat dan struktur data yang digunakan. Data numerik, kategorikal, ordinal, teks, citra, dan audio memiliki karakteristik yang berbeda sehingga memerlukan pendekatan representasi yang berbeda pula. Sebagai contoh, data kategorikal tidak dapat direpresentasikan secara langsung sebagai bilangan tanpa mempertimbangkan makna kategorinya, karena hal tersebut dapat menimbulkan interpretasi ordinal yang keliru oleh model. Sementara itu, data teks memerlukan pemetaan dari kata atau dokumen ke dalam ruang vektor berdimensi tinggi agar makna dan konteks bahasa dapat dipertahankan (Lu 2022).

Dalam beberapa tahun terakhir, pendekatan representasi data mengalami perkembangan pesat seiring dengan kemajuan *Deep Learning*. Representasi data tidak lagi sepenuhnya dirancang secara manual melalui *feature engineering* tradisional, tetapi dapat dipelajari secara otomatis melalui pendekatan *representation learning*. *Representation learning* memungkinkan model untuk mengekstraksi representasi tingkat tinggi langsung dari data mentah, seperti piksel citra atau urutan kata dalam teks. Pendekatan ini memungkinkan sistem menemukan struktur laten yang kompleks tanpa campur tangan manusia secara eksplisit (Bengio et al. 2017).

Pendekatan *representation learning* banyak digunakan dalam berbagai aplikasi modern, seperti *word embedding* pada pemrosesan bahasa alami dan *feature extraction* otomatis pada pengenalan citra. Penelitian mutakhir menunjukkan bahwa representasi vektor yang padat dan bermakna mampu meningkatkan performa model secara signifikan, khususnya pada data berdimensi tinggi. Representasi yang baik dapat mengurangi kompleksitas komputasi sekaligus meningkatkan stabilitas dan akurasi model (Cunningham and Ghahramani 2015).

Dalam konteks pendidikan dan penelitian, pemahaman mendalam tentang konsep representasi data menjadi sangat penting karena sering kali menjadi sumber kesalahan dalam penerapan Machine Learning. Banyak kegagalan model bukan disebabkan oleh algoritma yang tidak tepat, melainkan oleh representasi data yang tidak sesuai dengan karakteristik masalah. Oleh karena itu, penguasaan konsep representasi data merupakan kompetensi dasar yang wajib dimiliki oleh mahasiswa sebelum mempelajari algoritma pembelajaran mesin yang lebih lanjut.

Representasi data tidak hanya berkaitan dengan proses mengubah data mentah ke dalam bentuk numerik, tetapi juga mencakup pemahaman konseptual tentang bagaimana informasi direpresentasikan, asumsi apa yang melekat dalam fitur yang dipilih, serta bagaimana representasi tersebut memengaruhi kemampuan model dalam belajar dan melakukan generalisasi. Representasi data yang tepat akan membantu model *Machine Learning* menangkap pola yang relevan, mengurangi *noise*, dan menghasilkan prediksi yang lebih andal.

B. JENIS-JENIS DATA DALAM MACHINE LEARNING

Jenis data merupakan aspek mendasar yang harus dipahami sebelum melakukan representasi dan transformasi data dalam *Machine Learning*. Setiap jenis data memiliki karakteristik, struktur, dan perlakuan yang berbeda, sehingga kesalahan dalam mengidentifikasi jenis data dapat berdampak langsung pada pemilihan metode representasi, teknik *pre processing*, serta kinerja model secara keseluruhan. Pemahaman jenis data merupakan langkah awal yang krusial dalam proses data mining dan pembelajaran mesin (Han, Pei, and Tong 2022).

Secara umum, data dalam *Machine Learning* dapat diklasifikasikan menjadi beberapa jenis utama, yaitu data numerik, data kategorikal, data ordinal, data teks, dan data berdimensi tinggi. Data numerik adalah data berbentuk angka yang dapat berupa nilai diskrit maupun kontinu, seperti usia, jumlah transaksi, atau suhu. Data jenis ini relatif mudah diproses

oleh algoritma *Machine Learning* karena sudah berada dalam bentuk numerik, meskipun sering kali masih memerlukan penyesuaian skala melalui normalisasi atau standardisasi (James et al. 2023).

Data kategorikal merepresentasikan kelas atau kelompok tertentu yang tidak memiliki makna numerik secara langsung, misalnya jenis kelamin, warna, atau jurusan mahasiswa. Data ini tidak dapat langsung digunakan oleh sebagian besar algoritma *Machine Learning* sehingga perlu diubah ke dalam bentuk numerik melalui teknik *encoding*. Perlakuan yang tidak tepat terhadap data kategorikal dapat menimbulkan bias model karena algoritma dapat salah menafsirkan hubungan antar kategori (S. Aggarwal 2023).

Berbeda dengan data kategorikal nominal, data ordinal memiliki urutan atau tingkatan yang bermakna, seperti tingkat pendidikan atau skala kepuasan. Meskipun berbentuk kategori, data ordinal mengandung informasi urutan yang penting untuk dipertahankan dalam proses representasi. Oleh karena itu, teknik representasi data ordinal harus mampu mencerminkan hubungan tingkatannya tanpa memberikan jarak numerik yang keliru antar kategori (Kuhn and Johnson 2019).

Selain data terstruktur, *Machine Learning* juga banyak berhadapan dengan data tidak terstruktur, seperti data teks, citra, audio, dan video. Data teks merupakan salah satu jenis data tidak terstruktur yang paling umum digunakan, misalnya pada analisis sentimen, klasifikasi dokumen, dan sistem rekomendasi. Data teks tidak dapat diproses secara langsung dan harus diubah menjadi fitur numerik melalui teknik khusus seperti *Bag of Words*, TF-IDF, atau *embedding* (S. Aggarwal 2023).

Jenis data lain yang semakin sering ditemui adalah data berdimensi tinggi, yaitu data dengan jumlah fitur yang sangat besar, seperti citra digital, data genomik, atau data hasil sensor. Data berdimensi tinggi menimbulkan tantangan khusus yang dikenal sebagai curse of dimensionality, di mana kinerja model dapat menurun seiring bertambahnya jumlah fitur. Oleh karena itu, jenis data ini sering memerlukan teknik reduksi dimensi dan representasi yang lebih ringkas (Cunningham and Ghahramani 2015).

Pemahaman terhadap jenis-jenis data dalam *Machine Learning* menjadi landasan penting dalam menentukan strategi representasi dan transformasi data. Identifikasi jenis data yang tepat akan membantu praktisi dan mahasiswa memilih metode preprocessing yang sesuai,

meminimalkan kesalahan representasi, serta meningkatkan efektivitas dan keandalan model pembelajaran mesin.

C. REPRESENTASI DATA NUMERIK

Data numerik merupakan jenis data yang paling umum dan paling mudah diproses dalam *Machine Learning* karena sudah berada dalam bentuk angka yang dapat langsung digunakan dalam perhitungan matematis. Data numerik dapat berupa data diskrit, seperti jumlah transaksi atau banyaknya mahasiswa, maupun data kontinu, seperti suhu, berat badan, atau nilai waktu. Meskipun demikian, representasi data numerik tidak selalu bersifat trivial, karena perbedaan skala, distribusi, dan rentang nilai antar fitur dapat memengaruhi kinerja algoritma pembelajaran mesin secara signifikan.

Dalam Machine Learning, data numerik umumnya direpresentasikan sebagai vektor fitur dalam ruang berdimensi tertentu. Setiap fitur numerik merepresentasikan satu dimensi dalam ruang tersebut. Algoritma Machine Learning, seperti regresi linear, regresi logistik, dan *support vector machine*, mengasumsikan bahwa fitur numerik berada dalam skala yang sebanding (James et al. 2023). Apabila asumsi ini tidak terpenuhi, fitur dengan rentang nilai yang lebih besar dapat mendominasi proses pembelajaran dan menyebabkan model menghasilkan prediksi yang bias.

Oleh karena itu, representasi data numerik sering kali dikombinasikan dengan teknik transformasi skala, seperti normalisasi dan standardisasi. Normalisasi bertujuan memetakan nilai data ke dalam rentang tertentu, misalnya antara 0 dan 1, sedangkan standardisasi bertujuan mengubah distribusi data agar memiliki rata-rata nol dan simpangan baku satu. Transformasi skala sangat penting terutama untuk algoritma berbasis jarak dan gradien, seperti *k-nearest neighbor* dan *neural network* (Géron 2019).

Selain transformasi skala, representasi data numerik juga perlu mempertimbangkan distribusi data. Data numerik yang memiliki distribusi sangat miring atau mengandung pencilan *(outlier)* dapat menurunkan performa model. Dalam kasus seperti ini, transformasi non-linear seperti transformasi logaritmik atau *Box–Cox* sering digunakan untuk membuat distribusi data lebih mendekati normal. Transformasi distribusi dapat meningkatkan stabilitas model dan mempercepat proses konvergensi pada algoritma pembelajaran (Kuhn and Johnson 2019).

BAB 5 Supervised Learning: Konsep Dasar

A. Taqwa Martadinata, M.Kom.

A. PENGERTIAN SUPERVISED LEARNING

Supervised learning atau pembelajaran terawasi merupakan salah satu paradigma paling fundamental dalam bidang *machine learning* (pembelajaran mesin) yang berperan penting dalam pengembangan sistem kecerdasan buatan (*Artificial Intelligence* atau AI) modern. Pembelajaran mesin sendiri merupakan cabang dari ilmu komputer yang berfokus pada bagaimana sistem dapat belajar secara otomatis dari data untuk membuat prediksi atau keputusan tanpa harus diprogram secara eksplisit. Dalam konteks ini, *supervised learning* menjadi pendekatan utama karena kemampuannya dalam mempelajari hubungan matematis antara data masukan (*input*) dan keluaran (*output*) berdasarkan contoh-contoh berlabel yang telah tersedia (Syed dan Lokhande 2024).

Pendekatan *supervised learning* bekerja dengan memanfaatkan data berlabel, yaitu kumpulan data yang setiap elemennya terdiri dari pasangan antara fitur (*feature*) dan label (*target*). Misalnya, dalam kasus klasifikasi gambar, setiap gambar diberi label yang menunjukkan objek apa yang terdapat di dalamnya, seperti "anjing", "kucing", atau "burung". Model kemudian mempelajari pola dari pasangan data tersebut untuk dapat mengklasifikasikan gambar baru yang belum pernah dilihat sebelumnya. Tujuan utama dari *supervised learning* adalah membangun fungsi pemetaan dari ruang input ke ruang output sedemikian rupa sehingga model dapat melakukan prediksi dengan tingkat akurasi tinggi terhadap data baru.

Dalam prosesnya, *supervised learning* meniru cara manusia belajar dari pengalaman. Misalnya, ketika seseorang belajar mengenali objek tertentu seperti buah apel dan jeruk, ia diberikan contoh-contoh yang jelas mengenai kedua buah tersebut. Setelah melihat sejumlah contoh dan mengetahui labelnya, seseorang akan mulai memahami ciri khas masing-masing buah, seperti bentuk, warna, dan tekstur. Ketika kemudian

diberikan gambar buah lain yang belum pernah dilihat sebelumnya, ia mampu mengidentifikasinya dengan benar karena telah mempelajari pola visual yang membedakan apel dari jeruk. Analogi ini menggambarkan prinsip dasar *supervised learning*: model belajar dari data yang telah diberi label untuk kemudian dapat membuat prediksi terhadap data baru dengan cara mengenali pola yang telah dipelajari.

Secara konseptual, *supervised learning* bertujuan menemukan suatu fungsi matematis $f: X \rightarrow Y$, di mana Xmerupakan himpunan data masukan (fitur-fitur) dan Yadalah himpunan keluaran atau target yang diharapkan. Setiap pasangan data dilambangkan sebagai (x_i, y_i), di mana $x_i$merepresentasikan vektor fitur dan $y_i$merupakan label atau nilai sebenarnya. Model berusaha menghasilkan fungsi $f(x_i)$yang mendekati $y_i$untuk seluruh data pelatihan. Untuk mengukur seberapa jauh hasil prediksi dari nilai sebenarnya, digunakan sebuah ukuran yang disebut *loss function*. Fungsi ini menghitung selisih antara prediksi dan nilai aktual, dan nilai tersebut menjadi acuan seberapa baik model melakukan prediksi.

Tujuan utama pelatihan dalam *supervised learning* adalah meminimalkan *loss function* tersebut agar model semakin akurat. Proses minimisasi biasanya dilakukan menggunakan algoritma optimasi seperti *gradient descent*. *Gradient descent* bekerja dengan menghitung turunan (*gradient*) dari fungsi kehilangan terhadap parameter model, lalu memperbarui parameter tersebut ke arah yang menurunkan nilai kesalahan. Dengan melakukan pembaruan secara iteratif, model secara bertahap belajar menghasilkan prediksi yang semakin mendekati nilai sebenarnya hingga mencapai titik konvergensi, yaitu saat perubahan nilai kesalahan menjadi sangat kecil (Du, Li, dan Wang 2025).

Dalam konteks matematis dan komputasional, *supervised learning* merupakan proses pencarian parameter model yang paling tepat dalam suatu ruang hipotesis (*hypothesis space*). Ruang hipotesis ini berisi semua kemungkinan fungsi pemetaan yang dapat menjelaskan hubungan antara input dan output. Algoritma pembelajaran berupaya memilih fungsi terbaik berdasarkan data yang tersedia melalui proses pelatihan. Proses ini tidak hanya mengandalkan kemampuan komputasi, tetapi juga memerlukan pemahaman mendalam terhadap struktur data, jenis masalah yang dihadapi, serta pemilihan metrik evaluasi yang tepat.

Lebih jauh lagi, *supervised learning* dapat diterapkan dalam berbagai bentuk permasalahan yang berbeda. Dua kategori utamanya adalah **klasifikasi** dan **regresi**. Klasifikasi digunakan ketika target atau label bersifat kategorikal, misalnya membedakan apakah suatu email termasuk spam atau bukan, menentukan jenis bunga berdasarkan ukuran kelopak dan mahkota, atau mengenali wajah seseorang dari citra digital. Di sisi lain, regresi digunakan ketika target bersifat numerik atau kontinu, seperti memprediksi harga rumah berdasarkan lokasi dan luas tanah, memperkirakan suhu udara, atau menilai pertumbuhan ekonomi berdasarkan variabel makroekonomi tertentu.

Dalam praktiknya, terdapat berbagai algoritma populer yang digunakan dalam *supervised learning*, antara lain *Linear Regression*, *Decision Tree*, *Support Vector Machine (SVM)*, *Naïve Bayes*, *Random Forest*, dan *Artificial Neural Network*. Masing-masing algoritma memiliki karakteristik dan keunggulan tersendiri. *Linear Regression* dikenal karena kesederhanaan dan kemudahan interpretasinya, sedangkan *SVM* efektif dalam memisahkan kelas dengan batas yang optimal pada ruang berdimensi tinggi. *Random Forest* menggabungkan banyak *Decision Tree* untuk meningkatkan stabilitas dan akurasi prediksi, sementara *Neural Network* mampu mengenali pola non-linear yang kompleks, terutama dalam jumlah data besar. Pemilihan algoritma bergantung pada sifat data, tujuan analisis, serta kebutuhan interpretabilitas yang diharapkan (Nkemdilim, Uche, dan Okwara 2024).

Selain keunggulannya, *supervised learning* juga memiliki sejumlah tantangan. Salah satu tantangan utama adalah ketergantungan terhadap data berlabel dalam jumlah besar. Pelabelan data sering kali memerlukan intervensi manusia, yang dapat memakan waktu, biaya, dan tenaga, terutama dalam domain kompleks seperti pengenalan citra medis atau pemrosesan bahasa alami. Tantangan lainnya adalah kemungkinan terjadinya *overfitting*, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga kehilangan kemampuan untuk menggeneralisasi ke data baru. Oleh karena itu, dibutuhkan strategi seperti *cross-validation*, *regularization*, dan pemilihan parameter yang tepat untuk menjaga keseimbangan antara kompleksitas model dan kemampuan generalisasi.

Perkembangan *supervised learning* telah membawa dampak besar dalam berbagai bidang. Dalam bidang **kesehatan**, model *supervised*

learning digunakan untuk mendiagnosis penyakit, mendeteksi kanker dari citra medis, dan memperkirakan efektivitas pengobatan. Dalam **keuangan**, pendekatan ini diterapkan dalam analisis risiko kredit, deteksi penipuan, dan prediksi harga saham. Dalam **transportasi**, algoritma ini berperan penting dalam pengembangan kendaraan otonom yang mampu mengenali lingkungan dan membuat keputusan secara real-time. Bahkan dalam **pendidikan**, *supervised learning* dimanfaatkan untuk memprediksi kinerja siswa dan memberikan rekomendasi pembelajaran yang dipersonalisasi.

Kemajuan teknologi komputasi dan ketersediaan data besar (*big data*) turut mempercepat perkembangan *supervised learning*. Perkembangan arsitektur *deep learning* atau jaringan saraf dalam (*deep neural networks*) yang berbasis prinsip *supervised learning* memungkinkan sistem untuk mengekstraksi representasi fitur yang semakin kompleks. Teknik ini telah menjadi dasar bagi berbagai aplikasi mutakhir seperti pengenalan wajah, terjemahan otomatis, dan sistem rekomendasi cerdas (Jiang 2021).

Dengan demikian, *supervised learning* tidak hanya berperan sebagai salah satu pendekatan dalam pembelajaran mesin, tetapi juga menjadi pondasi utama dalam pengembangan kecerdasan buatan modern. Melalui kemampuan untuk belajar dari data historis, membangun model matematis yang merepresentasikan fenomena dunia nyata, dan melakukan generalisasi terhadap data baru, *supervised learning* telah membuka jalan bagi munculnya sistem-sistem cerdas yang dapat beradaptasi, belajar, dan memberikan solusi inovatif di berbagai sektor kehidupan manusia. Pemahaman mendalam terhadap konsep, prinsip matematis, serta penerapan *supervised learning* menjadi krusial bagi peneliti dan praktisi untuk menciptakan teknologi yang akurat, efisien, dan dapat diandalkan di era digital yang semakin kompleks.

B. SEJARAH DAN PERKEMBANGAN

Konsep *supervised learning* atau pembelajaran terawasi memiliki akar yang dalam dalam teori *statistical learning*, suatu cabang ilmu yang mempelajari bagaimana sistem dapat menarik kesimpulan atau membuat keputusan berdasarkan data yang mengandung ketidakpastian. Teori ini mulai terbentuk antara tahun 1950 hingga 1970-an, ketika para peneliti berusaha menghubungkan pendekatan statistik klasik dengan kemampuan mesin untuk belajar dari data. Perkembangan awal ini tidak

dapat dilepaskan dari kontribusi dua ilmuwan besar asal Rusia, **Vladimir Vapnik** dan **Alexey Chervonenkis**, yang memperkenalkan dua konsep fundamental dalam teori pembelajaran mesin: **Vapnik–Chervonenkis (VC) Dimension** dan prinsip **Empirical Risk Minimization (ERM)** (Du, Li, dan Wang 2025).

VC-Dimension memberikan dasar matematis untuk mengukur kapasitas suatu model dalam memisahkan atau mengklasifikasikan data dalam ruang fitur tertentu. Konsep ini menjelaskan seberapa kompleks suatu model dapat menjadi tanpa kehilangan kemampuan generalisasinya terhadap data baru. Sementara itu, prinsip ERM menjelaskan bahwa model pembelajaran yang baik adalah model yang mampu meminimalkan risiko empiris atau kesalahan rata-rata terhadap data pelatihan. Dengan kata lain, ERM menekankan bahwa pembelajaran yang efektif dapat dicapai dengan cara mengurangi *training error* seraya tetap menjaga kemampuan untuk beradaptasi pada data yang belum pernah dilihat sebelumnya. Kedua konsep ini menjadi fondasi utama bagi teori pembelajaran mesin modern, karena keduanya menyeimbangkan antara akurasi dan generalisasi, dua aspek yang menjadi kunci dalam keberhasilan model *supervised learning* (Vapnik 1995).

Pada masa 1950–1970-an, kajian *machine learning* masih sangat erat kaitannya dengan bidang matematika, statistik, dan teori informasi. Fokus penelitian pada masa itu lebih banyak diarahkan pada pengembangan kerangka teoretis yang dapat menjamin konsistensi dan konvergensi suatu algoritma terhadap distribusi data yang tidak diketahui. Misalnya, model seperti **Linear Regression** dan **Logistic Regression** dikembangkan untuk menemukan hubungan linear antara variabel masukan dan keluaran. Selain itu, **Perceptron** — model jaringan saraf sederhana yang diperkenalkan oleh **Frank Rosenblatt** pada tahun 1958 — menandai upaya awal untuk mensimulasikan cara kerja neuron biologis dalam mengenali pola. Walaupun pada awalnya terbatas pada data yang dapat dipisahkan secara linear, Perceptron menjadi inspirasi awal bagi banyak model *neural network* modern yang lebih kompleks.

Kajian Vapnik mengenai *Statistical Learning Theory* (SLT) pada dekade 1970–1980 membawa perubahan besar terhadap arah penelitian pembelajaran mesin. Jika sebelumnya pembelajaran mesin lebih bersifat statistik deskriptif, maka SLT memperkenalkan pendekatan berbasis data dan fungsi prediktif yang lebih formal. SLT memberikan penjelasan teoretis tentang bagaimana model dapat belajar dari data terbatas dan

BAB 6 REGRESI LINEAR DAN NON-LINEAR

Imam Halim Mursyidin, S.Kom., M.Kom.

A. KONSEP REGRESI

Regresi merupakan salah satu teknik fundamental dalam *Machine Learning* dan statistika yang digunakan untuk menganalisis serta memodelkan hubungan antara satu atau lebih variabel independen (fitur) dengan sebuah variabel dependen (target) yang bersifat kontinu. Tujuan utama regresi adalah memprediksi nilai numerik dan memahami pola hubungan yang terbentuk dari data.

Dalam konteks *Machine Learning*, regresi termasuk ke dalam kategori supervised learning, yaitu metode pembelajaran yang menggunakan data latih yang telah dilengkapi dengan label atau nilai target. Model regresi belajar dari pasangan data (X, y), di mana X merepresentasikan fitur-fitur masukan dan y adalah nilai target yang ingin diprediksi.

Regresi merupakan metode pembelajaran statistik yang mendasar untuk memodelkan dan menginterpretasikan hubungan antara variabel prediktor dan variabel respons, serta menjadi fondasi bagi pengembangan model prediksi yang lebih kompleks dalam *supervised learning*. (Hastie, Tibshirani, dan Friedman 2024)

Berbeda dengan klasifikasi yang bertujuan menentukan kelas atau kategori tertentu, regresi berfokus pada estimasi nilai kontinu. Sebagai contoh, permasalahan seperti memprediksi harga rumah berdasarkan luas bangunan dan lokasi, memperkirakan penjualan berdasarkan data historis, atau mengestimasi konsumsi energi berdasarkan kondisi cuaca merupakan contoh penerapan regresi. Berikut tujuan analisis regresi :

1. Mengetahui apakah variabel X berpengaruh terhadap variabel Y
2. Mengukur seberapa besar pengaruh variabel X terhadap Y
3. Memprediksi nilai Y berdasarkan nilai X
4. Menjelaskan pola hubungan antar variabel

Analisis regresi memiliki beberapa jenis yang digunakan sesuai dengan karakteristik data dan tujuan penelitian. Perbedaan jenis regresi ini didasarkan pada jumlah variabel bebas, bentuk hubungan antar variabel,

serta jenis data yang dianalisis. Pemilihan jenis regresi yang tepat akan membantu menghasilkan model yang akurat dan dapat diinterpretasikan dengan baik.

