PENGANTAR DATA SCIENCE

KONSEP, METODE, DAN IMPLEMENTASI

ISMAIL SETIAWAN, SETIANGGA FACHRURROZI, SIGIT SETYOWIBOWO, SUKMA PUSPITORINI, FERY PURNAMA, DWI ISMIYANA PUTRI, M. FAUZAN AZHARI, CYNTIA RIVATUNISA, SRI HANDAYANI, DETIN SOFIA, FATTACHUL HUDA AMINUDDIN, LANI NURLANI, RENY WAHYUNING ASTUTI, ANITA RATNASARI, MIFTAH FAROQ SANTOSO

PENGANTAR DATA SCIENCE : KONSEP, METODE, DAN IMPLEMENTASI

UU No. 28 tahun 2014 tentang Hak Cipta

Fungsi dan Sifat Hak Cipta Pasal 4

Hak cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Pelindungan Pasal 26

Ketentuan sebagaimana dimaksud dalam pasal 23, pasal 24, dan pasal 25 tidak berlaku terhadap:

i. Penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan infrmasi aktual;
ii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
iii. Penggandaan Ciptaan dan/atau produk hak Terkait hanya untuk keperluan pengjaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
iv. Penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

Sanksi Pelanggaran Pasal 113

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dala pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidanan dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp.100.000.000,00 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tenpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaiman 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).

PENGANTAR DATA SCIENCE : KONSEP, METODE, DAN IMPLEMENTASI

ISMAIL SETIAWAN, SETIANGGA FACHRURROZI, SIGIT SETYOWIBOWO, SUKMA PUSPITORINI, FERY PURNAMA, DWI ISMIYANA PUTRI, M. FAUZAN AZHARI, CYNTIA RIVATUNISA, SRI HANDAYANI, DETIN SOFIA, FATTACHUL HUDA AMINUDDIN, LANI NURLANI, RENY WAHYUNING ASTUTI, ANITA RATNASARI, MIFTAH FAROQ SANTOSO

PT. SURGA DIGITAL NUSANTARA

PENGANTAR DATA SCIENCE : KONSEP, METODE, DAN IMPLEMENTASI

Ismail Setiawan, Setiangga Fachrurrozi, Sigit Setyowibowo, Sukma Puspitorini, Fery Purnama, Dwi Ismiyana Putri, M. Fauzan Azhari, Cyntia Rivatunisa, Sri Handayani, Detin Sofia, Fattachul Huda Aminuddin, Lani Nurlani, Reny Wahyuning Astuti, Anita Ratnasari, Miftah Faroq Santoso

ISBN : 978-634-05-4067-3

Editor : Miftahul Jannah
Penyunting : Ilham Yani Putra
Desain Sampul : Raudatul Jannah

Penerbit
PT. Surga Digital Nusantara

Redaksi
Jalan Bathin Alam, Bengkalis Kota, Kec. Bengkalis, Kab. Bengkalis, Riau

Distributor Tunggal
PT. Surga Digital Nusantara
Jalan Bathin Alam, Bengkalis Kota, Kec. Bengkalis, Kab. Bengkalis, Riau

KATA PENGANTAR

السّلام عليكم ورحمةالله وبركاته

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga buku "Pengantar Data Science: Konsep, Metode, dan Implementasi" ini dapat diselesaikan dengan baik. Buku ini disusun sebagai sumber pembelajaran dan referensi bagi mahasiswa, akademisi, serta masyarakat yang ingin memahami konsep, metode, dan penerapan Data Science dalam berbagai bidang.

Perkembangan teknologi digital telah menjadikan data sebagai aset yang sangat berharga. Melalui pendekatan Data Science, data dapat diolah menjadi informasi dan pengetahuan yang mendukung pengambilan keputusan secara efektif. Oleh karena itu, buku ini membahas berbagai materi penting, mulai dari konsep dasar Data Science, data dan informasi, statistika, pemrograman, machine learning, data mining, hingga implementasi Data Science dalam berbagai sektor.

Penulis berharap buku ini dapat memberikan manfaat, memperluas wawasan, serta menjadi dasar yang kuat bagi pembaca dalam mempelajari dan mengembangkan kompetensi di bidang Data Science. Semoga kehadiran buku ini dapat turut mendukung peningkatan literasi dan pemanfaatan Data Science di lingkungan akademik maupun dunia kerja.

Bengkalis, Mei 2025

Penulis

DAFTAR ISI

BAB 1
KONSEP DASAR DATA SCIENCE

A. PENDAHULUAN

1. Latar belakang Ledakan Data (*Big Data)* di Era Digital

Kita saat ini hidup di tengah pusaran revolusi digital yang masif. Dalam dua dekade terakhir, transisi aktivitas manusia dari ranah fisik ke ekosistem digital telah memicu fenomena yang sering disebut sebagai ledakan data (data explosion). Setiap detik, miliaran bita data diproduksi dan direkam melalui interaksi kita dengan perangkat pintar, media sosial, transaksi e-commerce, sensor Internet of Things (IoT), hingga sistem informasi manajemen berbasis digital.

Di masa lalu, pengumpulan dan penyimpanan data merupakan proses yang mahal dan umumnya terbatas pada data yang terstruktur rapi. Namun, dengan kemajuan teknologi komputasi awan (cloud computing) dan penurunan drastis biaya infrastruktur penyimpanan, organisasi dan institusi kini mampu mengakumulasi data dalam skala Big Data. Skala ini dikarakteristikkan oleh Volume (ukuran data yang sangat besar), Velocity (kecepatan pergerakan dan penciptaan data yang tinggi), serta Variety (keberagaman format data, mulai dari teks, angka, hingga gambar dan video).

Kumpulan data raksasa ini ibarat lautan informasi; ia memiliki potensi nilai yang luar biasa. Namun, layaknya minyak mentah yang baru ditambang, data tersebut sama sekali tidak memiliki nilai guna jika hanya dibiarkan menumpuk di dalam pusat data tanpa diolah.

2. Mengapa *Data Science* menjadi bidang yang krusial saat ini

Keberadaan data yang melimpah melahirkan sebuah tantangan dan peluang baru: bagaimana mengubah lautan angka dan teks mentah tersebut menjadi wawasan (insights) yang bermakna dan dapat ditindaklanjuti? Di sinilah Data Science hadir sebagai jembatan solusi. Data Science menjadi sangat krusial karena ia membekali kita dengan metodologi ilmiah, algoritma komputasi, dan landasan statistik untuk mengekstrak pengetahuan dari tumpukan data tersebut.

Ada beberapa alasan mendasar mengapa bidang ini kini menjadi tulang punggung di hampir seluruh sektor industri dan riset:

a. Transisi dari Analisis Reaktif ke Prediktif dan Proaktif

   Secara tradisional, organisasi menggunakan data hanya untuk melihat apa yang sudah terjadi di masa lalu (analitik deskriptif), seperti membuat laporan akhir tahun. Data Science memungkinkan lompatan ke arah analitik prediktif. Sebagai contoh, dalam bidang pendidikan melalui Educational Data Mining, institusi tidak lagi sekadar merekap nilai di akhir semester. Dengan memodelkan rekam jejak akademis dan pola historis mahasiswa, algoritma dapat digunakan untuk memprediksi secara akurat risiko seorang mahasiswa mengalami dropout secara dini. Hal ini memungkinkan pihak kampus untuk melakukan intervensi proaktif dan memberikan pendampingan sebelum hal terburuk terjadi.

b. Otomatisasi dan Akurasi Skala Besar

   Berbagai sektor seperti kesehatan, finansial, dan bisnis dihadapkan pada kompleksitas masalah yang tidak lagi bisa diselesaikan secara manual. Model algoritma

dalam Data Science mampu mengotomatisasi proses pengambilan keputusan yang rumit, seperti deteksi transaksi penipuan (fraud), personalisasi rekomendasi produk, hingga interpretasi citra medis, dengan kecepatan dan tingkat akurasi yang jauh melampaui kemampuan pemrosesan manusia.

c. Pengambilan Keputusan Berbasis Bukti (Data-Driven Decision Making)

Intuisi, asumsi, atau sekadar tebakan (guesswork) tidak lagi memadai untuk bertahan di era modern. Para pemimpin, peneliti, dan pengambil kebijakan saat ini sangat bergantung pada hasil analisis data yang objektif dan teruji secara statistik. Data Science memastikan bahwa setiap strategi, alokasi sumber daya, dan kebijakan yang diambil memiliki landasan bukti analitik yang solid, sehingga dapat meminimalkan risiko kegagalan.

Singkatnya, Data Science bukan sekadar tren komputasi sesaat, melainkan sebuah kebutuhan fundamental di abad ke-21. Kemampuan untuk menguasai siklus data mulai dari proses ekstraksi, pemodelan matematis, hingga interpretasi algoritmik merupakan kunci utama untuk melahirkan inovasi dan memecahkan permasalahan kompleks di era yang sepenuhnya digerakkan oleh informasi ini.

B. DEFINISI *DATA SCIENCE*

1. Pengertian Data Science Menurut Para Ahli

Meskipun istilah Data Science terdengar modern, konsep dasarnya telah diperdebatkan dan dirumuskan oleh banyak pakar selama beberapa dekade. Untuk memahami fondasi ilmu ini, kita dapat merujuk pada beberapa definisi dari para ahli:

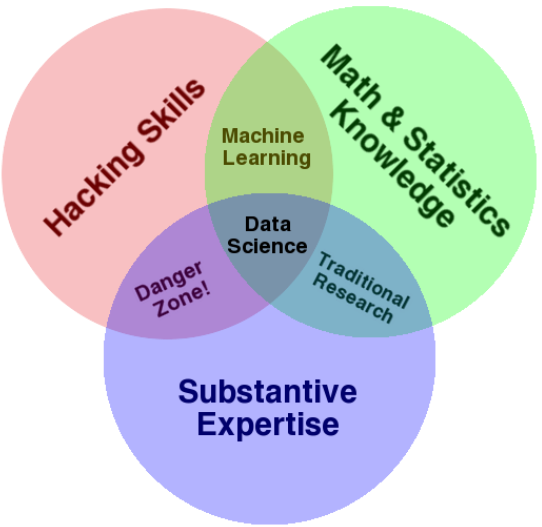
a. Chikio Hayashi (Hayashi, 1998) Salah satu pionir yang mempopulerkan istilah ini mendefinisikan Data Science sebagai sebuah konsep terpadu (sintesis) yang tidak hanya menyatukan statistika dan analisis data, tetapi juga mencakup metodologi serta hasil dari analisis tersebut. Menurut Hayashi, Data Science adalah ilmu yang bertujuan untuk memahami fenomena dunia nyata melalui data.
b. Vasant Dhar (Dhar, 2013) Seorang profesor dari Stern School of Business, mendefinisikan Data Science sebagai studi mengenai ekstraksi pengetahuan yang dapat digeneralisasi dari data (generalizable extraction of knowledge from data). Dhar menekankan pada pergeseran dari analitik tradisional menuju model prediktif yang kokoh.
c. Hal Varian (Varian, 2014) Chief Economist di Google, menyatakan bahwa kemampuan untuk mengambil data memahaminya, memprosesnya, mengekstrak nilai darinya, memvisualisasikannya, dan mengomunikasikannya akan menjadi keterampilan yang sangat penting di dekade mendatang.

Secara konsensus, Data Science dapat disimpulkan sebagai sebuah disiplin ilmu empiris yang memanfaatkan metode saintifik, algoritma komputasi, dan sistem arsitektur data untuk mengekstraksi wawasan berharga yang tersembunyi di balik data berskala besar.

2. *Data Science* Sebagai Perpaduan Antara Statistika, Matematika, Pemrograman, Dan Domain Knowledge

Data Science bukanlah disiplin ilmu yang berdiri sendiri, melainkan sebuah ilmu multidisipliner (irisan dari berbagai bidang keahlian). Konsep ini paling baik dijelaskan melalui

kerangka berpikir yang diperkenalkan oleh Drew Conway melalui Data Science Venn Diagram.



Gambar 1. Venn Diagram Data Science dari Drew Conway

Untuk menjadi seorang praktisi Data Science yang utuh, diperlukan sinergi dari ketiga elemen dasar berikut:

a. Matematika dan Statistika

Statistika memberikan kerangka kerja untuk menguji hipotesis, memahami distribusi data, dan memastikan bahwa temuan yang didapat memiliki signifikansi secara ilmiah, bukan sekadar kebetulan. Matematika (terutama aljabar linear dan kalkulus) menjadi fondasi dari cara kerja algoritma Machine Learning dan Deep Learning. Kombinasi ke dua ilmu tersebut ibarat "mesin" dari Data Science.

b. Pemrograman dan Ilmu Komputer (Keterampilan Teknis)

Mengingat volume data saat ini terlalu besar untuk dihitung secara manual, kemampuan menulis kode (seperti Python, R, atau SQL) sangat mutlak. Keterampilan ini mencakup manipulasi data,

perancangan algoritma, hingga pengelolaan basis data skala besar. Ibarat ingin bepergian ke suatu tempat maka kemampuan pemrograman adalah kendaraan untuk memproses data.

c. Pengetahuan Ranah (Domain Knowledge)

Tanpa pemahaman konteks bisnis atau penelitian, data hanyalah deretan angka kosong. Sebagai contoh, dalam penerapan Educational Data Mining, kemampuan teknis dalam menjalankan algoritma komputasi tingkat tinggi tidak akan memberikan hasil yang bermakna tanpa diimbangi pemahaman mendalam tentang ekosistem akademik. Seorang analis harus memahami regulasi kampus, dinamika kurikulum, serta faktor-faktor demografis mahasiswa untuk dapat mendefinisikan dan memprediksi risiko dropout dengan tepat sasaran.

3. Perbedaan *Data Science* dengan disiplin lain (Statistika Klasik, Ilmu Komputer, Business Intelligence)

Seringkali terdapat kebingungan mengenai batas-batas antara Data Science dan bidang terkait lainnya. Tabel berikut menjabarkan perbedaan prinsipil antara Data Science dengan Statistika Klasik, Ilmu Komputer, dan Business Intelligence (BI):

Tabel 1. Perbedaan Data Science dengan Disiplin Lain

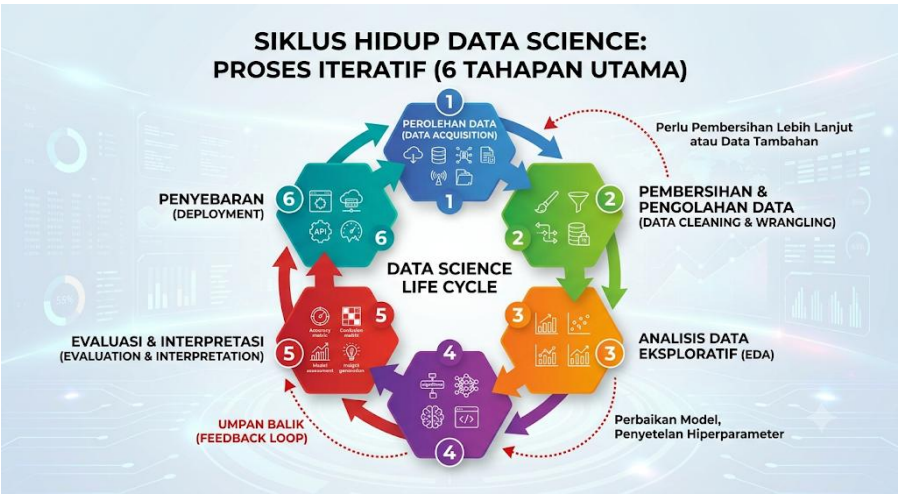
Aspek Pembeda	***Data Science***	**Statistika Klasik**	**Ilmu Komputer**	**Business Intelligence (BI)**
Fokus Utama	Analisis prediktif (memprediksi masa depan) dan penemuan pola tersembunyi.	Analisis inferensial (memahami populasi dari sampel) dan uji hipotesis.	Pengembangan perangkat lunak, arsitektur sistem, dan efisiensi algoritma komputasi.	Analisis deskriptif (melaporkan apa yang telah terjadi di masa lalu).

Karakteristik Data	Menangani data bervolume raksasa (Big Data), baik terstruktur maupun tidak terstruktur.	Umumnya menangani data terstruktur dengan ukuran sampel yang relatif lebih kecil dan terkontrol.	Menitikberatkan pada cara data disimpan, ditransmisikan, dan diproses oleh mesin.	Berfokus pada data historis terstruktur yang ada di dalam data warehouse.
Metode / Pendekatan	Machine Learning, Deep Learning, algoritma heuristik, metrik performa model (akurasi, dll).	Uji probabilitas, nilai-p (p-value), interval kepercayaan, distribusi probabilitas teoritis.	Pemrograman berorientasi objek, struktur data, rekayasa jaringan, dan keamanan logis.	Pembuatan dashboard, visualisasi matriks kinerja utama (KPI), Querying (SQL).
Pertanyaan yang Dijawab	"Berdasarkan pola ini, apa yang kemungkinan besar akan terjadi selanjutnya?"	"Apakah ada perbedaan yang signifikan secara statistik antara kelompok A dan B?"	"Bagaimana cara membangun sistem yang stabil untuk memproses perintah ini?"	"Berapa jumlah penjualan atau tingkat kehadiran pada kuartal lalu?"

Meskipun berbeda fokus, Data Science tidak berusaha menggantikan disiplin-disiplin tersebut. Sebaliknya, Data Science justru berdiri di atas bahu disiplin-disiplin klasik ini untuk menawarkan solusi komputasi cerdas yang lebih modern dan adaptif terhadap tantangan era digital.

C. RUANG LINGKUP DAN KOMPONEN UTAMA *DATA SCIENCE*

Secara operasional, Data Science bukanlah sebuah proses tunggal yang instan, melainkan sebuah siklus hidup (lifecycle) atau alur kerja (pipeline) yang berkesinambungan. Untuk mengubah data mentah menjadi wawasan (insight) atau sistem kecerdasan buatan yang beroperasi penuh, seorang praktisi harus melalui beberapa tahapan utama.



Gambar 2. Alur Kerja/Siklus Hidup Data Science

Ruang lingkup komprehensif dari Data Science mencakup enam komponen esensial berikut:

1. **Pengumpulan Data (Data Acquisition)**

   Tahap pertama bermula dari proses identifikasi, ekstraksi, dan pengumpulan data dari berbagai sumber. Data dapat berasal dari basis data internal perusahaan (seperti SQL), antarmuka pemrograman aplikasi (API), sensor IoT, hingga hasil penarikan data dari situs web (web scraping).

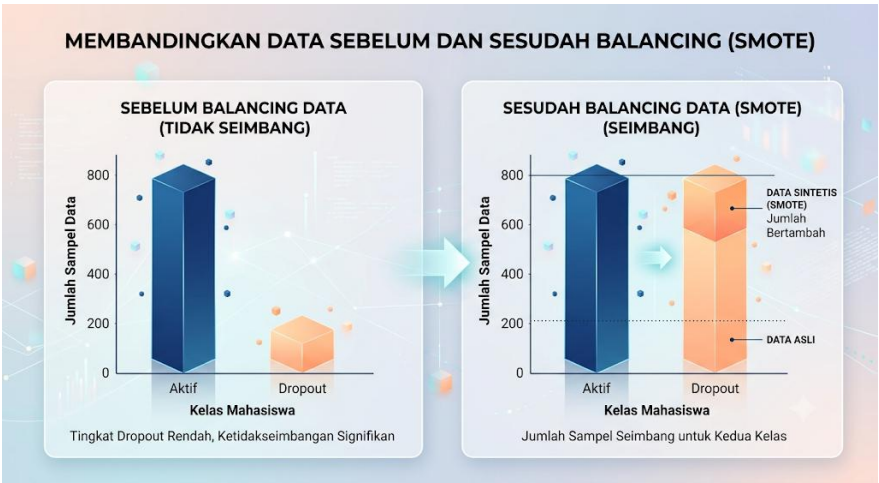
   Sebagai contoh, dalam penelitian Educational Data Mining

untuk memprediksi risiko mahasiswa dropout di program studi ilmu kesehatan, pengumpulan data dapat mencakup penarikan data longitudinal mahasiswa mulai dari jalur masuk, rekam jejak nilai tiap semester, hingga status akademik yang diekstraksi dari sistem informasi akademik universitas selama rentang waktu beberapa tahun (misalnya periode 2015–2024).

2. **Pemrosesan Dan Pembersihan Data (Data Cleaning & Wrangling)**

Sering kali dikatakan bahwa praktisi Data Science menghabiskan hingga 80% waktunya di tahap ini. Data mentah di dunia nyata umumnya tidak sempurna; sering kali mengandung nilai yang hilang (missing values), anomali (outliers), atau format yang tidak konsisten. Data wrangling adalah proses merapikan dan mentransformasi data mentah ke dalam format yang terstruktur.

Salah satu tantangan paling umum dalam tahap ini adalah disparitas kelas data (class imbalance). Misalnya, jumlah mahasiswa yang berstatus aktif jauh lebih banyak dibandingkan yang berstatus dropout. Jika data ini langsung dimodelkan, algoritma akan menjadi bias terhadap kelas mayoritas. Oleh karena itu, diperlukan teknik penanganan khusus, seperti penerapan metode SMOTE (Synthetic Minority Over-sampling Technique) untuk mensintesis data pada kelas minoritas, sehingga set data menjadi seimbang sebelum masuk ke tahap pemodelan.



Gambar 3. Ilustrasi Penanganan Disparitas Kelas (SMOTE)

3. **Eksplorasi Dan Visualisasi Data (Exploratory Data Analysis - EDA)**

Setelah data bersih, langkah selanjutnya adalah memahami pola fundamental di dalamnya melalui Exploratory Data Analysis (EDA). Pada tahap ini, pendekatan statistika deskriptif dan visualisasi grafik (seperti scatter plot, histogram, atau boxplot) digunakan untuk menemukan korelasi antar variabel.

EDA membantu analis merumuskan hipotesis. Misalnya, melalui visualisasi, peneliti mungkin menemukan bahwa terdapat penurunan drastis pada nilai indeks prestasi semester tiga yang berkorelasi kuat dengan keputusan mahasiswa untuk berhenti kuliah di semester berikutnya.

4. **Pemodelan (Machine Learning, Statistika Inferensial)**

Tahap ini merupakan inti komputasi dari Data Science. Berbekal data yang telah disiapkan, algoritma matematika dan Machine Learning dilatih untuk mengenali pola demi melakukan klasifikasi, regresi, atau klasterisasi.

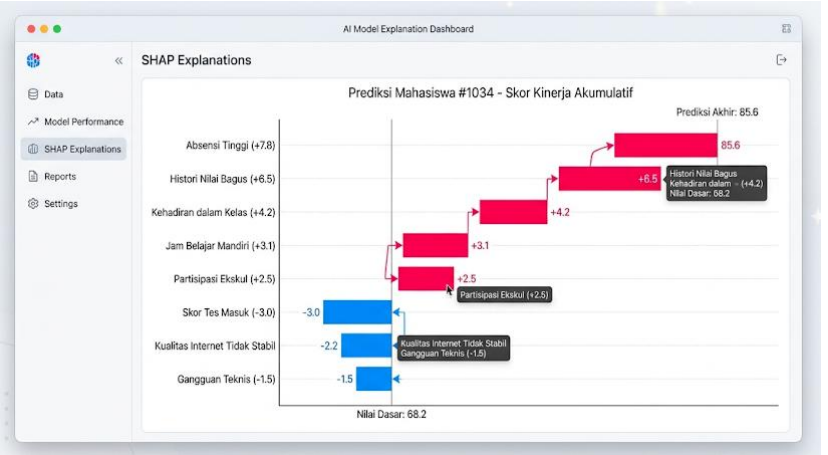
Pemilihan algoritma sangat bergantung pada jenis masalah.

Untuk kasus klasifikasi prediktif yang kompleks dengan data tabular, pemodelan sering kali memanfaatkan algoritma berbasis tree yang mutakhir dan berkinerja tinggi, seperti XGBoost atau LightGBM. Algoritma-algoritma ini mampu menangani hubungan non-linear antar variabel dengan sangat baik dan sering kali menjadi standar industri (state-of-the-art) dalam membangun pemodelan prediktif yang akurat.

5. Evaluasi Dan Interpretasi Hasil

Model yang telah dilatih tidak dapat langsung digunakan tanpa melalui pengujian yang ketat. Evaluasi dilakukan menggunakan teknik seperti cross-validation dan penyetelan ambang batas (threshold tuning) untuk memastikan model tidak hanya menghafal data (overfitting), melainkan mampu melakukan generalisasi pada data baru. Kinerja dievaluasi melalui metrik seperti Accuracy, Precision, Recall, atau F1-Score. Mengoptimalkan nilai Recall sering kali sangat krusial agar sistem tidak melewatkan deteksi kasus positif, seperti gagal mendeteksi mahasiswa yang sebenarnya berisiko tinggi.

Lebih dari sekadar metrik performa, interpretasi hasil (Explainable AI atau XAI) kini menjadi standar wajib, terutama untuk aplikasi pengambilan keputusan yang krusial. Model yang kompleks seperti XGBoost sering kali dianggap sebagai "kotak hitam" (black box). Oleh karena itu, metode seperti SHAP (SHAPley Additive exPlanations) digunakan. Dengan memanfaatkan SHAP summary plots dan waterfall plots, pengguna dapat memahami secara transparan fitur spesifik apa misalnya penurunan nilai praktikum atau status pembayaran yang paling memengaruhi keputusan model untuk satu kasus tertentu.



Gambar 4. Interpretasi Model (SHAP Waterfall Plot)

6. Deployment Dan Pemeliharaan Model

Langkah terakhir adalah deployment, yaitu mengimplementasikan model yang telah tervalidasi ke dalam lingkungan produksi (misalnya mengintegrasikannya ke dalam portal akademik web atau dasbor aplikasi). Tujuan akhirnya adalah agar model dapat memproses data baru secara real-time atau batch, serta memberikan luaran yang dapat diakses oleh pengguna akhir (seperti dosen pembimbing atau manajemen fakultas).

Setelah deployment, model juga memerlukan pemeliharaan (maintenance). Kinerja model harus terus dipantau karena seiring berjalannya waktu, karakteristik data bisa berubah (data drift), yang mengharuskan model untuk dilatih ulang menggunakan data terbaru guna mempertahankan akurasinya.

D. SIKLUS HIDUP *DATA SCIENCE* (*DATA SCIENCE LIFECYCLE*)

Sebuah proyek Data Science bukanlah sekumpulan aktivitas pemrograman yang dilakukan secara acak. Untuk memastikan bahwa hasil analitik relevan, akurat, dan dapat diimplementasikan, diperlukan sebuah metodologi yang terstruktur. Metodologi ini memandu praktisi mulai dari pemahaman masalah di dunia nyata hingga penggelaran model ke dalam sistem produksi.

1. CRISP-DM (Cross-Industry Standard Process for Data Mining)

Terdapat berbagai kerangka kerja dalam ilmu data, namun yang paling luas diadopsi sebagai standar emas di industri dan akademis adalah CRISP-DM. Diperkenalkan pertama kali pada akhir 1990-an, metodologi ini tetap relevan hingga era Deep Learning saat ini karena sifatnya yang tangguh, fleksibel, dan tidak bergantung pada perangkat lunak atau industri tertentu (industry-agnostic).

Kekuatan utama CRISP-DM terletak pada penekanannya terhadap iterasi. Proses analitik jarang berjalan linear dari awal hingga akhir; praktisi sering kali harus kembali ke tahap sebelumnya misalnya, kembali ke persiapan data setelah menyadari ada anomali saat fase pemodelan untuk menyempurnakan hasil.

2. Model Siklus Umum

Secara konseptual, siklus hidup Data Science melalui pendekatan CRISP-DM dijabarkan ke dalam enam fase berkesinambungan:

a. Pemahaman Bisnis/Masalah (Business Understanding)

Fase paling krusial di mana tujuan proyek

didefinisikan dari perspektif domain (bukan perspektif komputasi). Tujuannya adalah mengonversi masalah bisnis atau riset ke dalam formulasi teknis Data Science.

b. Akuisisi Data (Data Acquisition/Data Understanding)

   Proses mengumpulkan sumber data awal, mendeskripsikan volume dan formatnya, serta melakukan eksplorasi awal untuk memverifikasi apakah data tersebut cukup untuk menjawab masalah yang telah dirumuskan.

c. Persiapan Data (Data Preparation)

   Mengonsumsi waktu terbanyak dalam proyek, fase ini mencakup ekstraksi, transformasi, pembersihan (menangani nilai kosong), rekayasa fitur (feature engineering), dan penanganan ketidakseimbangan kelas (class imbalance).

d. Pemodelan (Modeling)

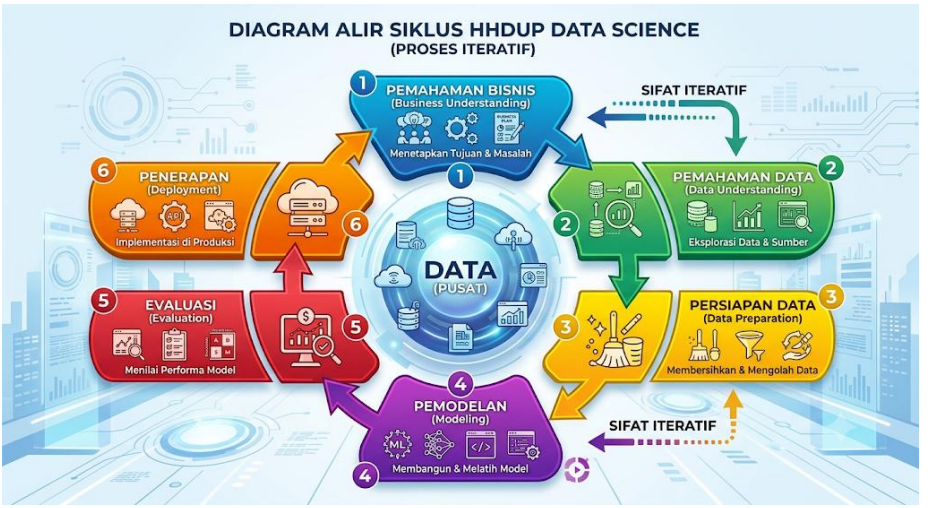
   Menerapkan berbagai teknik algoritma (seperti regresi, klasifikasi, atau clustering) pada data yang telah disiapkan. Model dieksperimenkan dengan kalibrasi parameter yang berbeda untuk mencari performa komputasi yang paling optimal.

e. Evaluasi (Evaluation)

   Mengevaluasi model tidak hanya berdasarkan metrik teknis statistika, tetapi juga memvalidasi apakah model tersebut benar-benar memecahkan masalah objektif yang ditetapkan pada fase pertama.

f. Penggelaran (Deployment)

   Mentransformasikan wawasan atau model menjadi langkah taktis operasional. Model diintegrasikan ke dalam infrastruktur IT (dashboard, API, aplikasi) agar dapat digunakan secara praktis oleh pengguna akhir.