B. REGRESI LINEAR

Di antara berbagai jenis regresi yang berkembang, regresi linear merupakan bentuk analisis regresi yang paling dasar, paling umum digunakan, dan menjadi fondasi bagi pengembangan metode regresi yang lebih kompleks. Regresi linear digunakan ketika hubungan antara variabel bebas dan variabel terikat diasumsikan berbentuk garis lurus (linear). Dalam konteks ini, perubahan pada variabel bebas diasumsikan akan menyebabkan perubahan yang sebanding pada variabel terikat. Regresi linear bertujuan untuk mengestimasi arah dan besarnya pengaruh variabel independen terhadap variabel dependen, serta digunakan secara luas dalam analisis prediksi dan pengujian hipotesis kuantitatif. (Ghozali, I. 2021)

Regresi linear secara umum dibedakan menjadi dua jenis utama, yaitu regresi linear sederhana dan regresi linear berganda. Regresi linear sederhana melibatkan satu variabel bebas dan satu variabel terikat, sedangkan regresi linear berganda melibatkan lebih dari satu variabel bebas yang secara bersama-sama memengaruhi satu variabel terikat. Meskipun berbeda dari segi jumlah variabel, kedua jenis regresi tersebut memiliki prinsip dasar yang sama, yaitu mencari garis terbaik yang dapat mewakili hubungan antara variabel-variabel yang dianalisis.

1. Regresi Linear Sederhana

Regresi linear sederhana adalah metode regresi yang digunakan untuk memodelkan hubungan linear antara satu variabel independen (X) dan satu variabel dependen (Y). Model ini mengasumsikan bahwa perubahan pada variabel X akan menyebabkan perubahan yang proporsional dan linier pada variabel Y.

Contoh kasus : Misalkan ingin mengetahui pengaruh jumlah jam belajar (X) terhadap nilai ujian (Y) mahasiswa.

Tabel 6. 1 Data contoh kasus regresi linear

Mahasiswa	Jam Belajar (X)	Nilai Ujian (Y)
1	2	65
2	4	70
3	6	75
4	8	85
5	10	90

Persamaan : y=a+bx

Keterangan:

a. x: variabel independen
b. y: variabel dependen
c. a: nilai $y$ketika $x = 0$, yang disebut sebagai intercept
d. b: koefisien regresi yang menunjukkan besarnya perubahan $y$akibat perubahan x

Tabel 6. 2 Perhitungan regresi linear sederhana

X	Y	X^2	XY
2	65	4	130
4	70	16	280
6	75	36	450
8	85	64	680
10	90	100	900
ΣX = 30	ΣY = 385	$\Sigma X^2 = 220$	ΣXY = 2440

Nilai koefisien regresi (b) dapat dihitung menggunakan rumus berikut:

$$b = \frac{N \sum xy - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2}$$

dengan keterangan:

a. N: jumlah data atau banyaknya sampel dalam dataset/tabel
b. $\sum xy$: jumlah hasil perkalian antara variabel $x$dan y

c. $\sum x$: jumlah seluruh nilai variabel independen
d. $\sum y$: jumlah seluruh nilai variabel dependen
e. $\sum x^2$: jumlah kuadrat dari variabel independen

$b = [5(2440) - (30)(385)] / [5(220) - (30)^2]$
$b = (12200 - 11550) / (1100 - 900)$
$b = 650 / 200 = 3{,}25$

Sedangkan nilai intercept (a) dapat dihitung menggunakan rumus berikut:

$$a = \bar{y} - b\bar{x}$$

dengan keterangan:
a. $\bar{y}$: nilai rata-rata dari variabel dependen
b. $\bar{x}$: nilai rata-rata dari variabel independen
c. b: koefisien regresi yang diperoleh dari Persamaan

$\bar{x} = 30 / 5 = 6$
$\bar{y} = 385 / 5 = 77$
$a = 77 - (3{,}25 \times 6)$
$a = 57{,}5$

Berdasarkan hasil perhitungan diperoleh persamaan regresi sebagai berikut:
$y = 57{,}5 + 3{,}25x$

Koefisien regresi bernilai positif, yang menunjukkan bahwa jumlah jam belajar berpengaruh positif terhadap nilai ujian mahasiswa. Setiap penambahan 1 jam belajar akan meningkatkan nilai ujian sebesar 3,25 poin. Suatu model dikategorikan sebagai regresi linear apabila memenuhi sejumlah persyaratan metodologis yang bertujuan untuk memastikan bahwa hasil estimasi yang diperoleh bersifat valid, Adapun persyaratan tersebut meliputi:

a. Jumlah sampel antara variabel dependent dan independent harus sama dan saling berpasangan, sehingga setiap nilai pada variabel independen memiliki satu nilai respon yang sesuai pada variabel dependen.
b. Hanya satu Variabel terikat Y, yang nilainya dipengaruhi oleh satu atau lebih variabel independen
c. Nilai residual, yaitu selisih antara nilai aktual dan nilai prediksi, terdistribusi secara normal. Asumsi ini penting untuk mendukung keabsahan pengujian

d. Model regresi linear yang baik harus terbebas dari pelanggaran asumsi klasik, yang meliputi:
   1) Multikolinearitas, yaitu adanya korelasi yang tinggi antarvariabel independen
   2) Autokorelasi, yaitu adanya korelasi antar residual pada pengamatan yang berbeda
   3) Heteroskedastisitas, yaitu kondisi varians residual yang tidak konstan pada seluruh tingkat variabel independen

2. Regresi Linear Berganda

Regresi linear berganda merupakan metode analisis statistik dan Machine Learning yang digunakan untuk memodelkan hubungan linear antara satu variabel dependen (Y) dengan dua atau lebih variabel independen (X_1, X_2, ..., X_n). Model ini dikembangkan untuk menggambarkan kondisi dunia nyata, di mana suatu variabel tidak hanya dipengaruhi oleh satu faktor, melainkan oleh beberapa faktor secara simultan. Dalam konteks Machine Learning, regresi linear berganda termasuk ke dalam metode supervised learning, karena proses pembelajaran model dilakukan menggunakan data latih yang telah memiliki nilai target.

Agar hasil analisis regresi linear berganda dapat dipercaya, beberapa asumsi dasar harus dipenuhi, yaitu:

a. Hubungan antara variabel bersifat linear
b. Nilai residual terdistribusi normal
c. Tidak terjadi multikolinearitas antar variabel independen
d. Tidak terdapat autokorelasi antar residual
e. Varians residual bersifat konstan (homoskedastisitas)

Contoh kasus : Mengetahui pengaruh **jam belajar (X_1)** dan **motivasi belajar (X_2)** terhadap **nilai ujian (Y)** mahasiswa. Bentuk persamaan regresi linear berganda:

$$Y = a + b_1X_1 + b_2X_2$$

Keterangan:

a. Y= Nilai ujian
b. X_1= Jam belajar
c. X_2= Motivasi belajar
d. a= Konstanta
e. b_1, b_2= Koefisien regresi

Misalkan diperoleh data dari *5* **mahasiswa** sebagai berikut:

Tabel 6. *3* Data contoh kasus regresi linear berganda

Mahasiswa	Jam Belajar (X_1)	Motivasi (X_2)	Nilai Ujian (Y)
1	2	3	68
2	4	4	78
3	6	4	84
4	8	5	94
5	10	5	100

Untuk memperoleh nilai konstanta dan koefisien regresi, diperlukan perhitungan tambahan seperti kuadrat dan hasil perkalian variabel.

Tabel 6.4 perhitungan regresi linear berganda

	X_1	X_2	Y	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
2	3	6	4	9	6	13	20	8
4	4	7	1	1	16	31	31	8
6	4	8	3	1	24	50	33	4
8	5	9	6	2	40	75	47	4
10	5	1	1	2	50	10	50	00
Σ			**2**	**9**	**13**	**27**	**18**	**20**
				1	**6**	**04**	**22**	

Jumlah nilai lainnya:

a. $\Sigma X_1 = 30$
b. $\Sigma X_2 = 21$
c. $\Sigma Y = 424$

Berdasarkan data di atas, diperoleh persamaan normal regresi linear berganda. Setelah dilakukan proses eliminasi dan substitusi, nilai konstanta dan koefisien regresi diperoleh sebagai berikut:

a. Konstanta $(a) = 50$

   Konstanta sebesar 50 menunjukkan bahwa apabila jam belajar dan motivasi dianggap tidak ada, maka nilai ujian mahasiswa diperkirakan sebesar 50.

b. Koefisien jam belajar $(b_1) = 3$

   Koefisien jam belajar sebesar 3 berarti setiap penambahan satu jam belajar akan meningkatkan nilai ujian mahasiswa rata-rata sebesar 3 poin, dengan asumsi motivasi belajar tetap.

c. Koefisien motivasi $(b_2) = 4$

   Koefisien motivasi sebesar 4 berarti setiap peningkatan satu tingkat motivasi belajar akan meningkatkan nilai ujian mahasiswa rata-rata sebesar 4 poin, dengan asumsi jam belajar tetap.

Dengan demikian, persamaan regresi linear berganda yang diperoleh adalah: Y=50+3X1+4X2

Berdasarkan hasil analisis regresi linear berganda, dapat disimpulkan bahwa jam belajar dan motivasi belajar sama-sama berpengaruh positif terhadap nilai ujian mahasiswa. Semakin lama mahasiswa belajar dan semakin tinggi motivasinya, maka nilai ujian yang diperoleh cenderung semakin tinggi.

3. Implementasi Regresi Linear Menggunakan Phyton

Implementasi Regresi Linear dalam machine learning menggunakan tool Python.

Step 1: Import library yang dibutuhkan

Library yang digunakan:

- **Pandas** untuk pengolahan data
- **NumPy** untuk operasi numerik
- **Matplotlib & Seaborn** untuk visualisasi
- **Scikit-learn** untuk machine learning (regresi linear)

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

Gambar 6. 1 Import library

Step 2: Menyiapkan dataset

```python
# Membuat dataset
data = {
    'JamBelajar': [2, 4, 6, 8],
    'NilaiUjian': [65, 70, 75, 85]
}

df = pd.DataFrame(data)
df
```

Gambar 6. 2 Menyiapkan dataset

Step 3: Analisis data

```python
df.describe()
```

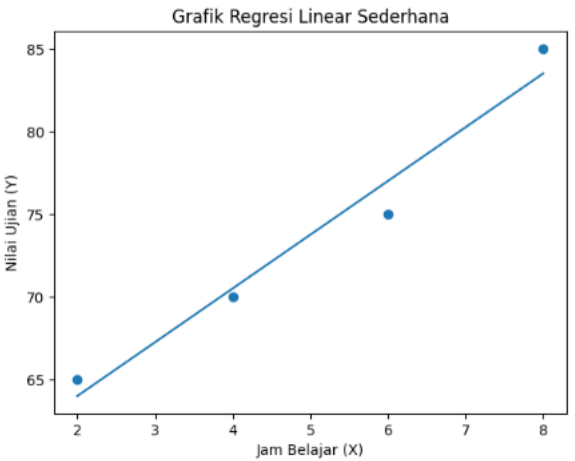
Gambar 6. 3 Hasil deskripftif

Dari hasil deskriptif dapat dilihat rentang jam belajar dan nilai ujian mahasiswa, serta nilai rata-rata masing-masing variabel.

```python
plt.title('Distribusi Nilai Ujian')
sns.histplot(df['NilaiUjian'], kde=True)
plt.show()
```

Gambar 6. 4 Distribusi nilai ujian

```python
plt.scatter(df['JamBelajar'], df['NilaiUjian'], color='lightcoral')
plt.title('Jam Belajar vs Nilai Ujian')
plt.xlabel('Jam Belajar')
plt.ylabel('Nilai Ujian')
plt.box(False)
plt.show()
```



Gambar 6. 5 Hubungan antara jam belajar dan nilai ujian

Grafik menunjukkan pola linear meningkat, sehingga regresi linear layak digunakan.

Step 4: Menentukan variabel independen dan dependen
X (independen): Jam Belajar
Y (dependen): Nilai Ujian

```python
X = df[['JamBelajar']]   # variabel independen
y = df['NilaiUjian']     # variabel dependen
```

Gambar 6. 6 Variabel Independen dan dependen

Step 5: Membagi data menjadi data latih dan data uji

Data dibagi:

a. 80% data latih
b. 20% data uji

```python
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=0
)
```

Gambar 6. 7 Membagi data

Step 6: Melatih model regresi linear

```python
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

```python
print("Intersep (a):", regressor.intercept_)
print("Koefisien regresi (b):", regressor.coef_[0])
```

Gambar 6. 8 Melatih model regresi linear

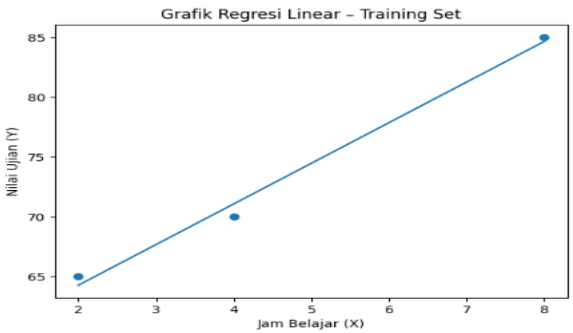
Step 7: Melakukan prediksi

```python
y_pred_test = regressor.predict(X_test)
y_pred_train = regressor.predict(X_train)
```

Gambar 6. 9 Prediksi

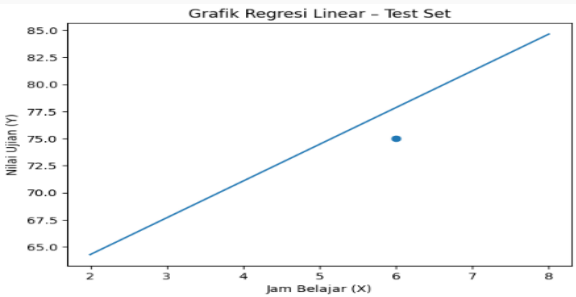
Step 8: Visualisasi hasil prediksi