Gambar 4. Model Siklus Umum Berdasarkan CRISP-DM

3. Contoh Sederhana Siklus Dalam Proyek Nyata

Untuk memberikan gambaran yang lebih konkret, mari kita telusuri penerapan keenam fase tersebut dalam sebuah studi kasus nyata: Sistem Prediksi Risiko Mahasiswa Dropout di Program Studi Ilmu Kesehatan.

a. Fase 1

Pemahaman Masalah: Pihak fakultas ingin menurunkan angka putus kuliah. Objektif analitiknya adalah membangun model prediktif yang dapat mengidentifikasi mahasiswa berisiko dropout sedini mungkin, sehingga intervensi akademis dapat dilakukan. (Catatan: Dalam merumuskan publikasi, nama sistem dibuat ringkas dan padat sesuai standar jurnal sasaran).

b. Fase 2

Akuisisi Data: Mengumpulkan data longitudinal mahasiswa dari rekam jejak akademik, demografi, hingga presensi, misalnya dari periode tahun 2015 hingga 2024. Data benchmark dari repositori standar, seperti UCI

Predict Students Dropout and Academic Success dataset, juga disiapkan sebagai basis komparasi model.

c. Fase 3

Persiapan Data: Data mentah dibersihkan. Karena jumlah mahasiswa yang dropout jauh lebih sedikit daripada yang lulus, terjadi ketidakseimbangan kelas. Teknik seperti SMOTE diaplikasikan untuk mensintesis data kelas minoritas. Dalam eksperimen, parameter diuji silang (cross-validation), misalnya membandingkan nilai parameter k-nearest neighbors dari k=1 hingga k=5 untuk menemukan distribusi paling optimal.

d. Fase 4

Pemodelan: Memanfaatkan algoritma klasifikasi. Untuk data tabular yang kompleks seperti riwayat akademis, algoritma ensemble tree seperti XGBoost atau LightGBM dilatih untuk memisahkan mahasiswa berisiko tinggi dan berisiko rendah.

e. Fase 5

Evaluasi: Performa dievaluasi menggunakan metrik seperti Recall untuk memastikan sistem tidak "kebocoran" dalam mendeteksi mahasiswa yang sebenarnya berisiko tinggi. Lebih lanjut, teknik Explainable AI (seperti SHAP values) digunakan agar keputusan algoritma bisa diinterpretasikan misalnya, model menemukan bahwa penurunan drastis pada indeks prestasi semester dua adalah faktor dominan risiko dropout.

f. Fase 6

Penggelaran: Model Machine Learning diintegrasikan ke dalam sebuah dasbor atau antarmuka pelaporan. Dosen pembimbing dapat memantau indikator risiko mahasiswa perwaliannya secara real-time dan

memberikan bimbingan proaktif berbasis data.

E. PERAN DAN PROFESI DALAM *DATA SCIENCE*

Seiring dengan meningkatnya kompleksitas data di berbagai institusi dan industri, Data Science tidak lagi dipandang sebagai pekerjaan satu orang (one-man show). Pemrosesan data dari bentuk mentah hingga menjadi sebuah sistem cerdas yang beroperasi penuh membutuhkan kolaborasi lintas disiplin. Oleh karena itu, profesi di bidang data telah berevolusi menjadi beberapa spesialisasi dengan peran, tanggung jawab, dan keahlian yang spesifik.

1. Data Scientist

Data Scientist adalah motor penggerak utama dalam pemodelan analitik tingkat lanjut. Peran ini berfokus pada analitik prediktif dan preskriptif berusaha menjawab pertanyaan "Apa yang akan terjadi di masa depan, dan tindakan apa yang harus diambil?". Seorang Data Scientist merancang eksperimen, melatih algoritma Machine Learning, dan melakukan evaluasi model yang mendalam. Sebagai contoh, dalam riset Educational Data Mining, seorang Data Scientist bertugas melatih algoritma prediktif berkinerja tinggi, seperti XGBoost atau LightGBM, untuk memproyeksikan risiko dropout mahasiswa. Lebih jauh lagi, praktisi ini juga dituntut untuk mampu menjelaskan logika di balik keputusan algoritma tersebut menggunakan teknik Explainable AI (XAI), seperti interpretasi nilai SHAP, agar hasil prediksinya dapat dipercaya secara akademis dan klinis.

2. Data Analyst

Jika Data Scientist berfokus pada masa depan, seorang Data Analyst bertugas membedah masa lalu. Peran ini berfokus pada analitik deskriptif dan diagnostik untuk menjawab pertanyaan

"Apa yang telah terjadi, dan mengapa hal itu terjadi?". Data Analyst mengambil data historis yang sudah ada, melakukan query, memvisualisasikan tren, dan menyusun laporan. Dalam konteks institusi pendidikan, seorang Data Analyst mungkin bertugas menganalisis tren penurunan rata-rata Indeks Prestasi Kumulatif (IPK) angkatan tertentu dan menyajikannya dalam bentuk grafik yang mudah dipahami oleh pemangku kebijakan kampus.

3. Data Engineer

Tanpa Data Engineer, peran Data Scientist dan Data Analyst tidak akan dapat berjalan secara optimal. Data Engineer bertindak sebagai arsitek dan pembangun infrastruktur data. Tanggung jawab utama mereka adalah membangun dan memelihara pipeline data (proses ETL: Extract, Transform, Load). Mereka memastikan bahwa data dari berbagai sumber (misalnya sistem informasi akademik, portal e-learning, dan basis data keuangan) diekstraksi dengan aman, diubah ke dalam format yang terstandarisasi, dan disimpan di dalam pusat data (data warehouse atau data lake) dengan arsitektur yang tahan banting serta dapat diakses dengan cepat.

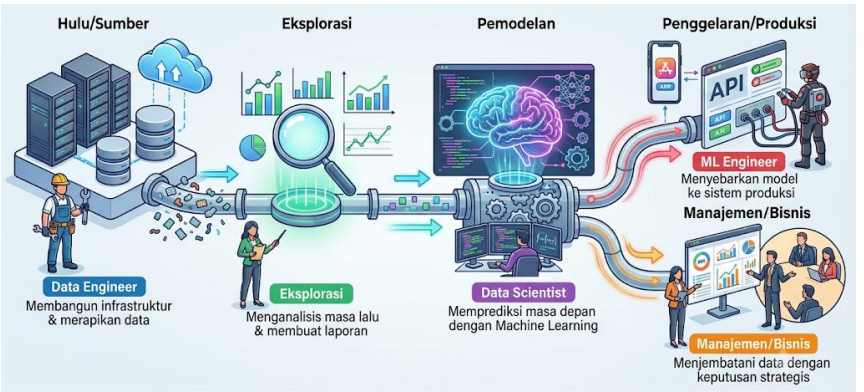
4. Machine Learning Engineer

Machine Learning Engineer (ML Engineer) adalah jembatan antara Data Science murni dan Software Engineering (Rekayasa Perangkat Lunak). Setelah Data Scientist berhasil menemukan dan melatih model terbaik di lingkungan eksperimen (misalnya di Jupyter Notebook), model tersebut tidak bisa dibiarkan begitu saja. Tugas ML Engineer adalah mengambil model tersebut dan melakukan deployment ke lingkungan produksi. Mereka memastikan bahwa model prediktif dapat melayani ribuan permintaan secara serentak (scalable), merancang antarmuka

pemrograman aplikasi (API), dan memantau performa model agar tidak mengalami penurunan akurasi seiring masuknya data-data baru di masa mendatang.

5. Business Intelligence Analyst

Business Intelligence Analyst memiliki peran yang beririsan dengan Data Analyst, namun dengan fokus yang jauh lebih tajam pada strategi dan pencapaian target metrik utama (KPI - Key Performance Indicators). Mereka bertugas menjembatani kesenjangan antara tim teknis data dan pimpinan eksekutif atau manajemen (misalnya dekanat atau rektorat). Fokus utama BI Analyst adalah merancang dashboard interaktif tingkat tinggi (menggunakan perangkat lunak seperti Tableau atau Power BI) yang memungkinkan pengambil keputusan untuk memantau kesehatan institusi atau bisnis secara real-time tanpa perlu memahami bahasa pemrograman sama sekali.



Gambar 5. Ekosistem Kolaborasi Profesi Data Science

6. Perbedaan tanggung jawab dan keterampilan yang dibutuhkan

Untuk memberikan pemahaman yang komprehensif mengenai perbedaan kelima peran di atas, tabel berikut merangkum distribusi tanggung jawab dan fokus keterampilan

(skills) yang umumnya dipersyaratkan:

Table 2. Perbedaan tanggung jawab dan keterampilan yang dibutuhkan

Profesi	Fokus Utama	Luaran Utama (Deliverables)	Keterampilan Inti (Core Skills)
Data Scientist	Memprediksi tren dan memecahkan masalah kompleks dengan probabilitas.	Model Machine Learning, pemahaman XAI, makalah/laporan riset.	Python/R, Statistika Lanjutan, Algoritma Pembelajaran Mesin, Data Mining.
Data Analyst	Menemukan wawasan awal dan tren dari data historis.	Laporan analitik, visualisasi data deskriptif.	SQL, Excel Tingkat Lanjut, Visualisasi Data, Statistika Dasar.
Data Engineer	Mengelola arsitektur aliran data dan pemeliharaan database.	Pipeline data (ETL), Data Warehouse, infrastruktur awan.	SQL Lanjutan, Java/Scala, Hadoop/Spark, Arsitektur Basis Data.
ML Engineer	Mengoptimalkan dan menyebarkan model ke sistem produksi.	API, model skalabel yang terintegrasi di aplikasi akhir.	Software Engineering, Cloud Computing (AWS/GCP), MLOps, C++/Python.
BI Analyst	Mengarahkan pengambilan keputusan manajerial melalui laporan matriks.	Dashboard interaktif, indikator kinerja utama (KPI).	Tableau, Power BI, SQL, Pemahaman Manajemen Strategis/Bisnis.

F. *DATA SCIENCE* VS *BIG DATA* VS *AI*

Dalam wacana teknologi modern maupun literatur populer, istilah *Data Science, Big Data, dan Artificial Intelligence (AI)* sering kali digunakan secara bergantian (interchangeably). Ketiganya

memang saling terkait erat dan menggerakkan revolusi industri 4.0, namun secara keilmuan, masing-masing memiliki definisi, batasan, dan fokus yang berbeda. Memahami hubungan dan tumpang tindih antara ketiga konsep ini sangat penting agar tidak terjadi miskonsepsi dalam perancangan arsitektur teknologi.

1. **Hubungan dan tumpang tindih ketiga konsep**

Untuk melihat benang merahnya, kita perlu mendefinisikan batas-batas dari masing-masing konsep terlebih dahulu:

a. Artificial Intelligence (AI)

AI adalah ranah ilmu komputer yang paling luas. Tujuannya adalah merancang sistem atau mesin yang mampu meniru kecerdasan dan fungsi kognitif manusia, seperti belajar, bernalar, pemecahan masalah, dan pengenalan pola. Di dalam AI, terdapat sub-bidang yang sangat krusial yaitu Machine Learning (ML). ML memungkinkan komputer untuk "belajar" dan meningkatkan akurasinya seiring waktu tanpa diprogram secara eksplisit untuk setiap skenario. Algoritma tangguh seperti XGBoost atau LightGBM adalah produk dari ranah Machine Learning ini.

b. Big Data

Jika AI adalah "mesin"-nya, maka Big Data adalah "bahan bakar"-nya. Big Data merujuk pada himpunan data yang ukurannya sangat masif (Volume), bertambah dengan sangat cepat (Velocity), dan memiliki format yang sangat bervariasi (Variety). Data berskala raksasa ini melampaui kapasitas perangkat lunak pemrosesan data tradisional dan membutuhkan arsitektur terdistribusi (seperti ekosistem Hadoop atau Apache Spark) untuk disimpan dan diolah.

c. Data Science

Data Science adalah disiplin ilmu multidisipliner yang bertindak sebagai jembatan pengekstraksi nilai. Ia memanfaatkan Big Data sebagai bahan baku, kemudian mengaplikasikan metode statistik dan algoritma AI/ML untuk memproses bahan baku tersebut guna menghasilkan wawasan yang dapat ditindaklanjuti (actionable insights).

2. Ilustrasi Integrasi dalam Kasus Nyata

Bayangkan sebuah proyek sistem peringatan dini di institusi pendidikan untuk memprediksi risiko mahasiswa dropout di program studi ilmu kesehatan dan keperawatan.

a. Big Data

Institusi mengumpulkan data historis mahasiswa selama satu dekade terakhir terdiri dari puluhan ribu baris data demografi, log aktivitas pembelajaran di portal akademik tiap detiknya, hingga transkrip nilai.

b. AI / Machine Learning

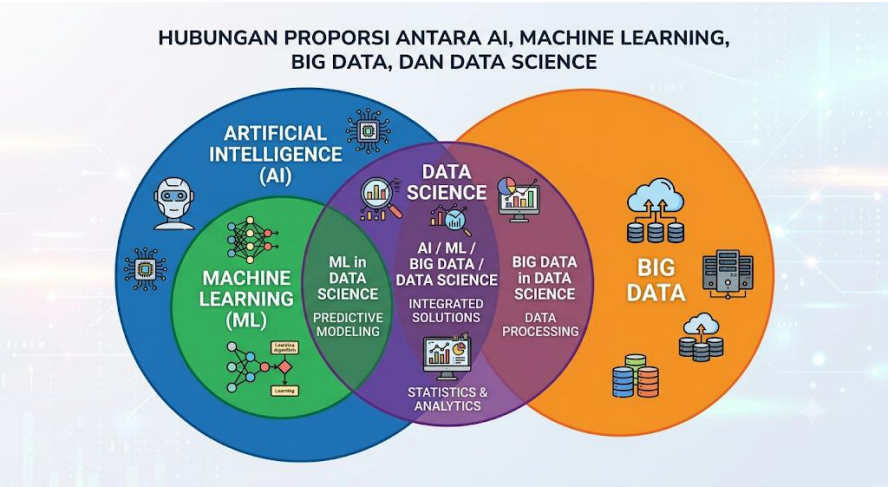
Algoritma berbasis tree (seperti XGBoost atau LightGBM) digunakan untuk membedah data tersebut dan memodelkan pola tersembunyi yang mengindikasikan seorang mahasiswa akan putus kuliah.

c. Data Science

Mengonsep keseluruhan proses di atas, melakukan rekayasa fitur data, mengevaluasi performa model komputasi, dan menerapkan Explainable AI (seperti SHAP values) agar keputusan algoritma dapat dipahami dan ditindaklanjuti oleh para dosen pembimbing.

Ketiga entitas tersebut tidak berdiri sendiri, melainkan saling beririsan. Data Science sering kali digambarkan sebagai irisan payung yang menaungi penerapan praktis dari Big Data dan

analitik prediktif Machine Learning.



Gambar 6. Hubungan konseptual antara Data Science, Big Data, dan Artificial Intelligence.

G. CONTOH PENERAPAN *DATA SCIENCE* DI BERBAGAI BIDANG

Sifat Data Science yang domain-agnostic (tidak terikat pada satu industri saja) membuatnya menjadi instrumen pemecahan masalah yang sangat fleksibel. Perpaduan antara algoritma Machine Learning, infrastruktur Big Data, dan pemahaman konteks spesifik telah mendisrupsi cara kerja di berbagai sektor. Berikut adalah beberapa contoh implementasi nyata *Data Science* di berbagai bidang:

1. **Kesehatan (diagnosis penyakit, analisis genom)**

   Di sektor kesehatan (healthcare) dan ilmu medis, Data Science bertindak sebagai asisten canggih bagi para tenaga medis.

   a. Diagnosis Penyakit

      Algoritma Computer Vision dan Deep Learning

digunakan untuk menganalisis citra medis (seperti foto Rontgen atau MRI) guna mendeteksi anomali seperti tumor atau sel kanker dengan tingkat akurasi yang menyamai atau bahkan melampaui dokter spesialis. Mengingat keputusan medis menyangkut nyawa manusia, penerapan teknik Explainable AI (XAI) seperti SHAP sangat krusial di bidang ini. Algoritma tidak boleh sekadar menjadi "kotak hitam"; dokter harus bisa melihat visualisasi fitur apa saja (misalnya kadar gula darah atau tekanan darah) yang paling memengaruhi prediksi risiko klinis seorang pasien.

b. Analisis Genom

Pemrosesan data DNA yang berukuran masif memungkinkan pengembangan personalized medicine, yaitu rancangan pengobatan atau dosis obat yang disesuaikan secara spesifik dengan profil genetik individu, bukan lagi sekadar resep standar umum.

2. Keuangan (deteksi fraud, credit scoring)

Industri perbankan dan fintech (teknologi finansial) adalah salah satu pengadopsi awal analitik data berskala besar untuk mengelola risiko dan keamanan.

a. Deteksi Fraud (Penipuan)

Algoritma deteksi anomali digunakan untuk memantau jutaan transaksi secara real-time. Jika ada pola transaksi kartu kredit yang tidak wajar (misalnya tiba-tiba digunakan di luar negeri dengan nominal besar), sistem akan langsung memblokirnya. Dalam kasus ini, data transaksi penipuan (fraud) biasanya sangat sedikit dibandingkan transaksi normal. Oleh karena itu, teknik penanganan kelas tidak seimbang seperti SMOTE sangat sering digunakan oleh analis data di bank sebelum

melatih model deteksinya.

b. Credit scoring

Bank tidak lagi menilai kelayakan kredit secara manual. Model klasifikasi digunakan untuk memprediksi probabilitas seorang nasabah akan gagal bayar (default) berdasarkan rekam jejak finansial, demografi, dan riwayat kredit sebelumnya.

3. E-commerce (sistem rekomendasi, prediksi churn)

Platform perdagangan elektronik seperti Amazon, Shopee, atau Tokopedia sangat bergantung pada data konsumen untuk mempersonalisasi pengalaman belanja.

a. Prediksi Churn

Konsepnya sangat mirip dengan memprediksi mahasiswa yang berisiko dropout di dunia akademik. Perusahaan e-commerce menggunakan riwayat interaksi pengguna untuk memprediksi pelanggan mana yang berisiko tinggi meninggalkan platform (beralih ke kompetitor) dalam waktu dekat. Dengan mengidentifikasi pelanggan ini lebih awal, perusahaan dapat memberikan intervensi proaktif, seperti mengirimkan voucer diskon khusus.

b. Sistem Rekomendasi

Menggunakan algoritma Collaborative Filtering atau Market Basket Analysis, sistem mempelajari pola belanja pengguna. Jika Pengguna A membeli laptop dan mouse, maka sistem akan merekomendasikan mouse kepada Pengguna B yang baru saja memasukkan laptop ke keranjang belanjanya.

4. Transportasi (optimasi rute, prediksi keterlambatan)

Aplikasi navigasi dan layanan ride-hailing (seperti Google

Maps, Gojek, atau Grab) adalah bentuk nyata implementasi Big Data di genggaman masyarakat.

a. Optimasi Rute

   Dengan memproses titik koordinat GPS dari jutaan pengguna secara real-time, algoritma dapat menganalisis kemacetan lalu lintas dan merekomendasikan rute tercepat dan terefisien untuk pengemudi logistik maupun pengguna umum.

b. Prediksi Keterlambatan

   Maskapai penerbangan dan perusahaan kereta api menggunakan data cuaca historis, jadwal lalu lintas udara, dan kondisi mesin (IoT sensors) untuk memprediksi probabilitas keterlambatan (delay). Hal ini memungkinkan perusahaan untuk menyesuaikan jadwal secara dinamis dan meminimalkan kerugian operasional.

5. Pemasaran (segmentasi pelanggan, kampanye personalisasi)

Metode pemasaran spray and pray (menyebarkan iklan secara acak dan berharap ada yang membeli) telah ditinggalkan berkat Data Science.

a. Segmentasi Pelanggan

   Menggunakan algoritma clustering (seperti K-Means), konsumen dikelompokkan ke dalam segmen-segmen spesifik berdasarkan kesamaan perilaku, rentang usia, daya beli, hingga ketertarikan produk.

b. Kampanye Personalisasi

   Berbekal data segmentasi tersebut, tim pemasaran dapat menyusun targeted campaign. Misalnya, alih-alih mengirimkan email promosi perlengkapan bayi ke seluruh basis pelanggan, model komputasi memastikan email tersebut hanya dikirimkan kepada segmen

pelanggan yang secara statistik memiliki probabilitas tertinggi untuk tertarik (misalnya, pengguna yang baru saja mencari artikel tentang kehamilan). Hal ini meningkatkan metrik Return on Investment (ROI) pemasaran secara drastis.

H. TANTANGAN DAN ISU ETIS DALAM *DATA SCIENCE*

Meskipun Data Science menawarkan potensi yang luar biasa untuk mendorong inovasi dan efisiensi, penerapannya di dunia nyata tidak luput dari berbagai tantangan besar. Sebagai sebuah disiplin ilmu yang mengolah informasi sensitif dan memengaruhi keputusan vital, seorang praktisi tidak hanya dituntut untuk menguasai metrik akurasi komputasi, tetapi juga harus peka terhadap isu moral dan batasan etis. Berikut adalah gambaran sekilas mengenai tantangan dan isu etis utama dalam Data Science:

1. Kualitas Dan Ketersediaan Data

Tantangan paling mendasar dalam setiap proyek data adalah pemenuhan prinsip "*garbage in, garbage out*" jika data yang digunakan sebagai masukan berkualitas buruk, maka model yang dihasilkan pun akan menghasilkan prediksi yang tidak akurat dan menyesatkan. Data di dunia nyata sering kali tidak lengkap, terfragmentasi, atau mengalami masalah ketidakseimbangan kelas (class imbalance) yang ekstrem.

Sebagai contoh, dalam pengklasifikasian kasus langka seperti mendeteksi kecurangan transaksi atau memprediksi mahasiswa yang berisiko dropout, jumlah sampel kasus positif (minoritas) biasanya jauh lebih sedikit dibandingkan kasus normal. Jika tantangan ketersediaan data ini tidak ditangani dengan manipulasi teknis yang tepat seperti pendekatan sintesis data

menggunakan metode SMOTE maka model matematika yang dibangun akan gagal mengenali pola minoritas tersebut karena terlalu condong pada kelas mayoritas.

2. Privasi Dan Keamanan Data

Data Science tumbuh subur di atas ketersediaan data berskala besar, yang sering kali mencakup informasi pribadi, rekam medis, riwayat finansial, hingga rekam jejak akademik individu. Isu etis muncul ketika data tersebut dikumpulkan, disimpan, atau diolah tanpa persetujuan eksplisit (informed consent) dari pemilik data.

Praktisi Data Science memiliki tanggung jawab hukum dan moral untuk menjaga kerahasiaan tersebut. Penerapan algoritma harus patuh pada regulasi pelindungan data global maupun nasional, seperti General Data Protection Regulation (GDPR) atau Undang-Undangan Pelindungan Data Pribadi (UU PDP). Pelanggaran terhadap aspek ini tidak hanya merusak reputasi institusi tetapi juga melanggar hak asasi individu atas privasi mereka.

3. Bias Dalam Data Dan Model

Sebuah model Machine Learning belajar dari data historis. Jika data historis yang digunakan untuk melatih algoritma mencakup keputusan masa lalu yang mengandung bias manusia baik secara sadar maupun tidak maka model tersebut akan mempelajari, mereplikasi, dan bahkan memperkuat bias tersebut secara otomatis.

Sebagai contoh, jika sebuah algoritma credit scoring otomatis dilatih menggunakan data historis yang secara tidak adil mendiskriminasi kelompok demografi atau wilayah tertentu, model tersebut akan terus menolak pengajuan kredit dari kelompok tersebut di masa depan. Menghilangkan bias (ensure

fairness) merupakan tantangan etis yang rumit karena memerlukan penilaian objektif untuk memastikan algoritma memperlakukan setiap entitas secara adil tanpa mengorbankan akurasi prediksi.

4. **Interpretabilitas Model (Black Box)**

Algoritma ensemble modern tingkat tinggi seperti XGBoost dan LightGBM, atau arsitektur Deep Learning, terkenal memiliki kemampuan akurasi prediktif yang sangat superior. Namun, algoritma ini memiliki struktur matematika yang sangat kompleks sehingga sering kali dikategorikan sebagai "kotak hitam" (black box) sebuah sistem yang diketahui masukan dan keluarannya, tetapi sangat sulit dipahami bagaimana proses internalnya dalam menghasilkan keputusan tersebut.

Secara etis, ketidakmampuan untuk menjelaskan alasan di balik keputusan model dapat berbahaya, terutama pada sektor dengan risiko tinggi seperti diagnosis penyakit di bidang kesehatan atau penentuan kebijakan akademik. Oleh karena itu, tantangan ini memicu perkembangan ranah Explainable AI (XAI). Penggunaan metode interpretasi seperti nilai SHAP (SHAPley Additive exPlanations) melalui visualisasi grafik kontribusi fitur menjadi jembatan krusial agar keputusan prediktif komputer dapat dipertanggungjawabkan, transparan, dan dapat diterima secara etis oleh para pengambil keputusan.

BAB 2
DATA DAN INFORMASI

Perkembangan teknologi digital telah mengubah cara individu, organisasi, dan pemerintah dalam menghasilkan, menyimpan, serta memanfaatkan data. Hampir seluruh aktivitas manusia saat ini meninggalkan jejak digital yang dapat direkam dan dianalisis, mulai dari transaksi keuangan, penggunaan media sosial, aktivitas e-commerce, hingga komunikasi melalui perangkat bergerak. Kondisi ini menyebabkan data menjadi salah satu aset strategis yang memiliki nilai ekonomi dan operasional yang sangat tinggi dalam berbagai sektor kehidupan modern (Laudon & Laudon, 2024; Sharda et al., 2024).

Dalam bidang Data Science, data berperan sebagai bahan dasar yang digunakan untuk menghasilkan informasi, pengetahuan, dan wawasan yang dapat mendukung proses pengambilan keputusan. Tanpa data yang memadai dan berkualitas, organisasi akan mengalami kesulitan dalam memahami kondisi yang sedang terjadi maupun memprediksi kejadian di masa mendatang. Oleh karena itu, pemahaman mengenai konsep data dan informasi menjadi fondasi utama sebelum mempelajari teknik analitik, machine learning, maupun kecerdasan buatan (Cao, 2020; Qamar & Raza, 2023).

Perusahaan modern tidak lagi hanya mengandalkan pengalaman atau intuisi dalam mengambil keputusan, tetapi juga memanfaatkan data dan informasi yang dihasilkan dari berbagai sistem digital untuk memperoleh keunggulan kompetitif dan meningkatkan efektivitas operasional (Marr, 2021; Provost & Fawcett, 2023).

Pemahaman yang baik mengenai hubungan antara data dan informasi sangat penting karena keduanya merupakan komponen utama dalam seluruh proses Data Science. Data yang dikumpulkan dari berbagai sumber perlu diolah, dianalisis, dan diinterpretasikan agar menghasilkan informasi yang relevan dan dapat digunakan untuk mendukung keputusan yang tepat. Oleh karena itu, bab ini membahas konsep data, informasi, jenis-jenis data, kualitas data, siklus pengolahan data, serta peran data dan informasi dalam Data Science (Saltz & Stanton, 2021; Plaue, 2023).

A. KONSEP DATA

1. Definisi Data

Data merupakan sekumpulan fakta, simbol, angka, teks, gambar, suara, atau representasi lain yang menggambarkan suatu objek, peristiwa, maupun aktivitas tertentu. Data bersifat mentah (raw data) dan belum memiliki makna yang jelas sebelum melalui proses pengolahan dan interpretasi. Dalam konteks sistem informasi, data menjadi bahan baku yang digunakan untuk menghasilkan informasi yang bermanfaat bagi pengguna (Stair & Reynolds, 2023; Laudon & Laudon, 2024).

Dalam ilmu Data Science, data dipandang sebagai representasi digital dari fenomena yang terjadi di dunia nyata. Data dapat berasal dari berbagai sumber seperti sensor, transaksi bisnis, survei, eksperimen, media sosial, maupun perangkat Internet of Things (IoT). Semakin lengkap dan berkualitas data yang tersedia, semakin besar peluang untuk menghasilkan analisis yang akurat dan bernilai guna (Qamar & Raza, 2023; Plaue, 2023).

Keberadaan data saat ini semakin melimpah karena hampir seluruh aktivitas manusia dilakukan melalui sistem digital. Setiap transaksi perbankan, pencarian informasi di internet,

penggunaan aplikasi seluler, maupun aktivitas pada media sosial menghasilkan data yang dapat direkam dan dianalisis. Fenomena ini menjadikan data sebagai aset yang sangat penting dalam mendukung inovasi dan pengambilan keputusan berbasis bukti (evidence-based decision making) (Marr, 2021; Kotu & Deshpande, 2023).

2. Contoh Data

Perhatikan data berikut : 85, 78, 90, 88, 80, 92, 74 dan seterusnya. Deretan angka tersebut hanya menunjukkan sekumpulan fakta tanpa konteks tertentu. Pengguna belum dapat mengetahui makna dari angka-angka tersebut sebelum diberikan penjelasan tambahan mengenai sumber dan tujuan data tersebut (Stair & Reynolds, 2023). Apabila diketahui bahwa angka-angka tersebut merupakan nilai ujian mahasiswa pada mata kuliah Pengantar Data Science, maka data mulai memiliki konteks yang memungkinkan untuk dilakukan analisis lebih lanjut. Dengan demikian, konteks menjadi faktor penting yang membedakan data dari informasi (Zins, 2023; Sengupta, 2023).

3. Karakteristik Data

Data memiliki sejumlah karakteristik yang membedakannya dari bentuk sumber daya lainnya, diantaranya :

a. Menggambarkan fakta atau kejadian yang dapat diamati, dicatat, dan diverifikasi. Karakteristik kedua adalah kemampuan untuk disimpan dalam berbagai media seperti basis data, data warehouse, maupun cloud storage sehingga dapat digunakan kembali di masa mendatang (Batini, 2022; Kimball & Ross, 2021).
b. Dapat diukur dan diproses menggunakan metode statistik maupun algoritma komputasi. Data yang telah dikumpulkan dapat diolah untuk menghasilkan berbagai

pola, hubungan, dan kecenderungan yang berguna dalam proses analisis. Kemampuan inilah yang menjadi dasar berkembangnya bidang Data Science dan Artificial Intelligence (Bruce et al., 2024; Provost & Fawcett, 2023).

c. Pada data modern, data memiliki karakteristik volume yang sangat besar, kecepatan pertumbuhan yang tinggi, dan format yang semakin beragam. Kondisi tersebut dikenal sebagai fenomena Big Data, yang mendorong organisasi untuk mengembangkan teknologi baru dalam pengelolaan dan analisis data (Kitchin, 2021; Delen, 2021).

B. JENIS-JENIS DATA

1. Berdasarkan Sifatnya

Data dapat dikelompokkan menjadi data kualitatif dan data kuantitatif. Data kualitatif merupakan data yang menggambarkan karakteristik, kategori, atau atribut tertentu tanpa menggunakan ukuran numerik secara langsung. Contohnya meliputi jenis kelamin, warna kendaraan, tingkat pendidikan, dan status pekerjaan seseorang (Bruce et al., 2024; Stair & Reynolds, 2023). Sebaliknya, data kuantitatif merupakan data yang dinyatakan dalam bentuk angka dan dapat diukur secara matematis. Data jenis ini sering digunakan dalam analisis statistik karena memungkinkan dilakukan perhitungan seperti rata-rata, median, varians, maupun regresi. Contohnya adalah umur, pendapatan, jumlah penjualan, dan nilai ujian mahasiswa (Plaue, 2023; Kotu & Deshpande, 2023).

2. Berdasarkan Sumbernya

Berdasarkan sumber perolehannya, data dibedakan menjadi data primer dan data sekunder. Data primer diperoleh secara

langsung dari sumber pertama melalui observasi, wawancara, survei, eksperimen, atau pengukuran lapangan. Data jenis ini umumnya memiliki tingkat relevansi yang tinggi karena dikumpulkan sesuai kebutuhan penelitian atau organisasi (Saltz & Stanton, 2021). Data sekunder merupakan data yang diperoleh dari pihak lain atau sumber yang telah tersedia sebelumnya, seperti laporan tahunan perusahaan, publikasi pemerintah, jurnal ilmiah, maupun basis data publik. Penggunaan data sekunder sering dilakukan karena lebih efisien dari segi waktu dan biaya dibandingkan pengumpulan data primer (Stair & Reynolds, 2023).

3. Berdasarkan Struktur

Dalam lingkungan Data Science, data dapat diklasifikasikan menjadi structured data, semi-structured data, dan unstructured data. Structured data memiliki format yang terorganisasi dan tersimpan dalam tabel dengan baris serta kolom yang jelas, sehingga mudah diakses menggunakan sistem basis data relasional (Kimball & Ross, 2021).

Semi-structured data merupakan data yang memiliki struktur sebagian, tetapi tidak mengikuti format tabel secara ketat. Contoh yang umum digunakan adalah XML dan JSON yang banyak digunakan dalam pertukaran data antar aplikasi modern (Kleppmann, 2022).

Unstructured data adalah data yang tidak memiliki struktur baku dan umumnya berbentuk teks bebas, gambar, audio, video, maupun konten media sosial. Saat ini sebagian besar data yang dihasilkan di internet termasuk dalam kategori unstructured data sehingga memerlukan teknik analisis yang lebih kompleks dibandingkan structured data (Marr, 2021; Kitchin, 2021).

C. KONSEP INFORMASI

Informasi merupakan hasil pengolahan data yang telah diberikan konteks, struktur, dan makna sehingga dapat digunakan untuk mendukung pengambilan keputusan. Berbeda dengan data yang masih berupa fakta mentah, informasi telah melalui proses analisis sehingga mampu memberikan pemahaman yang lebih baik mengenai suatu kondisi atau peristiwa tertentu (Laudon & Laudon, 2024; Stair & Reynolds, 2023).

Dalam sistem informasi, informasi berfungsi sebagai dasar dalam proses perencanaan, pengendalian, koordinasi, dan evaluasi organisasi. Informasi yang berkualitas akan membantu manajemen memahami kondisi organisasi secara lebih akurat sehingga keputusan yang diambil dapat menghasilkan dampak yang optimal (Sharda et al., 2024). Sebagai contoh, sekumpulan angka penjualan bulanan hanya merupakan data. Setelah dilakukan pengolahan dan diketahui bahwa penjualan mengalami peningkatan sebesar 15% setiap bulan, hasil tersebut berubah menjadi informasi yang dapat digunakan untuk menyusun strategi pemasaran dan perencanaan produksi (Provost & Fawcett, 2023).

D. KARAKTERISTIK INFORMASI YANG BERKUALITAS

Informasi yang baik harus memenuhi beberapa karakteristik utama, yaitu akurat, relevan, tepat waktu, lengkap, dan dapat dipercaya. Informasi yang akurat mencerminkan kondisi sebenarnya sehingga dapat mengurangi risiko kesalahan dalam pengambilan keputusan (Laudon & Laudon, 2024). Selain akurasi, informasi juga harus relevan dengan kebutuhan pengguna. Informasi yang sangat detail sekalipun tidak akan memberikan manfaat apabila tidak sesuai dengan kebutuhan pengambil

keputusan. Oleh karena itu, relevansi menjadi salah satu indikator utama kualitas informasi (Stair & Reynolds, 2023). Karakteristik lain yakni ketepatan waktu. Informasi yang terlambat diterima dapat kehilangan nilai manfaatnya karena kondisi yang digambarkan mungkin sudah berubah. Dalam lingkungan bisnis yang dinamis, informasi real-time menjadi semakin penting untuk mendukung respons yang cepat terhadap perubahan pasar (Sharda et al., 2024).

E. HUBUNGAN DATA DAN INFORMASI

Data dan informasi merupakan dua konsep yang saling berkaitan erat. Data berperan sebagai input yang akan diproses, sedangkan informasi merupakan output yang dihasilkan dari proses pengolahan tersebut. Hubungan ini menunjukkan bahwa kualitas informasi sangat dipengaruhi oleh kualitas data yang digunakan (Batini, 2022).

Hubungan antara data dan informasi sering dijelaskan menggunakan model DIKW (Data, Information, Knowledge, Wisdom). Model ini menggambarkan bagaimana data yang telah diolah menjadi informasi dapat berkembang menjadi pengetahuan dan akhirnya menghasilkan kebijaksanaan dalam pengambilan keputusan strategis (Sengupta, 2023; Zins, 2023).

F. KUALITAS DATA

Kualitas data merupakan faktor yang sangat menentukan keberhasilan implementasi Data Science. Data yang tidak akurat, tidak lengkap, atau tidak konsisten dapat menghasilkan informasi yang menyesatkan sehingga berdampak pada kualitas keputusan yang diambil organisasi (Batini, 2022).

Dimensi kualitas data meliputi accuracy, completeness,

consistency, timeliness, validity, dan uniqueness. Organisasi perlu memastikan bahwa data memenuhi seluruh dimensi tersebut agar hasil analisis dapat dipercaya dan digunakan sebagai dasar pengambilan keputusan yang efektif (Qamar & Raza, 2023).

G. PERAN DATA DAN INFORMASI DALAM DATA SCIENCE

Data dan informasi merupakan inti dari seluruh aktivitas Data Science. Seluruh tahapan mulai dari pengumpulan data, pembersihan data, eksplorasi data, pemodelan, hingga visualisasi bertujuan menghasilkan informasi dan wawasan yang bernilai bagi pengguna (Provost & Fawcett, 2023; Plaue, 2023).

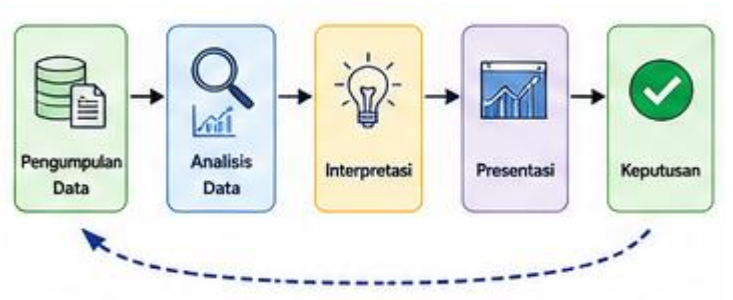
Keberhasilan suatu proyek Data Science tidak hanya ditentukan oleh kecanggihan algoritma yang digunakan, tetapi juga oleh kualitas data dan kemampuan dalam menerjemahkan hasil analisis menjadi informasi yang dapat dipahami oleh pengambil keputusan. Oleh karena itu, pemahaman terhadap konsep data dan informasi menjadi kompetensi dasar yang wajib dimiliki oleh seorang data scientist (Cao, 2020; Sharda et al., 2024).

BAB 3
DASAR STATISTIKA UNTUK DATA SCIENCE

Dalam lanskap bisnis yang kian kompetitif, statistika bukan sekadar instrumen teknis untuk pengolahan angka, melainkan fondasi bagi organisasi modern untuk bermigrasi dari pengambilan keputusan berbasis intuisi ke kebijakan yang berlandaskan bukti empiris. Sebagai ahli strategi, kita harus memandang statistika sebagai metodologi untuk mengelola ketidakpastian melalui akurasi yang terukur.(Febrieta & Fitriani, n.d.)(Bakti Siregar, Andi Pujo Rahadi, 2025)

Berdasarkan konteks teoretis dan praktis, statistika mencakup empat pilar utama:

a. Pengumpulan Data: Proses sistematis melalui survei, observasi, atau sampling terkontrol.
b. Analisis Data: Aplikasi teknik deskriptif dan inferensial untuk mengekstraksi pola numerik.
c. Interpretasi Data: Mengubah temuan matematis menjadi narasi logis mengenai tren dan hubungan antar variabel.
d. Presentasi Data: Komunikasi hasil melalui visualisasi yang memfasilitasi pemahaman lintas departemen.



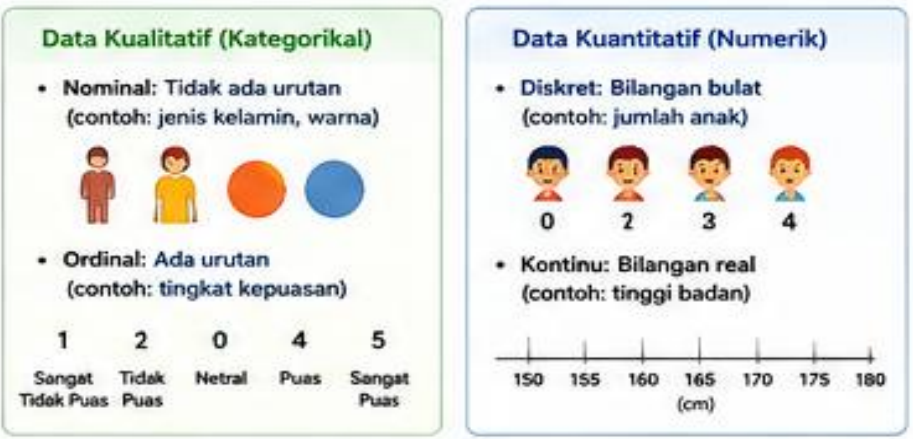
Gambar 1. Siklus Statistika dalam Data Science

Penggunaan metode matematis dan algoritmik ini secara langsung memitigasi risiko "kesalahan penilaian" yang sering muncul akibat asumsi subjektif. Dengan mengadopsi kerangka kerja ini, manajemen dapat membedakan antara fluktuasi acak dan tren pasar yang signifikan. Namun, penguasaan metodologi ini harus diawali dengan pemahaman mendalam terhadap taksonomi data yang dikelola.

A. TAKSONOMI DATA: KATEGORISASI DAN TINGKATAN PENGUKURAN

Ketepatan analisis ditentukan sejak tahap identifikasi jenis data. Kesalahan klasifikasi pada tahap ini akan mengakibatkan kegagalan metodologi seperti mencoba menghitung rata-rata dari data kategori, yang secara teoretis tidak bermakna dan menyesatkan secara strategis. (Agustina et al., 2025)

Berikut adalah klasifikasi data berdasarkan karakteristik pengukuran:

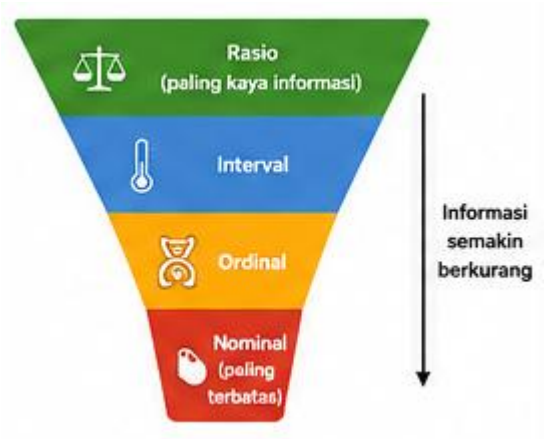


Gambar 2. Jenis Data: Kualitatif vs Kuantitatif

Pemilihan model bisnis sangat bergantung pada empat tingkatan pengukuran berikut:

Tabel 1. Tingkat Pengukuran

Tingkatan	Ciri-ciri	Operasi statistik	Contoh
Nominal	Kategori tanpa urutan	Frekuensi, Modus	Warna Mata
Ordinal	Ada urutan	Median, Persentil	Peringkat Kompetisi
Interval	Jarak tetap, tanpa nol mutlak	Mead, SD	Suhu Celcius
Rasio	Nol Mutlak	Rasio, Perkalian	Berat Badan, Tinggi



Gambar 3. Hierarki Tingkatan Pengukuran

B. STATISTIKA DESKRIPTIF : MENYEDERHANAKAN KOMPLEKSITAS INFORMASI

Statistika deskriptif adalah alat untuk "bercerita" mengenai kondisi performa saat ini tanpa melakukan generalisasi prematur. Ini memberikan potret objektif mengenai kesehatan operasional

perusahaan.(Sari et al., n.d.) (Cambell, 2025)

Ukuran Pemusatan Data (Central Tendency)

a. Mean (Rata-rata): Nilai rata-rata hitung, sangat efektif untuk data terdistribusi normal.
b. Median: Nilai tengah yang membagi data menjadi dua bagian. Median jauh lebih representatif daripada Mean ketika menghadapi data dengan outlier (pencilan) ekstrem, seperti distribusi gaji karyawan di mana gaji eksekutif puncaknya sering mendistorsi rata-rata.
c. Mode: Nilai dengan frekuensi tertinggi, krusial untuk analisis data kategorikal.

Ukuran Penyebaran Data (Dispersion)

a. Range: Selisih nilai maksimum dan minimum.

$$Range = Max - Min$$

b. Variance & Standard Deviation: Mengukur deviasi data dari rata-rata.

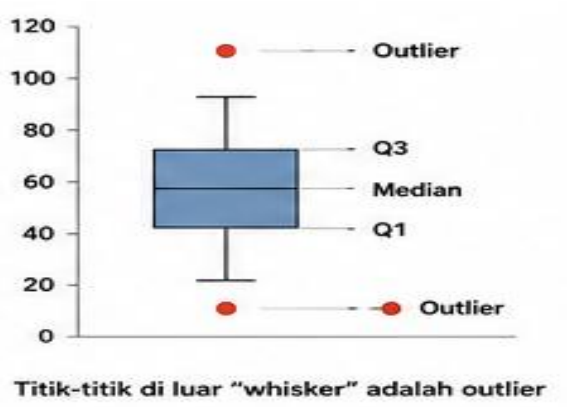
$$s^2 = \frac{\sum(x_1 - \bar{x})}{n-1} \, , s = \sqrt{s^2}$$

c. Pentingnya Koreksi Bessel (): Dalam menghitung varians sampel, kita menggunakan pembagi (bukan populasi). Hal ini dilakukan untuk memberikan estimasi yang lebih konservatif dan akurat guna menutupi ketidakpastian karena kita tidak mengamati seluruh populasi.
d. Quartile (Q1, Q2, Q3): Membagi data menjadi empat kuarter untuk mengidentifikasi sebaran dan ambang batas pencilan melalui Interquartile Range (IQR).

$$IQR = Q_3 - Q_1$$

Nilai $< Q_1 - 1{,}5 \; x \; IQR$

$> Q_3 + 1{,}5 \; x \; IQR$



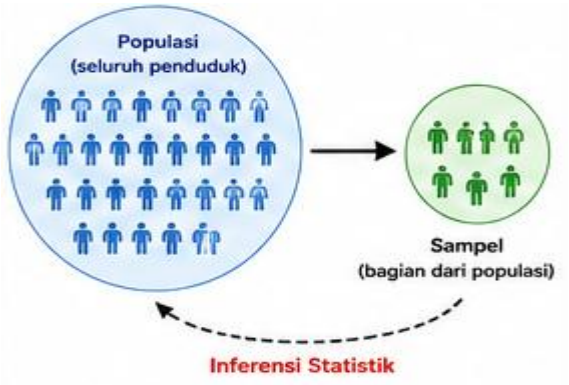
Gambar 5. Boxplot dengan Outlier

C. DINAMIKA POPULASI DAN SAMPEL DALAM PENELITIAN EMPIRIS

Keterbatasan sumber daya operasional membuat pengamatan terhadap seluruh Populasi seringkali mustahil dilakukan. Oleh karena itu, kita menggunakan Sampel representasi kecil yang diambil secara strategis untuk mengestimasi Parameter Populasi.(Agustina et al., 2025)

Tabel 2. Populasi, Sampel, Parameter dan Statistik

Istilah	Definisi	Contoh
Populasi	Seluruh individu / Benda yang ingin diteliti	Semua Penduduk Indonesia
Sampel	Bagian dari populasi yang mewakili	1.000 Penduduk
Parameter	Nilai numerik yang menggambarkan populasi	Rata-rata tinggi badan seluruh penduduk
Statistik Sampel	Nilai numerik yang dihitung dari sampel	Rata-rata tinggi badan 1.000 sampel

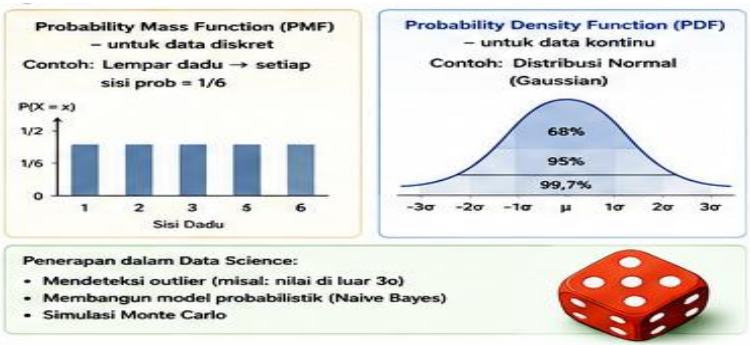


Gambar 6. Populasi vs Sampel

D. DISTRIBUSI PROBABILITAS: MEMAHAMI POLA DAN VARIABEL ACAK

Probabilitas memungkinkan kita melakukan simulasi skenario bisnis di bawah bayang-bayang ketidakpastian.

a. Probability Mass Function (PMF): Untuk variabel diskret (misal: probabilitas jumlah transaksi sukses).
b. Probability Density Function (PDF): Untuk variabel kontinu (misal: estimasi durasi penggunaan produk).



Gambar 7. Distribusi Probabilitas

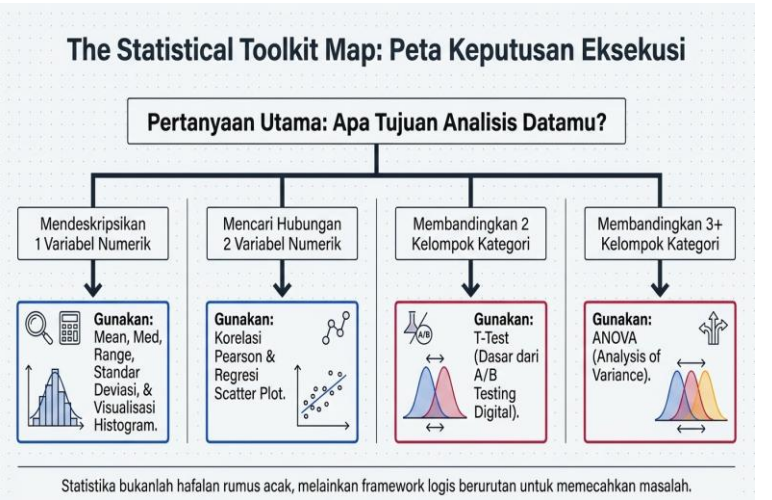
Distribusi Normal (Gaussian) vs. Skewed

Mayoritas fenomena alami mengikuti Distribusi Normal yang simetris (bell curve). Namun, data bisnis seringkali bersifat Skewed (miring). Contohnya, distribusi pendapatan biasanya Positive Skew (banyak yang berpendapatan rendah, sedikit yang sangat tinggi). Mengasumsikan data selalu normal saat kenyataannya miring akan menyebabkan kesalahan fatal dalam model Machine Learning.

E. STATISTIKA INFERENSIAL : MEMBANGUN KESIMPULAN BERBASIS BUKTI

Tahap ini adalah transisi dari sekadar mendeskripsikan masa lalu menuju penarikan kesimpulan yang memiliki bobot ilmiah untuk masa depan.(Bruce et al., n.d.). Teknik Utama dan Kegunaannya:

a. Uji Hipotesis:
    1) t-test: Menggunakan scipy.stats.ttest_ind untuk membandingkan rata-rata dua kelompok (misal: efektivitas dua metode penjualan).
    2) ANOVA: Membandingkan rata-rata lebih dari dua kelompok.
    3) Chi-Square: Menguji hubungan antara variabel kategorikal (misal: hubungan antara wilayah dan preferensi produk).
b. Interval Kepercayaan (Confidence Interval): Berfungsi sebagai "Margin of Safety" bagi eksekutif. Ini memberikan rentang di mana nilai parameter populasi sebenarnya kemungkinan besar berada.
c. Regresi dan Korelasi: Memodelkan hubungan antar variabel untuk prediksi nilai masa depan.



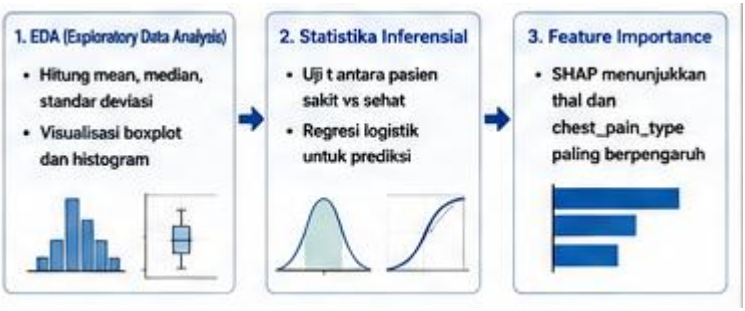
Gambar 8. Peta Keputusan

F. APLIKASI METODOLOGI STATISTIK DALAM TANTANGAN BISNIS

Transformasi data menjadi actionable insights dilakukan melalui aplikasi metodologi berikut:

a. Exploratory Data Analysis (EDA): Menggunakan teknik korelasi dan visualisasi untuk mengidentifikasi "key drivers" atau variabel utama yang paling memengaruhi hasil bisnis sebelum membangun model yang mahal.
b. Model Prediktif: Mengimplementasikan Regresi Linear melalui library scikit-learn untuk memprediksi pendapatan berdasarkan variabel seperti lokasi dan biaya pemasaran.
c. A/B Testing: Metode standar industri untuk memvalidasi apakah perubahan desain UI/UX atau strategi harga baru memberikan hasil yang lebih baik secara signifikan dibandingkan status quo.

d. Feature Importance: Teknik seperti Permutation Importance atau SHAP Values digunakan untuk menentukan variabel mana (misalnya: harga vs. kecepatan pengiriman) yang paling berkontribusi terhadap akurasi model prediksi.



Gambar 9. Penerapan Statistika dalam Data Science

BAB 4
DASAR MATEMATIKA DATA SCIENCE

Data Science adalah bidang interdisipliner yang mempertemukan statistika, matematika, ilmu komputer, dan pengetahuan domain dalam satu kerangka kerja yang utuh semuanya diarahkan pada satu tujuan: mengekstrak wawasan bermakna dari data. Pada konteks ini, Skiena (2017) menegaskan bahwa pemahaman terhadap dasar-dasar matematika bukan sekadar pelengkap, melainkan dasar yang tidak bisa diabaikan bagi yang ingin berkecimpung di bidang data science.

Bab ini membahas pilar matematika yang menjadi landasan Data Science, yaitu Statistika Deskriptif, Probabilitas, Aljabar Linear, dan Matriks. Setiap topik diurai secara sistematis mulai dari konsep teoritis hingga rumus matematis yang relevan dengan tetap mengacu pada kerangka yang dikembangkan oleh Deisenroth et al. (2020).

Setelah menyelesaikan bab ini, pembaca diharapkan tidak hanya mampu memahami konsep dasar dari masing-masing topik, tetapi juga dapat menerapkannya pada permasalahan nyata dan mengimplementasikan perhitungan.

A. STATISTIKA DESKRIPSTIF

Statistika Deskriptif menurut Bruce et al. (2020) adalah cabang statistika yang berfungsi untuk meringkas, menggambarkan, dan menyajikan data secara informatif. Tujuannya adalah memahami karakteristik umum suatu kumpulan data. Pada data science, statistika deskriptif digunakan dalam tahap

Exploratory Data Analysis (EDA) untuk mendapatkan gambaran awal mengenai distribusi, sebaran, dan pola dalam data sebelum dilakukan pemodelan lebih lanjut (VanderPlas, 2016). Untuk memahami karakteristik data secara lebih mendalam dalam tahap EDA, dalam statistika deskriptif dilakukan perhitungan pengukuran kecenderungan memusat dan pengkuran kecenderungan penyebaran (Agustina et al., 2025) (Kadir, 2022)

1. Ukuran Pemusatan Data

Ukuran pemusatan data atau tendensi sentral pada statistika deskriptif ditetapkan berdasarkan konsep nilai harapan, serta melalui pendekatan estimasi dan prediksi terhadap nilai yang merepresentasikan keseluruhan kumpulan data (dataset). Tiga ukuran yang sering digunakan untuk menentukan tendensi sentral antara lain mean(rata-rata), median (nilai tengah), dan modus (nilai yang paling sering muncul (Triola, 2022a)

a. Mean

Mean adalah seluruh nilai data dibagi dengan jumlah data observasi. Mean menggunakan seluruh informasi nilai dalam dataset, sehingga mampu memberikan gambaran umum mengenai kecenderungan data secara keseluruhan (Setiawan, 2021). Adapun mean dihitung dengan rumus:

$$\bar{x} = \frac{\sum X_i}{n}$$

Keterangan:

$\bar{x}$ = mean,

x_i = nilai dari data ke-i, dengan i=1, 2, 3, ……, n

n = banyaknya dataset

b. Median

Median adalah nilai tengah dari sebuah dataset setelah dilakukan pengurutan nilai dari kecil ke besar.

Median cocok digunakan untuk dataset yang memiliki data pencilan. Nilai median suatu dataset dapat ditentukan dengan memperhatikan jumlah data yang dimiliki.