```python
plt.scatter(X_train, y_train, color='lightcoral')
plt.plot(X_train, y_pred_train, color='firebrick')
plt.title('Jam Belajar vs Nilai Ujian (Training Set)')
plt.xlabel('Jam Belajar')
plt.ylabel('Nilai Ujian')
plt.legend(['Prediksi', 'Data Aktual'])
plt.box(False)
plt.show()
```



Gambar 6. 10 Grafik data latih

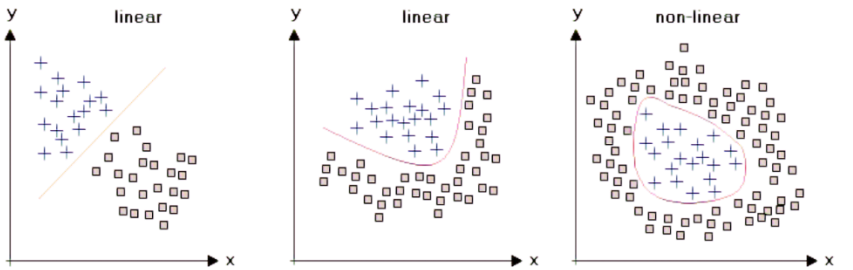
```python
plt.scatter(X_test, y_test, color='lightcoral')
plt.plot(X_train, y_pred_train, color='firebrick')
plt.title('Jam Belajar vs Nilai Ujian (Test Set)')
plt.xlabel('Jam Belajar')
plt.ylabel('Nilai Ujian')
plt.legend(['Prediksi', 'Data Aktual'])
plt.box(False)
plt.show()
```



Gambar 6. 11 Grafik data Uji

C. REGRESI NON-LINEAR

Regresi non-linear adalah metode analisis regresi yang digunakan ketika hubungan antara variabel bebas *X* dan variabel terikat *Y* tidak membentuk garis lurus. Artinya, perubahan *X* tidak menyebabkan perubahan *Y* secara konstan (tidak naik/turun dengan laju tetap), tetapi mengikuti pola kurva seperti melengkung, eksponensial, logaritmik, bertumbuh cepat di awal lalu melambat, atau pola lain yang tidak linear.



Gambar 6. 12 Perbedaan Linear dan Non-linear
Sumber: Website https://medium.com/

Namun, tidak semua fenomena dapat dijelaskan dengan hubungan linear. Pada kondisi tertentu, hubungan antar variabel bersifat tidak konstan dan menunjukkan pola pertumbuhan atau kejenuhan, sehingga diperlukan pendekatan regresi nonlinear (Supranto, J. 2019). Regresi linier lebih sederhana dan mudah diimplementasikan karena mengasumsikan hubungan linier langsung antara variabel. Regresi non-linier, di sisi lain, lebih kompleks, karena harus memperhitungkan kelengkungan atau pola non-linier lainnya dalam data.

Kosep dasar regresi non-linear sederhana maupun berganda sama-sama digunakan ketika hubungan antara variabel bebas dan variabel terikat tidak bersifat linear. Keduanya bertujuan memodelkan hubungan yang mengikuti kurva. Perbedaannya terletak pada jumlah variabel bebas (independent variable) yang digunakan dalam model.

1. Regresi Non-linear Sederhana

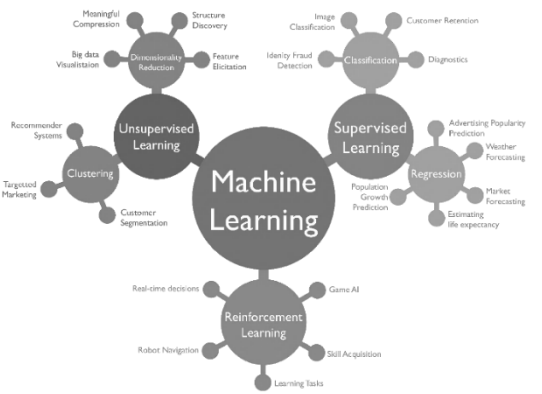
Regresi non-linear sederhana adalah regresi non-linear yang melibatkan satu variabel bebas (x) dan satu variabel terikat (y). Jenis regresi ini digunakan ketika suatu fenomena dipengaruhi oleh satu faktor utama yang hubungannya dengan variabel terikat bersifat tidak linear. Contoh model:

BAB 7 KLASIFIKASI DENGAN LOGISTIC REGRESSION DAN KNN

M. Rhifky Wayahdi

A. PENDAHULUAN

Dalam lanskap pembelajaran mesin (*machine learning*), algoritma klasifikasi memegang peranan vital dalam mengubah data menjadi keputusan cerdas. Dua metode yang paling fundamental dan sering menjadi rujukan utama dalam kategori supervised learning adalah *Logistic Regression* (LR) dan *k-Nearest Neighbors* (KNN). Literatur terkini menunjukkan bahwa LR dan KNN secara konsisten dipilih sebagai *baseline* (garis dasar) atau pembanding dalam berbagai domain aplikasi. Penerapannya meluas mulai dari penapisan dan diagnosis biomedis—seperti obesitas, diabetes, kanker payudara, aterosklerosis karotis, kesehatan tidur, dan halitosis—hingga klasifikasi non-medis yang mencakup pemodelan pendukung keputusan berbasis EEG, klasifikasi komposer musik, deteksi berita palsu (*fake news*), keamanan siber, penipuan kartu kredit, analitik pendidikan, hingga prediksi rekayasa perangkat lunak (Chang et al., 2025; Saveliev et al., 2025; Taşpınar & Çınar, 2023; Alfayez & Alazba, 2025; Wayahdi & Ruziq, 2025).



Gambar 7. 1 Taksonomi Machine Learning

Secara konseptual, LR dideskripsikan sebagai prosedur klasifikasi yang "andal dan terdefinisi dengan baik," yang dibingkai sebagai perluasan dari regresi untuk memodelkan kejadian atau ketidakjadian suatu peristiwa. Sementara itu, KNN dikenal sebagai metode nonparametrik yang "sederhana" secara konseptual, di mana label diberikan kepada instans yang tidak diketahui dengan memilih titik data terdekat (*neighbors*) berdasarkan metrik jarak tertentu. Kedua algoritma ini tidak hanya berdiri sendiri; studi sering mengevaluasi LR dan KNN secara berdampingan—kerap kali bersama algoritma tambahan seperti *Random Forest* (RF), *Support Vector Machine* (SVM), dan *XGBoost*—untuk mengukur *trade-off* dalam kinerja prediktif di bawah protokol evaluasi yang konsisten seperti validasi silang (*cross-validation*) dan validasi eksternal (Yun et al., 2023; Bhat et al., 2023).

Mempelajari bab ini memiliki urgensi yang tinggi bagi siapa pun yang mendalami ilmu data dan kecerdasan buatan. Meskipun algoritma kompleks terus bermunculan, LR dan KNN tetap menjadi fondasi yang tak tergantikan. Memahami kedua algoritma ini sangat penting karena mereka berfungsi sebagai standar emas untuk memvalidasi apakah model yang lebih rumit benar-benar diperlukan atau memberikan peningkatan kinerja yang signifikan. Tanpa pemahaman mendalam tentang bagaimana LR bekerja sebagai pengklasifikasi linier yang kuat dan bagaimana KNN beroperasi sebagai pengklasifikasi berbasis instans, seorang praktisi akan kesulitan membangun intuisi yang tepat dalam memilih model yang efisien berdasarkan batasan sumber daya dan karakteristik data yang dihadapi.

B. STUDI LITERATUR TERKAIT

Sebelum masuk ke pembahasan teknis implementasi, penting untuk merangkum temuan-temuan empiris dari berbagai literatur terkini mengenai karakteristik, kinerja, dan praktik terbaik penggunaan *Logistic Regression* dan *k-Nearest Neighbors*. Berikut adalah sintesis dari penelitian-penelitian tersebut:

Peran dan Karakteristik Model

Dalam dinamika penelitian pembelajaran mesin saat ini, *Logistic Regression* (LR) memegang peranan yang sangat sentral. Algoritma ini tidak sekadar dianggap sebagai metode dasar, melainkan sering ditempatkan sebagai pengklasifikasi utama dan tolok ukur (*benchmark*) standar untuk mengevaluasi efektivitas alternatif pembelajaran mesin lainnya yang lebih kompleks. Lebih jauh lagi, LR sering kali diadopsi

sebagai model dasar (*base learner*) yang andal di dalam struktur ansambel (*ensembles*) pada berbagai tugas klasifikasi (Deepaisarn et al., 2023). Fleksibilitas LR juga menjadi nilai tambah yang signifikan; penerapannya tidak terbatas pada tugas klasifikasi biner (dua kelas) semata, tetapi juga telah berhasil dioperasionalkan sebagai pengklasifikasi tujuan umum yang tangguh dalam konteks multi-kelas yang lebih rumit. Contoh nyata dari kapabilitas ini dapat dilihat pada klasifikasi status kesehatan yang membedakan kondisi "Tidak ada", "Sleep Apnea", dan "Insomnia", serta dalam domain seni untuk klasifikasi gaya komposer musik (Chang et al., 2025).

Di sisi lain spektrum, *k-Nearest Neighbors* (KNN) menawarkan pendekatan yang berbeda. KNN sering dikategorikan dalam literatur sebagai "metode pembelajaran mesin yang sangat baik" karena pendekatannya yang intuitif dan non-parametrik. Namun, di balik keunggulannya, KNN memiliki tantangan teknis tersendiri, terutama terkait pertimbangan kompleksitas ruang yang tinggi, yang sebanding dengan algoritma *Naive Bayes*. Aspek ini menjadi perhatian serius dalam skenario implementasi dunia nyata yang memiliki batasan sumber daya komputasi atau memori (Liu et al., 2022; Dam et al., 2022).

Pola Kinerja Empiris: LR vs KNN

Pertanyaan mendasar yang sering muncul adalah mengenai perbandingan performa kedua algoritma ini. Bukti empiris dari berbagai studi menunjukkan dominasi LR, di mana algoritma ini sering kali menunjukkan kinerja yang kuat, bahkan menjadi yang terbaik dalam set data tertentu dibandingkan kompetitornya.

a. Sebagai ilustrasi dalam bidang kesehatan masyarakat, khususnya pada prediksi obesitas, LR terbukti mencapai akurasi tertinggi sebesar 97%, angka yang secara signifikan mengungguli KNN dan algoritma lainnya dalam studi tersebut (Wong et al., 2022).
b. Keunggulan ini berlanjut pada diagnosis medis yang kritis. Pada prediksi kanker payudara berbasis citra medis, LR dilaporkan mampu mencapai akurasi 95%, melebihi kinerja KNN yang tertahan di angka 90% (Dinesh & Kalyanasundaram, 2022).
c. Dalam studi prediksi diabetes yang lebih luas, LR mencatatkan nilai diskriminasi yang impresif dengan *Area Under Curve* (AUC) sebesar 0,95. Hasil ini menunjukkan kinerja yang lebih baik daripada KNN, bahkan mengungguli model-model canggih seperti XGBoost dan SVM (Shu, 2024).

d. Demikian pula pada klasifikasi kesehatan tidur, LR mencatat tingkat keberhasilan klasifikasi sebesar 90,27%, sementara KNN harus puas dengan pencapaian 87,50% (Taşpınar & Çınar, 2023).

Meskipun demikian, KNN bukanlah tanpa taji. Algoritma ini tetap kompetitif dan bahkan unggul dalam kondisi spesifik. Misalnya, dalam tugas deteksi kelayakan konsumsi jamur (*edibility-detection*), KNN mampu mencapai akurasi sempurna 100% pada pengaturan parameter *k=1*. Walaupun hasil ini mengesankan, penulis studi memberikan peringatan keras mengenai risiko *overfitting* yang menyertai pengaturan parameter ekstrem tersebut (Gangu, 2022). Menariknya, pada kasus prediksi gaya hidup diabetes tahap awal, KNN justru menunjukkan superioritas dengan mencapai akurasi 89,6%, mengungguli LR yang mencatat 83,11% (Bhat et al., 2023). Temuan-temuan kontras ini menegaskan simpulan penting bahwa kinerja KNN sangat bergantung pada kalibrasi desain yang cermat, seperti penentuan ukuran lingkungan (*neighborhood size*) yang optimal (Beram & El-Kotory, 2024).

Alur Kerja Klasifikasi (*Workflow*)

Keberhasilan implementasi model pembelajaran mesin tidak hanya bergantung pada pemilihan algoritma, tetapi juga pada proses yang mendahuluinya. Implementasi yang sukses dari LR dan KNN mengikuti "template" alur kerja berbasis bukti yang melibatkan tahapan-tahapan kritis berikut:

a. **Pra-pemrosesan Data (*Data Preprocessing*):** Tahap ini merupakan fondasi yang tidak bisa ditawar. Baik ketika menangani data sinyal EEG yang kompleks, data operasional SCADA dari turbin angin, maupun data gaya hidup diabetes, pra-pemrosesan adalah prasyarat standar guna memastikan validitas dan kualitas model yang dihasilkan (Chang et al., 2025).
b. **Seleksi Fitur (*Feature Selection*):** Untuk meningkatkan efisiensi dan akurasi, pengurangan redundansi data sering diperkenalkan. Tujuannya adalah untuk meningkatkan daya diskriminasi model terhadap kelas-kelas yang ada. Contoh penerapan strategi ini termasuk penggunaan metode hibrida seperti *Genetic Algorithm-based KNN* (GA-KNN) atau mekanisme seleksi fitur khusus dalam mendeteksi penipuan kartu kredit, yang bertujuan mempertahankan hanya fitur relevan dan

meningkatkan kemampuan generalisasi model pada data baru (Yun et al., 2023; Mniai et al., 2023).

c. **Protokol Validasi (*Validation Protocol*):** Untuk menjamin objektivitas hasil, evaluasi model harus dilakukan dengan ketat. Standar emas dalam literatur mengandalkan teknik validasi silang (misalnya, *10-fold cross-validation*). Jika memungkinkan, validasi ini diperkuat dengan validasi eksternal (seperti menggunakan dataset multiregional pada studi prediksi operasi caesar) untuk menilai ketahanan (*robustness*) kinerja model di berbagai kondisi populasi (Chang et al., 2025).

Penggunaan Hibrida dan Ansambel

Perkembangan terkini dalam literatur pembelajaran mesin menunjukkan pergeseran paradigma yang signifikan, di mana *Logistic Regression* (LR) dan *k-Nearest Neighbors* (KNN) tidak lagi hanya dipandang sebagai alternatif yang saling menggantikan (*mutually exclusive*) atau sekadar kompetitor. Keduanya kini sering dilihat sebagai elemen komplementer yang dapat bekerja sama secara sinergis dalam strategi pemodelan ansambel untuk menutupi kelemahan masing-masing model tunggal. Salah satu manifestasi utama dari pendekatan ini adalah tren penggunaan *stacking classifier* yang cerdas, sebuah teknik di mana beberapa model dasar dilatih dan prediksi mereka digabungkan untuk menghasilkan keputusan akhir yang lebih akurat. Literatur menyoroti arsitektur yang memanfaatkan kekuatan gabungan dari algoritma seperti AdaBoost, KNN, dan LR, yang terbukti mampu meningkatkan stabilitas generalisasi model.

Selain arsitektur *stacking* standar, penerapan penggabungan heterogen dari berbagai jenis model juga semakin marak diterapkan untuk memecahkan masalah domain yang kompleks. Pendekatan ini menggabungkan algoritma dengan karakteristik pembelajaran yang berbeda untuk menangkap pola data yang lebih beragam dan mengurangi bias. Strategi penggabungan heterogen ini terbukti mampu meningkatkan kinerja prediksi secara signifikan dalam kasus-kasus spesifik yang menuntut presisi tinggi, seperti pada penyelesaian konflik penggabungan kode (*merge-conflict*) dalam rekayasa perangkat lunak dan klasifikasi diagnosis halitosis dalam bidang medis (Alfayez & Alazba, 2025; Saveliev et al., 2025).

BAB 8 DECESION TREE DAN RANDOM FOREST

Mifta Ardianti, S.T., M.Kom.

A. MENGAPA MODEL BERBASIS POHON/ TREE?

Secara alami, manusia cenderung mengambil keputusan dengan cara berpikir yang bercabang. Misalnya, seorang dosen dapat menilai kelulusan mahasiswa berdasarkan beberapa kondisi: jika nilai UAS ≥ 80 dan tingkat kehadiran ≥ 90%, maka mahasiswa tersebut lulus tanpa remedial; jika nilai UAS berada pada rentang 60–79 atau memiliki tugas yang sangat baik, mahasiswa dapat lulus dengan catatan atau remedial ringan; sedangkan jika nilai UAS rendah dan tingkat kehadiran buruk, maka hasilnya adalah tidak lulus. Pola berpikir seperti ini pada dasarnya mencerminkan struktur dari sebuah decision tree (Quinian, 1993), di mana setiap pertanyaan atau kondisi diwakili oleh cabang, dan setiap ujung cabang menghasilkan keputusan akhir atau kategori tertentu.

Model berbasis pohon layak mendapatkan perhatian khusus karena memiliki sejumlah keunggulan yang menjadikannya salah satu pendekatan paling intuitif dan praktis dalam machine learning (Loh, 2011). Pertama, model ini sangat intuitif dan mudah dijelaskan. Berbeda dengan jaringan saraf tiruan (*neural network*) yang sulit divisualisasikan, struktur pohon dapat digambarkan secara sederhana di papan tulis. Kita dapat dengan mudah menunjukkan bagaimana data mengalir melalui percabangan: "Jika IPK di bawah batas tertentu, masuk cabang kiri," atau "Jika tingkat kehadiran melebihi ambang batas, masuk cabang kanan." Visualisasi yang jelas seperti ini membuat pohon keputusan sangat efektif dalam konteks pembelajaran maupun komunikasi hasil analisis kepada pihak non-teknis.

Kedua, pohon keputusan mampu menerima fitur numerik maupun kategorikal tanpa memerlukan transformasi rumit (Hastie et al., 2009). Ia dapat memproses data seperti IPK, jumlah ketidakhadiran, atau umur (fitur numerik), sekaligus variabel seperti program studi, status beasiswa, atau jenis kelamin (fitur kategorikal). Kemampuan ini membuatnya sangat sesuai untuk data dunia nyata yang biasanya bersifat campuran.

Ketiga, model berbasis pohon mampu menangkap hubungan non-linear dan aturan kompleks. Banyak fenomena nyata tidak mengikuti hubungan linear sederhana (Maimon et al., 2015). Sebagai contoh, risiko mahasiswa untuk *dropout* mungkin tinggi bagi mereka dengan IPK rendah dan tingkat ketidakhadiran tinggi, namun tidak bagi mahasiswa dengan IPK sedang dan absensi baik. Pola interaksi seperti ini sulit ditangkap oleh model linear, tetapi dapat direpresentasikan secara alami oleh struktur bercabang dalam pohon keputusan.

Terakhir, pohon keputusan menjadi fondasi bagi metode ensemble yang sangat kuat, seperti Random Forest, Gradient Boosting Machine, XGBoost, dan LightGBM (Breiman, 2001). Semua algoritma tersebut dibangun dari kumpulan banyak pohon keputusan yang digabungkan dengan berbagai skema ensemble untuk menghasilkan model yang lebih stabil, akurat, dan tangguh terhadap overfitting. Namun demikian, penting disadari bahwa pohon keputusan juga memiliki keterbatasan. Model ini cenderung mudah mengalami overfitting, sensitif terhadap sedikit perubahan pada data, dan performanya bisa tidak stabil. Kelemahan-kelemahan inilah yang kemudian mendorong lahirnya model Random Forest sebagai solusi untuk meningkatkan kestabilan dan akurasi melalui pendekatan ensemble.

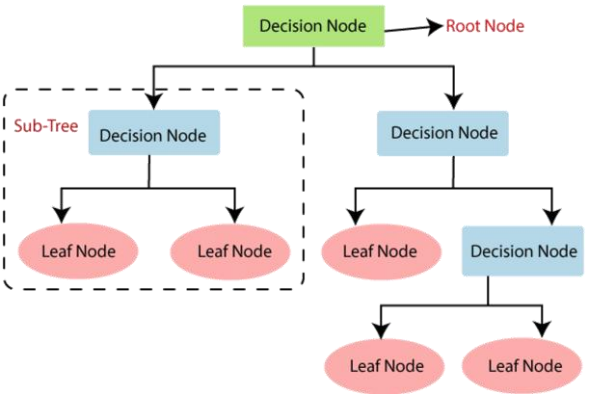
B. STRUKTUR DASAR DECISION TREE

Sebuah Decision Tree tersusun atas elemen-elemen utama berupa node dan cabang (edges) yang saling terhubung, membentuk struktur bercabang menyerupai pohon(Loh, 2011). Setiap node dalam pohon memiliki fungsi dan peran yang berbeda dalam proses pengambilan keputusan. Secara umum, struktur dasar pohon keputusan terdiri atas tiga komponen utama, yaitu root node (node akar), internal nodes (node internal), dan leaf nodes (node daun) (Han et al., 2012).

1. **Root Node (simpul akar**) merupakan node paling atas dalam pohon (tree) dan menjadi titik awal seluruh proses pembelajaran. Pada tahap ini, mewakili pertanyaan atau masalah yang ingin dipecahkan. Seluruh data training ditempatkan sebelum ada proses pemisahan atau split. Algoritma kemudian memilih satu fitur beserta kondisi yang dianggap paling 'informatif' untuk membagi data tersebut menjadi kelompok-kelompok yang homogen. Pemilihan fitur awal ini sangat penting, untuk menentukan arah dan efektivitas struktur pohon selanjutnya. Misalnya pada prediksi risiko dropout mahasiswa, algoritma mungkin menentukan bahwa

fitur 'jumlah ketidakhadiran' adalah variabel terbaik untuk dilakukan pemisahan/ spliting pertama.

2. **Internal Nodes (node internal),** adalah komponen kunci dari struktur decision tree. Setiap node internal memeriksa sebuah sebuah fitur dan memutuskan arah cabang berdasarkan hasil pengujian kondisi tersebut. Contohnya, sebuah node dapat menguji apakah $\text{IPK} \geq 30, jumlah\ absen > 3, atau\ status\ beasiswa =$ Ada. Hasil dari setiap pengujian ini direpresentasikan melalui cabang., biasanya dua cabang dalam kasus biner, misalnya cabang 'YA' dan 'TIDAK'. Proses ini berulang di setiap node internal, sehingga pohon tumbuh semakin kompleks seiring bertambahnya percabangan dan kedalaman.
3. **Leaf Nodes atau node daun** merupakan titik akhir dari phon yang tidak lagi mengalami pemisahan. Pada node ini, model memberikan prediksi akhir berdasarkan pola data yang mencapai titik tersebut. Pada kasus klasifikasi, leaf node akan menghasilkan label kategori, seperti 'Dropout' atau 'Tidak Dropout'. Sedangkan untuk kasus regresi, leafe node memberikan nilai numerik sebagai hasil prediksi, misalnya nilai IPK yang diperkirakan, omzet penjualan, atau jumlah tagihan pelanggan. Dengan demikian, setiap jalur dari akar (root) hingga daun (leaf) merepresentasikan satu aturan keputusan lengkap yang dapat dijelaskan dalam berntuk pernyataan IF-THEN.



Gambar 8. 1 Struktur Decision Tree

Sumber: Website https://www.almabetter.com/bytes/tutorials/data-science/decision-tree

Secara intuitif, proses membangun sebuah *Decision Tree* dapat dipandang sebagai serangkaian langkah pemecahan data yang berulang. Pertama, algoritma memilih fitur terbaik pada setiap tingkat untuk memisahkan data. Kedua, ia menentukan nilai ambang (*threshold*) bagi fitur numerik atau pemetaan kategori untuk fitur kategorikal. Ketiga, proses pemisahan ini terus diulang pada node-node yang baru terbentuk hingga tidak lagi menghasilkan peningkatan informasi yang signifikan, atau hingga batas kedalaman dan ukuran pohon tercapai. Apabila proses pemisahan dilakukan tanpa kendali, pohon dapat tumbuh terlalu besar dan kompleks, menyebabkan model menjadi *overfitting* terhadap data pelatihan. Untuk mencegah hal ini, konsep **impurity** dan **pruning** digunakan untuk mengukur tingkat homogenitas node serta memangkas cabang yang tidak memberikan manfaat berarti. Dengan cara inilah, *Decision Tree* dapat membangun struktur keputusan yang efisien, seimbang antara akurasi dan kesederhanaan.

C. IMPURITY: MENGUKUR KETIDAKHOMOGENAN DALAM NODE

Salah satu konsep penting dalam memahami bagaimana *Decision Tree* terbentuk adalah **impurity**, atau tingkat ketidakhomogenan dalam sebuah node. Sederhananya, impurity menggambarkan seberapa "campur" data di dalam node tersebut. Jika sebuah node berisi data dari berbagai kelas atau nilai target yang sangat beragam, maka node itu dikatakan **tidak murni** (impure). Sebaliknya, jika semua data dalam node memiliki label atau nilai yang sama, maka node tersebut dianggap **murni** (pure).