1) Jika data berjumlah ganjil, maka nilai median data tepat berada di tengah setelah data diurutkan atau dapat dicari dengan rumus

$$Me = \frac{X_{(n+1)}}{2}$$

Dimana n = jumlah data

Contoh 1

Diketahui urutan data: 65, 70, 75, 80, 85

Jumlah data (n) = 5, sehingga

$$Me = \frac{5+1}{2} = 3$$

Maka median terletak pada data urutan ke-3 yaitu 75

2) Jika data berjumlah genap, maka nilai median diperoleh dengan mencari nilai rata-rata dua nilai yang berada di tengah-tengah data setelah diurutkan. Rumusnya (Agustina et al., 2025)

$$Me = \frac{(X_{\frac{n}{2}} + X_{(\frac{n}{2})+1})}{2}$$

Contoh 2

Diketahui urutan data: 65, 70, 75, 80, 85, 90

Jumlah data (n) = 6, sehingga

$$Me = \frac{(X_{\frac{6}{2}} + X_{(\frac{6}{2})+1})}{2}$$

$$Me = \frac{(75 + 80)}{2}$$

$$Me = 77.5$$

Median diperoleh dari mean data ke-3 dan ke-4 yaitu 75 dan 80, maka Me=(75+80)/2 = 77.5

c. Modus

Modus dalam sebuah dataset adalah nilai yang paling sering muncul. Berbeda dari mean dan median, modus cocok digunakan untuk data kategorik maupun numerik dan membantu mengisi nilai yang hilang (missing value) atau tidak lengkap saat melakukan preprocessing data, yaitu dengan nilai atau kategori yang dominan pada data

Contoh 3:

Diketahui urutan data: 70, 72, 84, 85, 85, 90
Maka Modus (Mo) = 85
Jika terdapat 8 data: 70, 72, 84, 85, 85, 90, ?, 93
Maka data ke-7 yang hilang atau tidak lengkap dapat diisi dengan 85

2. Ukuran Penyebaran Data

Ukuran penyebaran digunakan untuk menunjukkan tingkat keberagaman data serta menggambarkan seberapa jauh nilai-nilai dalam dataset tersebar dari nilai pusatnya (Kadir, 2022). Tujuannya adalah untuk memahami karakteristik data serta tingkat homogenitas atau heterogenitas data, dan mendeteksi kemungkinan adanya nilai yang menyimpang (outlier) dari pola umum. Oleh karena itu, jika ukuran pemusatan data memberikan informasi mengenai lokasi pusat data, maka ukuran penyebaran memberikan gambaran tentang tingkat variasi data. Hal ini penting dalam proses analisis dan pemodelan data untuk mempertimbangkan pengambilan keputusan berbasis data (Sudaryono, 2021). Beberapa ukuran penyebaran yang dibahas dalam bagian ini antara lain:

a. Rentang (Range)

Rentang atau range menunjukkan penyebaran data yang dihitung dari selisih antara nilai maksimum dan nilai minimum pada dataset.

$$Range = X_{max} - X_{min}$$

Semakin besar nilai rentang, semakin luas penyebaran data, sedangkan semakin kecil nilai rentang menunjukkan bahwa data cenderung terkonsentrasi pada rentang nilai yang sempit.

Contoh 4:

Dataset A: 70, 72, 74, 75, 76 dimana Xmin=70 dan Xmax=76

Sehingga Range = 76 – 70 = 6, ini menunjukkan bahwa data relatif berdekatan atau tidak terlalu bervariasi.

Contoh 5

Dataset B: 40, 55, 70, 85,100 dimana Xmin=40 dan Xmax=100 sehingga Range = 100 – 40 = 60, menunjukkan bahwa data tersebar lebih luas dan memiliki variasi yang lebih tinggi.

b. Simpangan Baku (Standar Deviasi)

Simpangan baku atau standar deviasi digunakan untuk mengukur tingkat penyebaran atau variasi data terhadap nilai rata-ratanya (mean). Standar deviasi menunjukkan seberapa jauh penyimpangan nilai-nilai dalam dataset dari nilai pusat data. Nilai standar deviasi yang tinggi, juga akan mempengaruhi tingginya keberagaman. Demikian juga sebaliknya, semakin kecil nilai standar deviasi menunjukkan bahwa data cenderung terkonsentrasi di sekitar rata-ratanya (Triola, 2022). Nilai standar deviasi dapat dihitung dengan rumus baik untuk populasi maupun sample adalah

$$\sigma = \sqrt{\frac{\sum_{i-1}^{n}(x_i - \bar{x})}{N}}$$

Dimana $\boldsymbol{\sigma}$ = standar deviasi untuk populasi, N= jumlah dataset, xi = nilai data ke-i, dan $\bar{x}$= rata-rata populasi

$$s = \sqrt{\frac{\sum_{i-1}^{n}(x_i - \bar{x})}{n-1}}$$

Dimana s = standar deviasi untuk data sample, n= jumlah dataset sample

c. Varians (Variance)

Varians adalah kuadarat dari simpangan baku yaitu $\boldsymbol{\sigma}^2$ dan s^2. Pada data sains, varians digunakan untuk mengukur keragaman atau penyebaran data terhadap nilai mean-nya sehingga dapat dipahami karakteristik data sebelum dilakukan analisis lebih lanjut. Varians yang tinggi menunjukkan bahwa data memiliki heterogenitas yang besar, sedangkan varians yang rendah menunjukkan bahwa data cenderung homogen (James et al., 2022).

B. PROBABILITAS

Probabilitas adalah ukuran yang menyatakan tingkat kemungkinan terjadinya suatu kejadian. Nilai probabilitas berada pada rentang (0 1), di mana nilai 0 menunjukkan kejadian yang mustahil terjadi dan nilai 1 menunjukkan kejadian yang pasti terjadi. Menurut James et al. (2022), probabilitas merupakan salah satu konsep dasar yang banyak digunakan dalam data sains dan machine learning untuk memodelkan hubungan antar variabel, mengukur ketidakpastian, dan membangun model prediksi untuk mendukung decision maker dalam mengambil keputusan.

Probabilitas suatu kejadian A (P(A)), jika diketahui jumlah peristiwa yang mendukung kejadian A (N(A)) dan jumlah semua peristiwa dalam ruang sampel (N(S)), dicari dengan rumus

$$P(A) = \frac{N(A)}{N(S)}$$

Teknik probabilitas yang paling umum sering digunakan meliputi distribusi probabilitas, Naïve Bayes, dan regresi logistik. Distribusi probabilitas digunakan untuk memahami karakteristik data, sedangkan Naïve Bayes dan regresi banyak diterapkan untuk tugas klasifikasi dan prediksi pada berbagai bidang aplikasi data sains (Agustina et al., 2025)

a. Distribusi Probabilitas

Distribusi probabilitas digunakan untuk memahami pola data, mengukur ketidakpastian, mendeteksi data pencilan (outlier), serta mendukung proses analisis dan pemodelan prediktif (James et al., 2022). Salah satu distribusi yang banyak digunakan dalam data sains adalah distribusi normal, yaitu distribuasi variabel kontinyu dengan fungsi matematis

$$f(x) = \frac{e^{[\frac{x-\mu}{2\sigma}]}}{\sqrt{2\pi\sigma^2}}$$

Dimana (e, π) adalah konstanta, x=nilai distribusi variabel, μ = mean dari nilai distribusi variabel, dan σ = standar deviasi.

b. Naïve Bayes

Naïve Bayes digunakan untuk mengklasifikasikan data dalam kategori tertentu dan melakukan prediksi data baru kedepannya berdasarkan probabilitas kemunculannya. Algoritma ini menggunakan dasar Teorema Bayes dimana teorema ini digunakan untuk menghitung probabilitas suatu kejadian (hipotesis) berdasarkan informasi atau bukti (evidence) yang telah

diketahui sebelumnya. Pada data sains, teorema ini banyak digunakan untuk memperbarui prediksi dan mendukung proses klasifikasi berbasis probabilitas(Sutojo et al., 2011). Bentuk Teorema Bayes evidence tunggal E dan hipotesis tunggal H adalah

$$p(H|E) = \frac{p(E|H)\ x\ p(H)}{p(E)}$$

dengan (P(H|E)) adalah probabilitas hipotesis (H) jika evidence (E) terjadi, (P(E|H)) adalah probabilitas evidence (E) jika hipotesis (H) terjadi, sedangkan (P(H)) dan (P(E)) masing-masing merupakan probabilitas awal hipotesis (H) dan evidence (E) sebelum mempertimbangkan informasi lain (Sutojo et al., 2011).

c. Regresi

Regresi pada data science bertujuan untuk membangun suatu fungsi yang memodelkan data dengan meminimalkan selisih antara nilai prediksi dengan nilai sesungguhnya (Suyanto, 2018). Hal ini untuk mengetahui ada tidaknya korelasi antar variabel serta mengidentifikasi pengaruh suatu variabel terhadap variabel lainnya serta memprediksi nilai yang akan datang berdasarkan data historis. Salah satu bentuk regresi yang paling sederhana dan banyak digunakan adalah regresi linear sederhana dengan rumus

$$Y = \alpha + bX$$

dengan Y sebagai variabel dependen (terikat), X sebagai variabel independen (bebas), α sebagai konstanta, dan b sebagai koefisien regresi yang menunjukkan besarnya perubahan Y akibat perubahan X.

C. ALJABAR LINIER DAN MATRIKS

Aljabar linier merupakan cabang matematika yang mempelajari ruang vektor beserta operasi-operasi yang berlaku di dalamnya (Axler, 2024). Konsep ruang vektor menjadi landasan utama dalam aljabar linier karena menyediakan kerangka matematis untuk merepresentasikan berbagai objek, seperti vektor, matriks, dan fungsi. Pada ruang vektor berlaku operasi penjumlahan dan perkalian skalar yang memenuhi sifat-sifat tertentu, seperti komutatif, asosiatif, identitas, dan invers (Axler, 2024). Matriks, dalam aljabar linier juga digunakan untuk menyajikan data dimana matriks tidak hanya merepresentasikan hubungan linear antar data, tapi juga untuk menyimpan, mengorganisasi, dan mengolah data dalam jumlah besar sehingga memudahkan proses analisis serta visualisasi, (Putra Hatoguan & Musllim Karo Karo, 2025).

Penerapan matriks dapat ditemukan dalam berbagai bidang, seperti analisis data geospasial, transformasi koordinat, pengolahan citra digital, menjadi fondasi bagi berbagai metode kecerdasan buatan dan Machine Learning karena data, parameter model, serta proses komputasi umumnya direpresentasikan dalam bentuk vektor dan matriks.

BAB 5
PEMOGRAMANAN DATA SCIENCE

Pemrograman Data Science digunakan untuk mengumpulkan data, membersihkan data, melakukan analisis, membangun model prediksi, serta menyajikan hasil dalam bentuk visualisasi yang mudah dipahami(Grus, 2019)(Provost & Fawcett, 2013). Berbagai bahasa pemrograman dapat digunakan dalam Data Science, seperti Python, R, Java, dan SQL. Namun, Python menjadi salah satu bahasa pemrograman yang paling populer dan banyak digunakan oleh akademisi maupun praktisi karena sintaksnya yang sederhana, mudah dipelajari, serta didukung oleh ekosistem pustaka (library) yang sangat lengkap untuk pengolahan data dan machine learning(McKinney, 2022)(Vanderplas, 2023).

Kemampuan pemrograman dalam Data Science tidak hanya dibutuhkan oleh seorang data scientist, tetapi juga oleh peneliti, dosen, mahasiswa, analis data, maupun pengembang sistem yang bekerja dengan data. Dengan menguasai pemrograman Data Science, seseorang dapat melakukan analisis data secara lebih cepat, akurat, dan efisien dibandingkan dengan pengolahan data secara manual. Pemrograman juga memungkinkan proses analisis dilakukan secara otomatis dan dapat diterapkan pada dataset berukuran besar (Grus, 2019).

Pada bab ini akan dibahas konsep dasar pemrograman Data Science menggunakan Python, mulai dari pengenalan lingkungan pengembangan, dasar-dasar pemrograman Python, penggunaan pustaka untuk pengolahan data, teknik visualisasi data, hingga implementasi machine learning sederhana. Pembahasan dalam bab ini diharapkan dapat memberikan pemahaman dasar mengenai bagaimana pemrograman digunakan sebagai alat utama dalam proses Data Science sehingga pembaca dapat mengembangkan kemampuan analisis data secara mandiri dan sistematis..

A. PERAN PEMROGRAMAN DALAM DATA SCIENCE

Pemrograman memiliki peranan penting dalam setiap tahapan Data Science, antara lain:

a. Akuisisi data dari berbagai sumber.
b. Pembersihan dan transformasi data.
c. Analisis eksploratif data (Exploratory Data Analysis/EDA).
d. Pembuatan model machine learning.
e. Evaluasi model.
f. Penyajian hasil analisis dalam bentuk laporan dan visualisasi.

Proses tersebut umumnya mengikuti metodologi CRISP-DM yang terdiri atas(IBM, 2021):

a. Business Understanding

   Tahap awal yang bertujuan untuk memahami permasalahan, kebutuhan, dan tujuan bisnis atau penelitian yang ingin dicapai. Pada tahap ini ditentukan tujuan proyek, ruang lingkup pekerjaan, serta indikator keberhasilan yang akan digunakan.

b. Data Understanding

   Tahap pengumpulan dan pemahaman data yang akan digunakan dalam analisis. Aktivitas yang dilakukan meliputi pengumpulan data, eksplorasi data awal, identifikasi karakteristik data, serta menemukan masalah seperti data hilang, data duplikat, atau data yang tidak konsisten.

c. Data Preparation

   Tahap persiapan data agar siap digunakan dalam proses analisis dan pemodelan. Kegiatan yang dilakukan meliputi pembersihan data (data cleaning), transformasi data, penggabungan data, pemilihan atribut yang relevan,

serta penanganan nilai yang hilang atau tidak valid.

d. Modeling

Tahap pembangunan model analitik atau machine learning berdasarkan data yang telah dipersiapkan. Berbagai algoritma dapat digunakan sesuai kebutuhan, seperti regresi, klasifikasi, clustering, atau metode lainnya untuk menghasilkan pola dan prediksi dari data.

e. Evaluation

Tahap evaluasi untuk menilai kinerja model yang telah dibuat. Hasil model dibandingkan dengan tujuan awal proyek guna memastikan bahwa model mampu memberikan hasil yang akurat dan sesuai dengan kebutuhan pengguna atau organisasi.

f. Deployment

Tahap implementasi model ke dalam lingkungan nyata agar dapat digunakan oleh pengguna. Deployment dapat berupa integrasi ke aplikasi, dashboard, sistem informasi, atau penyajian laporan yang mendukung pengambilan keputusan berdasarkan hasil analisis data.

Metodologi ini menjadi salah satu standar yang banyak digunakan dalam proyek Data Science.

B. BAHASA PEMROGRAMAN UNTUK DATA SCIENCE

Beberapa bahasa pemrograman yang banyak digunakan dalam Data Science antara lain:

a. Python

Python merupakan bahasa pemrograman yang paling populer untuk Data Science karena:

1) Sintaks sederhana dan mudah dipelajari.
2) Memiliki komunitas yang besar.
3) Tersedia banyak library Data Science.

4) Mendukung machine learning dan deep learning.

b. R

Bahasa R banyak digunakan untuk:

1) Analisis statistik.
2) Visualisasi data.
3) Penelitian akademik.

c. SQL

SQL digunakan untuk:

1) Mengakses database.
2) Mengambil data dari sistem informasi.
3) Melakukan agregasi data.

C. LINGKUNGAN PENGEMBANGAN PYTHON

Sebelum memulai pemrograman Data Science, diperlukan lingkungan kerja (environment) yang mendukung penulisan, eksekusi, dan pengelolaan kode program. Beberapa lingkungan pengembangan yang umum digunakan antara lain Python, Anaconda, Jupyter Notebook, dan Google Colab.

a. Instalasi Python

Python dapat diunduh dan tahap instalasi dapat dilihat melalui: https://www.python.org

b. Instalasi Anaconda

Anaconda merupakan distribusi Python yang dirancang khusus untuk kebutuhan Data Science dan Machine Learning. Anaconda menyediakan berbagai library dan tools yang telah terintegrasi sehingga memudahkan proses instalasi.

Paket yang tersedia dalam Anaconda antara lain:

1) Python
2) Jupyter Notebook
3) NumPy
4) Pandas

5) Matplotlib
6) Scikit-Learn

Website: https://www.anaconda.com

c. Jupyter Notebook

Jupyter Notebook merupakan aplikasi interaktif berbasis web yang memungkinkan pengguna menulis, menjalankan, dan mendokumentasikan kode program dalam satu lingkungan kerja. Meskipun diakses melalui browser, Jupyter Notebook umumnya dijalankan pada komputer lokal pengguna.

Website : https://jupyter.org

d. Google Colab

Google Colaboratory atau Google Colab merupakan lingkungan pengembangan Python berbasis cloud yang disediakan oleh Google secara gratis. Google Colab memungkinkan pengguna menjalankan kode Python langsung melalui browser tanpa perlu melakukan instalasi Python maupun library Data Science pada komputer lokal.

Website: https://colab.research.google.com

D. LIBRARY UTAMA DATA SCIENCE

1. Numpy

NumPy digunakan untuk komputasi numerik.

Contoh:

```
import numpy as nmp
dt = nmp.array([75, 60, 80, 70, 45])

print("Rata-rata:", nmp.mean(dt))
print("Nilai maksimum:", nmp.max(dt))
```

Output:

```
Rata-rata: 66.0
Nilai maksimum: 80
```

2. Pandas

Pandas digunakan untuk manipulasi data tabular.

Contoh:

```
import pandas as pd

data = {
        'Nama': ['Andi', 'Budi', 'Citra'],
        'Umur': [20, 21, 19],
        'Nilai': [75, 80, 68]
}
df = pd.DataFrame(data)
print(df)
```

Output:

```
   Nama  Umur  Nilai
0  Andi    20     75
1  Budi    21     80
2 Citra    19     68
```

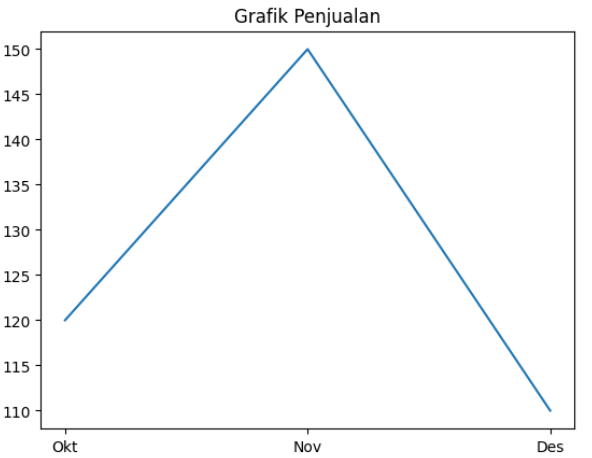
3. Matplotlib

Digunakan untuk visualisasi data.

Contoh :

```
import matplotlib.pyplot as mplt
bulan = ['Okt', 'Nov', 'Des']
penjualan = [120, 150, 110]

mplt.plot(bulan, penjualan)
mplt.title("Grafik Penjualan")

mplt.show()
```

Hasil Ouput



Gambar 1. Grafik Output Matplotlib

4. Scikit-Learning

Digunakan untuk membangun model machine learning seperti:

a. Klasifikasi
b. Regresi
c. Clustering
d. Dimensionality Reduction

E. PENGOLAHAN DATA MENGGUNAKAN PANDAS

Pandas menyediakan berbagai fungsi untuk mengelola dan menganalisis data secara efisien. Berikut beberapa operasi dasar pengolahan data menggunakan Pandas

Membaca File CSV

```
import pandas as pds

dfile = pds.read_csv('mahasiswa.csv')
print(dfile.head())
```

```
Menampilkan Informasi Dataset
print(dfile.info())

Menghapus Data Kosong
dfile = dfile.dropna()

Menghitung Statistik Dasar
print(dfile.describe())
```

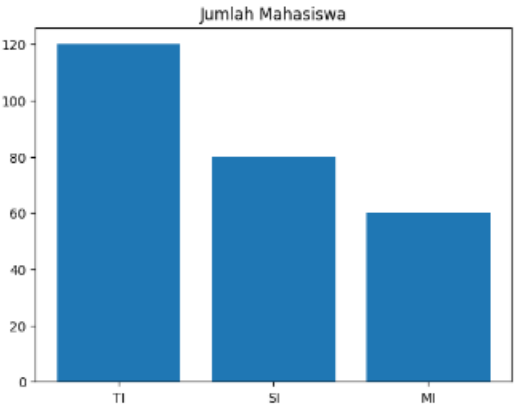
F. VISUALISASI DATA

Visualisasi data membantu memahami pola dan tren dalam data.

a. Grafik Batang

```
import matplotlib.pyplot as plt

jurusan = ['TI', 'SI', 'MI']
jumlah = [120, 80, 60]

plt.bar(jurusan, jumlah)
plt.title('Jumlah Mahasiswa')

plt.show()
```

Hasil Output

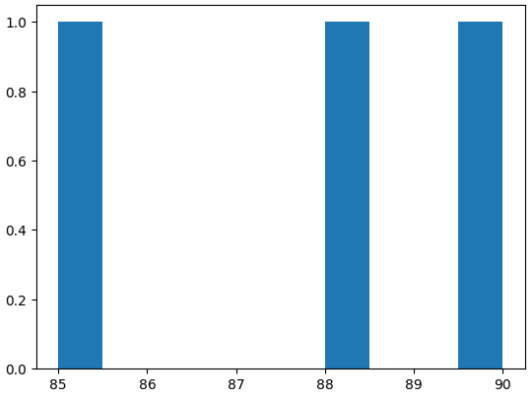


Gambar 2. Grafik Batang

b. Histogram

```
plt.hist(df['Nilai'])
plt.show()
```

Hasil Output

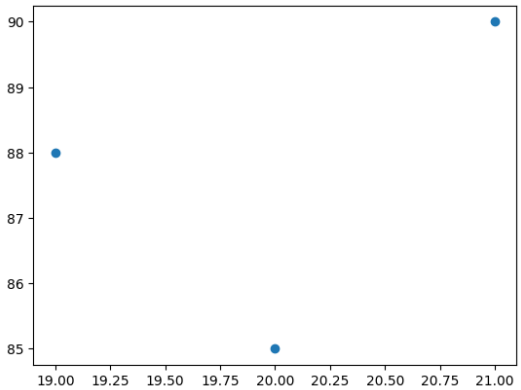


Gambar 3. Histogram Distribusi Nilai

c. Scatter Plot

```
plt.scatter(df['Umur'], df['Nilai'])
plt.show()
```

Hasil Output



Gambar 4. Diagram Pencar (Scatter Plot)

G. IMPLEMENTASI MACHINE LEARNING SEDERHANA

Regresi linear merupakan salah satu algoritma machine learning yang digunakan untuk memodelkan hubungan antara variabel independen (X) dan variabel dependen (Y). Tujuan regresi linear adalah menemukan garis terbaik yang menggambarkan hubungan antara kedua variabel tersebut. Sebagai contoh, semakin banyak waktu yang digunakan mahasiswa untuk belajar, maka biasanya nilai yang diperoleh juga akan meningkat. Hubungan tersebut dapat dimodelkan menggunakan regresi linear.

Tabel 1. Dataset Jam Belajar dan Nilai

Jam Belajar	Nilai
4	50
5	55
6	65
7	70
8	80

Berdasarkan data tersebut, model akan mempelajari hubungan antara jumlah jam belajar dan nilai yang diperoleh mahasiswa.

a. Pelatihan Model

Proses pelatihan (training) merupakan tahap di mana algoritma machine learning mempelajari pola yang terdapat dalam data. Pada Python, proses ini dapat dilakukan menggunakan library Scikit-Learn.

```
from sklearn.linear_model import LinearRegression
import pandas as pds

dt = {
'JamBelajar':[4,5,6,7,8], 'Nilai':[50,55,65,70,80]
}
```

```
dfile = pds.DataFrame(dt)

x = dfile[['JamBelajar']]
y = dfile['Nilai']

modelLR = LinearRegression()
modelLR.fit(x,y)
```

Pada program tersebut, variabel JamBelajar digunakan sebagai fitur (x), sedangkan variabel Nilai digunakan sebagai target (y). Fungsi fit() digunakan untuk melatih model berdasarkan data yang tersedia.

b. Prediksi

Setelah model berhasil dilatih, langkah berikutnya adalah melakukan prediksi terhadap data baru. Misalnya ingin diketahui nilai yang diperkirakan akan diperoleh mahasiswa yang belajar selama 9 jam.

```
prediksi = modelLR.predict([[9]])
print("Prediksi Nilai:", prediksi[0])
```

Output yang dihasilkan:

```
Prediksi Nilai: 86.5
```

Hasil tersebut menunjukkan bahwa model memperkirakan mahasiswa yang belajar selama 9 jam akan memperoleh nilai sekitar 86,5.

Dalam implementasi machine learning, terdapat beberapa istilah penting yang perlu dipahami:

a. Dataset

Dataset adalah kumpulan data yang digunakan dalam proses pembelajaran mesin. Dataset terdiri atas fitur (input) dan target (output) yang akan dipelajari oleh model.

b. Training

Training merupakan proses melatih model menggunakan data yang tersedia agar model dapat

mengenali pola dan hubungan antar variabel.

c. Testing

Testing adalah proses menguji performa model menggunakan data yang tidak digunakan saat pelatihan. Tujuannya untuk mengetahui kemampuan model dalam melakukan prediksi terhadap data baru.

d. Prediksi

Prediksi merupakan hasil keluaran model setelah menerima data masukan baru. Kualitas prediksi sangat dipengaruhi oleh kualitas data dan metode yang digunakan selama proses pelatihan.

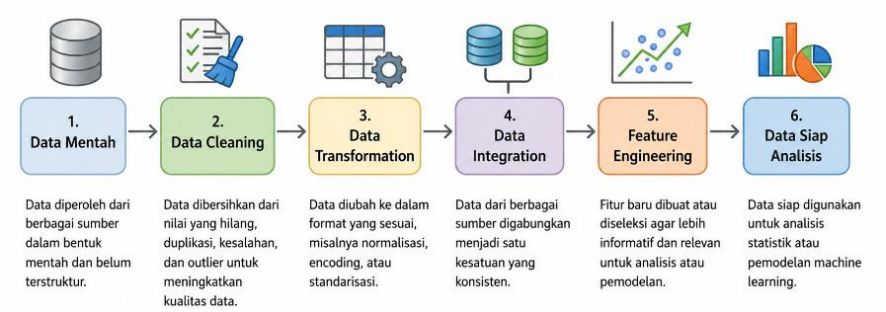
Pemrograman merupakan fondasi utama dalam Data Science. Python menjadi bahasa yang paling banyak digunakan karena didukung oleh berbagai pustaka seperti NumPy, Pandas, Matplotlib, dan Scikit-Learn. Dengan memanfaatkan bahasa pemrograman dan library tersebut, proses pengolahan data, visualisasi, hingga machine learning dapat dilakukan secara efektif dan efisien.

BAB 6
PENGOLAHAN DAN PERSIAPAN DATA

Pengolahan dan persiapan data (data processing dan data preparation) merupakan tahapan awal dalam Data Science yang bertujuan mengubah data mentah menjadi data yang siap dianalisis. Proses ini mencakup pengumpulan, integrasi, pembersihan, transformasi, dan penyusunan data agar memiliki kualitas yang baik serta sesuai dengan kebutuhan analisis (Ortiz et al., 2024).

Tahap persiapan data sangat penting karena kualitas data akan memengaruhi hasil analisis dan performa model yang dibangun. Data yang tidak lengkap, tidak konsisten, atau mengandung kesalahan dapat menghasilkan prediksi yang kurang akurat dan keputusan yang tidak tepat. Oleh karena itu, sebagian besar proyek Data Science menempatkan persiapan data sebagai salah satu aktivitas utama sebelum proses pemodelan dilakukan (Chicco et al., 2022).

Secara umum, pengolahan dan persiapan data mencakup pembersihan data (data cleaning), transformasi data, penanganan nilai hilang (missing values), serta standarisasi format data. Tahapan ini bertujuan menghasilkan dataset yang akurat, konsisten, dan siap digunakan untuk analisis statistik maupun machine learning (Steorts, 2023).



Gambar 1. Tahapan Pengolahan dan Persiapan Data dalam Data Science

A. KONSEP DASAR PENGOLAHAN DAN PERSIAPAN DATA

Pembersihan data (data cleaning atau data cleansing) merupakan proses identifikasi, perbaikan, transformasi, dan penghapusan data yang tidak akurat, tidak lengkap, tidak konsisten, atau duplikat sehingga menghasilkan dataset yang berkualitas dan siap digunakan untuk analisis maupun pemodelan machine learning. Dalam siklus Data Science, tahap ini menjadi bagian penting dari proses data preprocessing karena kualitas data secara langsung memengaruhi kualitas hasil analisis dan performa model yang dibangun. Berbagai penelitian menunjukkan bahwa sebagian besar permasalahan dalam proyek Data Science berasal dari kualitas data yang rendah, seperti nilai hilang (missing values), data duplikat, kesalahan format, inkonsistensi data, serta keberadaan nilai ekstrem (outlier) (Abdelaal et al., 2023).

Menurut Rahm dan Do (2000) pembersihan data merupakan proses yang bertujuan untuk meningkatkan kualitas data melalui identifikasi dan koreksi kesalahan, inkonsistensi, serta ketidaklengkapan data yang terdapat dalam suatu basis data. Proses ini menjadi semakin penting seiring meningkatnya volume,

variasi, dan kompleksitas data yang digunakan dalam berbagai aplikasi analitik. Data yang tidak dibersihkan dengan baik dapat menyebabkan kesalahan interpretasi, menurunkan kualitas informasi, dan mengurangi efektivitas proses pengambilan keputusan.

Berbagai penelitian dalam bidang kualitas data menjelaskan bahwa pembersihan data tidak hanya berfokus pada perbaikan kesalahan individual, tetapi juga mencakup upaya menjaga konsistensi data secara keseluruhan melalui standardisasi, validasi, dan integrasi data dari berbagai sumber. Dalam konteks Data Science modern, proses ini menjadi fondasi penting sebelum dilakukan eksplorasi data, analisis statistik, maupun pembangunan model machine learning. Dengan demikian, data cleaning dapat dipandang sebagai langkah awal untuk memastikan bahwa data yang digunakan benar-benar merepresentasikan kondisi yang sebenarnya.

Tabel 1. Permasalahan Data dan Teknik Penanganannya

Permasalahan Data	**Deskripsi**	**Teknik Penanganan**
Missing Values	Nilai data hilang atau kosong	Mean Imputation, Median Imputation, Mode Imputation, KNN Imputation
Data Duplikat	Data tercatat lebih dari satu kali	Deduplikasi (Remove Duplicates)
Outlier	Nilai yang sangat berbeda dari mayoritas data	IQR, Z-Score, Isolation Forest
Inkonsistensi Data	Format atapenulisan tidak seragam	Standardisasi dan Validasi Data
Kesalahan Tipe Data	Tipe data tidak sesuai atribut	Konversi dan Transformasi Data

Dalam praktiknya, terdapat beberapa permasalahan umum

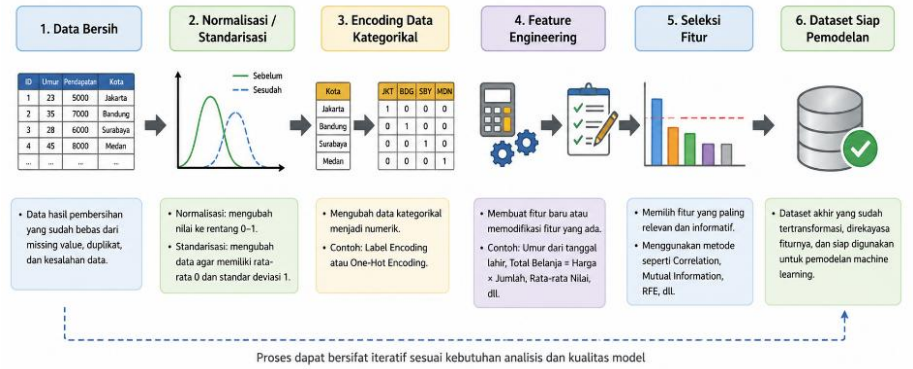
yang sering ditemukan pada dataset. Permasalahan tersebut perlu ditangani dengan teknik yang sesuai agar kualitas data dapat ditingkatkan sebelum digunakan untuk analisis lebih lanjut.

B. TRANSFORMASI DAN REKAYASA FITUR DATA

Transformasi data (data transformation) merupakan proses mengubah format, struktur, atau representasi data menjadi bentuk yang lebih sesuai untuk kebutuhan analisis dan pemodelan. Tahap ini dilakukan setelah proses data cleaning dengan tujuan meningkatkan kualitas data sehingga dapat diproses secara optimal oleh algoritma machine learning. Dalam praktik Data Science, data yang berasal dari berbagai sumber sering memiliki skala, format, dan tipe data yang berbeda sehingga memerlukan proses transformasi agar lebih konsisten dan mudah dianalisis. Menurut Koukaras dan Tjortjis (2025) transformasi data berperan penting dalam meningkatkan akurasi, interpretabilitas, dan performa model prediktif. Beberapa teknik transformasi yang umum digunakan meliputi normalisasi (normalization), standardisasi (standardization), encoding data kategorikal, transformasi logaritmik, serta reduksi dimensi menggunakan metode seperti Principal Component Analysis (PCA).

Selain transformasi data, proses yang tidak kalah penting adalah rekayasa fitur (feature engineering). Rekayasa fitur merupakan proses membentuk atribut baru atau memodifikasi atribut yang telah tersedia agar memiliki nilai informasi yang lebih tinggi. Pada Gambar 2 ditunjukkan bahwa proses ini dapat dilakukan dengan menghasilkan fitur baru, seperti umur dari tanggal lahir, total belanja dari harga dan jumlah pembelian, serta rata-rata nilai dari beberapa komponen penilaian. Pembentukan fitur tersebut bertujuan meningkatkan kualitas representasi data sehingga proses analisis dan pemodelan dapat menghasilkan

kinerja yang lebih optimal.



Gambar 2. Proses Transformasi dan Rekayasa Fitur Data

Chicco et al (2022) menyebutkan bahwa kualitas fitur yang digunakan sering kali memberikan pengaruh yang lebih besar terhadap performa model dibandingkan pemilihan algoritma yang digunakan. Melalui rekayasa fitur, data mentah dapat diubah menjadi informasi yang lebih bermakna sehingga model mampu mengenali pola dan hubungan antarvariabel secara lebih efektif. Contoh sederhana rekayasa fitur adalah pembentukan variabel umur dari tanggal lahir, pengelompokan kategori tertentu, atau penggabungan beberapa atribut menjadi fitur baru yang lebih informatif.

Perkembangan teknologi kecerdasan buatan telah mendorong munculnya berbagai pendekatan otomatis dalam proses transformasi dan rekayasa fitur. Penelitian Mumuni & Mumuni (2024) menjelaskan bahwa teknologi Automated Machine Learning (AutoML) mampu melakukan berbagai tahapan seperti normalisasi data, encoding, seleksi fitur, hingga pembentukan fitur baru secara otomatis. Pendekatan ini membantu meningkatkan efisiensi pengolahan data, terutama pada dataset berukuran besar dan kompleks. Dengan demikian, transformasi dan rekayasa fitur tidak hanya berfungsi meningkatkan kualitas data, tetapi juga

menjadi faktor penting yang menentukan keberhasilan analisis, akurasi model machine learning, serta kualitas pengambilan keputusan berbasis data.

C. IMPLEMENTASI PENGOLAHAN DATA DALAM DATA SCIENCE

Implementasi pengolahan data (data processing implementation) merupakan penerapan berbagai teknik, metode, dan perangkat lunak untuk mengubah data mentah menjadi informasi yang bernilai bagi proses analisis, pengambilan keputusan, maupun pengembangan model kecerdasan buatan (Artificial Intelligence/AI) dan Machine Learning. Dalam Data Science, pengolahan data mencakup serangkaian aktivitas mulai dari pengumpulan, integrasi, pembersihan, transformasi, hingga penyajian data agar siap digunakan pada tahap analisis lanjutan. Kualitas data menjadi faktor utama yang menentukan keberhasilan sistem AI dan Machine Learning karena data yang tidak siap digunakan dapat menurunkan akurasi model serta meningkatkan risiko kesalahan analisis (Ortiz et al., 2024). Oleh karena itu, implementasi pengolahan data berperan penting dalam memastikan data memiliki kualitas, konsistensi, dan reliabilitas yang memadai sebelum digunakan dalam proses analitik.

Dalam praktiknya, pengolahan data dilakukan menggunakan berbagai perangkat lunak dan teknologi seperti Microsoft Excel, SQL, Python, R, Apache Spark, maupun platform analitik berbasis cloud. Python menjadi salah satu bahasa pemrograman yang paling populer karena menyediakan berbagai pustaka seperti Pandas, NumPy, dan Scikit-learn yang mendukung proses pengolahan data secara efisien. Selain itu, perkembangan teknologi Big Data juga mendorong penggunaan kerangka kerja pemrosesan data berskala besar untuk menangani volume data yang terus meningkat.

Pemilihan teknik dan alat pengolahan data yang tepat dapat meningkatkan efisiensi proses analisis sekaligus menghasilkan dataset yang lebih siap untuk kebutuhan data mining dan machine learning (Koukaras & Tjortjis, 2025).

Pada proyek Data Science modern, implementasi pengolahan data umumnya menjadi bagian penting dalam siklus pengembangan solusi berbasis data. Sebagai contoh, pada sistem prediksi penjualan, data transaksi yang berasal dari berbagai sumber sering kali mengandung nilai kosong, data duplikat, format yang tidak seragam, maupun atribut yang belum sesuai dengan kebutuhan analisis. Melalui proses pengolahan data, berbagai permasalahan tersebut dapat diperbaiki sehingga menghasilkan dataset yang lebih terstruktur dan siap digunakan untuk membangun model prediksi. Shimaoka et al (2024) menjelaskan bahwa keberhasilan proyek Data Science sangat dipengaruhi oleh kualitas tahapan persiapan data dalam kerangka kerja CRISP-DM, karena data yang telah diproses dengan baik akan mempermudah proses pemodelan dan evaluasi. Meskipun berbagai teknologi otomatisasi telah berkembang pesat, pemahaman manusia terhadap konteks dan karakteristik data tetap menjadi faktor penting untuk memastikan hasil analisis yang akurat dan dapat dipercaya.

Tabel 2. Tahapan Pengolahan Data dan Tujuannya

Tahap	Tujuan
Data Collection	Mengumpulkan data dari berbagai sumber
Data Understanding	Memahami karakteristik dan kualitas data
Data Cleaning	Menghilangkan kesalahan, duplikasi, dan data tidak lengkap
Data Transformation	Mengubah data ke format yang sesuai untuk analisis
Feature Engineering	Membentuk fitur baru yang lebih informatif

Data Analysis / Modeling	Membangun model analitik atau machine learning
Evaluation	Mengukur kualitas dan performa model
Deployment	Mengimplementasikan model pada lingkungan nyata

Pengolahan data dalam Data Science merupakan proses yang sistematis dan saling terkait antar tahapan. Proses dimulai dari pengumpulan data (data collection) dan pemahaman data (data understanding) untuk mengenali karakteristik serta kualitas dataset yang tersedia. Selanjutnya dilakukan pembersihan data (data cleaning), transformasi data (data transformation), dan rekayasa fitur (feature engineering) guna menghasilkan data yang lebih bersih, konsisten, dan informatif. Data yang telah dipersiapkan kemudian digunakan pada tahap analisis dan pemodelan (data analysis/modeling), diikuti dengan evaluasi untuk mengukur performa model yang dihasilkan. Tahap terakhir adalah deployment, yaitu implementasi model pada lingkungan nyata sehingga hasil analisis dapat dimanfaatkan untuk mendukung pengambilan keputusan. Dengan demikian, setiap tahapan memiliki peran penting dalam memastikan data yang digunakan mampu menghasilkan informasi dan pengetahuan yang bernilai bagi organisasi.

D. TANTANGAN PENGOLAHAN DATA DI ERA BIG DATA

Perkembangan teknologi digital, Internet of Things (IoT), media sosial, cloud computing, dan kecerdasan buatan telah menyebabkan pertumbuhan data dalam jumlah yang sangat besar. Fenomena ini dikenal sebagai Big Data, yaitu kumpulan data yang memiliki karakteristik volume, velocity, dan variety yang melampaui kemampuan sistem pengolahan data tradisional. Dalam

konteks Data Science, tantangan pengolahan data tidak lagi hanya berkaitan dengan bagaimana memperoleh data, tetapi juga bagaimana memastikan data tersebut berkualitas, aman, dan dapat diproses secara efisien untuk menghasilkan informasi yang bernilai. Menurut Shahnawaz dan Kumar (2025), keberhasilan implementasi Big Data sangat dipengaruhi oleh kemampuan organisasi dalam mengelola data secara efektif sepanjang siklus pengolahannya.

Salah satu tantangan utama adalah volume data yang terus meningkat serta kecepatan data yang dihasilkan secara real-time dari berbagai sumber. Selain itu, keragaman format data juga menjadi kendala karena data modern tidak hanya berbentuk tabel terstruktur, tetapi juga berupa teks, gambar, video, dokumen, maupun data sensor. Kondisi ini menyebabkan proses integrasi, penyimpanan, dan analisis data menjadi lebih kompleks. Lalaoui et al (2025) menjelaskan bahwa organisasi memerlukan infrastruktur dan teknologi yang mampu mendukung pemrosesan data dalam skala besar agar informasi dapat diolah secara cepat dan akurat.

Tantangan lainnya berkaitan dengan kualitas data, privasi, keamanan informasi, serta tata kelola data (data governance). Data yang tidak lengkap, mengandung duplikasi, atau inkonsistensi dapat menurunkan kualitas hasil analisis. Di sisi lain, meningkatnya penggunaan data dalam berbagai sektor juga menuntut organisasi untuk menerapkan perlindungan data yang lebih baik dan mematuhi berbagai regulasi yang berlaku. Tata kelola data yang baik menjadi faktor penting dalam memastikan data dapat dimanfaatkan secara optimal sekaligus menjaga keamanan dan kepercayaan pengguna (Yukhno, 2024). Oleh karena itu, keberhasilan implementasi Data Science tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh kemampuan organisasi dalam mengelola data secara efektif, aman,

dan berkelanjutan.

Tabel 3. Tantangan Pengolahan Data di Era Big Data

Tantangan	Dampak
Volume Data Besar	Membutuhkan kapasitas penyimpanan dan komputasi yang tinggi
Velocity Data Tinggi	Memerlukan pemrosesan data secara real-time
Variety Data	Menyulitkan integrasi dan analisis data
Kualitas Data Rendah	Menurunkan akurasi analisis dan prediksi
Privasi dan Keamanan	Risiko kebocoran dan penyalahgunaan data
Data Governance	Kesulitan dalam pengelolaan dan kepatuhan regulasi
Infrastruktur dan SDM	Membutuhkan investasi teknologi dan tenaga ahli

Pengolahan data di era Big Data menghadapi berbagai tantangan yang tidak hanya berkaitan dengan jumlah data yang semakin besar, tetapi juga kecepatan pertumbuhan data, keragaman format, kualitas data, serta aspek keamanan dan tata kelola. Tantangan-tantangan tersebut menuntut organisasi untuk memiliki infrastruktur teknologi yang memadai, sumber daya manusia yang kompeten, serta strategi pengelolaan data yang efektif. Oleh karena itu, keberhasilan implementasi Data Science tidak hanya ditentukan oleh kemampuan dalam membangun model analitik atau machine learning, tetapi juga oleh kemampuan mengelola data secara terintegrasi sehingga informasi yang dihasilkan tetap akurat, aman, dan bernilai bagi pengambilan keputusan.

BAB 7
EKSPLORASI DAN VISUALISASI DATA

Segala aktivitas yang kita lakukan saat ini menghasilkan data mulai dari transaksi belanja, penggunaan media sosial, layanan kesehatan, pendidikan, hingga perangkat yang terhubung melalui internet. Data tersebut menyimpan berbagai macam informasi yang mungkin bisa memberikan nilai bagi individu maupun organisasi. Namun, keberadaan data dalam jumlah besar tadi tidak serta-merta langsung dapat menghasilkan wawasan yang bernilai karena data perlu dipahami, dianalisis, dan diinterpretasikan agar dapat mendukung proses pengambilan keputusan secara lebih efektif.

Dalam bidang ilmu Data Science, proses memahami data merupakan salah satu tahapan yang sangat penting sebelum melakukan analisis lanjutan. Pada tahap ini, Data Analyst atau Data Scientist berusaha untuk mengenali karakteristik data yang dimiliki, menemukan pola-pola tersembunyi, mengidentifikasi kemungkinan permasalahan pada data, hingga memperoleh gambaran awal mengenai informasi yang terdapat di dalamnya. Pemahaman yang baik terhadap data akan membantu kita untuk memastikan bahwa langkah-langkah analisis selanjutnya dilakukan berdasarkan informasi yang tepat.

Eksplorasi dan visualisasi data menjadi dua aktivitas yang sangat penting dalam proses memahami data. Eksplorasi data membantu kita untuk dapat memahami informasi awal data secara lebih mendalam, sedangkan visualisasi data bertujuan dalam menyajikan informasi tersebut ke dalam bentuk yang lebih mudah dipahami. Melalui eksplorasi dan visualisasi yang tepat, pola, tren,

hubungan antarvariabel, maupun anomali dalam data dapat terlihat dengan lebih cepat dibandingkan jika kita hanya melihat melalui deretan angka ataupun tabel.

Pada bab ini kita akan membahas konsep dasar eksplorasi dan visualisasi data sebagai bagian penting dalam alur kerja Data Science. Pembahasan mencakup konsep eksplorasi data, teknik-teknik eksplorasi yang umum digunakan, prinsip dasar visualisasi data dan berbagai jenis visualisasi beserta penggunaannya.

A. KONSEP EKSPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis atau EDA merupakan proses mengeksplorasi data untuk memperoleh pemahaman awal mengenai karakteristik, pola, hubungan, atau potensi permasalahan yang terdapat di dalamnya (Bruce et al., 2020). Fokus utama EDA bukan untuk menghasilkan model prediksi atau membuat kesimpulan akhir melainkan membantu analis memahami insight yang tersembunyi di balik kumpulan data yang sedang diamati.

Langkah ini sering kali dianggap sederhana namun kenyataannya di lapangan banyak keputusan analitis yang kurang tepat justru berawal dari pemahaman yang kurang mendalam terhadap data yang digunakan. Setelah data diperoleh seorang data analyst/scientist tidak langsung melakukan pemodelan, ia akan berusaha memahami kondisi data terlebih dahulu melalui berbagai pendekatan eksploratif sehingga dari proses inilah sering muncul berbagai temuan awal yang membantu menentukan arah analisis berikutnya. Dengan kata lain EDA berfungsi sebagai jembatan antara data mentah dan proses analisis yang lebih mendalam.

Dalam praktiknya, terdapat beberapa tujuan yang umumnya ingin dicapai melalui proses eksplorasi data. Yang pertama, memahami karakteristik dasar data. Pada tahap ini data

analyst/scientist berusaha mengenali bentuk umum dataset yang dimiliki mulai dari jenis data, ukuran data, dan gambaran distribusinya. Pemahaman awal ini menjadi fondasi bagi langkah-langkah berikutnya (D et al., 2024).

Tujuan kedua yaitu untuk menemukan pola dan kecenderungan yang muncul dari data. Tidak semua pola dapat terlihat secara langsung ketika data disajikan dalam bentuk tabel, kadang-kadang pola tersebut baru tampak setelah data diamati dari berbagai sudut pandang atau melakukan agregasi tertrntu.

Tujuan ketiga untuk mengidentifikasi anomali atau kondisi yang tidak biasa, contohnya seperti nilai-nilai ekstrem (outlier). Nilai ekstrem ini bisa muncul karena terdapat kesalahan pencatatan, atau bisa juga karena memang mencerminkan kejadian khusus di dunia nyata (McKinney, 2022).

Tujuan keempat, mendukung proses pengambilan keputusan dengan cara yang lebih sederhana. Hasil eksplorasi data sering kali menghasilkan wawasan awal yang bermanfaat bagi organisasi maupun individu bahkan sebelum analisis lanjutan dilakukan, sehingga proses eksplorasi sering kali sudah mampu menjawab sebagian pertanyaan yang diajukan.

Satu hal lagi yang perlu kita ingat adalah bahwa EDA bukan sekedar hal yang sifatnya teknikal. Di balik berbagai macam teknik dan tools yang bisa kita gunakan, terdapat proses berpikir kritis yang menjadi inti dari eksplorasi data. Seorang data analyst/scientist tidak hanya melihat output yang muncul, tapi juga berusaha memahami makna yang terkandung di baliknya. Untuk memahami makna hasil eksplorasi data dapat dibantu dengan melakukan business understanding terhadap domain dari data yang kita punya.

Melalui EDA, seorang data analyst/scientist dapat membangun pemahaman yang lebih kuat sebelum masuk ke proses analisis berikutnya. Pemahaman inilah yang nantinya akan

membantu menghasilkan analisis yang lebih akurat.

B. TEKNIK EKSPLORASI DATA

Eksplorasi data atau biasa disebut juga dengan Exploratory Data Analysis (EDA) dapat dilakukan dengan berbagai macam teknik sesuai dengan kondisi data yang kita punya. Salah satu pendekatan yang paling umum adalah dengan mengelompokkan analisis (eksplorasi) berdasarkan jumlah variabel pada data, yaitu analisis univariat, bivariat, dan multivariat.

1. Analisis Univariat

Analisis univariat merupakan bentuk eksplorasi yang berfokus pada satu variabel saja. Tujuan utamanya adalah memahami karakteristik suatu variabel secara individual tanpa mempertimbangkan hubungan dengan variabel lain (Bruce et al., 2020).

Melalui analisis univariat seorang data analyst/scientist dapat memperoleh wawasan awal mengenai data misalnya seperti gambaran umum (deskriptif), sebaran dan juga distribusi data. Selain untuk memahami kondisi data, analisis univariat juga digunakan untuk menemukan nilai yang tidak biasa atau pola distribusi yang unik. Temuan-temuan awal seperti ini dapat menjadi petunjuk untuk analisis yang lebih mendalam.

2. Analisis Bivariat

Jika analisis univariat berfokus pada satu variabel, maka analisis bivariat melibatkan dua variabel secara bersamaan. Fokus utama pendekatan ini adalah memahami apakah terdapat hubungan, keterkaitan, atau kecenderungan tertentu di antara kedua variabel tersebut (Bruce et al., 2020).

Sebagai contoh, seorang data analis ingin mengetahui apakah terdapat hubungan antara tingkat kepuasan pelanggan dengan

frekuensi pembelian produk, sehingga dalam kasus ini dua variabel yang akan dianalisis yaitu "tingkat kepuasan pelanggan" dan "frekuensi pembelian produk". Melalui analisis bivariat, berbagai hubungan semacam ini dapat diamati secara lebih jelas.

Analisis bivariat membantu kita untuk melihat data dari perspektif yang lebih luas. Karena ketika dua variabel diamati secara bersamaan sering kali muncul wawasan baru yang sebelumnya tidak terlihat saat masing-masing variabel dianalisis secara terpisah.

Berdasarkan tipe data pada masing-masing variabel nya analisis bivariat dapat dibedakan menjadi beberapa jenis yang dijelaskan pada tabel berikut.

Tabel 1. Kombinasi variabel pada analisis bivariat

Variabel 1	Variabel 2	Analisis	Teknik Analisis
Numerik	Numerik	Korelasi	Uji korelasi (pearson/spearman), visualisasi scatter plot
Kategorik	Kategorik	Asosiasi	Uji Chi-Square
Kategorik	Numerik	Uji perbandingan	Uji T-Test

3. Analisis Multivariat

Pada studi kasus yang lebih rumit, sesuatu biasanya tidak dipengaruhi oleh satu atau dua variabel saja. Banyak faktor dapat saling berinteraksi dan membentuk pola tertentu. Dalam kondisi seperti inilah analisis multivariat bisa digunakan (McKinney, 2022).

Analisis multivariat melibatkan lebih dari dua variabel secara bersamaan dengan tujuan memahami hubungan yang lebih kompleks pada data. Pendekatan ini membantu analis memperoleh gambaran yang lebih utuh mengenai fenomena yang sedang diamati. Sebagai contoh, tingkat penjualan suatu produk mungkin dapat dipengaruhi oleh harga, lokasi, promosi, kualitas

layanan, dan karakteristik pelanggan secara bersamaan. Jika setiap faktor dianalisis secara terpisah, sebagian informasi penting mungkin bisa saja terlewatkan. Analisis multivariat membantu melihat keterkaitan antarvariabel dalam satu konteks yang lebih menyeluruh.

Berdasarkan tujuannya analisis multivariat dapat dibedakan menjadi beberapa jenis yang dijelaskan pada tabel berikut.

Tabel 2. Analisis multivariat berdasarkan tujuannya

Tujuan	Deskripsi	Teknik Analisis
Menyederhanakan variabel data	Mereduksi banyak variabel menjadi sedikit komponen baru tanpa menghilangkan informasi penting.	Principal Component Analysis (PCA)
Pengelompokkan	Mengelompokkan data ke dalam beberapa grup berdasarkan kemiripan karakteristik.	K-Means / Hierarchical Clustering
Prediksi nilai numerik	Memprediksi nilai satu variabel terikat berdasarkan beberapa variabel bebas.	Regresi Berganda
Klasifikasi	Memprediksi kategori atau kelas suatu objek berdasarkan banyak karakteristik.	ANN, SVM, Naïve Bayes, dll

C. DASAR VISUALISASI DATA

Biasanya data ditampilkan dalam bentuk angka atau tabel. Kita mungkin akan merasa “pusing” ketika melihat ratusan bahkan ribuan baris data yang hanya disajikan dalam bentuk angka/tabel sehingga membuat informasi dalam data menjadi makin sulit ditemukan. Di sinilah visualisasi data mengambil peran penting.

Visualisasi data merupakan proses menyajikan data atau informasi dalam bentuk visual sehingga lebih mudah dipahami

oleh banyak orang. Bentuk visual tersebut dapat berupa grafik, diagram, peta, maupun representasi visual lainnya. Dengan bantuan visualisasi, informasi yang tersembunyi di balik barisan angka dapat terlihat lebih jelas dan lebih mudah diinterpretasikan (Wilke, 2020). Dalam konteks Data Science visualisasi tidak hanya digunakan untuk menyampaikan hasil analisis kepada orang lain, tetapi juga dapat digunakan sebagai alat bantu eksplorasi.

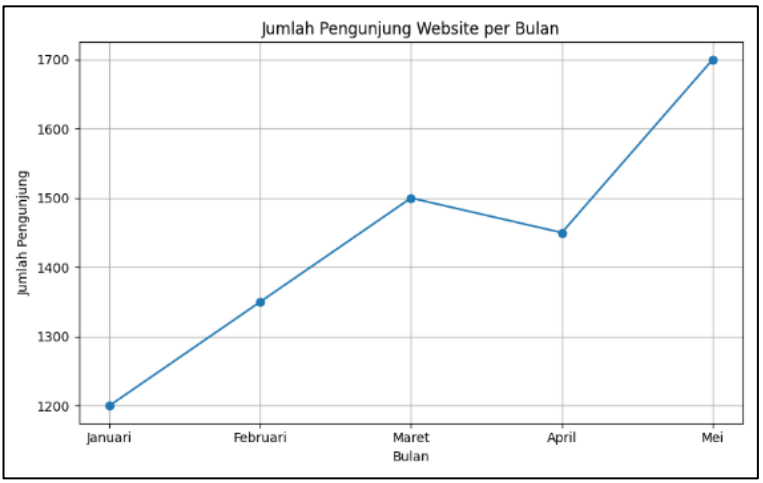
Hal lain yang harus kita perhatikan yaitu, data yang divisualisasikan harus merepresentasikan bentuk atau skala datanya. Kesalahan dalam pemilihan jenis visualisasi berdasarkan skala data akan membuat visualisasi justru jadi sulit untuk dipahami (Josyula et al., 2023).

D. JENIS-JENIS VISUALISASI DATA

Dalam praktik data science, visualisasiyang paling umum digunakan yaitu dalam bentuk grafik atau diagram. Berdasarkan karakter dan skala datanya terdapat beberapa jenis grafik/diagram antara lain yaitu diagram garis, diagram batang, histogram, diagram lingkaran, box plot, scatter plot.

1. Diagram Garis (line chart)

Diagram garis digunakan untuk menyajikan data yang mempunyai interval waktu yang teratur. Pada diagram garis, sumbu x menunjukkan interval waktu (contoh: hari, bulan tahun dll) sedangkan untuk sumbu y menunjukkan data hasil pengamatan yang memiliki skala numerik (interval atau rasio) dan bersifat kontinu. Data ini kemudian diurutkan berdasarkan sumbu x dan dihubungkan oleh sebuah garis (Healy, 2018).

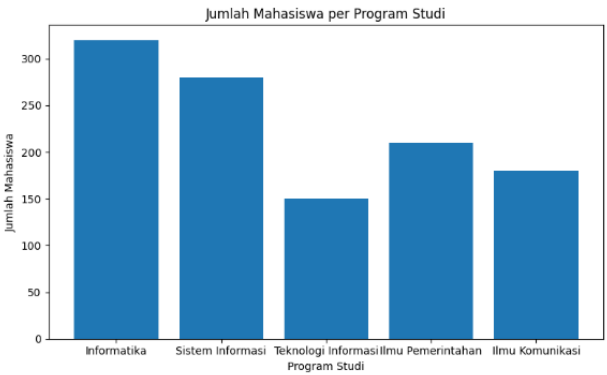


Gambar 1. Contoh visualisasi diagram garis

"Garis" pada diagram ini berfungsi untuk memperlihatkan tren atau perkembangan data dari waktu ke waktu. Diagram garis juga dapat digunakan untuk menampilkan dua tren pada dua data pengamatan sekaligus.

2. Diagram Batang (bar chart)

Diagram batang digunakan untuk menampilkan data dengan sifat kategorik (data yang bukan angka khususnya skala nominal). Sumbu x merepresentasikan setiap kategori data dengan jarak antar batang yang sama rata, sedangkan untuk sumbu y menunjukkan jumlah atau frekuensi dari masing-masing kategori tersebut (Healy, 2018).

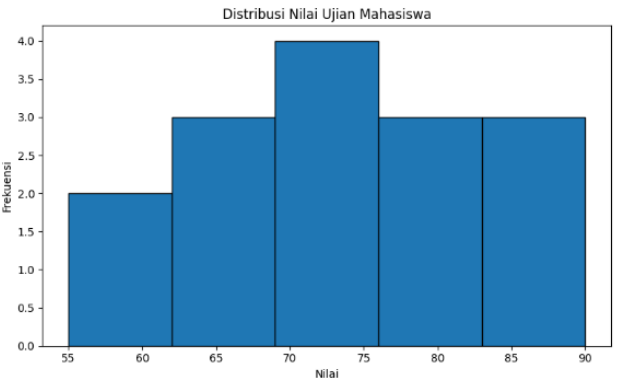


Gambar 2. Contoh visualisasi diagram batang

3. Histogram

Histogram digunakan untuk menampilkan distribusi atau sebaran dari suatu data. Sumbu x merepresentasikan bin disket atau interval dalam suatu pengamatan, sedangkan sumbu y merepresentasikan jumlah atau frekuensi dari pengamatan tersebut yang termasuk dalam setiap bin. Dengan kata lain kita bisa simpulkan bahwa pada histogram, data yang digunakan pada sumbu x dan y sama-sama bersifat numerik.

Secara visual histogram terlihat mirip dengan diagram batang. Hanya saja karena sumbu x pada histogram menggambarkan interval numerik, setiap batang (bar) pada histogram digambarkan tidak mempunyai celah (gap).

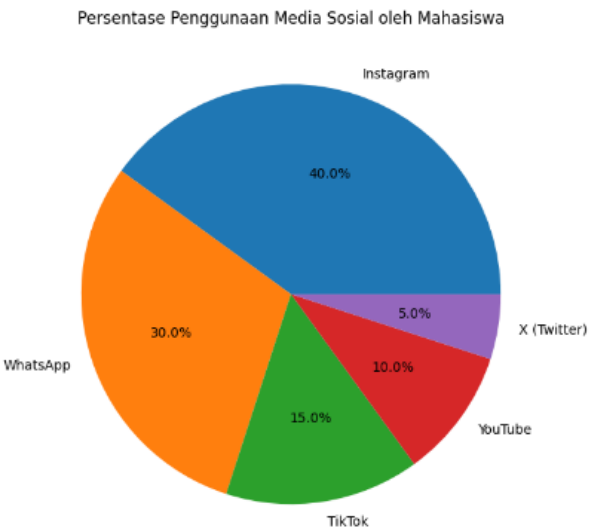


Gambar 3. Contoh visualisasi histogram

4. Diagram Lingkaran (pie chart)

Diagram lingkaran digunakan untuk menampilkan data yang bersifat kategorik (data bukan angka khususnya skala nominal). Kalau kita lihat secara sifat dan skala data, diagram lingkaran mempunyai kesamaan dengan diagram batang. Kita bisa menggunakan diagram lingkaran apabila kategori data yang ada tidak terlalu banyak, jika kategori datanya cukup banyak maka akan membuat slice (segmen) pada diagram menjadi terlalu banyak sehingga membuat data lebih sulit dibandingkan. Sehingga alternatif yang bisa digunakan ketika kategori data cukup banyak yaitu dengan menggunakan diagram batang (Knaflic, 2019).