Bayangkan kita memiliki data mahasiswa dengan dua label: "Lulus" dan "Tidak Lulus." Jika di dalam satu node terdapat 50% mahasiswa yang lulus dan 50% yang tidak lulus, maka node ini sangat tidak murni karena data di dalamnya masih bercampur rata antara dua kategori. Namun, jika seluruh mahasiswa dalam node tersebut adalah "Lulus," maka node itu sepenuhnya murni, karena tidak ada ketidakpastian dalam hasilnya. Tujuan utama Decision Tree saat membentuk strukturnya adalah membagi data sedemikian rupa sehingga setiap pemisahan membuat node-node yang dihasilkan menjadi lebih homogen — artinya, impurity-nya semakin kecil. Untuk mengukur impurity ini, terdapat beberapa fungsi matematis yang digunakan. Dua yang paling populer adalah Entropy dan Gini Indeks (Hastie et al., 2009). Keduanya berfungsi untuk memberi nilai kuantitatif pada tingkat ketidakhomogenan di dalam node.

- **Entropy** berasal dari teori informasi, dan digunakan untuk mengukur seberapa "acak" atau tidak pastinya isi node (Breiman, 2001). Nilai entropi akan tinggi jika data di dalam node terdiri dari kelas yang seimbang (misalnya 50% "Lulus" dan 50% "Tidak Lulus"), dan bernilai nol jika semua data dalam node berasal dari satu kelas saja. Sehingga, algoritma Decision Tree akan berusaha memilih pembagian (split) yang menurunkan entropi sebesar mungkin, proses ini disebut Information Gain, yaitu peningkatan kemurnian setelah pemisahan dilakukan.
- **Gini Index**, di sisi lain, adalah ukuran impurity yang lebih sederhana dan lebih cepat dihitung secara komputasi. Gini bernilai nol ketika node sepenuhnya murni, dan mencapai nilai maksimum ketika data di dalam node terbagi rata antara kelas-kelas yang ada. Dalam banyak implementasi modern seperti *CART* (Classification and Regression Tree) pada pustaka *scikit-learn*, Gini Index digunakan sebagai ukuran default karena efisien dan hasilnya sering kali mirip dengan Entropy.

Pada kasus regresi**,** di mana target berupa nilai numerik (misalnya prediksi harga rumah atau nilai IPK), impurity tidak lagi diukur dengan entropi atau Gini, melainkan dengan ukuran penyebaran seperti varians atau Mean Squared Error (MSE)**.** Node yang baik adalah node yang mampu menurunkan nilai varians atau MSE secara signifikan setelah dilakukan pemisahan. Secara umum, impurity membantu *Decision Tree* "belajar" bagaimana cara terbaik memecah data. Setiap kali algoritma mencoba sebuah kondisi pemisahan, ia akan menghitung perubahan impurity sebelum dan sesudah pemisahan itu. Jika impurity berkurang banyak, artinya pemisahan tersebut efektif dalam membuat data menjadi lebih homogen, sehingga layak dipilih (Quinian, 1993). Dengan cara ini, *Decision Tree* secara bertahap membangun struktur keputusan yang tidak hanya intuitif, tetapi juga efisien dalam memisahkan data berdasarkan pola yang paling bermakna.

D. CARA KERJA DAN JENIS-JENIS ALGORITMA DECISION TREE

Setelah memahami konsep impurity dan bagaimana sebuah *Decision Tree* berusaha memecah data menjadi kelompok yang lebih homogen, langkah berikutnya adalah mengenal bagaimana proses tersebut dilakukan secara algoritmik**.** Dalam sejarah pengembangan *machine learning*, terdapat beberapa varian algoritma *Decision Tree* yang terkenal dan sering digunakan hingga saat ini. Meskipun berbeda dalam detail

BAB 9 SUPPORT VEKTOR MACHINE (SVM)

Joni Karman, M.Kom.

A. PENGERTIAN SUPPORT VEKTOR MACHINE (SVM)

Salah satu algoritma machine learning yang digunakan untuk menyelesaikan permasalahan klasifikasi dan regresi adalah Support Vector Machine (SVM). SVM bekerja dengan mencari hyperplane terbaik yang mampu memisahkan data ke dalam dua kelas atau lebih dengan margin pemisah terbesar. Semakin besar margin antar kelas, semakin baik kemampuan model dalam melakukan generalisasi terhadap data baru.

Pada tahun 1990, Vladimir Vapnik pertama kali memperkenalkan SVM sebagai bagian dari teori pembelajaran statistik (statistical learning theory). Hingga saat ini, SVM dikenal sebagai algoritma yang dapat bekerja secara efektif pada data berdimensi tinggi (high-dimensional data) dengan jumlah sampel yang relatif kecil.

Pada model klasifikasi, SVM menggunakan fungsi kernel untuk memetakan data ke ruang fitur berdimensi lebih tinggi, sehingga data yang bersifat nonlinier dapat dioptimalkan menggunakan sebuah hyperplane.

B. FUNGSI HYPERLPLANE SVM

Hyperplane merupakan konsep kunci dalam algoritma Support Vector Machine (SVM) yang berfungsi sebagai bidang pemisah antara dua himpunan data. Dalam konteks SVM, hyperplane adalah batas keputusan yang membedakan dua kelas dalam ruang fitur. Pemilihan hyperplane yang optimal sangat penting karena dapat mempengaruhi akurasi model klasifikasi.

Dalam ruang berdimensi n, hyperplane adalah subruang berdimensi n−1. Sebagai contoh:

a. Dalam ruang dua dimensi (2D), hyperplane berbentuk garis lurus.
b. Dalam ruang tiga dimensi (3D), hyperplane berbentuk bidang datar (plane).

Secara matematis, hyperplane dapat dinyatakan dengan persamaan:

$$w \cdot x + b = 0$$

Di mana:
w adalah vektor berat (weight vector) yang menentukan arah hyperplane.
x adalah vektor fitur dari data.
b adalah bias yang menggeser hyperplane dari titik asal.

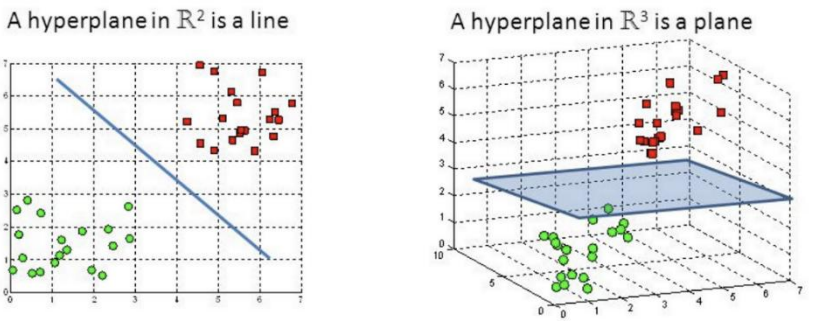
SVM meminimalkan ||w|| sambil mempertahankan margin yang lebih besar atau sama dengan 1:

$$y(i) * (w * x(i) + b) >= 1$$

Pada model klasifikasi, SVM menggunakan fungsi kernel untuk memetakan data ke ruang fitur berdimensi lebih tinggi, sehingga data yang bersifat nonlinier dapat dioptimalkan menggunakan sebuah hyperplane.

Untuk lebih memahami konsep ini, berikut adalah ilustrasi sederhana:

a. Gambar 1: Dalam ruang 2D, terdapat dua kelas data, misalnya titik merah dan biru. Hyperplane (garis) memisahkan kedua kelas ini.
b. Gambar 2: Dalam ruang 3D, hyperplane akan terlihat sebagai bidang datar yang memisahkan dua kelompok titik



Gambar 9. 1 Ilustrasi Hyperplane
Sumber: Website https://medium.com/

Untuk kasus nonlinier, SVM menggunakan pendekatan kernel untuk memetakan data ke ruang berdimensi lebih tinggi, sehingga data tersebut dapat dianalisis secara linier. Kernel yang umum digunakan adalah polynomial, Gaussian Radial Basis Function (RBF), dan sigmoid.

Kernel merupakan sebuah fungsi yang digunakan untuk mengubah data dari ruang berdimensi rendah ke ruang berdimensi lebih tinggi, sehingga pemisahan secara linier dapat dicapai. Beberapa jenis kernel yang umum digunakan adalah:

1. Kernel Linear
   Kernel linear adalah fungsi yang paling sederhana dan digunakan ketika data yang dianalisis sudah terpisah secara linear. Fungsi ini cocok untuk dataset dengan banyak fitur, di mana pemetaan ke ruang dimensi yang lebih tinggi tidak memberikan peningkatan kinerja yang signifikan.

   Persamaan:

   $$K(x, x_i) = x \cdot x_i$$

   Di mana x dan x_i adalah vektor fitur dari dua titik data.
   Kelebihan:kernel linear:
   a. Waktu komputasi cepat.
   b. Efektif untuk dataset besar dan berdimensi tinggi.
2. Kernel Polinomial
   Kernel polinomial digunakan untuk menganalisis data yang tidak dapat dianalisis secara linier dengan memanfaatkan fungsi polinomial dengan derajat tertentu. Derajat polinomial dapat disesuaikan untuk meningkatkan kemampuan pemisahan.
   Persamaan:

   $$K(x, x_i) = (x \cdot x_i + c)^d$$

   Di mana d merupakan derajat polinomial dan c adalah sebuah konstanta.
   Kelebihan:
   a. Dapat menangani interaksi non-linear antara fitur.
   b. Fleksibel dalam pengaturan derajat polinomial.
3. Kernel sigmoid
   Kernel sigmoid mirip dengan fungsi aktivasi sigmoid yang digunakan dalam jaringan saraf. Kernel ini dapat digunakan untuk klasifikasi non-linear tetapi kurang populer dibandingkan kernel lainnya.

Persamaan:

$$K(x, x_i) = \tanh(\alpha(x \cdot x_i) + c)$$

Di mana α dan c adalah parameter yang dapat disesuaikan.
Kelebihan:

a. Dapat digunakan dalam konteks neural networks.
b. Memungkinkan model untuk belajar dari interaksi non-linear antar fitur.

4. Kernel Gaussian RBF (radial Basis Function)
Salah satu kernel yang paling populer adalah Radial Basis Function (RBF) karena kemampuannya dalam menangani data nonlinier secara efektif. Kernel ini menggunakan fungsi eksponensial untuk memetakan data ke ruang berdimensi lebih tinggi.
Persamaan:

$$K(x, x_i) = \exp\left(-\frac{||x - x_i||^2}{2\sigma^2}\right)$$

Di mana σ adalah parameter yang mengontrol lebar dari fungsi Gaussian.
Kelebihan:

a. Sangat efektif untuk dataset dengan distribusi kompleks.
b. Dapat menangani kasus di mana tidak ada asumsi tentang distribusi data.

Kelebihan dan kekurangan SVM

Kelebihan	Kekurangan
Efektif dalam klasifikasi linear dan non-linear	Memerlukan normalisasi data
Fleksibel dengan berbagai pilihan kernel	Sering memerlukan tuning hyperparameter yang cermat
bisa menghasilkan model yang akurat	Kurang mudah diinterpretasi dibandingkan dengan model lain
Mengendalikan risiko overfitting dengan mengatur margin	Kurang efisien untuk dataset yang sangat besar

C. STUDI KASUS DUNIA NYATA PENGGUNAAN SVM

Studi kasus ini menggunakan dataset Pima Indians Diabetes. Tujuannya untuk memprediksi apakah seorang pasien berpotensi terkena diabetes berdasarkan fitur kesehatan. Dataset mencakup: Glucose, Blood

BAB 10 UNSUPERVISED LEARNING: Konsep Dasar

Muhammad Edya Rosadi, S.Kom, M.Kom.

A. PENDAHULUAN

Latar Belakang dan Motivasi

Dalam dunia *machine learning*, sebagian besar perhatian seringkali tertuju pada *supervised learning*, pendekatan yang menggunakan data berlabel untuk melatih model prediktif. Namun, realitas data di dunia nyata menunjukkan gambaran yang sangat berbeda: lebih dari 80-90% data di dunia tidak berlabel (Bishop, 2006; Murphy, 2012; Goodfellow et al., 2016). Situasi inilah yang menjadikan *unsupervised learning* sebagai pendekatan yang sangat penting dan relevan dalam era *big data* saat ini.

Unsupervised learning adalah cabang *machine learning* yang berfokus pada penemuan pola, struktur, atau hubungan tersembunyi dalam data tanpa menggunakan label atau target yang diketahui sebelumnya. Berbeda dengan *supervised learning* yang telah dibahas dalam bab sebelumnya, *unsupervised learning* tidak memiliki "guru" yang memberikan jawaban benar selama proses pembelajaran. Algoritma harus menemukan sendiri pola yang ada dalam data.

Pendekatan ini menjadi sangat penting karena: (1) biaya dan waktu untuk melabeli data seringkali sangat mahal; (2) label yang akurat mungkin tidak tersedia atau sulit diperoleh; (3) dapat menemukan pola yang tidak terduga yang memberikan wawasan berharga untuk pengambilan keputusan strategis. Contoh aplikasi nyata meliputi pengelompokan pelanggan e-commerce, deteksi anomali dalam transaksi keuangan, dan penemuan topik dalam dokumen tanpa label.

B. PENGERTIAN UNSUPERVISED LEARNING

Definisi Formal dan Informal

Secara formal, *unsupervised learning* adalah proses pembelajaran mesin di mana algoritma belajar dari data tanpa adanya label atau target variabel yang diketahui. Dalam notasi matematis, untuk dataset

$\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ dengan $x_i \in \mathbb{R}^d$, *unsupervised learning* bertujuan menemukan fungsi $f: \mathbb{R}^d \to \mathbb{R}^k$ yang memetakan data ke representasi bermakna atau mengungkapkan struktur tersembunyi. Berbeda dengan *supervised learning* yang menggunakan pasangan (x_i, y_i), *unsupervised learning* hanya bekerja dengan x_i tanpa informasi output.

Secara informal, *unsupervised learning* dapat diibaratkan seperti peneliti yang mengamati fenomena tanpa mengetahui pola sebelumnya, harus menemukan sendiri pola dan struktur dalam data. Algoritma bekerja dengan mengamati pola tanpa panduan label, mengelompokkan data (clustering), menemukan hubungan tersembunyi (association rules), dan mengidentifikasi struktur tidak terlihat (dimensionality reduction). Pendekatan ini adalah alat untuk eksplorasi data dan penemuan pengetahuan (*knowledge discovery*) yang memungkinkan memahami data tanpa pengetahuan sebelumnya tentang strukturnya (Tan, Steinbach, & Kumar, 2016; Aggarwal, 2015).

Karakteristik Utama

Unsupervised learning memiliki tiga karakteristik utama yang membedakannya dari pendekatan pembelajaran mesin lainnya:

- **Tidak Ada Label atau Target Variable**

  Karakteristik paling fundamental adalah tidak adanya label atau variabel target yang diketahui. Berbeda dengan *supervised learning* yang memiliki label untuk setiap contoh, *unsupervised learning* hanya bekerja dengan data input tanpa informasi output, sehingga algoritma harus bekerja tanpa panduan eksplisit tentang apa yang harus dicari.

- **Fokus pada Discovery (Penemuan Pola)**

  Tujuan utama adalah menemukan pola, struktur, atau hubungan tersembunyi dalam data, bukan memprediksi nilai tertentu. Proses ini sering disebut sebagai *exploratory data analysis* yang mengungkapkan informasi tidak terlihat secara langsung.

- **Evaluasi Subjektif dan Berguna untuk Eksplorasi**

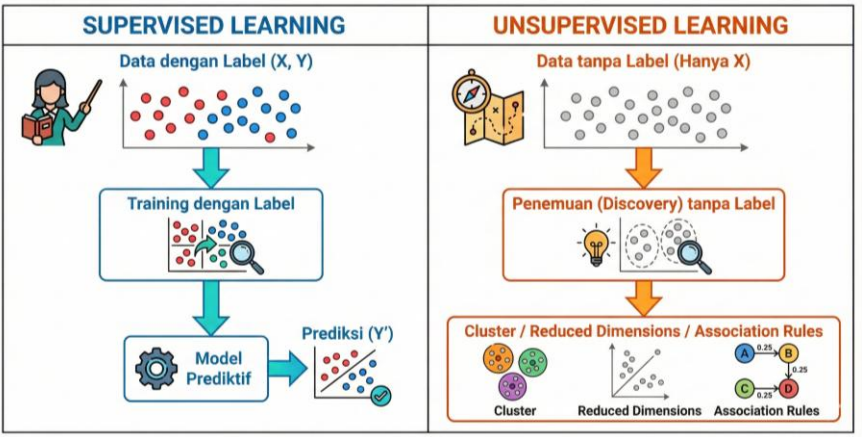
  Karena tidak ada label sebagai acuan, evaluasi menjadi lebih subjektif dan menantang, tidak ada metrik objektif seperti akurasi yang dapat langsung digunakan. Evaluasi bergantung pada interpretasi manusia, validasi domain expert, atau metrik internal. Pendekatan

ini sangat berguna dalam tahap awal analisis data untuk mendapatkan wawasan awal sebelum analisis lebih terarah.

Perbedaan Fundamental dengan Supervised Learning

Perbedaan utama terletak pada:

i. **Input data**: *supervised learning* memerlukan data berlabel (X, Y), sedangkan *unsupervised learning* hanya X tanpa label;
ii. **Tujuan**: *supervised learning* fokus pada prediksi/klasifikasi, sedangkan *unsupervised learning* fokus pada penemuan pola dan eksplorasi data;
iii. **Evaluasi**: *supervised learning* menggunakan metrik objektif (akurasi, MSE), sedangkan *unsupervised learning* memerlukan metrik internal atau validasi domain expert.



Gambar 10. 1 Diagram Konseptual Unsupervised Learning

C. PENGGUNAAN UNSUPERVISED LEARNING

Skenario Penggunaan

Unsupervised learning cocok digunakan dalam empat skenario utama:

- **Data Tidak Berlabel dan Sulit Dilabeli**: Ketika proses pelabelan memerlukan biaya tinggi, waktu lama, atau tidak memungkinkan. Contoh: mengelompokkan jutaan tweet berdasarkan topik tanpa label kategori emosi.
- **Eksplorasi Data Awal**: Untuk memahami karakteristik dasar data sebelum analisis lebih terarah. Contoh: mengidentifikasi segmen

pelanggan yang tidak terduga sebelum merancang strategi pemasaran (Provost & Fawcett, 2020).

- **Preprocessing dan Penemuan Pola**: Sebagai langkah preprocessing (misalnya, PCA untuk reduksi dimensi) atau untuk mengungkapkan pola tersembunyi yang tidak terpikirkan sebelumnya. Contoh: mengidentifikasi pola pembelian tidak biasa dalam transaksi keuangan.
- **Anomaly Detection**: Mendeteksi anomali ketika tidak memiliki contoh lengkap dari semua jenis anomali. Contoh: deteksi fraud dalam transaksi keuangan, intrusi jaringan, atau cacat produk.

Kelebihan dan Keterbatasan

- **Kelebihan**: (1) Tidak memerlukan data berlabel, hemat biaya dan waktu secara signifikan, sangat berharga di era *big data*; (2) Dapat menemukan pola tidak terduga yang tidak terpikirkan oleh manusia, menghasilkan wawasan baru dan inovatif; (3) Berguna untuk eksplorasi data awal ketika belum ada hipotesis jelas tentang struktur data; (4) Dapat digunakan untuk preprocessing (misalnya, dimensionality reduction) sebelum *supervised learning*.
- **Keterbatasan**: (1) Sulit dievaluasi karena tidak ada ground truth, evaluasi subjektif dan memerlukan validasi domain expert; (2) Hasil bisa subjektif dengan interpretasi yang bervariasi antar analis; (3) Membutuhkan domain knowledge untuk interpretasi yang bermakna; (4) Tidak ada jaminan menemukan pola bermakna, algoritma selalu menghasilkan output bahkan tanpa pola nyata dalam data.

Kapan Tidak Cocok Digunakan

Unsupervised learning tidak cocok digunakan dalam dua situasi utama:

- **Ketika label tersedia dan berkualitas baik**, menggunakan *supervised learning* akan memberikan hasil lebih baik dan lebih mudah dievaluasi, sementara *unsupervised learning* akan membuang informasi berharga dalam label;
- **Ketika tujuan jelas adalah prediksi/klasifikasi atau interpretasi sangat kritis**, *supervised learning* lebih tepat untuk prediksi, dan untuk aplikasi kritis (seperti diagnosis medis) dengan evaluasi jelas lebih aman daripada interpretasi subjektif *unsupervised learning*.

BAB 11 K-MEANS DAN HIERARCHICAL CLUSTERING

Ahmad Khusaeri, M.Kom.

A. PERSIAPAN DATA

Sebelum kita dapat menjalankan algoritma K-Means atau Hierarchical Clustering, langkah pertama dan terpenting adalah mempersiapkan data. Dalam dunia Data Science, sering dikatakan bahwa 80% pekerjaan adalah membersihkan dan menyiapkan data, sedangkan 20% sisanya adalah pemodelan (Anaconda, 2020). Khusus untuk clustering, persiapan data menjadi lebih vital karena metode ini sangat bergantung pada pengukuran jarak (seperti Jarak Euclidean). Variabel dengan skala yang besar (misalnya: Gaji dalam jutaan) akan mendominasi variabel dengan skala kecil (misalnya: Umur dalam puluhan), sehingga "membutakan" algoritma terhadap pola yang sebenarnya.

1. EKSPLORASI DATA AWAL (EDA)

Sebelum melakukan transformasi apa pun, kita perlu memahami karakteristik dataset melalui Exploratory Data Analysis (EDA).

a. Tipe Data: Pastikan data yang digunakan untuk K-Means adalah data numerik (Interval atau Rasio). Data kategorikal (seperti "Pria/Wanita" atau "Kota") harus dikonversi terlebih dahulu menjadi angka (encoding), namun perlu hati-hati karena K-Means menganggap angka memiliki urutan dan jarak.
b. Distribusi Data: Melihat apakah data terdistribusi normal atau miring (skewed).

2. PENANGANAN MISSING VALUES & OUTLIERS

a. Missing Values (Data Hilang)
Algoritma K-Means standar tidak dapat memproses data yang kosong (NaN/Null). Jika ada perhitungan jarak yang melibatkan nilai kosong, hasilnya akan error.
   i. Solusi 1 (Penghapusan): Jika jumlah data kosong sedikit (<5%), baris data tersebut bisa dihapus (Handayani, 2022).

ii. Solusi 2 (Imputasi): Mengisi kekosongan dengan nilai Rata-rata (Mean) atau Nilai Tengah (Median) dari kolom tersebut.

b. Outliers (Pencilan)
K-Means sangat sensitif terhadap outliers karena algoritma ini menggunakan Mean (Rata-rata) untuk menentukan pusat klaster (Centroid) (Salman et al., 2025). Satu nilai ekstrem bisa menarik posisi centroid menjauh dari kelompok utamanya.
   i. Deteksi: Gunakan Boxplot atau Z-Score.
   ii. Penanganan: Hapus outlier jika itu adalah kesalahan input, atau gunakan transformasi data untuk meredam efeknya.

3. TRANSFORMASI DATA (FEATURE SCALING)

Ini adalah langkah wajib dalam clustering. Bayangkan kita ingin mengelompokkan pelanggan berdasarkan dua fitur:

- Usia: Rentang 18 - 60 tahun.
- Gaji: Rentang 3.000.000 - 20.000.000 rupiah.

Tanpa penskalaan, perbedaan jarak pada fitur "Gaji" (jutaan) akan menenggelamkan perbedaan pada fitur "Usia". Algoritma akan menganggap Usia tidak penting. Oleh karena itu, kita harus menyamakan skalanya. Dua metode paling umum adalah (Allorerung et al., 2024):

a. Min-Max Normalization

Mengubah data sehingga semua nilai berada dalam rentang 0 hingga 1.

$$Xnew = \frac{X - Xmin}{Xmax - Xmin}$$

Kapan digunakan: Jika kita ingin rentang yang pasti dan data tidak memiliki outlier ekstrem.

b. Z-Score Standardization

Mengubah data sehingga memiliki Rata-rata (μ) = 0 dan Standar Deviasi (σ) = 1.

$$Xnew = \frac{X - \mu}{\sigma}$$

Kapan digunakan: Ini adalah metode yang lebih disarankan untuk K-Means, karena lebih tahan terhadap outlier dibandingkan Min-Max.

4. KONSEP PENGUKURAN JARAK (DISTANCE METRICS)

Bagaimana kita menentukan bahwa Data A "mirip" dengan Data B? Kita mengukurnya dengan jarak.

a. Euclidean Distance (Jarak Garis Lurus)
Ini adalah metode default pada K-Means. Mengukur jarak garis lurus terpendek antara dua titik dalam ruang Euclidean.Rumus untuk dua titik p dan q dalam ruang n-dimensi (Kusuma & Oktavianto, 2022):

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (qi - pi)^2}$$

b. Manhattan Distance (Taxicab Distance)
Mengukur jarak berdasarkan pergerakan tegak lurus (seperti taksi yang melewati blok-blok gedung di kota). Jarak ini adalah jumlah selisih absolut dari koordinat.

$$d(p,q) = \sum_{i=1}^{n} |qi - pi|$$

Manhattan seringkali lebih baik digunakan jika data memiliki dimensi yang sangat tinggi (banyak kolom/variabel).

B. K-MEANS CLUSTERING

1. LOGIKA & ALGORITMA K-MEANS

Setelah data dibersihkan dan dinormalisasi, K-Means bekerja dengan filosofi sederhana: "Benda-benda yang mirip akan berkumpul di sekitar pusat yang sama". Secara matematis, K-Means bertujuan meminimalkan Objective Function yang disebut SSE (Sum of Squared Errors) atau inersia:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} || Xi^{(j)} - Cj ||^2$$

Dimana:

- k = Jumlah klaster.

- xi = Titik data.
- cj = Pusat klaster (Centroid) ke-j.
- $||...||^2$ = Jarak Euclidean kuadrat.

Langkah-langkah Algoritma (The Algorithm)(Saputra & Nataliani, 2021):

a. Inisialisasi: Tentukan jumlah K. Pilih K titik acak sebagai centroid awal.
b. Assignment (Pengelompokan): Untuk setiap data, hitung jaraknya ke semua centroid. Masukkan data tersebut ke klaster milik centroid terdekat.
c. Update (Pembaruan Pusat): Hitung ulang posisi centroid. Posisi baru adalah rata-rata (mean) dari koordinat semua anggota di klaster tersebut.

$$Cj = \frac{1}{|Sj|} \sum_{Xi \epsilon Sj} Xi$$

d. Iterasi: Ulangi langkah 2 dan 3 hingga posisi centroid tidak berubah (konvergen).

2. MASALAH INISIALISASI & SOLUSI (K-MEANS++)

Salah satu kelemahan fatal K-Means standar adalah Inisialisasi Acak. Jika kita kurang beruntung memilih titik awal yang berdekatan, hasil klaster bisa menjadi buruk dan terjebak di Local Optima. Di dunia industri (seperti di Scikit-Learn Python), kita jarang menggunakan random murni. Kita menggunakan metode K-Means++. Memilih centroid pertama secara acak, tetapi centroid berikutnya dipilih berdasarkan probabilitas jarak terjauh dari centroid yang sudah ada. Ini memastikan penyebaran awal yang lebih baik.

3. STUDI KASUS: PERHITUNGAN MANUAL

Mari kita simulasikan cara kerja algoritma dengan data sederhana 2 Dimensi.

Dataset (4 Data Pelanggan):

- A (1, 1)
- B (2, 1)

- C (4, 3)
- D (5, 4)

Tujuan: Kelompokkan menjadi **K=2** klaster.

Iterasi 1:

1. **Inisialisasi:** Misalkan kita pilih secara acak **A (1,1)** sebagai Centroid 1 (c1) dan **B (2,1)** sebagai Centroid 2 (c2).
    - c1 = (1, 1)
    - c2 = (2, 1)
2. **Hitung Jarak (Euclidean) setiap data ke c1 dan c2:**
    - **Data A (1,1):** Jarak ke c1=0, Jarak ke c2=1. → Masuk **Klaster 1**.
    - **Data B (2,1):** Jarak ke c1=1, Jarak ke c2=0. → Masuk **Klaster 2**.
    - **Data C (4,3):**
        - Ke c1: $\sqrt{(4-1)^2 + (3-1)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.6$
        - Ke c2: $\sqrt{(4-2)^2 + (3-1)^2} = \sqrt{4+4} = \sqrt{8} \approx 2.8$
        - Karena $2.8 < 3.6$, C masuk **Klaster 2**.
    - **Data D (5,4):**
        - Ke c1: $\sqrt{(5-1)^2 + (4-1)^2} = \sqrt{16+9} = \sqrt{25} \approx 5.0$
        - Ke c2: $\sqrt{(5-2)^2 + (4-1)^2} = \sqrt{9+9} = \sqrt{18} \approx 4.2$
        - Karena $4.2 < 5.0$, D masuk **Klaster 2**.
3. **Hasil Kelompok Iterasi 1:**
    - Klaster 1: {A}
    - Klaster 2: {B, C, D}
4. **Update Centroid Baru:**
    - **Baru c1:** Rata-rata dari {A} → Tetap **(1, 1)**.
    - **Baru c2:** Rata-rata dari {B, C, D}:
        - $x_{baru} = (2+4+5) / 3 = 11/3 \approx 3.67$

- $y_{baru} = (1+3+4) / 3 = 8/3 \approx 2.67$
- Baru c2 = **(3.67, 2.67)**.

Centroid bergeser! Maka kita lanjut ke Iterasi 2.

Iterasi 2:

Kita hitung ulang jarak semua titik ke centroid baru (c1 di (1,1) dan c2 di (3.67, 2.67)).

- Jika dihitung, **Data B (2,1)** sekarang akan lebih dekat ke **c1 (1,1)** daripada ke c2 (3.67, 2.67).
    - Jarak B ke c1: 1.
    - Jarak B ke c2: $\sqrt{(2-3.67)^2 + (1-2.67)^2} \approx 2.3$
- Maka B pindah ke Klaster 1.

Hasil Kelompok Iterasi 2:

- Klaster 1: {A, B}
- Klaster 2: {C, D}

Posisi Centroid akan diupdate lagi. Iterasi berlanjut hingga tidak ada data yang berpindah klaster. Secara visual, Klaster 1 adalah kelompok "Kiri Bawah" dan Klaster 2 adalah "Kanan Atas".

4. MENENTUKAN JUMLAH K OPTIMAL

a. Elbow Method (Metode Siku)

Metode ini membandingkan nilai SSE (Sum of Squared Error) untuk berbagai nilai K.

- Jika K bertambah, SSE pasti turun (karena klaster makin kecil dan rapat).
- Kita cari titik di mana penurunan SSE mulai melambat drastis (membentuk siku). Itulah K optimal.

b. Silhouette Score

Mengukur seberapa "tepat" sebuah objek berada di klasternya.

- Rumus: $S = \frac{b - a}{max\,(a,b)}$
    - a: Jarak rata-rata ke teman se-klaster (Intra-cluster).
    - b: Jarak rata-rata ke klaster tetangga terdekat (Inter-cluster).
- **Interpretasi:**
    - Mendekati +1: Sangat Bagus (Terpisah jauh).
    - 0: Berhimpitan (*Overlapping*).
    - Negatif: Salah kamar (Salah masuk klaster).

C. HIERARCHICAL CLUSTERING

1. PENDEKATAN: AGGLOMERATIVE VS DIVISIVE

Berbeda dengan K-Means yang langsung mempartisi data ("memotong kue"), Hierarchical Clustering membangun struktur hubungan antar data secara bertahap.

Ada dua pendekatan utama:

a. Agglomerative (Bottom-Up)

Ini adalah metode yang paling umum digunakan.

- **Filosofi:** "Bersatu kita teguh."
- **Langkah Awal:** Setiap titik data dianggap sebagai satu klaster individu. Jika kita memiliki 100 data, maka kita mulai dengan 100 klaster.
- **Proses:** Pada setiap langkah, dua klaster yang paling mirip (jarak terdekat) digabungkan (*merge*) menjadi satu klaster baru.
- **Akhir:** Proses berlanjut hingga semua data menyatu menjadi satu klaster raksasa.

b. Divisive (Top-Down)

Metode ini jarang digunakan dalam praktik karena kompleksitas komputasinya sangat tinggi.

- **Filosofi:** "Pecah belah dan kuasai."
- **Langkah Awal:** Semua data dianggap berada dalam satu klaster besar.
- **Proses:** Klaster besar dipecah secara bertahap menjadi klaster-klaster yang lebih kecil.

2. VISUALISASI UTAMA: DENDROGRAM

Hasil dari Hierarchical Clustering bukanlah sekadar label kelompok, melainkan sebuah diagram pohon yang disebut **Dendrogram**.

Cara Membaca Dendrogram:

1. **Sumbu X (Horizontal):** Mewakili titik-titik data (sampel).
2. **Sumbu Y (Vertikal):** Mewakili **jarak** (dissimilarity) di mana penggabungan klaster terjadi.
3. **Garis Vertikal:** Semakin tinggi garis vertikal sebelum bertemu garis horizontal (cabang), semakin besar perbedaan (jarak) antara kedua klaster yang digabungkan tersebut.

3. LINKAGE METHODS (METODE PENGGABUNGAN)

Kunci dari algoritma Agglomerative adalah pertanyaan: *"Bagaimana kita mengukur jarak antara Klaster A (yang berisi 5 titik) dengan Klaster B (yang berisi 3 titik)?"*. Karena klaster memiliki banyak titik, kita butuh aturan main yang disebut **Linkage**:

a. Single Linkage (Nearest Neighbor)

- **Aturan:** Jarak antar klaster ditentukan oleh jarak **terdekat** antara anggota klaster A dan anggota klaster B.
- **Karakteristik:**
    - Bisa mendeteksi klaster dengan bentuk tidak beraturan (non-spherical).

BAB 12 NEURAL NETWORK DASAR

Novi Lestari, M.Kom.

A. PENGERTIAN NEURAL NETWORK

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence/* AI) dalam beberapa dekade terakhir telah membawa perubahan besar dalam berbagai bidang kehidupan manusia. Mulai dari sistem rekomendasi pada layanan digital, deteksi penyakit pada dunia kesehatan, otomatisasi industri, hingga analisis data pada pertanian modern, seluruhnya tidak terlepas dari peran penting Machine Learning sebagai fondasi utama. Di antara berbagai pendekatan di dalam Machine Learning, *Neural Network* atau jaringan saraf tiruan menjadi salah satu teknologi yang paling berpengaruh dan revolusioner.

Neural Network (jaringan saraf tiruan) adalah model komputasi yang terinspirasi dari cara kerja otak manusia dalam memproses informasi. Model ini terdiri atas kumpulan unit pemroses sederhana yang disebut neuron. Neural Network merupakan pemroses paralel terdistribusi yang tersusun dari banyak unit sederhana. Setiap unit mampu melakukan perhitungan sederhana, namun ketika digabungkan, jaringan dapat mempelajari pola kompleks.

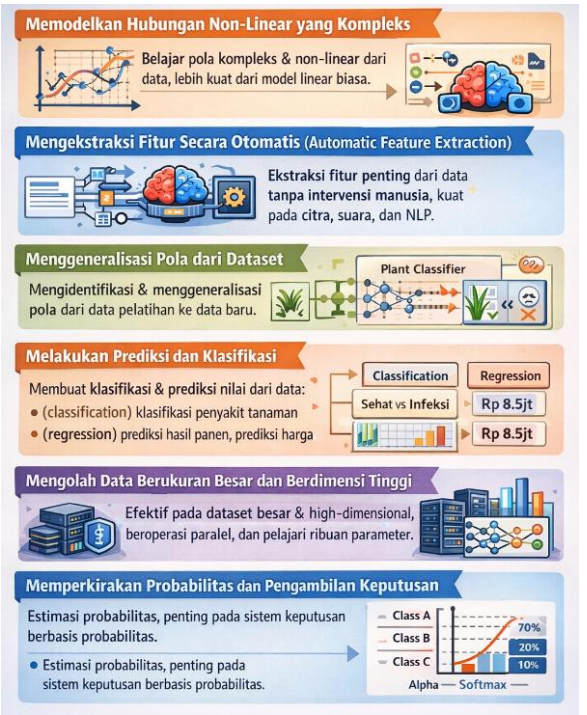
Neural Network mampu meniru sebagian kecil cara kerja otak manusia dalam mengolah informasi. Ia dapat belajar dari data, mengenali pola kompleks, serta menghasilkan prediksi yang sebelumnya mustahil dicapai oleh model tradisional. Para ahli seperti Goodfellow, Bengio, Courville, Bishop, dan Chollet telah memperkenalkan berbagai konsep fundamental yang memperkokoh posisi Neural Network sebagai pendekatan utama dalam pengembangan algoritma cerdas saat ini.

B. FUNGSI NEURAL NETWORK

Di era perkembangan teknologi yang semakin pesat, kebutuhan akan sistem yang mampu memahami, mempelajari, dan mengenali pola dari data menjadi semakin krusial. Mulai dari perangkat pintar di rumah, aplikasi digital yang kita gunakan setiap hari, hingga sistem analitik

dalam berbagai sektor industri, semuanya membutuhkan metode komputasi yang dapat bekerja secara adaptif dan otomatis. Dalam konteks inilah *Neural Network* atau jaringan saraf tiruan memainkan peran fundamental.

Sebagai salah satu pilar utama dalam bidang *Machine Learning* dan *Artificial Intelligence*, Neural Network dirancang untuk meniru cara kerja otak manusia dalam memproses informasi. Model ini bukan hanya berfungsi sebagai alat komputasi biasa, tetapi sebagai sistem cerdas yang mampu mempelajari hubungan non-linear yang kompleks, mengenali pola tersembunyi, mengekstraksi fitur secara otomatis, serta melakukan prediksi atau klasifikasi dengan tingkat akurasi yang tinggi. Para ahli seperti Goodfellow, Bengio, Courville, Bishop, dan Géron menegaskan bahwa kekuatan utama Neural Network terletak pada kemampuannya melakukan *representation learning*, yaitu belajar langsung dari data tanpa memerlukan perancangan fitur secara manual.



Gambar 12. 1 Fungsi Neural Network

Berikut Fungsi Neural Network Dasar :

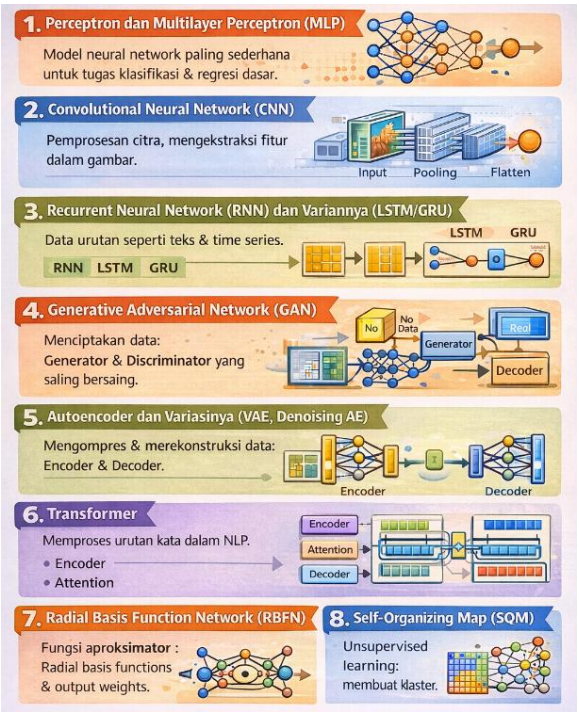
1. Memodelkan Hubungan Non-Linear yang Kompleks. Menurut Goodfellow dkk., fungsi utama neural network adalah *universal function approximator*, yaitu mampu mempelajari hubungan yang sangat kompleks dan non-linear dari data. Artinya, NN dapat mempelajari pola yang tidak bisa ditangani model linear biasa.
2. Mengekstraksi Fitur Secara Otomatis (Automatic Feature Extraction). Chollet menjelaskan bahwa jaringan saraf modern memiliki kemampuan *representation learning*, yakni mengekstraksi fitur penting dari data tanpa intervensi manusia. Inilah alasan neural network sangat kuat di citra, suara, dan NLP.
3. Menggeneralisasi Pola dari Dataset. Bishop menyebut bahwa neural network berguna untuk *pattern recognition*, yaitu mengidentifikasi pola dari data pelatihan dan menggeneralisasikannya pada data baru. Contoh: membedakan daun padi sehat vs terinfeksi.
4. Melakukan Prediksi dan Klasifikasi. Andrew Ng menekankan bahwa fungsi utama neural network adalah *supervised learning*:
   - Klasifikasi (classification)
   - Prediksi nilai (regression)

   Contoh:
   - Klasifikasi penyakit tanaman
   - Prediksi hasil panen
   - Prediksi harga
5. Mengolah Data Berukuran Besar dan Berdimensi Tinggi. Géron menjelaskan bahwa neural network sangat efektif untuk dataset besar karena:
   - Mampu beroperasi secara parallel
   - Dapat mempelajari ribuan hingga jutaan parameter
   - Cocok untuk big data dan high-dimensional data
6. Memperkirakan Probabilitas dan Pengambilan Keputusan. Neural Network digunakan untuk estimasi probabilitas, misalnya output *softmax* untuk klasifikasi. Ini penting pada sistem keputusan berbasis probabilitas (decision-making systems).

C. JENIS-JENIS NEURAL NETWORK

Dalam perkembangan kecerdasan buatan, neural network telah berevolusi menjadi berbagai arsitektur yang masing-masing dirancang untuk menyelesaikan tipe permasalahan tertentu. Para pakar seperti **I**an

Goodfellow, Yoshua Bengio, Yann LeCun, Christopher Bishop, hingga François Chollet, memberikan kontribusi besar dalam menjelaskan dan mengembangkan variasi neural network tersebut. Dari berbagai jenis neural network yang dikembangkan para ahli, bahwa setiap jenisnya memiliki fungsi dan kekhususan tertentu dalam menyelesaikan masalah. Mulai dari pengenalan gambar, analisis data berurutan, generasi data sintetik, hingga pemetaan data kompleks, neural network terus berkembang dan menjadi pilar utama dalam teknologi kecerdasan buatan modern. Keanekaragaman arsitektur ini menjadi landasan penting bagi para peneliti, pendidik, dan praktisi untuk memilih model yang tepat sesuai kebutuhan.



Gambar 12. 2 Jenis-Jenis Neural Network

Berikut jenis-jenis neural network :

1. Perceptron dan Multilayer Perceptron (MLP). Model Neural Network paling dasar adalah Perceptron, yang diperkenalkan oleh Frank Rosenblatt. Perceptron berfungsi sebagai unit dasar pengambil keputusan berbasis linear. Seiring kebutuhan terhadap pemrosesan pola semakin kompleks, berkembanglah Multilayer Perceptron (MLP)

yang terdiri dari beberapa lapisan tersembunyi. Menurut Goodfellow dan Bishop, MLP mampu mempelajari hubungan non-linear dan digunakan secara luas untuk tugas klasifikasi maupun regresi.

2. Convolutional Neural Network (CNN). CNN adalah jaringan yang dirancang untuk memproses data spasial seperti gambar. LeCun memperkenalkan konsep ini melalui studi pengenalan tulisan tangan (MNIST). CNN bekerja dengan melakukan ekstraksi fitur otomatis melalui operasi konvolusi. Buku *Deep Learning* menyatakan bahwa CNN sangat efektif menangani masalah penglihatan komputer—misalnya deteksi objek, pengenalan wajah, dan segmentasi gambar.
3. Recurrent Neural Network (RNN) dan Variannya (LSTM/GRU). RNN dirancang untuk memproses data berurutan seperti teks, suara, atau sinyal waktu. Namun, RNN klasik sering mengalami masalah *vanishing gradient*. Untuk mengatasinya, dikembangkan LSTM dan GRU, yang lebih stabil dalam mempelajari dependensi jangka panjang. Menurut Hochreiter & Schmidhuber, LSTM sangat efektif digunakan dalam NLP, speech recognition, dan machine translation.
4. Generative Adversarial Network (GAN). GAN merupakan jaringan yang bekerja dengan dua model: generator dan discriminator. Goodfellow sebagai pencetus GAN menjelaskan bahwa metode ini mampu menghasilkan data baru yang realistis, misalnya gambar wajah sintetis, data augmentasi, atau peningkatan resolusi citra. GAN menjadi salah satu terobosan terbesar dalam dunia deep learning karena kemampuannya menciptakan data baru mirip dengan data asli.
5. Autoencoder dan Variasinya (VAE, Denoising AE). Autoencoder merupakan neural network yang belajar merepresentasikan data ke dalam bentuk kompresi. Hinton menjelaskan bahwa autoencoder sangat penting untuk *feature learning*. Sementara itu, Variational Autoencoder (VAE) yang diperkenalkan Kingma & Welling memungkinkan generasi data baru (mirip GAN) namun berbasis pendekatan probabilistik.
6. Transformer. Transformer adalah arsitektur yang merevolusi NLP. Vaswani memperkenalkan konsep *attention mechanism* yang memungkinkan model fokus pada bagian penting dalam urutan data. Transformer menjadi pondasi model besar seperti BERT, GPT, dan T5. Chollet juga menekankan peran transformasi atensi dalam meningkatkan efisiensi pembelajaran sekuens.
7. Radial Basis Function Network (RBFN). RBF Network menggunakan fungsi basis radial untuk memetakan data ke ruang fitur. Bishop menjelaskan bahwa RBFN sangat efektif untuk klasifikasi non-linear

BAB 13 DEEP LEARNING: ARSITEKTUR DAN APLIKASI

Budi Berlinton Sitorus S.T, M.Sc

A. PENGERTIAN DEEP LEARNING

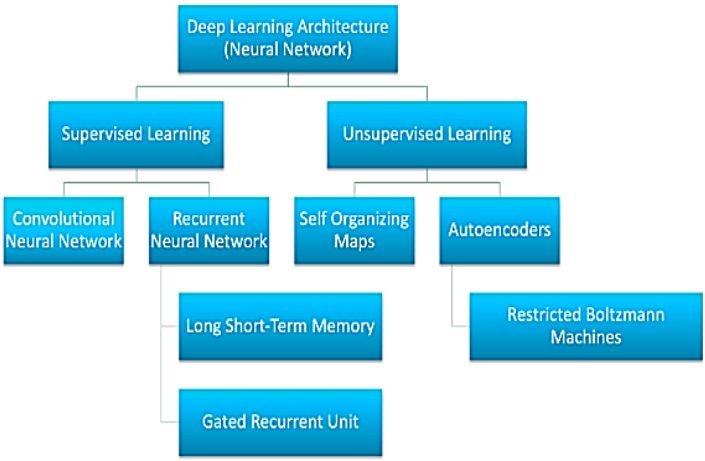
Istilah deep learning tidak terlepas dari machine learning. Sebelum membahas deep learning, ada baiknya pemahaman tentang machine learning didalami. Menurut Alpaydin, dalam buku Introduction to machine learning (ML), ML adalah melakukan pemrograman terhadap komputer untuk mengoptimalkan sebuah kriteria kinerja dengan menggunakan data contoh atau pengalaman sebelumnya. Terdapat model yang akan mendefinisikan parameter-parameter, dan pembelajaran adalah aksi untuk menjalankan program komputer dengan tujuan mengoptimasi parameter-parameter dari model dengan menggunakan data pelatihan atau pengalaman sebelumnya. Modelnya dapat bersifat prediktif untuk membuat prediksi atau dapat juga bersifat deskriptif untuk memperoleh pengetahuan dari dari atau dapat bersifat keduanya.

Deep learning (DL) merupakan salah satu tipe dari ML. Menurut salah satu situs dari vendor komputer terbesar IBM, DL merupakan himpunan bagian dari ML yang diinisiasi oleh jaringan-jaringan saraf lapisan jamak yang desainnya terinspirasi dari struktur otak manusia. Model-model DL memperkuat hampir semua perkembangan terkini dari artificial intelligence (AI) , mulai dari computer vision, generative AI hingga mobil-mobil kendali mandiri dan robotika.

B. ARSITEKTUR DEEP LEARNING

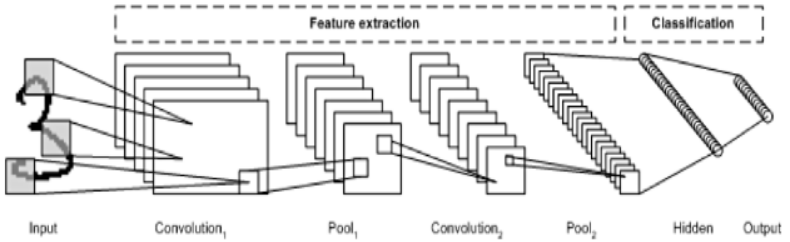
Dalam subbab ini, pembahasan arsitektur DL akan mengacu pada klasifikasi arsitektur yang ditulis dalam situs resmi IBM seperti yang ditunjukkan pada gambar 13.1. Pada gambar tersebut, arsitektur dibagi menjadi dua kelompok, yaitu Supervised learning dan Unsupervised learning. Masing-masing mempunyai cabang dua , baik supervised maupun unsupervised. Supervised learning dibagi menjadi Convolutional Neural Network (CNN) dan Recurrent Neural Network (RNN), sedangkan unsupervised learning dibagi menjadi Self Organizing Maps

(SOM) dan Autoencoder (AE), dan restricted Boltzman machine (RBM) sebagai bagian dari AE..



Gambar 13. 1 Klasifikasi Arsitektur Deep Learning
Sumber: https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

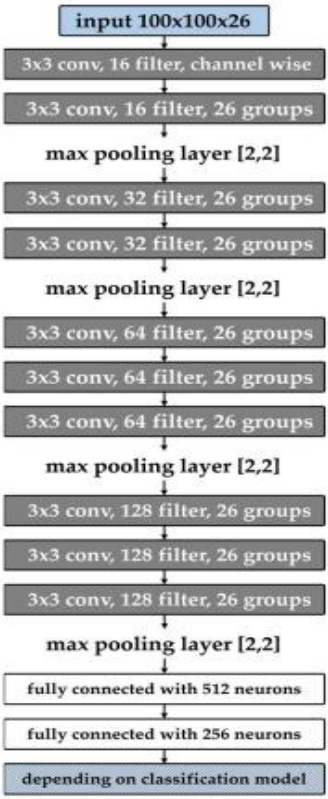
Dalam klasifikasi pertama, supervised learning, arsitektur yang dibahas adalah yang sering dikenal dengan istilah Convolutional Neural Networks (CNN). CNN merupakan sebuah jaringan saraf lapisan jamak yang secara biologi diinspirasi oleh korteks penglihatan binatang. Arsitektur ini secara khusus bermanfaat dalam aplikasi pemrosesan gambar. Arsitektur ini merupakan ide dari Yann LeCun. Saat itu, arsitektur berfokus ada pengenalan karakter tulisan tangan, seperti interpretasi kode pos. Sebagai sebuah jaringan dalam, lapisan-lapisan awal bertugas mengenali fitur seperti pinggiran-pinggiran, dan lapisan berikutnya akan melakukan rekombinasi ulang fitur-fitur menjadi atribut-atribut tingkat tinggi sebagai masukan. Arsitektur CNN LeNet dibuat dari beberapa lapisan yang mengimplementasi fitur ekstraksi kemudian klasifikasi. Gambar arsitektur ini ditunjukkan pada gambar 13.2 di halaman selanjutnya.



Gambar 13. 2 Lapisan Arsitektur CNN
Sumber: https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

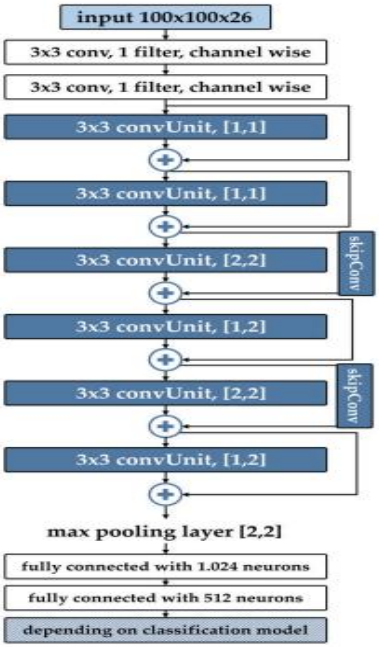
Pada gambar 13.2 dapat dilihat, gambar yang ada di bagian paling kiri, dibagi menjadi beberapa bagian reseptif. Bagian ini yang akan diberikan ke lapisan konvolusi. Lapisan konvolusi ini akan mengekstrak fitur-fitur dari gambar masukan. Langkah selanjutnya adalah pooling. Pada tahapan pooling ini dilakukan pengurangan dimensi dari fitur-fitur yang diekstrak melalui down-sampling dan di saat yang sama, tahapan pooling mempertahankan informasi yang paling penting, khususnya max pooling. Tahapan konvolusi dan pooling lain dilakukan, yang kemudian dikirim ke perseptron lapisan jamak yang terhubung secara penuh. Lapisan keluaran akhir dari jaringan ini adalah kumpulan titik-titik yang akan mengidentifikasi fitur-fitur dari gambar. Jaringan dilatih dengan menggunakan metode back-propagation.

Terdapat beberapa arsitektur CNN yang akan dibahas yaitu Vgg19, ResNet, InceptionV3, and MobileNet-V2. CNN merupakan arsitektur DL yang dikembangkan oleh kelompok Visual Geometry di universitas Oxford pada tahun 2014. Model VGG-19 ini memiliki ciri sebagai berikut : terdapat 19 lapisan, ada keseragaman ukuran saringan konvolusi 3×3 yang melalui jaringan.. Arsitektur turunan VGG-19 atau sering disebut juga Vgg19-derivate, seperti ditunjukkan pada gambar 13.3 di halaman selanjutnya, menggunakan empat kelompok dari lapisan-lapisan konvolusi sekuensial, masing-masing diikuti oleh sebuah lapisan Rectified Linear Unit atau ReLU. Pada arsitektur awal vgg-19, sekumpulan lapisan-lapisan konvolusi dihilangkan berkaitan dengan resolusi rendah dari masukan matriks. Selanjutnya, jumlah dari saringan-saringan tiap lapisan konvolusi dikurangi.



Gambar 13. 3 Arsitektur VGG-19
Sumber: https://doi.org/10.3390/a16040209

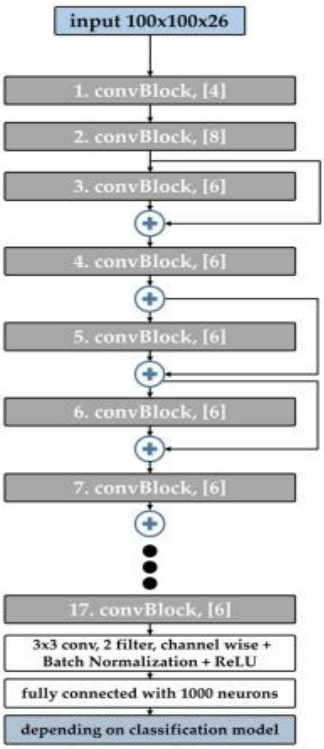
Jaringan VGG-19 terdiri dari 19 lapisan bobot dengan komposisi : enam belas lapisan konvolusi dan tiga lapisan-lapisan yang saling terkoneksi penuh, serta lima lapisan max-pooling dan sebuah lapisan final softmax untuk klasifikasi. Lapisan konvolusi memanfaatkan saringan dengan ukuran 3×3, menjadi sebuah jalur dan landasan yang sama, yang memelihara resolusi spasial dari gambar masukan melalui jaringan. Max-pooling dilakukan menggunakan jendela ukuran 2×2 dengan sebuah jalur berukuran dua, antar blok konvolusi untuk mengurangi dimensi spasial dari peta-peta fitur.



Gambar 13. 4 Arsitektur ResNet
Sumber: https://doi.org/10.3390/a16040209