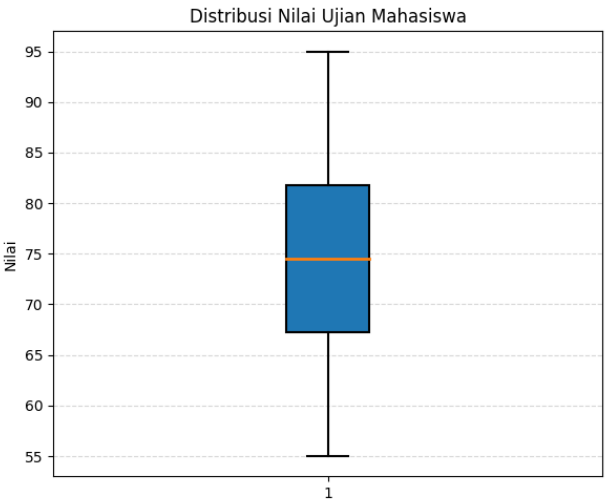
Umumnya diagram lingkaran menampilkan angka dalam bentuk persentase dengan nilai total dari seluruh slice yaitu 100%. Kemudian tiap slice akan diurutkan berdasarkan ukuran atau besaran persentasenya. Tabel dibawah ini merupakan contoh data sederhana yang sesuai untuk divisualisasikan kedalam bentuk diagram lingkaran (konteks: data persentase penggunaan sosial media oleh mahasiswa).



Gambar 4. Contoh visualisasi diagram lingkaran

5. Box Plot

Box plot (box and whisker plot) merupakan visualisasi yang digunakan untuk menggambarkan distribusi data numerik secara ringkas. Melalui box plot, kita dapat melihat informasi penting seperti nilai minimum, kuartil pertama (Q1), median, kuartil ketiga (Q3), nilai maksimum, serta keberadaan outlier (nilai ekstrem).

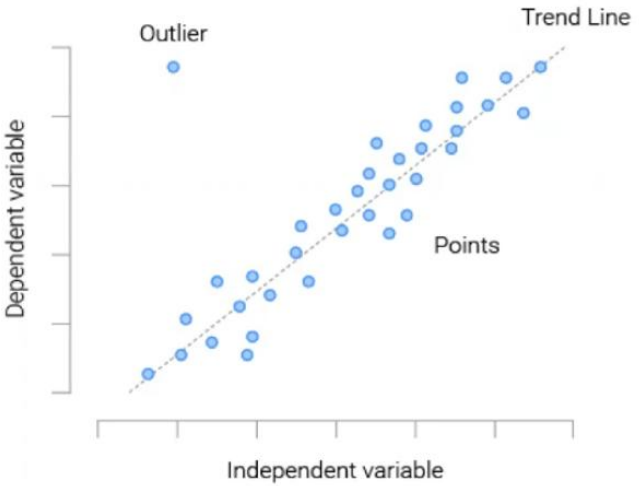


Gambar 5. Contoh visualisasi box plot

6. Scatter Plot

Scatter plot digunakan untuk menampilkan hubungan antar dua variabel data yang bersifat numerik. Scatter plot akan menunjukkan tren maupun distribusi dari data dan juga dapat melihat nilai ekstrem (outlier).

Secara sifat dan karakter data, scatter plot mempunyai kesamaan dengan box plot. Perbedaan mendasar antara dua jenis visualisasi ini adalah jika pada box plot hanya merepresentasikan satu variabel data, maka pada scatter plot merepresentasikan hubungan antar dua variabel data sekaligus.



Gambar 6. Contoh visualisasi scatter plot

BAB 8
BASIS DATA DAN BIG DATA

A. KONSEP DASAR BASIS DATA

Basis data adalah kumpulan data berkaitan yang merepresentasikan fakta-fakta bermakna dari dunia nyata, seperti nama, nomor telepon, dan alamat. Beberapa sifat utamanya: mencerminkan aspek dunia nyata (miniworld) sehingga setiap perubahan nyata tercermin di dalamnya; bersifat koheren dan terstruktur secara logis (bukan sekadar data acak); serta dibangun untuk tujuan dan pengguna tertentu. Ukurannya pun beragam, dari daftar sederhana hingga katalog jutaan entri. Untuk mengelola basis data digunakan DBMS (Database Management System), yaitu perangkat lunak yang menangani empat fungsi utama: pendefinisian (menentukan tipe dan struktur data), pembangunan (menyimpan data secara fisik), manipulasi (pencarian, pembaruan, dan pembuatan laporan), serta berbagi data antar banyak pengguna dan aplikasi.

DBMS menawarkan berbagai keuntungan mendasar bagi pengelolaan data organisasi. Dari sisi efisiensi, DBMS mengendalikan redundansi dengan menyimpan data hanya di satu tempat, menggunakan struktur indeks untuk mempercepat pemrosesan kueri, serta menyediakan fasilitas backup dan pemulihan otomatis saat terjadi kegagalan sistem. Dari sisi keamanan, DBMS dilengkapi subsistem kontrol akses yang memungkinkan DBA mengatur hak tiap pengguna, sekaligus menerapkan batasan integritas mulai dari validasi tipe data hingga keterkaitan antar rekaman. Selain itu, DBMS mendukung

penyimpanan persisten untuk objek program yang kompatibel dengan bahasa seperti C++ dan Java, mampu merepresentasikan hubungan data yang kompleks, serta menyediakan beragam antarmuka pengguna mulai dari bahasa kueri, formulir, hingga bahasa alami sesuai tingkat keahlian pengguna. Lebih jauh, DBMS deduktif bahkan dapat menyimpulkan informasi baru dari data yang ada, sementara DBMS aktif mampu memicu tindakan otomatis berdasarkan kondisi yang telah ditetapkan

B. ARSITEKTUR DAN PEMODELAN BASIS DATA

Arsitektur DBMS juga berkembang dari sistem terpusat berbasis mainframe tunggal menjadi arsitektur yang lebih terdistribusi seiring turunnya harga perangkat keras. Pada arsitektur klien/server dasar, lingkungan jaringan dengan banyak perangkat terhubung ditangani oleh server-server bertugas khusus. Arsitektur ini kemudian berkembang menjadi dua tingkat, di mana antarmuka dan aplikasi berada di sisi klien sementara server menangani kueri dan transaksi dengan SQL sebagai pemisah standar di antara keduanya. Untuk kebutuhan aplikasi web yang lebih kompleks, ditambahkan lapisan menengah berupa server aplikasi atau web yang menyimpan aturan bisnis dan menjadi perantara antara klien dan server basis data, membentuk arsitektur tiga tingkat yang kini umum digunakan.

Pemodelan konseptual adalah fase yang sangat penting dalam merancang aplikasi basis data yang berhasil. Model Entity-Relationship (ER) adalah model data konseptual tingkat tinggi yang populer. Model ini dan variasinya sering digunakan untuk perancangan konseptual aplikasi basis data, dan banyak alat perancangan basis data menggunakan konsep-konsepnya

C. BASIS DATA RELASIONAL (SQL)

Basis data yang berbasis model relasional meliputi MySQL, MS-SQL Server, Oracle Database, dan lain-lain; masing-masing mendukung SQL sebagai bahasa kuerinya. Basis data SQL, karena sifatnya yang berorientasi relasional, memiliki keunggulan dalam hal skalabilitas vertikal dan konsistensi yang kuat. Namun, karena konsistensi diprioritaskan dalam basis data SQL, sistem manajemen basis data harus melakukan banyak pekerjaan untuk mempertahankan keadaan yang konsisten, yang tentunya akan mengorbankan kinerja

1. Structured Query Language (SQL)

SQL (Structured Query Language) adalah bahasa komputer yang dirancang untuk berinteraksi dengan basis data relasional, awalnya dikembangkan oleh IBM dengan nama SEQUEL sebelum dipersingkat menjadi SQL karena alasan merek dagang. SQL bekerja melalui DBMS yang memproses permintaan pengguna terhadap data, mencakup operasi kueri, penyimpanan, hingga pengelolaan data dalam berbagai konteks mulai dari inventaris, penggajian, dan penjualan di lingkungan perusahaan, hingga catatan kontak dan data keuangan pribadi. Proses pengiriman permintaan ke basis data dan penerimaan hasilnya inilah yang disebut sebagai kueri basis data, sesuai dengan namanya Structured Query Language.

Fungsionalitas SQL sangat luas dan mencakup seluruh aspek pengelolaan data. Dari sisi struktur, SQL memungkinkan pengguna mendefinisikan organisasi dan hubungan antar data, sekaligus mendukung pengambilan dan manipulasi data berupa penambahan, penghapusan, maupun pengubahan data yang tersimpan. Di samping itu, SQL menangani kontrol akses untuk melindungi data dari akses tidak sah, mengoordinasikan berbagi

data antar pengguna secara bersamaan agar tidak terjadi konflik pembaruan, serta menjaga integritas data dari kerusakan akibat pembaruan yang tidak konsisten maupun kegagalan system

2. **Data Definition Language (DDL) dan Data Manipulation Language (DML)**

DBMS harus menyediakan bahasa dan antarmuka yang tepat untuk setiap kategori pengguna. Jenis-jenis bahasa yang didukung DBMS meliputi:

a. Bahasa Definisi Data (DDL): digunakan oleh DBA dan perancang basis data untuk mendefinisikan skema konseptual dan internal.
b. Bahasa Manipulasi Data (DML): menyediakan serangkaian operasi atau bahasa untuk pengambilan, penyisipan, penghapusan, dan modifikasi data.
c. Bahasa Definisi Penyimpanan (SDL): menentukan skema internal.
d. Bahasa Definisi Tampilan (VDL): menentukan tampilan pengguna dan pemetaannya ke skema konseptual.

SQL adalah contoh bahasa basis data komprehensif yang merupakan kombinasi dari DDL, VDL, dan DML, serta pernyataan untuk spesifikasi batasan, evolusi skema, dan fitur-fitur lainnya.

D. BASIS DATA NON-RELASIONAL (NOSQL)

Teknologi utama pendukung Big Data mencakup Hadoop untuk pemrosesan data terdistribusi, NoSQL untuk penanganan data semi-terstruktur dan tidak terstruktur, In-Memory Database (IMDB) untuk akses data yang lebih cepat melalui penyimpanan di memori, Massively Parallel Processing (MPP) untuk pemrosesan paralel skala besar, serta Cloud Computing yang menyediakan layanan fleksibel dan hemat biaya. Di antara teknologi tersebut,

NoSQL (Not Only SQL) menjadi salah satu yang paling berkembang pesat dipelopori oleh perusahaan seperti Amazon, Google, dan LinkedIn dengan basis data terkemuka seperti MongoDB, Cassandra, CouchDB, dan HBase. NoSQL hadir dalam empat jenis utama (key-value store, document store, column store, dan graph store), dirancang untuk fleksibel dan cepat dengan mengurangi overhead konsistensi sehingga mampu menyimpan data dalam berbagai format secara terdistribusi, meskipun sebagai konsekuensinya NoSQL mungkin tidak sepenuhnya mendukung transaksi ACID sehingga berpotensi menimbulkan inkonsistensi data.

E. BASIS DATA BERORIENTASI GRAF

Basis data graf, yang telah lama ada namun baru matang berkat Big Data dan Kecerdasan Buatan, merepresentasikan pergeseran paradigma dari pengodean imperatif ke deklaratif di mana struktur esensial dipindahkan dari kode ke data sehingga elemen-elemen data muncul dalam konteks yang menyederhanakan analisis dan pemrosesan. Berbeda dengan basis data relasional yang kaku namun sangat teroptimasi (seperti Fortran), basis data graf bersifat dinamis dan fleksibel (seperti C++ atau Java), sekaligus merupakan sintesis antara dunia berorientasi objek dan relasional yang mampu menggambarkan jaringan objek kompleks dengan hanya menangkap relasi-relasi esensial. Secara teknis, basis data graf menyimpan data dalam bentuk graf G = (V, E), yaitu himpunan simpul (V) yang merepresentasikan entitas dan sisi (E) yang merepresentasikan relasi antar entitas beserta properti-propertinya, (Wang, R., & Yang, Z. 2017).

F. PERBANDINGAN SQL DAN NOSQL

Tabel. Perbandingan SQL dan NOSQL

Aspek	SQL	NoSQL
Model	Relasional. Menyimpan data dalam tabel.	Non-relasional. Menyimpan data dalam dokumen JSON, pasangan key-value, kolom, atau graf.
Data	Setiap rekaman memiliki atribut yang sama. Cocok untuk data terstruktur.	Menawarkan fleksibilitas sehingga setiap rekaman dapat memiliki atribut yang berbeda. Cocok untuk data semi-terstruktur.
Skema	Skema tetap (fixed schema).	Skema dinamis atau fleksibel.
Transaksi	Mendukung transaksi ACID.	Bergantung pada solusi yang digunakan.
Konsistensi & Ketersediaan	Konsistensi kuat yang ditegakkan, diprioritaskan di atas ketersediaan dan kinerja.	Konsistensi, ketersediaan, dan kinerja dapat disesuaikan untuk memenuhi kebutuhan aplikasi (Teorema CAP).
Kinerja	Kinerja insert dan update bergantung pada kecepatan I/O disk.	Kinerja dapat dimaksimalkan dengan mengurangi konsistensi.
Skala	Skalabilitas vertikal (scale-up).	Skalabilitas horizontal (scale-out).

G. KONSEP DAN KARAKTERISTIK BIG DATA

Manusia dan robot terikat untuk bekerja secara koaktif dalam lingkungan kolaboratif pada era Industry 5.0 untuk menghasilkan hasil bernilai tambah menggunakan kecerdasan yang diperoleh dari pengalaman masa lalu, yaitu Big Data (BD). Industry 5.0 tidak hanya bertujuan untuk mengotomatisasi robot cerdas agar bekerja sebagai tim, tetapi juga bertujuan untuk mengotomatisasi manusia dan mesin dalam lingkungan kolaboratif

yang sangat sinergis, memungkinkan peningkatan Kualitas Produk (QoP), Kualitas Pengalaman (QoE), dan Kualitas Hidup (QoL).

Populasi internet dunia tumbuh secara signifikan dari tahun ke tahun; per Mei 2026 nternet telah menjangkau sekitar 73,8% dari total populasi dunia, yang setara dengan lebih dari 6,12 miliar pengguna aktif. unia menciptakan data dalam jumlah besar setiap harinya. IDC (sebuah perusahaan riset teknologi) memperkirakan bahwa data terus tumbuh pada tingkat 24.4% hingga 25.1% setiap tahun, Menurut Statista, total jumlah data yang diprediksi akan diciptakan, ditangkap, disalin, dan dikonsumsi secara global pada tahun 2026 adalah 230 hingga 240 zettabyte, angka yang diproyeksikan akan tumbuh menjadi 258- 284 zettabyte pada tahun 2027.

Kepemilikan digital atas aset virtual fidelitas tinggi pada Web 3.0 meningkat secara eksponensial, yang menjadi indikasi bahwa nilai ekonomi dari dunia virtual digital yang menggunakan teknologi blockchain dan metaverse akan meningkat secara signifikan di tahun-tahun mendatang, menjanjikan pertumbuhan ekonomi yang jauh lebih besar, yang pada gilirannya mengarah pada generasi BD dengan volume yang sangat besar. Berdasarkan laporan DOMO 'Data Never Sleeps 6.0', lebih dari setengah lalu lintas web dunia berasal dari smartphone, dan diprediksi bahwa lebih dari 5,78 hingga 7,52 bahtera/miliar orang memiliki akses ke smartphone pada tahun 2026. Kita dapat menyimpulkan dengan aman bahwa lebih dari 90% BD yang ada di dunia saat ini dihasilkan dalam beberapa tahun terakhir.

H. DEFINISI BIG DATA

Istilah Big Data pertama kali diperkenalkan oleh Gartner Group pada 2008, ketika sistem manajemen data yang umum masih berbasis model relasional. Beberapa definisi awal mendekati

Big Data secara komparatif: McKinsey Institute mendefinisikannya sebagai data yang tidak dapat diproses oleh teknologi manajemen data tradisional, sementara Jacobs menyebutnya sebagai data yang ukurannya memaksa pencarian metode baru di luar metode lazim yang ada. (Isitor, E., & Stanier, C. 2010)

Namun pendekatan komparatif ini memiliki keterbatasan mendasar. Seiring berkembangnya teknologi, sistem relasional yang disempurnakan kini mampu menangani volume data yang jauh lebih besar, dan kemampuan Business Intelligence Big Data semakin terintegrasi ke dalam sistem BI relasional tradisional, sehingga batas antara keduanya menjadi semakin kabur. Pendekatan ini juga bergantung pada konsensus tentang apa yang dianggap "tradisional" atau state-of-the-art tanpa mendefinisikan istilah-istilah tersebut secara jelas, serta cenderung mendefinisikan Big Data hanya dalam kerangka fungsionalitas relasional yang justru membatasi pemahaman tentang Big Data itu sendiri (Jain, V. K. 2018)

I. KARAKTERISTIK 5V: VOLUME, VELOCITY, VARIETY, VERACITY, DAN VALUE

Tabel 1. I. Karakteristik 5V

Karakterisitik	Definisi
Velocity (Kecepatan)	Kecepatan di mana jumlah data yang sangat besar dihasilkan, dikumpulkan, dan dianalisis.
Volume (Jumlah)	Jumlah data yang sangat besar yang dihasilkan setiap menit.
Value (Nilai)	Nilai atau kegunaan dari data tersebut.
Variety (Keragaman)	Berbagai jenis data dalam format yang berbeda-beda.
Veracity (Verasitas):	Kualitas dan tingkat kepercayaan dari data.
Variability (Variabilitas)	Banyaknya dimensi data terkait berbagai jenis dan sumber data.

Validity (Validitas)	Akurasi dan kebenaran data terkait tujuan penggunaannya.
Vulnerability (Kerentanan)	Adanya risiko pelanggaran dan masalah keamanan data.
Volatility (Volatilitas)	Interval waktu yang diperlukan untuk penyimpanan dan pengambilan informasi.
Visualisation (Visualisasi)	Penyajian data dalam waktu respons yang wajar.

J. SUMBER DALAM BIG DATA

1. Data Tidak Terstruktur (Machine-Generated)
   a. Citra Satelit

      Meliputi rekaman data cuaca hingga sistem pemantauan berbasis satelit milik pemerintah (seperti tampilan visual pada Google Earth).
   b. Informasi Ilmiah

      Mencakup hasil pemindaian seismik, pencatatan kondisi atmosfer, serta data penelitian fisika energi tinggi.
   c. Dokumentasi Visual

      Segala bentuk foto dan rekaman video, termasuk yang dihasilkan oleh kamera keamanan (CCTV), sistem pengawasan, dan pemantau lalu lintas.
   d. Data Radar dan Sonar

      Meliputi pemetaan seismik untuk kebutuhan otomotif, meteorologi, hingga eksplorasi oseanografi.

2. Data Tidak Terstruktur Besutan Manusia (Human-Generated)
   a. Komunikasi Internal Perusahaan: Seluruh teks yang tersebar di dalam dokumen kerja, catatan log, hasil survei, hingga korespondensi email. Dokumen korporat seperti ini menyumbang porsi yang sangat besar

terhadap total data teks global saat ini.

b. Aktivitas Media Sosial: Konten yang diproduksi secara masif oleh pengguna platform digital seperti Facebook, X (Twitter), YouTube, LinkedIn, dan Flickr.
c. Data Perangkat Seluler: Informasi yang datang dari aktivitas smartphone, seperti pesan teks (SMS/chat) dan riwayat koordinat lokasi.
d. Konten Situs Web: Seluruh informasi tidak terstruktur yang disajikan di berbagai platform internet, mulai dari Wikipedia hingga media berbagi visual seperti Instagram.

K. TANTANGAN DAN PELUANG PENGELOLAAN BIG DATA

Pengelolaan Big Data menghadapi tantangan berlapis dari sisi data (volume yang terus meningkat, keragaman sumber dan format, kebutuhan analitik real-time, serta persoalan kualitas, keakuratan, privasi, dan keamanan), sisi proses (mulai dari akuisisi, penyaringan, pembersihan, integrasi, dan agregasi data hingga pemodelan, analisis kompleks, dan interpretasi hasil), serta sisi manajemen (pengelolaan siklus hidup data, ketersediaan platform dan teknologi yang sesuai, hingga kebutuhan SDM dengan keterampilan khusus). Selain itu, tantangan kritis lainnya meliputi kurangnya skalabilitas akibat manajemen infrastruktur yang buruk baik on-premise maupun cloud, banyaknya aplikasi/alat analitik yang belum memanfaatkan transformasi, analisis, dan visualisasi data secara optimal, serta isu keamanan dan privasi yang menjadi tantangan paling mendesak—terutama ketika data dikelola pihak ketiga sehingga menimbulkan kekhawatiran terhadap kerahasiaan informasi organisasi. (Saini, P. 2016)

Di balik berbagai tantangan tersebut, Big Data memberikan manfaat besar bagi berbagai sektor: dalam bisnis untuk pelaporan,

peramalan, dan pengambilan keputusan; dalam pemerintahan untuk mendukung transparansi dan peningkatan layanan publik; dalam kesehatan untuk menurunkan biaya layanan sekaligus meningkatkan kualitas perawatan; serta dalam penelitian, di mana ketersediaan data dalam jumlah besar membuka peluang lahirnya metode ilmiah baru berbasis komputasi intensif data. (United Nations Global Pulse. 2012)

L. INFRASTRUKTUR DAN TEKNOLOGI BIG DATA

Teknologi utama yang mendukung Big Data meliputi Hadoop, NoSQL, In-Memory Database (IMDB), Massively Parallel Processing (MPP), dan Cloud Computing. Hadoop memungkinkan pemrosesan data secara terdistribusi, NoSQL mendukung data semi-terstruktur dan tidak terstruktur, IMDB mempercepat akses data melalui penyimpanan di memori, MPP memungkinkan pemrosesan paralel skala besar, sedangkan cloud computing menyediakan layanan yang fleksibel dan hemat biaya. (Venkatraman, R., & Venkatraman, S. 2019)

M. ARSITEKTUR DAN ANALIS BIG DATA

Arsitektur Big Data dan pemrosesan Big Data sebagai arsitektur yang dapat diskalakan yang mendukung persyaratan pemrosesan data dengan volume tinggi yang mungkin memiliki berbagai format data dan mungkin mencakup akuisisi dan pemrosesan data dengan velocity tinggi, sedangkan Analisis big data adalah proses mengekstrak wawasan yang bermakna dari big data, seperti pola tersembunyi, korelasi yang tidak diketahui, tren pasar, dan preferensi pelanggan, yang dapat membantu organisasi membuat keputusan bisnis yang tepat.

N. HADOOP

Hadoop adalah kerangka kerja sumber terbuka yang memungkinkan pemrosesan terdistribusi kumpulan data besar secara paralel menggunakan server industri standar yang ekonomis. Arsitekturnya terdiri dari beberapa komponen utama: HDFS sebagai lapisan penyimpanan, MapReduce untuk menjalankan fungsi analitik, YARN untuk manajemen kluster dan penjadwalan aplikasi, serta Spark di atas HDFS untuk analisis mendalam menggunakan algoritma machine learning. Meski hemat biaya dan andal untuk pemrosesan data skala besar, Hadoop sebagai teknologi sumber terbuka tetap membawa risiko terkait keberlangsungan pemeliharaan dan dukungan jangka panjang.

O. BATCH PROCESSING DAN REAL-TIME PROCESSING

Pemrosesan batch dan real-time memiliki trade-off yang saling berkebalikan pada empat dimensi utama. Dari sisi efisiensi versus kecepatan, batch mengutamakan efisiensi komputasi melalui penjadwalan sumber daya, sementara real-time memprioritaskan respons cepat meskipun mengorbankan efisiensi di mana ukuran batch kecil dalam real-time dapat menurunkan throughput hingga 5x. Dari sisi biaya versus kompleksitas, batch menawarkan infrastruktur yang lebih murah, sedangkan real-time membutuhkan infrastruktur lebih kompleks dengan redundansi dan pemantauan, sehingga tingkat utilisasinya sering hanya 20-30% akibat provisi kapasitas puncak. Dari sisi skalabilitas versus ketepatan waktu, batch unggul untuk dataset masif dengan skalabilitas mendekati linear, sementara real-time membutuhkan perencanaan kapasitas cermat untuk mempertahankan latensi rendah, meski pengaturan ukuran batch dinamis dapat

mengurangi latensi 3-10x. Terakhir, dari sisi kelengkapan data versus ketepatan waktu, batch bekerja dengan dataset lengkap untuk analisis menyeluruh, sedangkan real-time hanya menggunakan tampilan parsial yang berpotensi kehilangan konteks, dengan pendekatan window-based yang dapat menimbulkan tingkat kesalahan 5-15% dibandingkan pemrosesan batch.

P. VISUALISASI BIG DATA

Visualisasi data adalah kunci keberhasilan Big Data. Kecuali keluaran akhir dalam bentuk yang dapat diterima oleh orang-orang yang membutuhkan data untuk dianalisis, seluruh usaha Big Data tidak akan ada nilainya. Laporan yang sangat besar atau grafik rumit yang jarang dipahami orang tidak akan menghasilkan pengambilan keputusan atau tindakan yang bermakna. Terdapat berbagai alat visualisasi yang mencakup dasbor manajemen dan platform visualisasi data komersial yang menghasilkan grafik dan diagram yang menarik untuk komunikasi yang jelas dan ringkas guna mendapatkan wawasan data.

Q. INTEGRASI BASIS DATA DAN BIG DATA DALAM DATA SCIENCE

Big Data dan Data Science merupakan dua bidang yang saling terintegrasi secara menyeluruh dan saling melengkapi secara fungsional dalam membentuk ekosistem analitik modern: Big Data yaitu kumpulan data bervolume masif baik terstruktur maupun tidak terstruktur dari berbagai sumber seperti sensor, perangkat IoT, jaringan sosial, log transaksi, hingga video dan audio dikumpulkan dan disimpan dalam data lake dalam berbagai format, kemudian diproses secara paralel menggunakan ekosistem

Hadoop, Spark, dan Storm sebelum dipindahkan ke data warehouse; dari sinilah Data Science, sebagai bidang interdisipliner konvergensi matematika, statistika, algoritma terapan, rekayasa teknologi, dan keahlian domain, mengambil peran dengan menerapkan machine learning, deep neural networks, dan analisis statistika untuk mengungkap pola dan korelasi tersembunyi, mengekstrak pengetahuan, memprediksi hasil, serta menemukan solusi optimal atas masalah bisnis kompleks; dan pada akhirnya, Business Intelligence melengkapi rangkaian ini dengan menerjemahkan seluruh hasil analisis menjadi laporan dan dashboard yang siap digunakan manajer dan eksekutif sehingga ketiganya bersama-sama mengubah data mentah menjadi actionable insights yang bernilai strategis bagi pengambilan keputusan bisnis yang lebih baik dan lebih cepat.

BAB 9
MACHINE LEARNING DASAR

A. PENGERTIAN MACHINE LEARNING

Machine learning adalah cabang kecerdasan buatan yang membuat komputer belajar dari data. Machine learning digunakan untuk menangani dan memprediksi data yang sangat besar dengan cara mempresentasikan data-data tersebut dengan algoritma (Wilson & Anwar, 2024). Komputer tidak hanya menjalankan instruksi tetap tapi membentuk pola dari contoh yang diberikan. Model kemudian menggunakan pola tersebut untuk menyelesaikan data baru.

Model machine learning menggunakan data untuk membuat prediksi, klasifikasi, atau keputusan berdasarkan pola yang dipelajari (Liu et al., 2026). Model machine learning membutuhkan data latih agar algoritma dapat membangun hubungan antara fitur dan target

Machine learning memiliki tiga komponen utama. Komponen pertama adalah data. Data menyediakan contoh bagi model. Komponen kedua adalah algoritma. Algoritma mengatur cara model belajar. Komponen ketiga adalah evaluasi. Evaluasi mengukur kualitas hasil model. Sebagai contoh, perguruan tinggi memiliki data mahasiswa. Data tersebut berisi IPK, jumlah SKS, presensi, status pembayaran, dan aktivitas organisasi. Model machine learning dapat mempelajari hubungan antara variabel tersebut dan status kelulusan. Program studi dapat memakai hasil model untuk mendeteksi mahasiswa yang berisiko terlambat lulus.

Contoh lain terdapat pada bidang bisnis. Perusahaan memiliki data pelanggan. Data tersebut berisi riwayat transaksi, frekuensi pembelian, dan nilai pembelian. Model machine learning dapat memprediksi pelanggan yang berpotensi berhenti memakai layanan. Perusahaan dapat memakai hasil tersebut untuk menyusun strategi retensi pelanggan.

B. PERBEDAAN PEMROGRAMAN TRADISIONAL DAN MACHINE LEARNING

Pemrograman tradisional menggunakan aturan yang ditulis langsung oleh programmer. Programmer menentukan logika. Komputer menjalankan logika. Program menghasilkan keluaran berdasarkan aturan tersebut.