Jaringan saraf residual atau juga dikenal dengan istilah ResNet , merupakan arsitektur DL. Lapisan-lapisan mempelajari fungsi-fungsi residual dengan referensi, terhadap masukan-masukan lapisan. Arsitektur ini dikembangkan pada tahun 2015 untuk keperluan pengenalan gambar. Istilah koneksi residual mengacu pada motif arsitektur spesifik $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})+\mathbf{x}$, dimana $\mathbf{f}$ merupakan sebuah modul jaringan saraf acak.

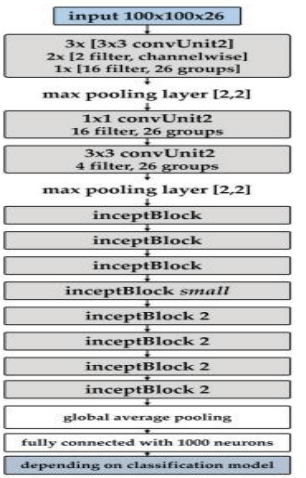
Gambar 13.4 diatas menunjukkan arsitektur ResNet. Gambar tersebut merupakan salah satu varian dari arsitektur ResNet. Arsitektur ini menggunakan convUnits, yang masing-masing terdiri dari sebuah konvolusi , normalisasi batch, aktivasi ReLU, lapisan konvolusi, dan juga lapisan normalisasi batch. Ukuran saringan adalah 3×3 untuk lapisan-lapisan konvolusi dan nilai-nilai dalam tanda kurung menunjukkan jumlah dari saringan dan jalur dari lapisan.



Gambar 13. 5 Arsitektur MobileNetV2
Sumber: https://doi.org/10.3390/a16040209

Arsitektur selanjutnya adalah MobileNet. Arsitektur ini satu rumpun dengan arsitektur CNN yang dirancang untuk klasifikasi gambar, deteksi obyek, dan pekerjaan-pekerjaan computer vision lainnya. Arsitektur ini dirancang dengan ukuran yang kecil, latensi yang rendah, serta konsumsi tenaga yang rendah. Arsitektur ini cocok untuk perangkat-perangkat dengan inferensi on-device dan juga batasan energi seperti HP dan sistem-sistem tertanam. Gambar 13.5 diatas, menunjukkan arsitektur MobileNet versi 2.

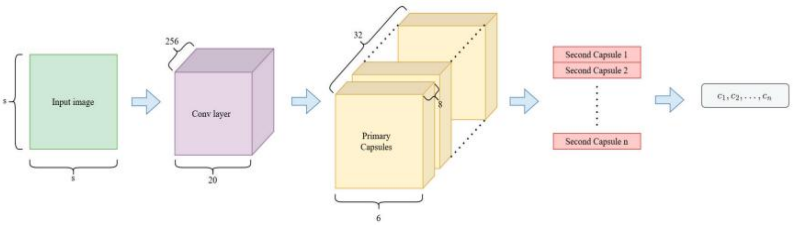
Arsitektur tersebut adalah turunan dari arsitektur MobileNet versi 2. Sebuah convBlock terdiri dari sebuah kanal konvolusi dengan enam saringan, normalisasi batch, sebuah lapisan ReLU, sebuah lapisan kanal konvolusi dengan jumlah saringan yang tetap atau sesuai dengan masukan ke convBlock, dan sebuah lapisan normalisasi batch. Terdapat total tujuh belas blok yang disusun satu dengan yang lain dengan jalur-jalur pengabaian parsial. Arsitektur ini dikembangkan oleh Google, dengan menggunakan bahasa Python. Versi terakhir yang dikembangkan adalah MobileNetV4 yang dipublikasi pada bulan September 2024. Pada versi ke-4 ini, arsitektur telah menyertakan fitur yang disebut multi-query attention. Arsitektur ini juga menggabungkan residu dan leher botol yang telah diinversi, yang sebelumnya ada di versi ke-3.



Gambar 13. 6 Arsitektur InceptionV3
Sumber: https://doi.org/10.3390/a16040209

Arsitektur keempat yaitu Inception. Arsitektur ini juga merupakan keluarga dari CNN. Diperkenalkan oleh para periset di Google pada tahun 2014 dengan nama GoogLeNet lalu kemudian diubah namanya menjadi Inception. Pada arsitektur diatas, lapisan-lapisan pertama digantikan oleh blok-blok convUnit2, yang masing-masing disusun dari sebuah konvolusi, normalisasi batch dan lapisan ReLU, dengan parameter-parameter yang ditunjukkan dalam kurung kotak. Sehubungan dengan resolusi yang lebih rendah dari masukan , dua buah blok inception terakhir yang lebih besar dihilangkan.

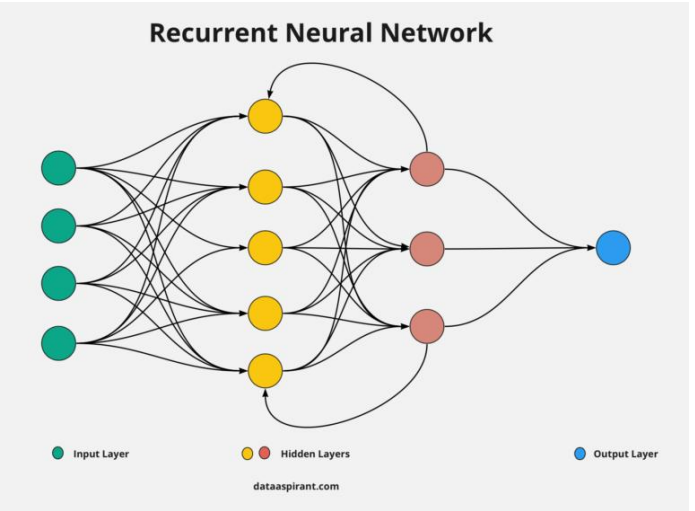
Pada paragraf sebelumnya telah dibahas arsitektur yang tergolong dalam CNN. Selanjutnya akan dibahas salah satu contoh arsitektur yang diambil dari penelitian periset dan telah dipublikasi di jurnal yang dikenal oleh publik. Arsitektur yang dibahas adalah arsitektur dari penelitian Hollosi, Ballagi & Pozna , dengan judul Simplified Routing Mechanism for Capsule Networks. Arsitektur yang digunakan disebut dengan arsitektur jaringan kapsul. Arsitektur yang digunakan ini merupakan arsitektur yang diusulkan oleh Sabour dkk. tahun 2017 yang dipublikasi pada konferensi ke-31 Neural Information Processing Systems (NIPS) di Amerika. Setelah lapisan masukan, arsitektur ini memiliki tiga komponen utama yaitu lapisan convolutional, lalu lapisan kapsul utama dan lapisan kapsul sekunder. Masing-masing ditandai dengan warna-warna berbeda. Warna hijau merupakan masukan, ungu lapisan convolutional, kuning lapisan utama kapsul, merah lapisan kedua kapsul, dan abu-abu merupakan keluaran atau prediksi. Dalam versi aslinya, arsitektur yang digunakan Sabour dkk menggunakan satu input dengan ukuran atau juga ditulis dengan format $32 \times 32 \times 1$-sized input tensor.



Gambar 13. 7 Arsitektur Jaringan Kapsul
Sumber: https://doi.org/10.3390/a16070336,

Dalam risetnya, Hollosi, Ballagi & Pozna, melakukan sedikit modifikasi dengan menggunakan bentuk masukan berupa satu masukan ukuran 28x28 atau $28 \times 28 \times 1$, satu masukan ukuran 48x48 atau $48 \times 48 \times 1$ dan tiga masukan ukuran 32x32 atau $32 \times 32 \times 3$. Dalam riset ini dibandingkan antara arsitektur klasik jaringan saraf konvolusi dengan jaringan kapsul yang mengarah ke potensi baru dari jaringan kapsul . Dan dalam riset ini juga ditunjukkan bahwa algoritma dynamic routing algorithm pada arsitektur ini terlalu kompleks dan waktu pelatihannya menjadikan jaringan ini sulit membangun jaringan yang kompleks dan dalam, namun di saat yang bersamaan jaringan kapsul dapat mencapai efisiensi yang saat baik. Gambar 13.6 menunjukkan arsitektur jaringan kapsul.

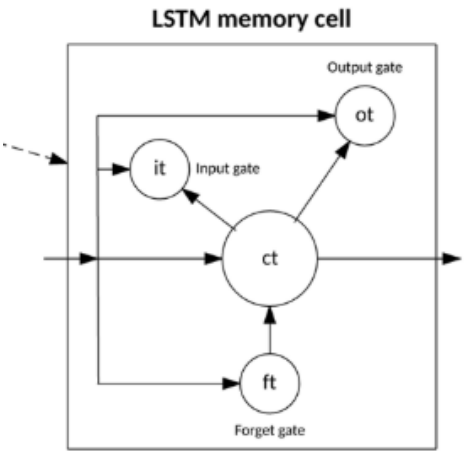
Masih dalam klasifikasi pertama, supervised learning, arsitektur berikutnya adalah Recurrent neural networks atau RNN. Perbedaan utama dengan jaringan lapisan jamak adalah jika pada lapisan jamak semua hasil akan dikirim ke lapisan selanjutnya , maka pada RNN , arsitektur ini mungkin memiliki koneksi yang akan dikirim kembali ke lapisan tertentu atau bahkan lapisan yang sama, jadi ada umpan balik. Umpan balik ini mengijinkan RNN untuk memelihara memori dari masukan-masukan sebelumnya dan masalah-masalah model saat tersebut. Gambar 13.7 menunjukkan arsitektur RNN tersebut.



Gambar 13. 8 Arsitektur Recurrent Neural Network
Sumber: https://dataaspirant.com/how-recurrent-neural-network-rnn-works/

Terdapat dua arsitektur dalam arsitektur RNN yaitu jaringan long short-term memory (LTSM) dan gated recurrent unit (GRU). LSTM dibuat pada tahun 1997 oleh Hochreiter dan Schimdhuber, dan perkembangannya cukup pesat dalam berbagai aplikasi. LSTM ini dapat ditemukan pada produk-produk yang digunakan sehari-hari seperti smartphone. IBM juga menerapkan LSTM pada konfigurasi salah satu fitur conversational speech recognition.

LSTM dimulai dari arsitektur jaringan saraf berbasis neuron dengan memperkenalkan konsep sel memori. Sel memori dapat mempertahankan nilainya untuk jangka waktu pendek atau panjang sebagai sebuah fungsi untuk masukan-masukannya, dan mengijinkan sel untuk mengingat apa yang penting dan apa yang tidak penting dari nilai terakhir yang dihitung. Gambar 13.8 menunjukkan sel LSTM. Sel memori terdiri dari tiga gerbang yang mengendalikan bagaimana informasi mengalir kedalam dan keluar dari sel. Gerbang masukan mengendalikan kapan informasi baru dapat masuk ke dalam memori. Gerbang forget mengendalikan kapan sebuah informasi yang ada, dilupakan serta mengijinkan sel untuk mengingat data baru. Dan pada akhirnya, gerbang keluaran mengendalikan kapan informasi yang ada di dalam sel , digunakan sebagai keluaran. Sel juga memiliki bobot yang akan mengendalikan tiap gerbang. Algoritma pelatihan yang digunakan , umumnya BPTT, bertujuan untuk mengoptimasi bobot-bobot ini berdasarkan dari kesalahan keluaran jaringan.



Gambar 13. 9 LSTM memory cell
Sumber: https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

Pada tahun 2014, penyederhanaan dari LSTM diperkenalkan dengan istilah gated recurrent unit (GRU). Model ini memiliki dua gerbang, yaitu dengan menghilangkan gerbang keluaran yang ada di model sebelumnya, LSTM. Gerbang tersebut adalah gerbang update dan reset. Gerbang update menunjukkan seberapa banyak konten dari sel yang harus

BAB 14 CONVOLUTIONAL NEURAL NETWORK (CNN)

Dr. Muhamad Akbar, S.T., M.IT.

A. PENDAHULUAN

Bab ini memperkenalkan Convolutional Neural Network (CNN), sebuah arsitektur *deep learning* revolusioner yang mendominasi bidang *Computer Vision* modern. CNN dirancang secara spesifik untuk memproses data yang memiliki topologi grid, seperti gambar (2D), sinyal waktu (1D), dan data spasial (3D)(Ipinlaye, 2022). Berbeda dengan Jaringan Saraf Tiruan (*Artificial Neural Network/ANN*) tradisional yang rentan terhadap kompleksitas data spasial, CNN memanfaatkan operasi konvolusi untuk mengekstrak fitur secara hierarkis dan adaptif, menjadikannya solusi utama untuk tugas-tugas pengenalan citra, deteksi objek, dan segmentasi.

B. KONSEP DASAR DAN PRINSIP KERJA CNN

CNN didasarkan pada konsep untuk meniru cara kerja korteks visual pada otak, di mana neuron-neuron merespons rangsangan hanya pada wilayah tertentu yang disebut *receptive field.*

Definisi Formal

CNN adalah Jaringan Saraf Tiruan yang menggunakan operasi matematika konvolusi sebagai pengganti operasi perkalian matriks linear umum pada setidaknya salah satu lapisannya. Tujuannya adalah untuk secara otomatis mempelajari representasi spasial (fitur) dari input tanpa memerlukan ekstraksi fitur manual.

Karekteristik CNN

Desain struktural efektifitas CNN terletak pada tiga prinsip:

1. *Shared Weights* (Pembagian Bobot): Sebuah filter atau kernel digunakan berulang-ulang di seluruh input. Hal ini mengurangi jumlah parameter yang harus dipelajari model secara drastis dan memberikan properti invariansi translasi (ketahanan terhadap pergeseran posisi objek).

2. *Local Receptive Fields* (Bidang Reseptif Lokal): Setiap *neuron* dilapisan konvolusi hanya terhubung ke wilayah kecil dan terlokalisasi dari input sebelumnya. Fitur kompleks dibentuk secara hierarkis dari kombinasi fitur lokal yang lebih sederhana.

3. *Spatial Downsampling* (*Pooling*): Penggunaan lapisan *pooling* secara berkala untuk mengurangi dimensi spasial (*width* dan *height*) dari data. Ini mengurangi beban komputasi dan membuat representasi yang diekstraksi lebih invarian terhadap variasi posisi objek yang kecil

C. KOMPONEN ARSITEKTUR CNN

Arsitektur CNN secara umum tersusun dari tiga jenis lapisan utama yang tersusun secara berurutan sebagai berikut (Alzubaidi et al., 2021):

Lapisan Konvolusi (Convolutional Layer)

Lapisan ini adalah blok bangunan fundamental CNN.

- **Operasi Konvolusi:** Proses di mana sebuah matriks kecil, yang disebut **Kernel ($\mathbf{W}$) atau Filter**, digeser (*stride*) melintasi tensor input. Pada setiap lokasi, operasi perkalian *element-wise* antara Kernel dan *patch* input dilakukan, lalu hasilnya dijumlahkan.

- **Feature Map ($\mathbf{Z}$):** Hasil dari operasi konvolusi adalah **Feature Map**, yang menunjukkan intensitas di mana fitur yang dideteksi oleh Kernel berada pada input.

- **Fungsi Aktivasi:** *Feature map* kemudian dilewatkan melalui fungsi non-linear (umumnya **ReLU**, $f(x) = \max(0, x)$) untuk memperkenalkan non-linearitas, yang penting agar model dapat mempelajari pemetaan yang kompleks.

Lapisan *Pooling* (*Pooling Layer*)

Lapisan *pooling* melakukan operasi *downsampling* pada *feature map* untuk mengurangi ukuran spasialnya.

- **Max Pooling:** Jenis yang paling umum. Ia mengambil nilai maksimum dari elemen-elemen yang berada dalam *window*

pooling. Ini membantu mengurangi jumlah parameter sambil mempertahankan informasi fitur yang paling menonjol.

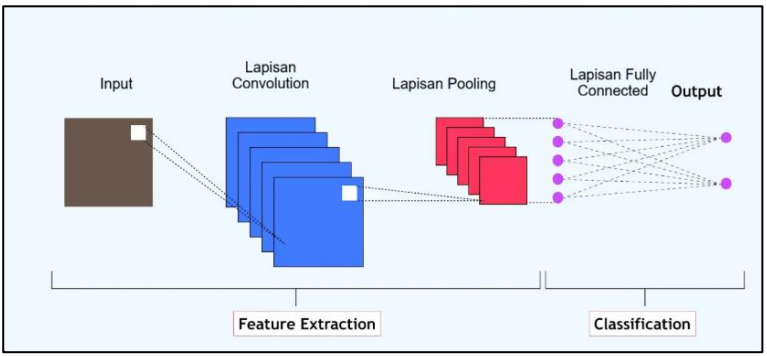
- **Average Pooling:** Mengambil nilai rata-rata dari elemen-elemen dalam *window*.

Lapisan *Fully Connected* (FC Layer)

Setelah rangkaian lapisan Konvolusi dan *Pooling*, data harus diubah bentuknya (*flatten*) menjadi vektor 1D. Vektor ini kemudian diumpankan ke Lapisan *Fully Connected* (FC), di mana setiap neuron terhubung ke semua neuron dilapisan sebelumnya, mirip dengan ANN tradisional. Lapisan ini bertanggung jawab untuk melakukan **klasifikasi** atau **regresi** berdasarkan fitur yang telah diekstrak oleh lapisan-lapisan sebelumnya.

- **Lapisan Output:** Lapisan FC terakhir biasanya menggunakan fungsi aktivasi **Softmax** untuk tugas klasifikasi multikelas, menghasilkan distribusi probabilitas:

$$ P(y=i|\mathbf{x}) = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{K} e^{\mathbf{z}_j}}$$



Gambar 14. 1 Komponen Arsitektur CNN
Sumber : (Gurucharan, 2025)

D. DATA TOPOLOGI GRID DALAM CNN

Data topologi grid (seperti gambar, deret waktu, atau volume 3D) memiliki dua properti penting yang dieksploitasi secara optimal oleh CNN, tidak seperti Jaringan Saraf Tiruan (*Artificial Neural Network/ANN*) tradisional:

Pola Lokal yang Signifikan (*Locality*)

- Pada data grid, nilai-nilai (piksel, titik data) yang berdekatan secara spasial atau temporal cenderung memiliki **korelasi yang kuat** dan membentuk **pola lokal** yang bermakna.
- **Contoh:** Dalam gambar, piksel yang berdekatan membentuk tepi, sudut, atau tekstur.
- **CNN memanfaatkan ini** melalui **Kernel (Filter)**. Filter ini hanya beroperasi pada *local receptive field* (area lokal kecil) dari input, sehingga hanya fokus pada ekstraksi pola lokal tersebut.

Invariansi Translasi (*Translation Invariance)*

- Pola atau fitur yang sama dapat muncul dilokasi yang berbeda dalam data grid. Misalnya, mata dapat berada dikanan atas atau kiri atas wajah.
- **CNN memanfaatkan ini** melalui **Shared Weights**. Filter yang mendeteksi pola tertentu (misalnya, garis vertikal) akan menggunakan bobot yang sama diseluruh input. Ini berarti jika pola tersebut bergeser di-input, filter yang sama tetap dapat mendeteksinya.
- ANN tradisional yang *fully connected* harus mempelajari pola yang sama berulang kali disetiap lokasi, yang sangat tidak efisien dan memerlukan jauh lebih banyak data.

Perbedaan implementasi terletak pada **dimensi filter (*kernel*)** dan **arah pergeseran (*sliding*)** filter melintasi data input.

Tabel 14. 1 Implementasi Dimensi Konvolusi 1D, 2D, 3D

Dimensi	Konvolusi 1D	Konvolusi 2D	Konvolusi 3D
Data Input	Vektor (Deret Waktu, Teks *Embedding*).	Matriks (Gambar Grayscale/RGB).	Kubus (Video, Data Medis Volumetrik).
Dimensi Filter	$(k \times 1)$	$(k_h \times k_w)$	$(k_d \times k_h \times k_w)$
Arah Pergeseran	Sepanjang 1 sumbu (panjang).	Sepanjang 2 sumbu (tinggi dan lebar).	Sepanjang 3 sumbu (kedalaman,

Dimensi	Konvolusi 1D	Konvolusi 2D	Konvolusi 3D
			tinggi, dan lebar).
Output	Vektor 1D (Deret Waktu Baru).	Matriks 2D (*Feature Map*).	Kubus 3D (*Feature Map* Volumetrik).

Konvolusi 1D diterapkan pada data sekuensial atau deret waktu. Filter bergerak hanya sepanjang satu sumbu.

- **Input Data:** Vektor L (panjang) $\times C_{\text{in}}$ (jumlah *channel* input).
- **Filter:** Memiliki lebar k dan kedalaman C_{in}.
- **Aplikasi Nyata:**
    - **Pemrosesan Bahasa Alami (NLP):** Mengklasifikasikan urutan kata (dinyatakan sebagai *word embeddings*) untuk tugas analisis sentimen atau klasifikasi teks.
    - **Pemrosesan Sinyal:** Analisis data sensor, EKG, atau *time series* keuangan.

Konvolusi 2D adalah bentuk CNN yang paling umum, dirancang untuk memproses citra statis. Filter bergerak sepanjang dua sumbu: tinggi dan lebar.

- **Input Data:** Tensor H (tinggi) $\times W$ (lebar) $\times C_{\text{in}}$ (jumlah *channel*, misal 3 untuk RGB).
- **Filter:** Memiliki ukuran $k_h \times k_w$ dan kedalaman C_{in}. Filter harus mencakup seluruh kedalaman *channel* input.
- **Aplikasi Nyata:**
    - **Klasifikasi Citra:** Mengidentifikasi konten utama gambar (misalnya, Dog, Car, Cat).
    - **Deteksi Objek:** Menentukan lokasi objek dalam gambar.

BAB 15 MACHINE LEARNING UNTUK NATURAL LANGUAGE PROCESSING

Muhammad Irvai, M.Kom.

A. PENGANTAR NATURAL LANGUAGE PROCESSING (NLP)

1. Definisi dan Ruang Lingkup NLP

Natural Language Processing (NLP) atau pemrosesan bahasa alami merupakan salah satu bidang penting dalam kecerdasan buatan yang berfokus pada bagaimana komputer dapat memahami, mengolah, dan menghasilkan bahasa manusia secara alami. NLP menjadi semakin relevan karena bahasa manusia memiliki karakteristik yang kompleks, ambigu, serta tidak selalu terstruktur. Untuk itu, diperlukan pendekatan komputasional yang mampu mengubah bahasa alami menjadi bentuk representasi yang dapat diproses oleh mesin. Di Indonesia, penelitian di bidang NLP mengalami perkembangan pesat, terutama pada topik sentiment analysis, klasifikasi teks, dan chatbot, seperti yang ditunjukkan dalam berbagai penelitian terbaru (Purwarianti, 2018) (Wibowo, A., 2017) (Mahendra, R., Wibisono, A., & Adriani, 2018). Dukungan teknologi modern seperti *deep learning* dan *word* embeddings juga semakin memperkuat kemampuan sistem NLP dalam memahami konteks bahasa Indonesia secara lebih akurat (Purwarianti, A., & Crisdayanti, 2019) (Al-Aufi, M., & Santoso, 2020).

Secara sederhana, NLP dapat didefinisikan sebagai seperangkat teknik, algoritma, dan model yang memungkinkan komputer memahami makna teks atau ucapan manusia. Pemahaman ini tidak hanya terbatas pada membaca teks secara literal, tetapi juga mencakup interpretasi konteks, gaya bahasa, hubungan antar kata, hingga maksud yang terkandung di dalamnya.

Ruang lingkup NLP sangat luas dan mencakup berbagai tahapan mulai dari *text preprocessing* (pembersihan teks), representasi teks menjadi angka, hingga pembelajaran pola menggunakan model-machine learning. Di tingkat lanjutan, NLP juga mencakup

pemahaman semantik, analisis sintaksis, dan generasi bahasa alami (*natural language generation*).

NLP sendiri merupakan area multidisiplin yang memadukan berbagai bidang ilmu, seperti:

a. ***Linguistik***, untuk memahami struktur, morfologi, dan sintaks bahasa.
b. **Ilmu komputer**, khususnya algoritma dan struktur data.
c. ***Machine Learning***, untuk membangun model yang mampu belajar pola dari data.
d. **Statistika dan probabilitas**, untuk mendukung analisis dan prediksi berdasarkan data bahasa.
e. ***Cognitive science***, yang membantu memahami bagaimana manusia memproses bahasa.

Bahasa manusia memiliki sifat dinamis selalu berkembang dari waktu ke waktu. Munculnya kosakata baru, perubahan gaya komunikasi, slang, hingga perkembangan budaya mempengaruhi bagaimana NLP harus dirancang. Di era digital, jumlah data teks semakin meningkat secara eksponensial. Pesan singkat, komentar media sosial, artikel berita, dokumen administrasi, hingga percakapan digital menjadi sumber informasi yang sangat besar. Data dalam jumlah masif ini membutuhkan teknik yang terstruktur agar dapat diproses dan dimanfaatkan secara efektif. Di sinilah NLP menjadi sangat relevan.

Secara keseluruhan, definisi dan ruang lingkup NLP dapat disimpulkan sebagai upaya sistematis untuk memungkinkan komputer "mengerti" bahasa manusia, mengolahnya dengan teknik-teknik komputer, dan menggunakannya kembali untuk menghasilkan keluaran yang bermakna, relevan, dan berguna bagi pengguna.

2. Hubungan NLP dalam Machine Learning

Perkembangan NLP modern tidak dapat dipisahkan dari kemajuan Machine Learning (ML). Sebelum era machine learning, pendekatan NLP lebih banyak mengandalkan aturan eksplisit (*rule-based system*). Pendekatan tersebut memerlukan daftar aturan bahasa yang ditulis oleh ahli linguistik, misalnya aturan tata bahasa, daftar kata dasar, atau pola tertentu dalam sebuah kalimat. Walaupun metode ini cukup efektif dalam beberapa kasus, rule-based NLP memiliki

keterbatasan karena sulit menangani variasi bahasa yang sangat banyak.

Kemunculan Machine Learning membawa perubahan besar dalam pendekatan NLP. Dengan machine learning, komputer tidak lagi bergantung pada aturan yang ditulis secara manual, melainkan **belajar pola dari data bahasa**. Artinya, semakin banyak data yang diberikan, semakin baik kemampuan model dalam memahami teks.

Hubungan antara NLP dan machine learning dapat dipahami melalui tiga aspek berikut:

a. Representasi Teks sebagai Data Numerik

Machine learning bekerja menggunakan data numerik. Sementara itu, bahasa manusia berbentuk teks. NLP menyediakan jembatan dengan cara mengubah teks menjadi representasi angka melalui teknik-teknik seperti Bag-of-Words (BoW), TF-IDF, dan embedding. Tanpa representasi numerik, model ML tidak dapat memproses teks.

b. Pembelajaran Pola dari Data Teks

Dalam konteks NLP, machine learning berperan untuk "belajar" hubungan atau pola dari data. Misalnya:

1) Pada ***sentiment analysis***, model ML mempelajari apakah suatu kalimat bersentimen positif atau negatif.
2) Pada ***text classification***, model mempelajari kategori dokumen berdasarkan isi teks.
3) Pada ***spam detection***, model dapat membedakan antara pesan normal dan pesan spam.

c. Kemampuan Generalisasi Model

Keunggulan *machine learning* terletak pada kemampuannya melakukan generalisasi. Artinya, model dapat membuat prediksi terhadap teks baru yang belum pernah dilihat sebelumnya. Ini yang membuat model ML lebih unggul daripada rule-based system, karena tidak perlu mengatur aturan untuk setiap kemungkinan variasi bahasa.