Machine learning menggunakan pendekatan yang berbeda. Programmer menyediakan data dan algoritma. Algoritma mempelajari pola dari data. Model menghasilkan aturan internal melalui proses pelatihan. Tabel 9.1 menunjukkan perbedaan pemrograman tradisional dan machine learning

Tabel 9.1 Perbedaan Pemrograman Tradisional dan Machine Learning

Aspek	**Pemrograman Tradisional**	**Machine Learning**
Masukan utama	Data dan aturan	Data dan label/pola
Peran manusia	Manusia menulis aturan	Manusia menyiapkan data dan memilih algoritma
Proses	Program menjalankan aturan	Model mempelajari pola
Keluaran	Hasil berdasarkan aturan tetap	Prediksi atau keputusan berdasarkan pola
Contoh	Rumus gaji otomatis	Prediksi risiko mahasiswa terlambat lulus

Pemrograman tradisional cocok untuk masalah yang memiliki aturan jelas. Machine learning cocok untuk masalah yang memiliki pola kompleks. Masalah seperti deteksi spam, analisis sentimen, prediksi risiko, dan rekomendasi produk lebih cocok diselesaikan dengan machine learning.

C. JENIS-JENIS MACHINE LEARNING

Machine learning memiliki beberapa jenis pembelajaran. Setiap jenis memiliki karakteristik data, tujuan, dan metode yang berbeda.

1. Supervised Learning

Supervised learning adalah pembelajaran mesin yang menggunakan data berlabel. Data berlabel memiliki pasangan antara masukan dan keluaran. Model mempelajari hubungan antara masukan dan keluaran. Model kemudian memprediksi keluaran untuk data baru. Supervised learning mencakup tugas klasifikasi dan regresi karena model mempelajari target dari data berlabel (Liu et al., 2026). Klasifikasi memprediksi kelas atau kategori. Regresi memprediksi nilai numerik. Contoh klasifikasi adalah prediksi email spam atau bukan spam. Contoh lain adalah prediksi mahasiswa lulus tepat waktu atau tidak tepat waktu. Contoh regresi adalah prediksi harga rumah. Contoh lain adalah prediksi jumlah penjualan, nilai ujian, atau lama studi mahasiswa.

2. Unsupervised Learning

Unsupervised learning adalah pembelajaran mesin yang menggunakan data tanpa label. Model mencari pola tersembunyi dari data. Model tidak menerima jawaban benar pada saat pelatihan. Unsupervised learning sering digunakan untuk clustering dan reduksi dimensi karena model dapat mencari struktur data tanpa target berlabel (Liu et al., 2026). Clustering

mengelompokkan data berdasarkan kemiripan. Reduksi dimensi menyederhanakan data yang memiliki banyak variabel. Contoh clustering adalah pengelompokan pelanggan berdasarkan pola pembelian. Contoh lain adalah pengelompokan mahasiswa berdasarkan pola akademik dan aktivitas belajar.

3. Semi-Supervised Learning

Semi-supervised learning adalah pembelajaran mesin yang menggunakan sedikit data berlabel dan banyak data tidak berlabel. Pendekatan ini berguna ketika pelabelan data membutuhkan biaya, waktu, atau tenaga ahli.

Model awal dilatih dengan data berlabel. Model kemudian memberi label sementara pada data tidak berlabel. Label sementara tersebut disebut pseudo-label. Data dengan tingkat keyakinan tinggi dapat ditambahkan ke data latih.

Pendekatan semi-supervised learning membantu peneliti memanfaatkan data tidak berlabel ketika jumlah data berlabel masih terbatas (Zhou et al., 2024). Namun, peneliti harus memeriksa kualitas pseudo-label agar model tidak belajar dari label yang salah.

4. Reinforcement Learning

Reinforcement learning adalah pembelajaran mesin yang menggunakan konsep agen, lingkungan, aksi, dan reward. Agen melakukan aksi pada lingkungan. Lingkungan memberikan reward. Agen belajar memilih aksi yang menghasilkan reward terbaik.

Reinforcement learning digunakan pada robotika, permainan, optimasi rute, dan sistem kendali. Model belajar melalui pengalaman. Model tidak selalu menerima jawaban benar sejak awal. Pada Tabel 9.2 tersaji rangkuman jenis-jenis machine learning yang dapat difahami perbedaan dari masing-masing

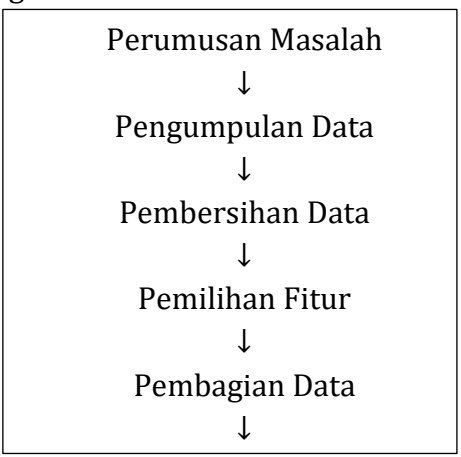
jenis machine learning.

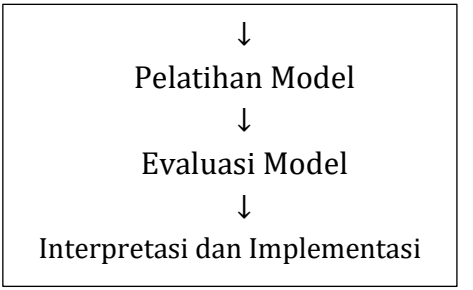
Tabel 9.2 Jenis-Jenis Machine Learning

Jenis Pembelajaran	Data Yang Digunakan	Tujuan	Contoh Metode
Supervised Learning	Berlabel	Prediksi kelas atau angka	Logistic regression, decision tree, SVM
Unsupervised Learning	Tanpa label	Menemukan pola atau kelompok	Logistic regression, decision tree, SVM
Semi Supervised Learning	Sedikit data berlabel dan banyak data tidak berlabel	Menemukan pola atau kelompok	Pseudo-labeling, co-training, self - training
Reinforcement Learning	Data interaksi agen dan lingkungan	Memilih aksi terbaik	Q-learning, policy gradient

D. ALUR KERJA MACHINE LEARNING

Machine learning membutuhkan alur kerja yang sistematis. Alur kerja yang baik membantu peneliti menghasilkan model yang valid dan dapat digunakan, alur kerja machine learning dapat diilustrasikan di gambar 9.1





Gambar 9.1 Alur Kerja Machine Learning

Pada gambar 9.1 menunjukkan peneliti merumuskan masalah pada tahap awal. Masalah harus jelas, terukur, dan sesuai dengan kebutuhan pengguna. Pertanyaan penelitian harus dapat dijawab dengan data.

Contoh pertanyaan yang umum adalah "bagaimana meningkatkan layanan mahasiswa?". Pertanyaan tersebut masih terlalu luas. Pertanyaan yang lebih tepat adalah "apakah model machine learning dapat memprediksi mahasiswa yang berisiko terlambat lulus berdasarkan data akademik dan aktivitas mahasiswa?".

Tahap selanjutnya peneliti mengumpulkan data dari sumber yang relevan. Data dapat berasal dari basis data, survei, sensor, aplikasi, dokumen, atau sistem transaksi. Data harus sesuai dengan tujuan analisis. Data yang baik harus lengkap, akurat, konsisten, dan relevan. Kualitas data menjadi faktor penting karena data yang buruk dapat menurunkan kinerja, keadilan, ketahanan, keamanan, dan skalabilitas model machine learning (Zhou et al., 2024).

Tahap ketiga, peneliti membersihkan data sebelum proses pelatihan. Data mentah sering memiliki nilai kosong, duplikasi, format tidak seragam, dan nilai ekstrem. Pembersihan data membantu model mempelajari pola yang lebih tepat. Data-centric AI (Artificial Intellegence) menekankan perbaikan data sebagai strategi penting dalam peningkatan kualitas sistem AI dan machine learning (Zha et al., 2023). Oleh karena itu, peneliti perlu

memeriksa nilai kosong, data duplikat, label salah, dan data tidak konsisten sebelum model dilatih.

Tahap keempat, peneliti memilih fitur yang relevan. Fitur adalah variabel masukan yang digunakan oleh model. Fitur yang baik membantu model mengenali pola. Fitur yang tidak relevan dapat menambah gangguan. Contoh fitur untuk prediksi kelulusan mahasiswa adalah IPK, jumlah SKS, presensi, dan aktivitas akademik. Fitur yang tidak berhubungan sebaiknya tidak digunakan. Pemilihan fitur harus mengikuti pemahaman domain.

Tahap kelima, peneliti membagi data menjadi data latih dan data uji. Data latih digunakan untuk melatih model. Data uji digunakan untuk mengukur kemampuan model pada data baru. Rasio pembagian data yang sering digunakan adalah 80:20 atau 70:30. Rasio 80:20 berarti 80 persen data digunakan untuk pelatihan dan 20 persen data digunakan untuk pengujian. Peneliti perlu memisahkan data latih dan data uji dengan benar karena kesalahan pembagian data dapat menyebabkan data leakage (Liu et al., 2026) . Data leakage adalah kebocoran informasi dari data yang seharusnya tidak boleh diketahui model saat proses pelatihan dan membuat hasil evaluasi terlihat baik, tetapi model dapat gagal saat digunakan pada data nyata.

Tahap keenam, peneliti melatih model dengan algoritma tertentu. Algoritma mempelajari pola dari data latih. Model menyimpan pola dalam bentuk parameter atau struktur keputusan. Setiap algoritma memiliki karakteristik yang berbeda. Linear regression cocok untuk prediksi angka. Logistic regression cocok untuk klasifikasi sederhana. Decision tree mudah dijelaskan. Random forest lebih stabil. SVM cocok untuk pemisahan kelas. K-means cocok untuk pengelompokan.

Tahap terakhir, peneliti mengevaluasi model dengan metrik yang sesuai. Masalah klasifikasi membutuhkan accuracy, precision, recall, F1-score, dan confusion matrix. Masalah regresi

membutuhkan MAE, MSE, RMSE, dan R-squared. Pemilihan metrik evaluasi harus mengikuti tujuan tugas, biaya kesalahan, dan kondisi data karena satu metrik tunggal dapat memberi kesimpulan yang menyesatkan. Evaluasi model supervised learning harus mempertimbangkan karakteristik dataset, desain validasi, ketidakseimbangan kelas, biaya kesalahan, dan tujuan operasional model (Liu et al., 2026).

Setelah selesai melakukan evaluasi model, peneliti menafsirkan hasil model setelah evaluasi selesai. Peneliti menjelaskan arti hasil kepada pengguna. Pengguna membutuhkan penjelasan agar hasil model dapat dipakai dalam keputusan nyata. Organisasi perlu memantau model setelah implementasi. Data dapat berubah dari waktu ke waktu. Perubahan data dapat menurunkan kinerja model. Oleh karena itu, organisasi perlu melakukan pemantauan, pembaruan, dan evaluasi berkala.

E. METODE DASAR MACHINE LEARNING

Bagian ini membahas beberapa metode dasar yang sering digunakan dalam data science yaitu:

1. Linear Regression

Linear regression digunakan untuk memprediksi nilai numerik. Model mencari hubungan linier antara variabel masukan dan variabel keluaran. Contoh penggunaan linear regression adalah prediksi harga rumah berdasarkan luas tanah. Contoh lain adalah prediksi nilai ujian berdasarkan jumlah jam belajar. Apapun rumus Linear regression tersaji di persamaan 9.1.

$$\boldsymbol{Y = \alpha + bX}$$

Keterangan:

Y = nilai prediksi

α = konstanta
b = koefisien
X = variabel masukan

2. Logistic Regression

Logistic regression digunakan untuk klasifikasi. Model menghasilkan probabilitas kelas. Model kemudian mengubah probabilitas menjadi label kelas. Model logistic regression dapat memprediksi probabilitas suatu kejadian melalui fungsi logistik. Jika nilai p lebih besar dari 0,5, model dapat memberi label kelas 1. Jika nilai p lebih kecil atau sama dengan 0,5, model dapat memberi label kelas 0. Rumus logistic regression dapat menggunakan fungsi log-odds dalam regresi logistik adalah cara mengubah probabilitas menjadi angka yang tak terbatas (−∞ hingga +∞) dengan mengambil logaritma dari rasio peluang (odds). Rumus fungsi log-odds tersaji di persamaan 9.2.

$$\boldsymbol{z} = \textit{log-odds} = \textit{logit}\ (\boldsymbol{x}) = \boldsymbol{\beta_0} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{P}} \boldsymbol{\beta_j}.\boldsymbol{x_j}$$

Keterangan:

$\boldsymbol{\beta_0}$ (intercept) = nilai log-odds ketika semua fitur = 0 (setelah standardisasi).

3. Decision Tree

Decision tree adalah algoritma yang membentuk struktur seperti pohon keputusan. Setiap simpul berisi kondisi. Setiap cabang berisi hasil dari kondisi. Setiap daun berisi keputusan. Decision tree mudah dipahami oleh pembaca nonteknis. Model ini dapat menunjukkan alasan keputusan. Namun, decision tree dapat mengalami overfitting jika pohon terlalu dalam.

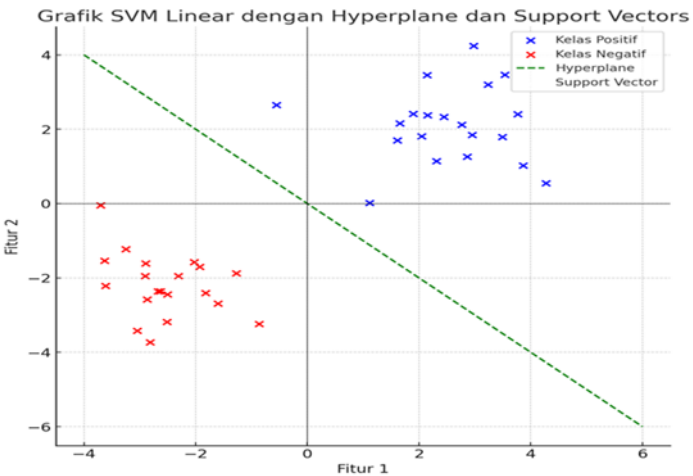
4. Random Forest

Random forest adalah metode ensemble yang memakai

banyak decision tree. Setiap pohon menghasilkan prediksi. Model menggabungkan prediksi dari banyak pohon. Hasil gabungan menjadi prediksi akhir. Random forest biasanya lebih stabil daripada decision tree tunggal. Model ini dapat menangani data yang kompleks. Namun, model ini lebih sulit dijelaskan daripada decision tree tunggal.

5. Support Vector Machine

Support Vector Machine atau SVM adalah algoritma yang mencari garis atau bidang pemisah terbaik antar kelas. Garis atau bidang tersebut disebut hyperplane. SVM memilih hyperplane dengan margin terbesar. Pada gambar 9. 1 diperlihatkan sebagai garis lurus yang berada di tengah-tengah antara dua kelas data. Garis ini memisahkan dua kelas secara optimal dan jarak antara hyperplane dengan objek-objek data berbeda dengan kelas yang berdekatan (terluar) yang diberi tanda bulat kosong dan positif. Dalam SVM objek data terluar yang paling dekat dengan hyperplane disebut support vector.



Gambar 9.2 Hyperplane yang memisahkan dua kelas positif (+1), negatif

6. K-Nearest Neighbors

K-Nearest Neighbors atau KNN adalah algoritma yang memprediksi kelas berdasarkan tetangga terdekat. Model menghitung jarak antara data baru dan data latih. Model memilih K data terdekat. Kelas mayoritas menjadi hasil prediksi. KNN mudah dipahami. Namun, KNN dapat menjadi lambat jika jumlah data sangat besar. KNN juga sensitif terhadap skala fitur. Oleh karena itu, peneliti sering melakukan normalisasi data sebelum menggunakan KNN.

7. Naive Bayes

Naive Bayes adalah algoritma klasifikasi berbasis Teorema Bayes. Model menghitung probabilitas kelas berdasarkan fitur. Model ini sering digunakan pada klasifikasi teks. Contoh penggunaan Naive Bayes adalah deteksi spam. Model menghitung peluang sebuah email masuk kelas spam berdasarkan kata-kata yang muncul. Model kemudian menentukan kelas email tersebut.

8. K-Means Clustering

K-means adalah algoritma clustering yang membagi data ke dalam K kelompok. Model menentukan pusat kelompok yang disebut centroid. Model menempatkan setiap data ke centroid terdekat. K-means cocok untuk segmentasi pelanggan, pengelompokan mahasiswa, dan analisis pola transaksi. Rumus jarak Euclidean yang sering digunakan tersaji di persamaan 9.3.

$$\boldsymbol{d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}$$

Keterangan:

$\boldsymbol{d(A,B)}$ = jarak antara dua titik A $\boldsymbol{A(x_1, y_1)}$ dan $\boldsymbol{B(x_2, y_2)}$

BAB 10
DATA MINING DAN ANALITIK DATA

Perkembangan teknologi informasi yang pesat telah menghasilkan ledakan data dalam berbagai bidang kehidupan. Setiap hari, miliaran transaksi digital, interaksi media sosial, rekaman sensor, dan aktivitas pengguna menghasilkan data dalam jumlah yang sangat besar. Data tersebut, jika tidak diolah dengan tepat, hanya akan menjadi beban penyimpanan tanpa memberikan nilai tambah yang berarti. Di sinilah peran data mining dan analitik data menjadi sangat krusial dalam ekosistem data science.

Data mining merupakan bagian inti dari proses Knowledge Discovery in Databases (KDD), yaitu proses non-trivial untuk mengidentifikasi pola yang valid, baru, berpotensi berguna, dan dapat dipahami dari kumpulan data (Fayyad et al., 1996). Berbeda dengan tahap pengolahan dan persiapan data yang telah dibahas pada bab sebelumnya, bab ini berfokus pada bagaimana teknik-teknik tersebut diintegrasikan secara sistematis dalam sebuah pipeline data mining yang utuh mulai dari pemilihan metodologi, penerapan algoritma yang tepat, hingga interpretasi hasil untuk menghasilkan pengetahuan yang dapat ditindaklanjuti.

Bab ini membahas konsep dasar data mining, metodologi standar industri, teknik dan algoritma utama, serta implementasi melalui studi kasus yang konkret. Dengan memahami bab ini, pembaca diharapkan mampu menentukan pendekatan data mining yang sesuai dan mengintegrasikannya ke dalam alur kerja data science secara menyeluruh.

A. KONSEP DASAR DATA MINING

1. Definisi dan Sejarah Singkat

Data mining, atau yang juga dikenal sebagai Knowledge Discovery in Databases (KDD), secara sederhana dapat didefinisikan sebagai proses non-trivial untuk mengidentifikasi pola-pola valid, baru, berpotensi berguna, dan dapat dipahami dari data besar (Fayyad et al., 1996). Dalam literatur modern, kedua istilah ini sering digunakan secara bergantian meskipun secara teknis KDD merujuk pada keseluruhan proses sedangkan data mining adalah tahap intinya (Han et al., 2022). Istilah "non-trivial" mengandung makna bahwa proses tersebut bukan sekadar pencarian langsung atau komputasi sederhana, melainkan memerlukan inferensi dan penemuan pola yang tidak terlihat secara eksplisit.

Sejarah data mining dapat ditelusuri sejak era 1960-an ketika para ilmuwan statistika mulai mengembangkan metode analisis data eksploratif. Namun, istilah "data mining" baru populer pada tahun 1990-an seiring dengan berkembangnya basis data berukuran besar dan meningkatnya kebutuhan industri untuk mengolah data tersebut secara otomatis. Konferensi internasional pertama yang secara khusus membahas KDD diadakan pada tahun 1989, dan sejak saat itu bidang ini berkembang pesat menjadi disiplin ilmu tersendiri yang menjadi tulang punggung data science modern.

2. Posisi Data Mining dalam Ekosistem Data Science

Memahami posisi data mining di antara bidang-bidang terkait sangat penting untuk menghindari kerancuan konsep. Statistika berfokus pada inferensi populasi dari sampel menggunakan model probabilistik dan pengujian hipotesis. Machine learning yang fondasinya telah dibahas pada

sebelumnya berfokus pada pengembangan algoritma yang dapat belajar dari data untuk membuat prediksi atau keputusan. Sementara data mining memanfaatkan algoritma dari kedua bidang tersebut sebagai alat, dengan tujuan spesifik menemukan pola tersembunyi dari kumpulan data besar dalam konteks bisnis atau penelitian yang nyata.

3. Jenis Tugas Data Mining

Secara umum, tugas dalam data mining dibagi menjadi dua kategori besar. Tugas deskriptif (descriptive tasks) bertujuan mengkarakterisasi sifat-sifat umum data, mencakup analisis asosiasi, klasterisasi, dan karakterisasi kelas. Tugas prediktif (predictive tasks) melakukan inferensi pada data untuk menghasilkan prediksi, mencakup klasifikasi dan regresi (Han et al., 2022). Pemahaman tentang jenis tugas ini menjadi dasar dalam pemilihan algoritma dan metodologi yang tepat.

4. Proses Knowledge Discovery in Databases (KDD)

Proses KDD menggambarkan alur lengkap dari data mentah hingga pengetahuan yang berguna. Fayyad et al. (1996) mendefinisikan KDD sebagai proses yang terdiri dari lima tahap:

a. Seleksi, memilih data yang relevan dari basis data;
b. Pra-pemrosesan, menghapus noise dan data tidak konsisten serta mengintegrasikan berbagai sumber data;
c. Transformasi, mengubah data ke bentuk yang sesuai untuk proses mining;
d. Data mining, menerapkan algoritma untuk mengekstrak pola; serta
e. Interpretasi dan evaluasi, menginterpretasikan pola yang ditemukan menjadi pengetahuan yang dapat dipahami. Detail teknis tahap pra-pemrosesan telah dibahas pada sebelumnya.

B. METODOLOGI CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu metodologi yang paling banyak digunakan dalam proyek data mining dan data science. Dikembangkan pada pertengahan hingga akhir 1990-an sebagai proyek yang sebagian didanai oleh European Commission melalui program ESPRIT, CRISP-DM dibangun oleh konsorsium yang terdiri dari DaimlerChrysler, SPSS, NCR, dan OHRA. Metodologi ini memberikan kerangka kerja yang fleksibel, tidak bergantung pada alat maupun domain industri tertentu, sehingga dapat diterapkan secara luas di berbagai bidang (Wirth & Hipp, 2000).

CRISP-DM terdiri dari enam fase yang saling berkaitan dan bersifat iteratif, yaitu:

a. Pemahaman Bisnis (Business Understanding) yaitu Berfokus pada pemahaman tujuan dan kebutuhan proyek dari perspektif bisnis, kemudian mengubahnya menjadi definisi masalah data mining yang jelas dan terukur.
b. Pemahaman Data (Data Understanding) yaitu Mencakup pengumpulan data awal, eksplorasi data secara statistika deskriptif dan visual sejalan dengan teknik eksplorasi yang dibahas pada bab sebelumnya serta identifikasi masalah kualitas data.
c. Persiapan Data (Data Preparation) yaitu Mencakup seleksi, pembersihan, konstruksi, dan integrasi data hingga menjadi dataset final yang siap dimodelkan. Detail teknis tahap ini telah dibahas pada bab sebelumnya.
d. Pemodelan (Modeling) yaitu Berbagai teknik pemodelan dipilih dan diterapkan sesuai karakteristik masalah. Parameter model dikalibrasi untuk mendapatkan

performa yang optimal. Pada fase ini sering kali diperlukan iterasi kembali ke fase persiapan data.

e. Evaluasi (Evaluation) yaitu Model dievaluasi secara menyeluruh untuk memastikan bahwa model tersebut memenuhi tujuan bisnis yang telah ditetapkan, sebelum masuk ke tahap deployment.
f. Deployment yaitu Pengetahuan yang diperoleh diimplementasikan ke dalam sistem nyata, mulai dari laporan sederhana hingga pipeline analitik otomatis yang berjalan secara berkala.

Salah satu keunggulan CRISP-DM bersifat siklus yaitu hasil dari fase deployment menjadi masukan untuk siklus proyek berikutnya, mendorong perbaikan berkelanjutan pada model dan pengetahuan yang dihasilkan. Inilah yang membedakan CRISP-DM dari proses analisis data biasa yang bersifat linear.

C. TEKNIK DAN ALGORITMA DATA MINING

Pada bagian ini, fokus diarahkan pada konteks penerapan algoritma tersebut dalam kerangka data mining terkait dengan ketepatan teknik yang digunakan, bagaimana memilih di antara beberapa algoritma, dan hal-hal praktis yang perlu diperhatikan dalam implementasinya.

1. Klasifikasi

Klasifikasi adalah tugas data mining prediktif yang bertujuan memprediksi kelas atau kategori dari suatu instance data berdasarkan atribut-atributnya. Model dibangun menggunakan data berlabel, kemudian digunakan untuk memprediksi kelas data baru (Tan et al., 2019).

Dalam konteks data mining, pemilihan algoritma klasifikasi tidak hanya didasarkan pada akurasi semata, tetapi juga

mempertimbangkan interpretabilitas model, ukuran dataset, serta kecepatan komputasi. Decision Tree dipilih ketika interpretabilitas tinggi diperlukan karena menghasilkan aturan keputusan yang dapat dibaca langsung. Naive Bayes dipilih untuk dataset berukuran kecil dengan waktu komputasi yang sangat singkat. Support Vector Machine (SVM) cocok untuk data berdimensi tinggi dan ketika batas keputusan yang kompleks diperlukan. Untuk detail cara kerja setiap algoritma, pembaca dapat merujuk pada bab sebelumnya.

2. Klasterisasi

Klasterisasi adalah teknik deskriptif yang mengelompokkan objek-objek data ke dalam cluster berdasarkan kemiripan fitur tanpa menggunakan label kelas yang telah ditentukan sebelumnya (unsupervised learning). Teknik ini berguna ketika tujuannya adalah memahami struktur alami dari data, bukan memprediksi kelas tertentu.

K-Means efisien untuk dataset besar dengan jumlah cluster yang sudah diketahui, namun sensitif terhadap outlier dan mengasumsikan cluster berbentuk sferis. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) lebih unggul ketika bentuk cluster tidak beraturan dan keberadaan noise perlu diidentifikasi secara otomatis (Ester et al., 1996). Pemilihan antara keduanya bergantung pada karakteristik data dan tujuan analisis.

3. Analisis Asosiasi

Analisis aturan asosiasi (association rule mining) bertujuan menemukan hubungan yang menarik antara variabel-variabel dalam dataset besar. Aplikasi klasiknya adalah market basket analysis yaitu menemukan pola seperti "pelanggan yang membeli produk A cenderung juga membeli produk B."

Dua metrik utama dalam aturan asosiasi adalah support yaitu seberapa sering itemset muncul dalam dataset dan confidence yaitu seberapa sering aturan terbukti benar. Algoritma Apriori memanfaatkan prinsip anti-monoton untuk memangkas ruang pencarian, sedangkan FP-Growth menggunakan struktur data FP-tree yang lebih efisien sehingga menghindari pembangkitan kandidat yang tidak perlu menjadikannya lebih scalable untuk dataset besar (Han et al., 2022).

4. Deteksi Anomali

Deteksi anomali bertujuan mengidentifikasi pola data yang menyimpang secara signifikan dari perilaku normal (Chandola et al., 2009). Teknik ini memiliki aplikasi penting dalam deteksi penipuan (fraud detection), pemantauan jaringan, dan diagnosa medis (Chandola et al., 2009). Pendekatan yang umum mencakup metode berbasis statistika (z-score, IQR), metode berbasis jarak seperti Local Outlier Factor (LOF), serta metode berbasis machine learning seperti Isolation Forest (Liu et al., 2008). Deteksi anomali tergolong unik karena dapat bersifat supervised apabila label anomali tersedia maupun unsupervised apabila tidak ada label (Chandola et al., 2009).

D. PRA-PEMROSESAN DATA DALAM KONTEKS DATA MINING

Teknik-teknik pra-pemrosesan data secara lengkap telah dibahas pada bab lain sebelumnya. Pada bagian ini, pembahasan difokuskan pada pertimbangan spesifik yang relevan dalam pipeline data mining, khususnya aspek-aspek yang langsung memengaruhi kualitas model yang dihasilkan.

Prinsip "garbage in, garbage out" berlaku sepenuhnya dalam data mining yaitu data berkualitas rendah akan menghasilkan

model yang tidak dapat diandalkan sekalipun algoritma yang digunakan sangat canggih. Oleh karena itu, keputusan pra-pemrosesan harus selalu dikaitkan dengan tujuan mining dan karakteristik algoritma yang akan digunakan. Misalnya, algoritma berbasis jarak seperti K-Means dan KNN sangat sensitif terhadap perbedaan skala fitur, sehingga normalisasi atau standardisasi menjadi wajib. Sebaliknya, Decision Tree tidak memerlukan normalisasi karena bekerja berdasarkan threshold nilai.

Penanganan dataset tidak seimbang (imbalanced dataset) merupakan pertimbangan kritis yang sering diabaikan. Ketika satu kelas memiliki jumlah sampel yang jauh lebih banyak, model klasifikasi cenderung bias ke kelas mayoritas dan menghasilkan akurasi yang menyesatkan. Teknik Synthetic Minority Over-sampling Technique (SMOTE) mengatasi masalah ini dengan membangkitkan sampel sintetis untuk kelas minoritas berdasarkan interpolasi antar tetangga terdekat (Chawla et al., 2002). Pilihan lain mencakup undersampling pada kelas mayoritas atau penggunaan algoritma yang mendukung class weight.

Reduksi dimensi, seperti Principal Component Analysis (PCA), penting untuk dataset dengan jumlah fitur yang sangat banyak guna mengurangi kompleksitas komputasi dan menghindari curse of dimensionality kondisi di mana performa algoritma justru menurun seiring bertambahnya jumlah fitur.