3. Contoh Aplikasi NLP

NLP telah digunakan dalam berbagai aspek kehidupan sehari-hari. Hampir setiap aplikasi digital yang berhubungan dengan teks atau suara

memanfaatkan teknik NLP secara langsung maupun tidak langsung. Berikut beberapa contoh aplikasi NLP yang paling umum dan relevan:

a. Sentiment Analysis

Sentiment analysis digunakan untuk mengetahui apakah suatu teks memiliki sentimen positif, negatif, atau netral. Aplikasi ini sangat populer dalam analisis media sosial, seperti komentar pengguna, ulasan produk, dan opini publik.

Gambar 15. 1 Aplikasi Sentiment Analysis For Netflixs App
https://share.google/images/AHoqUlFTQHvc6E6PW

Contoh penggunaannya:

1) Analisis komentar pelanggan terhadap produk.
2) Analisis respon masyarakat terhadap kebijakan pemerintah.
3) Pemantauan reputasi merek di media sosial.

b. Text Classification

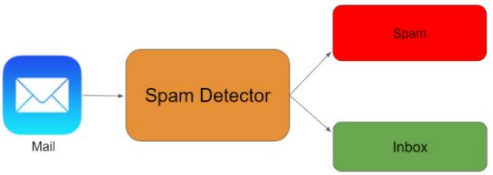
Text classification adalah proses pemberian label otomatis pada teks berdasarkan topik atau kategori tertentu.

Contohnya:

1) Klasifikasi artikel berita berdasarkan topik (politik, olahraga, ekonomi).
2) Pengelompokan artikel jurnal berdasarkan tema penelitian.
3) Klasifikasi email kantor berdasarkan jenis layanan atau permohonan.

c. Spam Detection

NLP digunakan dalam sistem penyaringan email untuk memisahkan pesan penting dari pesan spam. Model machine learning mempelajari pola kata atau frasa yang sering muncul di email spam, sehingga dapat melakukan deteksi otomatis.



Gambar 15. 2 Spam Detection
Sumber: Website https://share.google/images/1ErLt6bMeDThOD1Ti

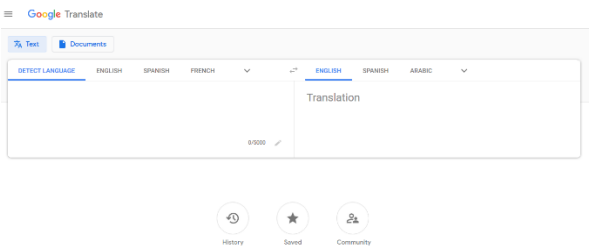
d. Chatbot dan Virtual Assistant

Chatbot modern menggunakan NLP untuk memahami pertanyaan pengguna dan memberikan jawaban relevan. Sistem seperti Google Assistant, Siri, atau chatbot layanan pelanggan memanfaatkan NLP untuk memahami bahasa natural, memproses maksud (intent), dan menghasilkan respons yang sesuai.

Gambar 15. 3 Chatbot
Sumber: Website https://share.google/images/CitF4bthgi9sEwnME

e. Machine Translation

Penerjemah otomatis seperti Google Translate atau DeepL menggunakan NLP untuk menerjemahkan teks dari satu bahasa ke bahasa lain. Pada tingkat dasar, model machine learning mempelajari hubungan antar kata dan struktur bahasa di dua bahasa berbeda.



DAFTAR PUSTAKA

Al-Aufi, M., & Santoso, H. B. (2020). Implementasi NLP untuk Klasifikasi Dokumen Bahasa Indonesia Menggunakan Deep Learning. Jurnal Teknologi Informasi Dan Ilmu Komputer, 7(4), 623–631.

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

Aggarwal, C. C. 2022. Neural networks and deep learning: A textbook (2nd ed.). Springer.

Aggarwal, Charu C, and others. 2018. 10 Neural Networks and Deep Learning. Springer.

Aggarwal, Puneet Kumar, Parita Jain, Jaya Mehta, Riya Garg, Kshirja Makar, and Poorvi Chaudhary. 2021. "Machine Learning, Data Mining, and Big Data Analytics for 5G-Enabled IoT." In Blockchain for 5G-Enabled IoT: The New Wave for Industrial Automation, Springer, 351–75.

Aggarwal, Sakshi. 2023. "Machine Learning Algorithms, Perspectives, and Real-World Application: Empirical Evidence from United States Trade Data."

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207-216.

Aisyah, B. N., & Gunawan, I. (2024). Penerapan Machine Learning Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree. Seminar Nasional Hasil Penelitian & Pengabdian Masyarakat Bidang Ilmu Komputer, 1–6.

Alexandropoulos, Stamatios Aggelos N., Sotiris B. Kotsiantis, and Michael N. Vrahatis. 2019. "Data Preprocessing in Predictive Data Mining." Knowledge Engineering Review 34. doi:10.1017/S026988891800036X.

Alfayez, R., & Alazba, A. (2025). Merge conflict prediction using feature selection and stacking heterogeneous ensembles: An empirical investigation. Journal of Software Evolution and Process, 37(9).

Alifiani, R., & Rahman, R. (2019). Penerapan Kecerdasan Buatan untuk Mendeteksi Plagiat dalam Tugas Akademik. Jurnal Teknologi Informasi Dan Ilmu Komputer, 6(2), 152–160.

Aliyah, A., Waseso, B., Kurniabudi, K., Meilano, R., Riza, F., Jinan, A., Agus Oka Gunawan, I. M., Muis, A., Noor Intan, D., Sembiring, A., Trisnawan, A. B., & Fachruddin, F. (2025). Dasar-Dasar Cybersecurity (1st ed.). Faaslib Serambi Media. https://faaslibsmedia.com/

Allorerung, P. P., Erna, A., & Bagussahrir, M. (2024). Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit . JISKA (Jurnal Informatika Sunan Kalijaga), 9(3).

Alnuaimi, A M, dan M A Albaldawi. 2024. "Applications of supervised machine learning in predictive modeling." Journal of Artificial Intelligence Research 9(2): 112–28.

Alpaydin, E. 2014. Introduction to Machine Learning, 3rd ed, MIT Press, London England

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Astuti, S. (2021). Pemanfaatan Teknologi Kecerdasan Buatan dalam Manajemen Pembelajaran. Jurnal Pendidikan Dan Teknologi Informasi, 8(1), 37–45.

Baichtal, Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. Pearson.

Baladram, R, S Kumar, dan A Prasad. 2020. "Support Vector Machines: Theory and Applications in Pattern Recognition." International Journal of Computational Intelligence 14(3): 245–62.

Bangar, S. (2022a, June 24). AlexNet Architecture Explained. Medium. https://medium.com/@siddheshb008/alexnet-architecture-explained-b6240c528bd5

Bangar, S. (2022b, June 28). VGG-Net Architecture Explained. Medium. https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f

Bangar, S. (2022c, July 5). Resnet Architecture Explained. Medium. https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d

Barreca, Daniele. 2018. "A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems." SIGKDD Explorations 3(1).

Baty, A. (2023). An educational approach to Higgs boson hunting using machine learning classification algorithms on ATLAS Open Data. Journal of Advanced Research in Natural and Applied Sciences, 9(3), 560–576.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

Bengio, Yoshua, Ian Goodfellow, Aaron Courville, and others. 2017. 1 Deep Learning. MIT press Cambridge, MA, USA.

Beram, R., & El-Kotory, A. (2024). Simulation-based assessment of classification methods: Statistical models vs. machine learning algorithms. The Egyptian Statistical Journal, 68(1), 91–124.

Bhat, S., Selvam, V., & Ansari, G. (2023). Predicting life style of early diabetes mellitus using machine learning technique. International Journal of Computing, 22(3), 345–351.

Bickel, S., Goetz, S., & Wartzack, S. (2023). Detection of Plausibility and Error Reasons in Finite Element Simulations with Deep Learning Networks. Algorithms, 16(4), 209. https://doi.org/10.3390/a16040209

Bintoro, P., Ratnasari, R., Wihardjo, E., Putri, I. P., & Asari, A. (2024). Pengantar Machine Learning (1st ed.). PT. Mafy Media Literasi Indonesia.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324

Brownlee, Jason. 2020. "Tour Data Preparation Techniques for Machine Learning." Marchine Learning Mastery.

Budi Raharjo. (2021). Pembelajaran Mesin (Machine Learning). Yayasan Prima Agus Teknik.

C Hollósi, J., Ballagi, Á., & Pozna, C. R. .2023. Simplified Routing Mechanism for Capsule Networks. Algorithms, 16(7), 336. https://doi.org/10.3390/a16070336,

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27.

Chai, Christine P. 2020. "The Importance of Data Cleaning: Three Visualization Examples." Chance 33(1): 4–9.

Chang, Y., Seong, Y., & Yi, S. (2025). Supervised classification model for neural correspondence to environmental uncertainty in multiple cue judgment system with decision support. Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Chollet, F. (2021). Deep learning with Python (2nd ed.). Manning Publications.

Chollet, F. 2018. Deep learning with Python. Manning Publications.

Chollet, F. 2021. Deep learning with Python (2nd ed.). Manning Publications.

Cieslak, Matthew C., Ann M. Castelfranco, Vittoria Roncalli, Petra H. Lenz, and Daniel K. Hartline. 2020. "T-Distributed Stochastic Neighbor Embedding (t-SNE): A Tool for Eco-Physiological Transcriptomic Analysis." Marine Genomics 51. doi:10.1016/j.margen.2019.100723.

Cunningham, John P, and Zoubin Ghahramani. 2015. "Linear Dimensionality Reduction: Survey, Insights, and Generalizations." The Journal of Machine Learning Research 16(1): 2859–2900.

Dam, T., Roggeveen, L., Diggelen, F., Fleuren, L., Jagesar, A., Otten, M., & Beudel, M. (2022). Predicting responders to prone positioning in mechanically ventilated patients with COVID-19 using machine learning. Annals of Intensive Care, 12(1).

Das K. , 2020, How Recurrent Neural Network (RNN) Works, https://dataaspirant.com/how-recurrent-neural-network-rnn-works/

Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. Harvard Business Review.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

Deepaisarn, S., Chokphantavee, S., Chokphantavee, S., Prathipasen, P., Buaruk, S., & Sornlertlamvanich, V. (2023). NLP-based music processing for composer classification. Scientific Reports, 13(1).

Dinesh, P., & Kalyanasundaram, P. (2022). Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest, and decision tree to measure accuracy. ECS Transactions, 107(1), 12681–12691.

Dolphins R. , 2020, LSTM Networks : A Detailed Explanation, https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9/

Du, C, X Li, dan Y Wang. 2025. "Optimization techniques in supervised learning: A mathematical perspective." Computational Intelligence Letters 18(1): 77–95.

Elwirehardja, G. N., Suparyanto, T., & Pardamean, B. (2023). Pengenalan Konsep Machine Learning untuk Pemula (1st ed.). INSTIPER Press.

Emi Susilowati, Pradhana Edi Kresnha, Aulia Syifa, Noviarum Widyasmara L, Amelia Tri Hapsari, Muhammad Faizal, Andri Nurhadi, Fernanda Awalia, Yudo Witni Prasetyo, & Yusup Hidayat Winata. (2020). Pembelajaran Mesin: Teori dan Studi Kasus. Canting Mas Anyar.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, 96(34), 226-231.

Firdawanti, Aulia Rizki, Hafidlotul Fatimah Ahmad, and Nur Agustiani. 2025. "Analisis Perbandingan Kinerja Algoritma K-Means Dan K-Medoids Dengan Reduksi Dimensi PCA Pada Indikator Kesehatan Dan Sosial." Bulletin of Computer Science Research 5(5): 1235–47.

Firmansyah, Muhammad Raihan, and Yani Parti Astuti. 2024. "Stroke Classification Comparison with KNN through Standardization and

Normalization Techniques." Advance Sustainable Science, Engineering and Technology 6(1). doi:10.26877/asset.v6i1.17685.

Gallatin, Kyle, and Chris Albon. 2023. Machine Learning with Python Cookbook. " O'Reilly Media, Inc."

Gangu, S. (2022). Edibility detection of mushroom using logistic regression and PCA. International Journal of Advanced Research in Computer Science, 13(3), 30–34.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Géron, A. 2019. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Géron, A. 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Géron, Aurélien. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition . O'Reilly Media, Inc.

Géron, Aurélien. 2019. O'Reilly Media, Inc. Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd Edition.

Géron, Aurélien. 2022. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc."

Ghozali, I. (2021) Aplikasi Analisis Multivariate dengan Program IBM SPSS 26. Semarang: Badan Penerbit Universitas Diponegoro, pp. 112–118.

Goldberg, Yoav. 2018. "Neural Network Methods for Natural Language Processing." Computational Linguistics 44(1). doi:10.1162/COLI_r_00312.

Gonzalez, Rafael C, and Richard E Woods. 2018. Pearson Education Digital Image Processing (4th Ed).

Goodfellow, I. 2020. Generative adversarial networks. Communications of the ACM, 63(11), 139–144.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gurucharan, M. K. (2025, August 7). Basic CNN Architecture Explained: Key Layers and Workflow. upGrad Blog. https://www.upgrad.com/blog/basic-cnn-architecture/

Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6), 773–786. https://doi.org/10.1016/j.patrec.2007.12.011

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Han, Jiawei, Jian Pei, and Hanghang Tong. 2022. Data Mining: Concepts and Techniques, Fourth Edition Data Mining: Concepts and Techniques, Fourth Edition. doi:10.1016/C2013-0-18660-6.

Han, Jiawei., Kamber, Micheline., & Pei, Jian. (2012). Data mining : concepts and techniques. Elsevier/Morgan Kaufmann.

Handayani, F. (2022). Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar. Jurnal Teknologi Dan Informasi, 12(1). https://doi.org/10.34010/jati.v12i1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York, NY: Springer New York. doi: 10.1007/978-0-387-84858-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2024). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Edisi diperluas). Springer.

Hastie, Trevor, Robert Tibshirani, Gareth James, and Daniela Witten. 2021. "An Introduction to Statistical Learning (2nd Ed.)." Springer texts 102.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193-218.

Ichsan, C. (2025). Dasar-Dasar Machine Learning: Teori, Algoritma, dan Implementasi. CV. Green Publisher Indonesia.

Ipinlaye, O. O. (2022, November 10). Convolutional Neural Network Dimensions & Model Performance. Paperspace by DigitalOcean Blog. https://blog.paperspace.com/convolutional-neural-network-dimensions-model-performance/

Isabelle Guyon, Andr´e Elisseeff. 2016. "An Introduction to Variable and Feature Selection." Procedia Computer Science 94.

Iwan Nurhidayat, A., & Fatrianto, D. (2021). Prediksi Kinerja Akademik Mahasiswa Menggunakan Machine Learning dengan Sequential Minimal Optimization untuk Pengelola Program Studi.

J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy KNearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. "Statistical Learning." In An Introduction to Statistical Learning: With Applications in Python, Springer, 15–67.

Jiang, Z. 2021. "Deep learning foundations and supervised learning integration." International Journal of Machine Intelligence 15(1): 33–49.

Jollife, Ian T., and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065). doi:10.1098/rsta.2015.0202.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202.

Jurafsky, D., & Martin, J. H. 2023. Speech and language processing (3rd ed. draft). Stanford University.

Jurafsky, Daniel, and James H Martin. 2023. "RNNs and LSTMs (Ch 9 of Speech and Language Processing, Jurafsky and Martin)." Speech and Language Processing.

Keerthi, S. S., & Gilbert, E. G. (2002). Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning, 46, 351-360.

Kuhn, Max, and Kjell Johnson. 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models Feature Engineering and Selection: A Practical Approach for Predictive Models. doi:10.1201/9781315108230.

Kurniawati, putri. 2017. "Data Mining: Concepts and Techniques Han, j. and Kamber, m. 4th Edition." Universitas Nusantara PGRI Kediri 01.

Kusuma, A. P., & Oktavianto, A. D. (2022). Analisis Metode Euclidean Distance dalam Menentukan Koordinat Peta pada Alamat Rumah. Jurnal Teknologi Dan Manajemen Informatika, 8(2), 108–115. https://doi.org/10.26905/jtmi.v8i2.8871

LeCun, Y., Bengio, Y., & Hinton, G. 2021. Deep learning. Nature.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. "A Survey on Deep Learning in Medical Image Analysis." Medical Image Analysis 42. doi:10.1016/j.media.2017.07.005.

Liu, G., Zhao, H., Fang, F., Liu, G., Xu, Q., & Nazir, S. (2022). An enhanced intrusion detection model based on improved kNN in WSNs. Sensors, 22(4), 1407.

Liu, Q., He, Q., & Shi, Z. (2008). Extreme support vector machine classifier. In Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings 12 (pp. 222-233). Springer Berlin Heidelberg.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2020). Understanding and enhancement of internal clustering validation measures. IEEE Transactions on Cybernetics, 50(3), 1234-1246. https://doi.org/10.1109/TCYB.2019.2904802

Loh, W. (2011). Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 1(1), 14–23. doi: 10.1002/widm.8

Lu, Xiaolei. 2022. "Machine Learning for Text, by Charu C. Aggarwal, New York, Springer, 2018. ISBN 9783319735306. XXIII + 493

Pages." Natural Language Engineering 28(4). doi:10.1017/s1351324920000637.

M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine learning techniques in cognitive radios" IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136–1159, Oct. 2012

Madhavan S., Jones M.T., 2024, Deep learning architectures : The rise of artificial intelligence. https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/.

Mahendra, R., Wibisono, A., & Adriani, M. (2018). Enhancing Indonesian NLP with Word Embeddings for Named Entity Recognition. Journal of Physics: Conference Series, 978.

Mahesh, Batta. Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR) ISSN: 2319-7064

Maimon, O. Z. ., & Rokach, Lior. (2015). Data Mining With Decision Trees: Theory And Applications (2nd Edition) : Theory and Applications. World Scientific Publishing Company.

McCarthy, Richard V., Mary M. McCarthy, and Wendy Ceccucci. 2022. Applying Predictive Analytics Applying Predictive Analytics. doi:10.1007/978-3-030-83070-0.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction | Https://Arxiv.Org/Abs/1802.03426v2." arXiv [PREPRINT].

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations. ArXiv.

Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2022. "Deep Learning-Based Text Classification." ACM Computing Surveys 54(3). doi:10.1145/3439726.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Mitra Novitri Waruwu, Yulisman Zega, Ratna Natalia Mendrofa, & Yakin Niat Telaumbanua. (2024). Implementasi Algoritma Machine Learning Untuk Deteksi Performa Akademik Mahasiswa. TEKNIMEDIA: Teknologi Informasi dan Multimedia, 5(2), 181–188.

Mniai, A., Tarik, M., & Jebari, K. (2023). A novel framework for credit card fraud detection. IEEE Access, 11, 112776–112786.

Muhammad Hermawan, B., Abdurahman Hakim, M., Arifin, R., & Puspitasari, N. (2024). Pemanfaatan Artificial Intelligence, Khususnya Machine Learning dan Deep Learning System dalam Pendidikan. Seminar Nasional Amikom Surakarta (SEMNASA), 1–10.

Müller, Andreas C., and Sarah Guido. 2015. O'Reilly Media, Inc. Introduction to Machine Learning with Python and Scikit-Learn.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Murphy, K. P. (2022). Machine Learning: A Probabilistic Perspective. MIT Press.

Muzakir, A., Adi, K., & Kusumaningrum, R. (2024). Penerapan Konsep Machine Learning & Deep Learning (Pendekatan Ekspansi Semantik untuk Klasifikasi Ujaran Kebencian) (1st ed.). UNDIP Press.

Nkemdilim, E, C Uche, dan K Okwara. 2024. "Comparative analysis of supervised learning algorithms for predictive analytics." Data Science Review 7(4): 211–30.

Parasa, G., Arulselvi, M., & Razia, S. (2023). Comparative Analysis of VGG and ResNet for the Prediction of Rice Leaf Disease. 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), 191–199. https://doi.org/10.1109/ICIRCA57980.2023.10220897

Paul, D, dan S Das. 2023. "Ensemble learning approaches: Random forest and gradient boosting revisited." Data Science and Engineering Journal 9(4): 299–315.

Permana, A. A., S, W., Santoso, L. W., Wibowo, G. W. N., Wardhani, A. K., Rahmaddeni, R., Wahidin, A. J., Yliastuti, G. E., Elisawati, E., Wijayanti, R. R., & Abdurrasyid, A. (2023). Machine Learning (1st

ed.). PT. Global Eksekutif Teknologi. www.globaleksekutifteknologi.co.id

Pratiwi, R. D. (2020). Peran Kecerdasan Buatan dalam Pengembangan Kurikulum Pendidikan Abad 21. Jurnal Ilmiah Pendidikan Fisika Al-Biruni, 9(1), 13– 23

Provost, F., & Fawcett, T. (2013). Data Science for Business.O'Reilly Media.

Provost, F., & Fawcett, T. (2020). Data science for business: What you need to know about data mining and data-analytic thinking (2nd ed.). O'Reilly Media.

Purba, F. N., & Azizah, R. A. (2021). Implikasi Kecerdasan Buatan dalam Privasi dan Keamanan Data Siswa. Jurnal Informatika Mulawarman, 16(1), 32–40.

Purwarianti, A. (2018). Pemrosesan Bahasa Alami (Natural Language Processing). Informatika.

Purwarianti, A., & Crisdayanti, M. (2019). Improving Indonesian Sentiment Analysis Using Hybrid Word Embedding. Proceedings of the 2019 International Conference on Asian Language Processing (IALP).

Quinian, Ross. (1993). C4.5 : programs for machine learning. Morgan Kaufmann.

R. S. Sutton, "Introduction: The Challenge of Reinforcement Learning", Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992 [8] P. Harrington, "Machine Learning in action", Manning Publications Co., Shelter Island, New York, 2012

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336), 846-850.

Raschka, S., Liu, Y., & Mirjalili, V. 2022. Machine learning with PyTorch and Scikit-Learn. Packt Publishing.

Rawat, Waseem, and Zenghui Wang. 2017. "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review." Neural Computation 29(9). doi:10.1162/NECO_a_00990.

Ren, L, dan J Wang. 2023. "Advances in weakly supervised learning: Theory, methods, and applications." Artificial Intelligence Review 56(2): 1247–81.

Rifky, S., Yani, A., & Cahyani, D. (2023). Implementasi Manajemen PTKIS Berbasis Pondok Pesantren (Studi di STISHK Kuningan). Jurnal Manajemen Pendidikan Dasar, Menengah Dan Tinggi [JMP-DMT], 4(4), 406– 411. https://doi.org/10.30596/jmp-dmt.v4i4.16090

Riska Rismaya, Dwi Yuniarto, & David Setiadi. (2025). Penerapan Algoritma Machine Learning dalam Prediksi Prestasi Akademik Mahasiswa. Router: Jurnal Teknik Informatika Dan Terapan, 3(1), 15–23. https://doi.org/10.62951/router.v3i1.389

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A Primer in Bertology: What We Know about How Bert Works." Transactions of the Association for Computational Linguistics 8. doi:10.1162/tacl_a_00349.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015

Salman, M. D., Rahmaddeni, R., Pratama, N. R., A, M. N. F., Setiawan, A. A., Zalianti, F., & Huda, I. B. (2025). Perbandingan Kinerja Algoritma Clustering K-Means dan K-Medoids dalam Pengelompokan Sekolah di Provinsi Riau Berdasarkan Ketersediaan Sarana dan Prasarana. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5(3), 797–806. https://doi.org/10.57152/malcom.v5i3.1950

Santoso, I., Manurung, A. M., & Subhiyakto, E. R. (2025). Comparison of ResNet-50, EfficientNet-B1, and VGG-16 Algorithms for Cataract Eye Image Classification. Journal of Applied Informatics and Computing, 9(2), 284–294. https://doi.org/10.30871/jaic.v9i2.8968

Santoso, Teguh Joseph. Aplikasi AI Machine Learning dan dalam bisnis. Yayasan Prima Agus Teknik

Saputra, E. A., & Nataliani, Y. (2021). Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means. Journal of Information Systems and Informatics, 3(3), 424–439. https://doi.org/10.51519/journalisi.v3i3.164

Saputri, E. (2025). Teknik dan aplikasi data mining di Indonesia: tinjauan literatur satu dekade (2015-2024). IT-Explore: Jurnal Penerapan Teknologi Informasi Dan Komunikasi, 4(2). https://doi.org/10.24246/itexplore.v4i2.2025

Satoto, B. D., Khotimah, B. K., & Iswati, I. (2015). Pengelompokan wilayah madura berdasar indikator pemerataan pendidikan menggunakan partition around medoids dan validasi adjusted random index. Journal of Information Systems Engineering and Business Intelligence, 1(1), 17. https://doi.org/10.20473/jisebi.1.1.17-24

Saveliev, M., Volchek, A., Lavrenova, G., Malay, O., Grevtsev, M., & Jahatspanian, I. (2025). Determination of halitosis by exhaled breath analysis using semiconductor metal oxide sensors and chemometric methods. Journal of Chemometrics, 39(2).

Sayal, Anu, Janhvi Jha, Veethika Gupta, Ashulekha Gupta, Omdeep Gupta, Minakshi Memoria, and others. 2023. "Neural Networks and Machine Learning." In 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), , 58–63.

Shabbir, Javid. 2021. "An Introduction to Statistical Learning with Applications in R." Statistical Theory and Related Fields.

Shu, Y. (2024). Diabetes prediction based on KNN, XGBoost, SVM and LR model. Applied and Computational Engineering, 104(1), 91–95.

Sihombing, Oloan, Edwin Sitanggang, Erik Luis, and Kevin Wilmar Winata. 2023. "72 Analisis Spare Part Harbour Tag Pada Divisi Workshop Menggunakan Algoritma Knn Min-Max Scaling." Jurnal TEKINKOM 6(1).

Siti Maesaroh, Roy Mubarak, Lukman Hakim, Imam Yunianto, Siti Mutmainah, Hadi Santoso, Wiranti Sri Utami, Agung Yuliyanto Nugroho, Khairunnas, Oleh Soleh, Rahmat Oktavian, Syamsir Alam, Solihin, Bayu Waseso, Mohamad Yusuf, & Yuni Roza. (2024).

Pembelajaran Mesin dan Kecerdasan Buatan: Teori dan Aplikasi Praktis). Sada Kurnia Pustaka.

Supranto, J. (2019) Statistik: Teori dan Aplikasi. Jakarta: Erlangga, pp. 245–260.

Syed, F, dan S Lokhande. 2024. "Supervised learning: Concepts, challenges, and applications." AI and Data Analytics Journal 12(3): 45–62.

Szeliski, Richard. 2018. "Computer Vision: Algorithms and Applications, 2nd Edition." In Springer,.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to Data Mining (2nd ed.). Pearson.

Taşpınar, Y., & Çınar, İ. (2023). Prediction of sleep health status, visualization and analysis of data. ICAT.

Trisnawan, A. B. (2025). Pemanfaatan Big Data dalam Sistem Informasi untuk Pengambilan Keputusan Strategis. JISED: Journal of Information System and Education Development, 3(3), 39–43.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

Vapnik, Vladimir N. 1995. The Nature of Statistical Learning Theory. New York: Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems.

W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0

Wang, G. (2008, September). A survey on training algorithms for support vector machine classifiers. In 2008 Fourth international conference on networked computing and advanced information management (Vol. 1, pp. 123-128). IEEE.

Wardhani, R., & Budiarto, R. (2018). Pemanfaatan Teknologi Kecerdasan Buatan dalam Evaluasi Pembelajaran Berbasis Online. Jurnal Pendidikan Vokasi, 8(2), 255– 266.

Wayahdi, M. R., & Ruziq, F. (2025). Predicting smartphone addiction levels with K-nearest neighbors using user behavior patterns. Jurnal Teknik Informatika (Jutif), 6(5), 3379–3391.

Wibowo, A., & A. (2017). Penerapan Natural Language Processing untuk Analisis Sentimen pada Media Sosial. Jurnal Informatika, 1(2), 45–53.

Wicaksana, Arif, and Tahar Rachman. 2018. 3 Angewandte Chemie International Edition, 6(11), 951–952. Pattern Recognition and Machine Learning Christopher M . Bishop.

Wijoyo, A., Saputra, A. Y., Ristanti, S., Sya'ban, S. R., Amalia, M., & Febriansyah, R. (2024). Pembelajaran Machine Learning. OKTAL: Jurnal Ilmu Komputer dan Science, 3(2), 375–380.

Wiratama, I. K. (2021). Pemanfaatan Kecerdasan Buatan dalam Pendidikan: Tantangan dan Etika. Jurnal Teknologi Pendidikan, 24(1), 15–25.

Wong, J., Yamaguchi, M., Nishi, N., Araki, M., & Wee, L. (2022). Predicting overweight and obesity status among Malaysian working adults with machine learning or logistic regression: Retrospective comparison study. JMIR Formative Research, 6(12), e40404.

Yu, Teng To, Chun Yuan Chen, Tai Hsi Wu, and Yu Chen Chang. 2023. "Application of High-Dimensional Uniform Manifold Approximation and Projection (UMAP) to Cluster Existing Landfills on the Basis of Geographical and Environmental Features." Science of the Total Environment 904. doi:10.1016/j.scitotenv.2023.167013.

Yulianto, A. B., & Suryadi, D. (2020). Pemanfaatan Kecerdasan Buatan dalam Sistem Pembelajaran Jarak Jauh. Jurnal Teknologi Pendidikan, 22(2), 125–134

Yun, K., He, T., Zhen, S., Quan, M., Yang, X., Man, D., ... & Han, X. (2023). Development and validation of explainable machine-learning models for carotid atherosclerosis early screening. Journal of Translational Medicine, 21(1).

Zhang, H. 2023. "Statistical learning theory and its impact on modern machine learning." Journal of Machine Learning Perspectives 5(2): 101–18.

Zhang, Q., Yang, L. T., Chen, Z., & Li, P. 2021. A survey on deep learning for big data. Information Fusion.

Zhang, Y. (2012). Support vector machine classification algorithm and its application. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3 (pp. 179-186). Springer Berlin Heidelberg.

Zheng, Alice, and Amanda Casari. 2018. "Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists." O'Reilly.

Zhou, Zhi-Hua. 2021. Machine Learning. Springer nature.

TENTANG PENULIS

NADA ARINA ROMLI, M.I.KOM

Nada Arina Romli, S.I.Kom., M.I.Kom. lahir di Bandung, 14 September 1991. Nada menempuh pendidikan S-1 Komunikasi jurusan Public Relations di Universitas Padjadjaran, serta pendidikan S-2 Komunikasi Konsentrasi Public Relations di Universitas Padjadjaran. Saat ini Nada merupakan seorang pengajar prodi Ilmu Komunikasi, Fakultas Ilmu Sosial, Universitas Negeri Jakarta. Sebelum menjadi dosen, Nada berkecimpung sebagai praktisi di bidang perbankan dan financial technology. Nada pernah bekerja sebagai Sales Management Asst Manager di Bank Sahabat Sampoerna, kemudian menjabat sebagai CDD & EDD Compliance di Bank Standard Chartered Indonesia, serta terakhir sebagai marketing communication Asst Manager di PT Futuready Insurance Broker, part of Aegon Worldwide Group. Nada memiliki minat pada kajian komunikasi pemasaran, komunikasi bisnis, gender, new media, komunikasi digital. Pada tahun 2019, Nada meraih gelar CPR (Certified Public Relations) pada bidang stratgeic public relations dan media relations dan juga sebagai asesor kompetensi pada bidang public relations. Sebelumnya pada tahun 2017 meraih gelar sebagai Junior Public Relations Certification. Dan pada tahun 2019 meraih gelar sebagai asesor dalam bidang public relations.

AHMAD BUDI TRISNAWAN, S.T., M.KOM.

Seorang penulis dan dosen tetap Prodi Sistem Informasi pada Universitas Mahakarya Asia. Lahir di Jakarta, 7 Maret 1992. Penulis merupakan anak pertama dari dua bersaudara dari pasangan Bapak Sutrisno dan Ibu Murti. menamatkan pendidikan dasar dan menengah di Kota Tangerang, setelah lulus dari SMK Negeri 2 Kota Tangerang, kemudian melanjutkan pendidikan pada program Sarjana (S1) di Universitas Satya Negara Indonesia dengan Program Studi Teknik Informatika dan Program Pasca Sarjana (S2) di Universitas Budi Luhur dengan Program

Studi Ilmu Komputer. Penulis memiliki dua (2) buah hati yakni Ardanu Fatih Trisnawan dan Bahira Freya Trisnawan dari pasangan Budi Lestiarini, S.E. Penulis sudah menulis dan menerbitkan beberapa buku di Hadla Media Informasi sebagai sarana penuangan tulisan. Penulis bisa dihubungi via email: abudit75@gmail.com.

EKA PRASETYA ADHY SUGARA, S. T., M. KOM.

Penulis lahir di Palembang pada tanggal 24 April 1982. Penulis menamatkan pendidikan SD, SMP dan SMA di Palembang. Setelah lulus dari SMA Negeri 5 Palembang, penulis melanjutkan kuliah di Program Studi Teknik Mesin Universitas Gadjah Mada pada tahun 2000. Penulis mulai mengajar pada tahun 2009 sebagai dosen luar biasa dan bergabung menjadi dosen tetap di Progam Studi Desain Komunikasi Visual Politeknik Palcomtech (yang kemudian bergabung dengan STMIK menjadi Institut Teknologi dan Bisnis Palcomtech pada 2022). Penulis berkesempatan untuk melanjutkan studi program magister di Program Studi Teknik Informatika Universitas Bina Darma Palembang pada 2013. Pada tahun 2024, penulis dialih tempatkan di Program Studi Sistem Informasi dengan bidang keilmuan di Desain Grafis, Aplikasi Multimedia, Augmented Reality dan Virtual Reality serta Pengembangan Game. Saat ini, penulis memiliki ketertarikan di bidang Kecerdasan Buatan, Machine Learning, Deep Learning dan Generative AI. Beberapa buku yang penah ditulis diantaranya berjudul "Penganggaran untuk Taman Pendidikan Al-Qur'an (TPA)" yang merupakan luaran hibah PKM pada tahun 2019, "Cara Praktis Membangun Aplikasi Mobile" pada tahun 2021, book chapter "Visualisasi Data" pada buku "Data Science" pada tahun 2022 dan bab "Pemrosesan Data" pada buku ajar "Organisasi dan Arsitektur Komputer" pada tahun 2024 serta bab "Desain Antarmuka Pengguna (UI/UX)" pada buku ajar "Rekayasa Perangkat Lunak: Prinsip, Praktik, dan Teknologi Modern" pada tahun 2025.

A.TAQWA MARTADINATA, S.KOM., M.KOM

Penulis lahir di Lubuklinggau, Sumatera Selatan bulan Oktober 1995. Penulis menamatkan pendidikan dasar dan menengah di Lubuklinggau, setelah lulus dari SMK Negeri 3 Lubuklinggau melanjutkan kuliah S1 di STMIK Musi Rawas Jurusan Teknik Informatika.

Kemudian Pada tahun 2020 lulus S2 di universitas Bina Darma Palembang Program studi Magister Teknik Informatika. Memulai karir sebagai IT sejak tahun 2017 sebagai Technikal Support, Saat ini penulis aktif sebagai Dosen dan Konsultan di Bidang Ilmu Komputer.

Saat ini sebagai dosen di Universitas Bina Insan Lubuklinggau, mengajar untuk mata Kuliah **Pemrograman & Pengembangan Perangkat Lunak** (Rekayasa Perangkat Lunak, Pemrograman Lanjut, Pemrograman Web, Pemrograman Web Lanjutan, Pemrograman Web Berbasis Framework Lanjut, Pemrograman Perangkat Bergerak, Pemrograman Berbasis Platform, Pengantar Basis Data, Basis Data Berbasis Objek). **Jaringan & Keamanan Komputer** (Jaringan Komputer & Komunikasi Data, Pengantar Jaringan, Keamanan Informasi dan Jaringan, Pemrograman Jaringan, Praktek Kriptografi). **Kecerdasan Buatan & Sains Data** (Text Mining, Pengenalan Pola, Komputer Vision, Robotika Mobile). **Infrastruktur & Teknologi Modern (**Komputasi Awan (Cloud Computing), Teknologi IOT (Internet of Things), Sistem Informasi Geografis (SIG), Pengantar Teknologi Informasi)

Penulis juga aktif membuat video pembelajaran, salah satunya tentang pemrograman yang dapat di lihat di channel youtube penulis di ***https://www.youtube.com/martadinata.***

IMAM HALIM MURSYIDIN, S.KOM., M.KOM.

Seorang penulis dan dosen Prodi Sistem Informasi pada Universitas Islam Syekh Yusuf Tangerang. Pendidikan telah ditempuh program Sarjana (S1) Universitas Budi Luhur Prodi Sistem Informasi dan Program Pasca Sarjana (S2) di Universitas Budi Luhur prodi Ilmu Komputer. Selain itu penulis bekerja sebagai praktisi IT Auditor, *Compliance* dan *Risk* management di perusahaan swasta bidang *financial technology*. Dengan pengalaman lebih dari 7 tahun di bidang ini, penulis telah menggabungkan pengetahuan akademis dan praktisi untuk memberikan wawasan yang mendalam tentang pentingnya keamanan informasi dalam era digital. Sebagai IT Auditor penulis memiliki pengalaman langsung dalam menerapkan dan mengaudit sistem manajemen keamanan informasi yang dihadapi organisasi.

M. RHIFKY WAYAHDI, M.KOM.

Penulis lahir di Medan, 05 Februari 1993, merupakan anak pertama dari tiga bersaudara. Penulis merupakan alumni Program Sarjana (S-1) di Universitas Potensi Utama pada Jurusan Sistem Informasi dan lulus tahun 2015. Penulis melanjutkan studi Program Magister (S-2) Teknik Informatika di Universitas Sumatera Utara dan lulus tahun 2019. Kemudian saat ini Penulis sedang melanjutkan Pendidikan Doktoral (S-3) Ilmu Komputer di Universitas Sumatera Utara mulai 2024 sampai sekarang (*on-going*). Berkarir sebagai dosen dimulai dari tahun 2020 di Universitas Battuta. Penulis juga merupakan praktisi dalam pengembangan aplikasi (*software*), sebagai *founder* di PT. Technology Laboratories Indonesia.

MIFTA ARDIANTI, S.T., M.KOM

Seorang penulis dan dosen tetap Prodi Sistem Informasi Telkom University Bandung. Lahir di Gadingrejo, 27 Mei 1994. Penulis merupakan anak pertama dari pasangan Bapak Hanriadi dan Ibu Helmiyati. Pendidikan telah ditempuh program Sarjana (S1) UIN Sunan Gunung Djati Bandung Prodi Teknik Informatika dan Program Pasca Sarjana (S2) di Universitas Diponegoro prodi Sistem Informasi. Penulis mengampu mata kuliah Matematika Diskrit, Pengujian dan Implementasi Sistem dan Data Mining. Selain menjadi pengajar, penulis juga aktif dalam menjalankan tri dharma penelitian dan pengabdian masyarakat.

JONI KARMAN, M.KOM.

.Seorang penulis dan dosen tetap Prodi Sistem Informasi pada Universitas Bina Insan. Lahir di Lubuklinggau, 10 Oktober 1986. Penulis merupakan anak Ketiga dari tiga bersaudara dari pasangan Bapak Suwardi dan Ibu Ikah. Pendidikan telah ditempuh program Sarjana (S1) Universitas Bina Darma Prodi Sistem Informasi dan Program Pasca Sarjana (S2) di Universitas Bina Darma prodi Teknik Informatika. Penulis memiliki dua (2) buah hati Muhannad Sulthan Tsaqif dan M. Uwais Al Linggowi dari pasangan Mutma Innah, SPd. Gr. Berikut judul buku yang telah ditulis dan terbitkan: Sistem Informasi Geografis Berbasis Android dan Sistem Informasi Geografis Berbasis Web.

MUHAMMAD EDYA ROSADI, S.KOM., M.KOM.

Dosen Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin (UNISKA Banjarmasin).

Menyelesaikan pendidikan Sarjana (S1) Teknik Informatika di UNISKA Banjarmasin (S.Kom.) dan Magister (S2) Teknik Informatika di Universitas Dian Nuswantoro Semarang (M.Kom.). Aktif melakukan penelitian dan publikasi ilmiah dalam bentuk buku serta artikel pada jurnal internasional dan nasional bereputasi. Berkomitmen untuk terus berkontribusi dalam pengembangan ilmu pengetahuan, khususnya di bidang teknologi informasi..

AHMAD KHUSAERI, M.KOM.

Seorang penulis dan dosen tetap Prodi Sistem Informasi pada Universitas Singaperbangsa KArawang. Lahir di Karawang, 24 Januari 1997. Penulis merupakan anak Pertama dari empat bersaudara dari pasangan Bapak Nanang Supriatna dan Ibu Eti Hernawati. Pendidikan telah ditempuh program Sarjana (S1) Universitas Singaperbangsa Karawang Prodi Teknik Informatika dan Program Pasca Sarjana (S2) di Institut Pertanian Bogor prodi Ilmu Komputer. Penulis memiliki seorang buah hati Azkiya Almira dari pasangan Qonita Robiyatul Adawiyah. Berikut judul buku yang telah ditulis dan terbitkan: Multimedia dengan Adobe Photoshop dan CorelDraw.

NOVI LESTARI, M.KOM.

Seorang penulis dan dosen tetap Prodi Rekayasa Sistem Komputer pada Universitas Bina Insan. Lahir di Lubuklinggau, 21 Agustus 1987. Penulis merupakan anak Kedua dari enam bersaudara dari pasangan Bapak Rasman Simanjuntak dan Ibu Siti Aisyah. Pendidikan telah ditempuh Program Sarjana (S1) Sekolah Tinggi Manajemen dan Ilmu Komputer (STMIK) Musi Rawas Prodi Sistem Komputer dan Program Pasca Sarjana (S2) di Universitas Bina Darma Palembang Prodi Magister Teknik Informatika. Penulis memiliki tiga (3) buah hati Anindita Prameswari, Adnan Giri Warabrata dan Adistia Pramesti dari pasangan Satrianansyah. Berikut judul buku ajar yang telah ditulis dan diterbitkan : Jaringan Komputer : Dari Teori Dasar Hingga Jaringan Nirkabel

.

BUDI BERLINTON SITORUS S.T, M.SC

Penulis lahir di Jakarta, DKI Jakarta. Penulis menamatkan pendidikan dasar dan menengah di Jakarta. Setelah lulus dari SMA St. Antonius melanjutkan kuliah S1 di STT Telkom Bandung Jurusan Informatika, kemudian melanjutkan S2 di Greenwich University, Inggris, Jurusan Distributed Computer Systems. Pada tahun 2001 lulus S2 dan memulai karir sebagai dosen paruh waktu di Universitas Bina Nusantara tahun 2002. Saat ini sebagai dosen homebase di Universitas Multimedia Nusantara

DR. MUHAMAD AKBAR, S.T.,M.IT

Seorang penulis dan dosen tetap Prodi Teknik Informatika pada Universitas Bina Insan Lubuklinggau. Lahir di Bandung, 17 Februari 1972. Penulis merupakan anak Kedua dari empat bersaudara dari pasangan Bapak Aidi Sani dan Ibu Sofiatin. Pendidikan telah ditempuh program Sarjana (S1) ST.INTEN, Bandung Prodi Teknik Informatika, Program Magister (S2) di CURTIN University, Australia Barat Prodi Internet Studies dan Program Pasca Sarjana (S3) di Universitas Sriwijaya prodi Teknik Ilmu Teknik BKU Teknik Informtika. Penulis memiliki dua (2) buah hati Kayla Aurelia dan Davina Ramadhani dari pasangan Nahdia.

MUHAMMAD IRVAI, M.KOM.

adalah seorang dosen di Program studi Informatika Fakultas Ilmu Teknik, Universitas Bina Insan, Lubuklinggau, Indonesia. Beliau meraih gelar sarjana Teknik Informatika dari STMIK Musi Rawas Kota Lubuklinggau pada tahun 2017. Kemudian menyelesaikan studi magister Teknik Informatika di Universita Bina Dharma pada tahun 2021. Bidang penelitian beliau meliputi pengembangan aplikasi, kriptografi dan pembelajaran mesin.

TENTANG EDITOR

NURHADI, S.KOM., M.KOM

lahir di Bekasi, Jawa Barat bulan Nopember 1978. Beliau menamatkan pendidikan dasar dan menengah di Bekasi, setelah lulus dari STM Negeri 1 Bekasi (sekarang SMK Neger 1 Cikarang Barat) melanjutkan kuliah D3 di STMIK Pranata Indonesia Jurusan Managemen Informatika.

kemudian melanjutkan S1 di STMIK Pranata Indonesia Jurusan Sistem informasi. Pada tahun 2015 lulus S2 di universitas Budi Luhur Jakarta Program studi Magister Ilmu Komputer.

Beliau banyak mengikuti workshop tentang editor dan penulis buku yang di selenggarakan oleh lembaga pemerintahan maupun lembaga swasta. Saat ini sebagai dosen jurusan sistem informasi di STMIK Pranata Indonesia, mengajar untuk mata Kuliah Bahasa Pemrograman 1 dan Bahasa Pemrograman 2. Penulis juga aktif membuat video pembelajaran, salah satunya tentang pemrograman visual basic dan SQL server yang dapat di lihat di channel youtube penulis di ***https.//bit.ly/PDMVBSQL.***