E. ANALITIK DATA

Analitik data (data analytics) merupakan proses berbasis data yang bertujuan menghasilkan pengetahuan dan mendukung pengambilan keputusan bisnis melalui teknik-teknik statistika, machine learning, dan visualisasi (Provost & Fawcett, 2013). Jika data mining berfokus pada penemuan pola tersembunyi dalam data, analitik data mencakup spektrum yang lebih luas yaitu mulai

dari pelaporan data historis hingga pemberian rekomendasi tindakan berbasis prediksi (Evans, 2019; Davenport & Harris, 2007).

Terdapat empat tingkatan analitik data yang menggambarkan kompleksitas dan nilai yang dihasilkan (Evans, 2019). Analitik deskriptif (descriptive analytics) menjawab pertanyaan "apa yang terjadi?" dengan merangkum data historis melalui statistika deskriptif dan visualisasi berkaitan erat dengan teknik yang dibahas pada bab lain sebelumnya. Analitik diagnostik (diagnostic analytics) menjawab "mengapa hal itu terjadi?" dengan menggali lebih dalam untuk menemukan akar penyebab suatu kejadian. Analitik prediktif (predictive analytics) menjawab "apa yang akan terjadi?" menggunakan model dari machine learning dan statistika. Analitik preskriptif (prescriptive analytics) menjawab "apa yang harus dilakukan?" dengan memberikan rekomendasi tindakan berbasis hasil prediksi (Evans, 2019; Davenport & Harris, 2007)

Keempat tingkatan ini bersifat hierarkis yaitu setiap tingkatan membangun di atas tingkatan sebelumnya. Dalam praktiknya, sebagian besar organisasi masih berada di tingkat deskriptif dan baru mulai bergerak menuju analitik prediktif. Visualisasi data yang telah dibahas secara mendalam pada lain sebelumnya memainkan peran penting di setiap tingkatan karena memungkinkan komunikasi temuan yang kompleks secara intuitif kepada pemangku kepentingan non-teknis.

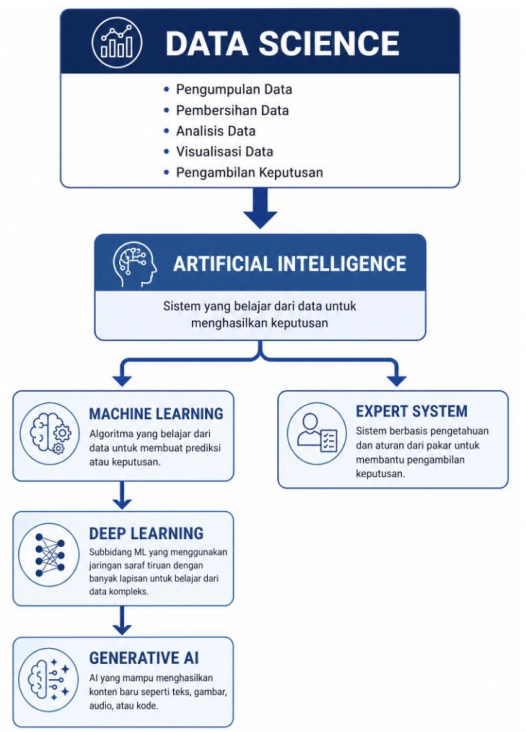
BAB 11
DEEP LEARNING DAN ARTIFICIAL INTELLIGENCE

A. KONSEP DASAR *ARTIFICIAL INTELLIGENCE*

Artificial Intelligence (AI) atau kecerdasan buatan adalah bagian dari ilmu komputer yang berfokus pada pengembangan sistem yang bisa meniru kemampuan kognitif manusia, seperti belajar, berpikir, memahami informasi, memecahkan masalah, dan mengambil keputusan. AI saat ini menjadi salah satu teknologi utama di era transformasi digital karena dapat mengotomatisasi banyak tugas yang dulu hanya bisa dilakukan manusia.

Artificial Intelligence adalah bidang yang mempelajari tentang agen cerdas (intelligent agents), yaitu sistem yang mampu mengamati lingkungan, memproses informasi, dan bertindak secara rasional untuk mencapai tujuan tertentu. Jadi dapat disimpulkan bahwa AI tidak hanya berkaitan dengan kemampuan komputasi, tetapi juga kemampuan sistem untuk bertindak secara cerdas berdasarkan data yang tersedia (Russell, 2021).

Perkembangan teknologi AI sekarang ini sangat tergantung pada kemajuan di bidang Data Science. Data Science memberikan data, cara menganalisisnya, dan metode memproses informasi, sedangkan AI menggunakan hasil dari proses tersebut untuk menciptakan sistem yang bisa belajar sendiri secara otomatis. Dengan kata lain, Data Science berperan sebagai dasar yang menguatkan pembuatan Artificial Intelligence. Berikut adalah diagram hubungan Artificial Intelligence dalam ekosistem data science yang terlihat pada gambar.



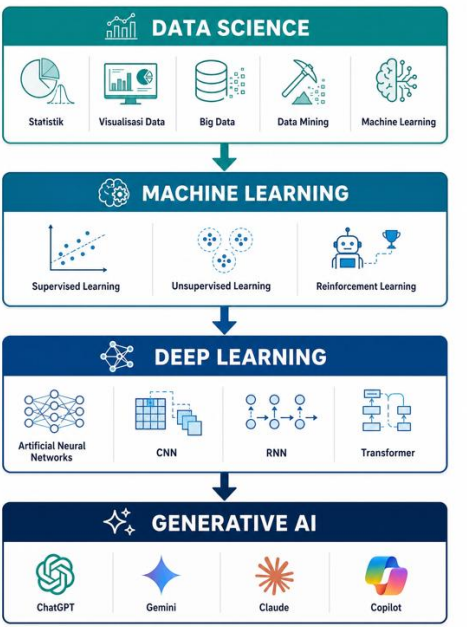
Gambar 1. Hubungan AI dalam Ekosistem Data Science

Pada Gambar 1 terlihat bahwa data science menjadi fondasi utama bagi pengembangan Artificial Intelligence. Data yang dikumpulkan, dibersihkan, dan dianalisis melalui proses Data Science kemudian digunakan oleh sistem AI untuk belajar dan membuat keputusan. Machine Learning adalah salah satu pendekatan utama dalam bidang AI, sedangkan Deep Learning merupakan pengembangan lebih lanjut dari Machine Learning yang menggunakan jaringan saraf tiruan untuk proses pembelajaran. Teknologi AI Generative seperti chatGPT dan gemini merupakan contoh terbaru dari Deep Learning berbasis model bahasa besar (Large Language Models).

B. HUBUNGAN DATA SCIENCE, MACHINE LEARNING, DAN DEEP LEARNING

Istilah data science, machine learning, deep learning, dan artificial intelligence sering digunakan secara bergantian, meskipun masing-masing memiliki makna dan cakupan yang berbeda. Data Science adalah bidang yang berfokus pada pengelolaan dan analisis data. Artificial Intelligence merupakan bidang yang bertujuan menciptakan sistem cerdas. Machine Learning adalah bagian dari kecerdasan buatan yang memungkinkan sistem belajar dari data yang diberikan, sedangkan Deep Learning merupakan bagian dari Machine Learning yang menggunakan jaringan saraf tiruan dengan banyak lapisan (deep neural networks) (Geron, 2022).

Artificial intelligence, machine learning, dan deep learning membentuk hirarki konseptual yang saling berkaitan. Bidang data science bertindak sebagai disiplin ilmu yang beririsan dengan ketiganya, karena data science merupakan studi holistik tentang ekstraksi wawasan dan pengetahuan dari data, yang mencakup prinsip-prinsip statistik, manipulasi data (data engineering), dan pemahaman domain bisnis (Provost, 2013). Adapun hierarki antara data science, artificial intelligence dan juga deep learning seperti terlihat pada Gambar 2.



Gambar 2. Hubungan hierarki AI, ML, dan DL

Pada Gambar 2 di atas terlihat bahwa, deep learning merupakan bagian dari Machine Learning, sedangkan Machine Learning merupakan salah satu pendekatan yang digunakan dalam Artificial Intelligence. Seluruh teknologi tersebut memanfaatkan data yang dikelola melalui proses Data Science. Oleh karena itu, Data Science sering disebut sebagai fondasi utama bagi pengembangan AI modern. Sedangkan chatgpt, gemini, claude dan copilot dikategorikan sebagai generative AI karena memiliki kemampuan untuk menciptakan konten baru seperti teks, kode, gambar, atau hasil analisis berdasarkan instruksi menggunakan prompt dari pengguna. Fungsi AI Generative tidak hanya digunakan untuk mencari informasi, namun juga sebagai alat berfikir logis dan penalaran yang mampu memproses informasi untuk menghasilkan jawaban yang kontekstual.

Berdasarkan penjelasan sebelumnya, dapat disimpulkan

bahwa data science dan AI saling berkaitan dan saling melengkapi sehingga tidak dapat dipisahkan satu sama lain. Ilmu Data Science bertugas mengumpulkan, membersihkan, dan menyusun data agar memiliki kualitas yang baik, sedangkan AI menggunakan teknik Machine Learning (ML) dan Deep Learning (DL) untuk memanfaatkan data tersebut dalam mempelajari pola, membuat prediksi, serta menghasilkan solusi yang lebih cerdas. Dengan kata lain, data yang baik adalah "bahan bakar" bagi AI, sementara AI adalah "mesin" yang mengubah data menjadi informasi dan keputusan yang bernilai. Tanpa adanya Data Science, AI tidak memiliki data yang layak untuk dipelajari, dan tanpa AI, Data Science akan lebih sulit mengolah data dalam jumlah besar untuk menemukan pola yang kompleks. Oleh karena itu, keduanya bekerja secara sinergis untuk menghasilkan analisis dan inovasi yang lebih efektif.

C. PERKEMBANGAN APLIKASI KECERDASAN BUATAN DALAM EKOSISTEM DIGITAL

Pada beberapa tahun terakhir, artificial intelligence telah mengalami perubahan mendasar mulai dari tahapan penelitian hingga menjadi aplikasi nyata yang digunakan dan diakses oleh masyarakat luas. Munculnya berbagai platform aplikasi asisten AI menunjukan adanya pergeseran ke arah spesialisasi fungsi, setiap pengembang besar berusaha mengoptimalkan teknologi mereka untuk kebutuhan pengguna yang beragam, mulai dari kebutuhan riset akademik, pembelajaran interaktif, pelayanan, analisis, diagnosis, deteksi, tools generator, sistem pemasaran produk, hingga peningkatan produktivitas perusahaan. Untuk memahami posisi dan kemampuan masing-masing platform AI yang terkenal saat ini, Tabel 1.1 di bawah ini menyajikan klasifikasi berdasarkan pengembang, peran utama, serta model inti yang digunakan.

Tabel 1. Klasifikasi AI dan Kapabilitas Apllikasi Generative AI

Nama Aplikasi	Pengembang	Peran utama dalam ekosistem	Model inti yang digunakan
ChatGPT	OpenAI	Asisten serbaguna, penalaran logis, dan kreativitas multimodal.	GPT-04 /GPT-5.2
Gemini	Google	Integrasi layanan awan, analisis konteks besar, dan produktivitas Workspace	Gemini 1.5 Pro / Gemini 3
Claude	Anthropic	Penulisan naratif alami, keamanan AI (Constitutional AI), dan koding	Claude 3.5 Sonnet / Opus 4.6
Copilot	Microsoft	Pendamping produktivitas perkantoran dan integrasi sistem operasi.	GPT-4/ GPT-5

D. IMPLEMENTASI DEEP LEARNING DALAM KEHIDUPAN NYATA

1. Deep Learning pada Natural Languange Processing (NLP)

Natural Language Processing (NLP) adalah bagian dari cabang ilmu Artificial Intelligence yang bertujuan agar komputer dapat mengerti, menganalisis, dan menciptakan bahasa yang digunakan oleh manusia. Perkembangan di bidang Deep Learning, khususnya arsitektur Transformer, telah meningkatkan kemampuan pemrosesan bahasa alami (NLP) dalam memahami konteks bahasa secara lebih akurat

dibandingkan metode tradisional (Vaswani, 2017).

Model NLP modern seperti ChatGPT, Gemini, dan Claude menggunakan Large Language Models (LLMs) yang dilatih menggunakan miliaran kata yang berasal dari berbagai sumber data. Model ini mampu menjawab pertanyaan, menerjemahkan bahasa, membuat ringkasan teks, hingga menghasilkan konten secara otomatis (Brown, 2020). Implementasi Deep Learning dalam kehidupan sehari-hari, seperti:

a. Chatbot layanan pelanggan
b. Mesin Penerjemah otomatis (Google Translate)
c. Analisis sentimen media sosial
d. Pembuatan ringkasan dokumen
e. Asisten virtual berbasis AI

Manfaat dari penerapan deep learning di atas adalah: (1) Mempercepat proses pengolahan informasi berbasis teks, (2) Meningkatkan kualitas layanan kepada pelanggan (3) Membantu dalam pengambilan keputusan berdasarkan opini publik.

2. Deep Learning pada Sistem Rekomendasi

Sistem rekomendasi (recommendation systems) merupakan aplikasi kecerdasan buatan yang bertujuan memberikan saran produk, layanan, atau konten yang sesuai dengan selera dan preferensi pengguna. Deep Learning memungkinkan sistem mengenali pola perilaku pengguna secara lebih kompleks dibandingkan metode rekomendasi konvensional. Menurut (Zhang, 2019) Deep Learning mampu memanfaatkan data interaksi pengguna, riwayat pencarian, serta karakteristik produk untuk menghasilkan rekomendasi yang lebih personal dan relevan. Adapun penerapan deep learning pada sistem rekomendasi, yakni:

a. Netflix merekomendasikan film dan serial
b. Spotify merekomendasikan lagu dan podcast.

c. Shopee dan Tokopedia merekomendasikan produk
d. TikTok menampilkan konten pada halaman For You.

Manfaat yang didapatkan meliputi: (1) meningkatkan pengalaman pengguna, (2) Membantu pengguna menemukan konten yang relevan, (3) Meningkatkan penjualan dan keterlibatan pengguna.

3. Deep Learning pada Bidang Kesehatan

Sektor kesehatan adalah salah satu bidang yang paling banyak menggunakan teknologi Deep Learning. Teknologi ini membantu dalam menganalisis citra medis, mendeteksi dan mengenali penyakit, serta membantu dokter dalam proses menentukan diagnosis. Model pembelajaran mendalam menghasilkan tingkat keakuratan yang hampir sama dengan dokter spesialis dalam mendeteksi beberapa jenis penyakit kulit melalui analisis citra medis dengan dokter spesialis dalam mengidentifikasi berbagai jenis penyakit kulit berdasarkan analisa gambar medis citra medis (Esteva, 2017). Adapun penerapan deep learning pada bidang kesehatan, yakni:

a. Deteksi kanker melalui citra radiologi
b. Analisis hasil CT Scan dan MRI.
c. Diagnosis penyakit mata berbasis citra retina.
d. Prediksi risiko penyakit jantung.

Adapun manfaat yang didapatkan meliputi: (1) Mempercepat proses diagnosis, (2) Membantu tenaga medis dalam pengambilan keputusan, (3) Mengurangi potensi kesalahan manusia.

4. Deep Learning pada Bidang Pendidikan

Dalam bidang pendidikan, Deep Learning digunakan untuk menciptakan sistem pembelajaran yang lebih personal, adaptif, dan berbasis data. Teknologi ini membantu menganalisis cara

belajar siswa, sehingga proses mengajar bisa disesuaikan dengan kebutuhan setiap orang. Menurut (UNESCO, 2023) AI memiliki kemampuan untuk meningkatkan akses, kualitas, dan efektivitas pendidikan dengan cara mengotomatisasi tugas administratif, memberikan pembelajaran yang disesuaikan dengan kebutuhan siswa, serta dukungan akademik berbasis teknologi. Contoh implementasi deep learning pada bidang pendidikan, yakni:

a. Sistem pembelajaran adaptif (adaptive learning).
b. Chatbot akademik untuk konsultasi mahasiswa
c. Penilaian otomatis tugas dan kuis.
d. Prediksi keberhasilan dan risiko putus studi mahasiswa
e. Asisten pembelajaran berbasis Generative AI.

Manfaat yang didapatkan meliputi: (1) Mendukung pembelajaran yang lebih personal (2) Membantu dosen dalam proses evaluasi (3) Meningkatkan efisiensi pengelolaan pembelajaran. Adapun implementasi deep learning di atas rapat dirangkum pada Tabel 2.

Tabel 2. Implementasi Deep Learning

Bidang	Contoh Implementasi
NLP	ChatGPT, Google Translate, Chatbot
Sistem Rekomendasi	Netflix, Spotify, Shopee
Kesehatan	Deteksi kanker, analisis MRI
Pendidikan	Adaptive Learning, Chatbot Akademik

Berdasarkan penjelasan di atas, Deep Learning telah menjadi teknologi utama yang mendukung berbagai aplikasi AI modern. Implementasinya tidak hanya terbatas pada bidang teknologi informasi, tetapi juga melibatkan bahasa alami, sistem rekomendasi, kesehatan, dan pendidikan. Kemampuan Deep Learning dalam mengenali pola yang kompleks menjadikannya salah satu teknologi paling penting dalam pengembangan solusi berbasis Data Science dan Artificial Intelligence pada era digital saat ini.

BAB 12
IMPLEMENTASI DATA SCIENCE

Bab-bab sebelumnya membahas Data Science dari sisi konsep dan metode: statistika, pemrograman, pengolahan data, hingga pemodelan dan kecerdasan buatan. Semua itu baru bernilai ketika model benar-benar dioperasikan untuk mendukung keputusan. Sesuai judul buku ini, Konsep, Metode, dan lmplementasi, bab ini menem pati pos isi sebagai muara, yaitu tahap yang mengubah hasil eksperimen menjadi sistem yang berjalan, andal, dan dipakai pengguna.

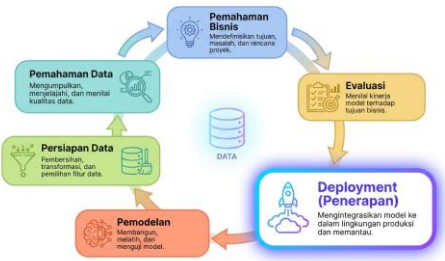
lmplementasi Data Science adalah proses menjadikan model dan hasil analisis sebagai bagian dari sistem yang berjalan berkelanjutan. Davenport dan Patit (2012) menegaskan bahwa keunggulan organisasi ditentukan oleh kemampuan mengubah data menjadi keputusan bernilai, bukan sekadar memilikinya. Provost dan Fawcett (2013) pun memandang Data Science sebagai prinsip untuk mengekstraksi pengetahuan dari data bagi kepentingan praktis. Bab ini membahas kedudukan

implementasi dalam siklus hidup proyek, arsitektur dan pipeline, strategi penyebaran model, MLOps, pemantauan dan pemeliharaan, tata kelola data, serta studi kasus penerapannya.

A. KEDUDUKAN IMPLEMENTASI DALAM SIKLUS HIDUP PROYEK

lmplementasi bukan tahap yang berdiri sendiri, melainkan kelanjutan dari seluruh rangkaian kerja Data Science. Kerangka yang umum dipakai adalah CRlSP-DM dengan enam fase:

pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan atau deployment (Shearer, 2000). Seperti diuraikan pada Bab 5, lima fase pertama berfokus membangun model, sedangkan fase penerapan justru kerap menjadi titik terlemah. Wirth dan Hipp (2000) menilai kekuatan CRISP-DM terletak pada sifatnya yang dapat diu lang lintas industri dan teknologi.



Gambar 1. Siklus CRISP-DM dengan penekanan pada fase penerapan (deployment).

Banyak proyek berhenti di fase evaluasi karena dianggap selesai saat akurasi tercapai, padahal nilai sesungguhnya muncul setelah model dioperasikan. Keberhasilan implementasi ditentukan oleh beberapa faktor utama:

a. Kejelasan tujuan bisnis, sehingga model dapat ditelusuri ke persoalan nyata yang ingin diselesaikan.
b. Kesiapan data operasional yang setara kualitas dan formatnya dengan data pelatihan.
c. lntegrasi dengan sistem yang ada agar model menyatu dengan aplikasi atau proses bisnis.
d. Mekanis me pemeliharaan yang disiapkan sejak awal, bukan setelah masalah muncul.

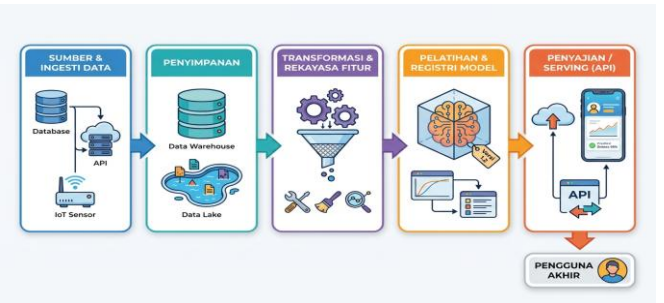
Karena itu, implementasi menuntut kolaborasi data scientist, data engineer, software engineer, dan pemangku kepentingan bisnis. Sinergi inilah yang membedakan prototipe laboratorium dari solusi yang andal di lingkungan nyata.

B. ARSITEKTUR DAN PIPELINE IMPLEMENTASI

lmplementasi bertumpu pada arsitektur yang mengalirkan data dari sumber ke model lalu ke pengguna. Alur ini disebut pipeline data, kelanjutan dari tahap pengolahan dan persiapan data pada Bab 6 hingga penyajian hasil. McKinney (2017) menyebut ekosistem Python dengan Pandas dan NumPy sebagai fondasi pipeline yang efisien dan dapat diulang.

1. Komponen Utama Pipeline

a. Lapisan sumber dan ingesti data dari basis data, API, sensor, atau berkas, secara batch maupun streaming, sering dibantu Apache Airflo w.
b. Lapisan penyimpanan: data warehouse untuk data terstruktur dan data lake untuk data mentah berbagai format.
c. Lapisan transformasi, tempat data dibersihkan dan direkayasa menjadi fitur (feature engineering).
d. Lapisan pelatihan dan registri model, misalnya dengan MLflow, agar versi model terlacak dan dapat direproduksi.
e. Lapisan penyajian (serving) yang mengekspos prediksi melalui antarmuka, umumnya berupa API.



Gambar 2. Arsitektur pipeline implementasi, dari sumber data hingga pengguna akhir.

2. Pilihan Lingkungan Penyebaran

Organisasi dapat memilih lingkungan on-premise, cloud, atau hibrida sesuai kebutuhan keamanan, skala, dan biaya. Cloud menawarkan elastisitas sumber daya, kontainer seperti Docker mengemas modet bersama dependensinya, dan orkestrator seperti Kubernetes mengelolanya secara otomatis sehingga modet berjalan konsisten dari komputer pengembang hingga server produksi.

C. STRATEGI PENYEBARAN MODEL (MODEL DEPLOYMENT)

Penyebaran model adalah inti implementasi, yaitu menjadikan model dapat diakses untuk memprediksi data baru. Geron (2019) menegas kan keputusan arsitektur penyebaran sama pentingnya dengan pemilihan algoritma karena menentukan pengalaman pengguna dan biaya operasional. Pilihannya bergantung pada tuntutan kecepatan respons dan volume data.

a. Prediksi Batch

   Mode batch memproses sekumpulan besar data secara berkala, misalnya menghitung skor seluruh pengguna setiap malam. Sederhana dan hemat sumber daya, cocok bila respons seketika tidak diperlukan.

b. Prediksi Daring (Real-Time/Online)

   Mode daring membungkus model sebagai layanan API yang memberi prediksi langsung saat dipanggil, misalnya rekomendasi produk. Mode ini menuntut latensi rendah dan ketersediaan tinggi sehingga infrastrukturnya lebih kompleks.

c. Penyebaran pada Perangkat Tepi (Edge) dan Tertanam

   Sebagian model ditempatkan langsung di perangkat pengguna seperti ponsel atau sensor loT agar berlatensi

sangat rendah dan tetap berfungsi tanpa koneksi tetap. Untuk model deep learning berukuran besar sebagaimana dibahas pada Bab 11, diperlukan teknik kompresi model serta pertimbangan kebutuhan memori dan prosesor.

Tabel 1. Perbandingan strategi penyebaran model.

Aspek	**Batch**	**Real-Time (API)**	**Edge/Tertanam**
Latensi	Tinggi (berkala)	Rendah (seketika)	Sangat rendah (lokal)
Pola eksekusi	Terjadwal, banyak data	Per permintaan	Di perangkat pengguna
Contoh	Skor risiko harian	Rekomendasi daring	Deteksi wajah di ponsel
Kelebihan	Sederhana, hemat sumber daya	Respons cepat, interaktif	Luring, menjaga privasi
Kekurangan	Hasil tidak seketika	Infrastruktur kompleks	Sumber daya terbatas

d. Strategi Perilisan yang Annan

Untuk menekan risiko, perilisan model baru sebaiknya bertahap. Canary release memperkenalkan model ke sebagian kecil pengguna lebih dulu, shadow deployment menjalan kan nya paralel tanpa memengaruhi keputusan nyata, dan pengujian A/B memasti kan model baru benar-benar lebih baik sebelum diberlakukan penuh.

Tabel 2. Strategi perilisan model yang aman.

Strategi	**Cara Kerja**	**Manfaat Utama**
Canary release	Dirilis ke sebagian kecil pengguna lebih dulu	Membatasi dampak bila terjadi kesalahan
Shadow deployment	Berjalan paralel tanpa memengaruhi keputusan	Membandingkan kinerja secara aman

A/B testing	Dua model dibandingkan pada dua kelompok	Mengukur peningkatan secara statistik

e. Contoh Potongan Kode Penyajian Model

Sebagai gambaran teknis, potongan kode berikut menyajikan model yang telah disimpan: pertama sebagai layanan API dengan Flask untuk prediksi daring kedua sebagai prediksi batch atas sekumpulan data.

```
# Menyajikan model sebagai API (prediksi
daring)
from flask import Flask, request, jsonify
import joblib

app = Flask(__name__)
model = joblib.load("model.pkl")

@app.route("/prediksi", methods=["POST"])
def prediksi():
    data = request.get_json()
    hasil = model.predict([data["fitur"]])
    return jsonify({"prediksi": int(hasil[0])})

if __name__ == "__main__":
    app.run(host="0.0.0.0", port=5000)
```

Kode Program 1. Menyajikan model sebagai API dengan Flask.

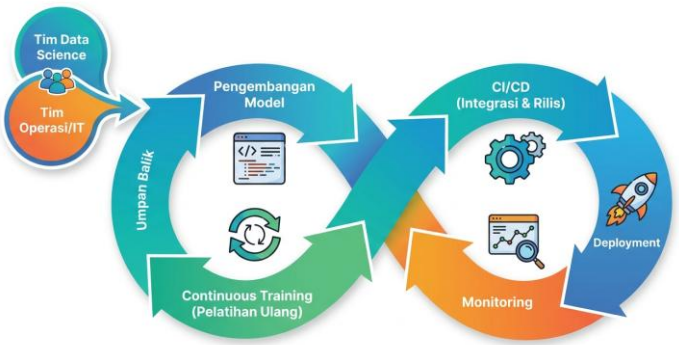
```
# Prediksi batch terhadap banyak data sekaligus
import pandas as pd, joblib

df = pd.read_csv("data_baru.csv")
```

```
model = joblib.load("model.pkl")
df["skor"] =
model.predict(df.drop(columns=["id"]))
df.to_csv("hasil_skor.csv", index=False)
```

Kode Program 2. Menjalankan prediksi secara batch.

D. MLOPS: MENGOPERASIONALKAN MODEL SECARA BERKELANJUTAN

Saat jumlah model bertambah, organisasi memerlukan disiplin untuk mengelola siklus hidupnya, yaitu MLOps (Machine Learning Operations): penerapan prinsip Devops pada sistem pembelajaran mesin. Kreuzberger, Kuhl, dan Hirschl (2023) mendefinisikannya sebagai paduan praktik, budaya kolabo rasi, dan perangkat untuk mengotomatis kan serta mengoperasionalkan sistem machine learning secara andal dari pen gembangan hingga produksi.



Gambar 3. Siklus MLOps yang memadukan tim Data Science dan tim Operasi/IT.

Prinsip pokok MLOps yang relevan dengan implementasi adalah:

a. Continuous Integration dan Continuous Delivery (Cl/CD), mengotomatiskan pengujian dan perilisan kode serta model agar perubahan cepat dan aman diterapkan.
b. Continuous Training (CT), melatih ulang model secara otomatis saat ada data baru atau kinerja menurun.
c. Versioning data, fitur, dan model agar hasil dapat direproduksi, misalnya dengan MLflow.
d. Otomatisasi pipeline untuk mengurangi intervensi manual sekaligus memperkecil kesalahan manusia.
e. Dengan MLOps, model diperlakukan sebagai produk hidup yang terus dipantau dan diperbarui, sehingga inovasi analitik sampai ke pengguna lebih cepat dan konsisten.

E. PEMANTAUAN, PEMELIHARAAN, DAN UTANG TEKNIS

lmplementasi tidak berakhir saat model dirilis. Kinerja model menurun seiring waktu karena karakteristik data nyata terus berubah, fenomena yang disebut model drift atau data drift. Paleyes, Urma, dan Lawrence (2022) menemukan bahwa sebagian besar kesulitan penerapan machine learning justru muncul pasca-deployment, terutama pada pemantauan dan pemeliharaan.

Sculley dkk. (2015) mengingatkan bahwa sistem machine learning rentan menumpuk utang teknis tersembunyi (hidden technical debt) karena, selain memiliki persoalan perangkat lunak biasa, menghadapi ketergantungan data tak terdeklarasi dan keterikatan antarkomponen. Maka implementasi yang sehat menyertakan praktik berikut:

a. Pemantauan kinerja model melalui metrik akurasi dan

indikator bisnis secara berkala.

b. Pemantauan kualitas data masukan agar tetap konsisten dengan asumsi pelatihan.
c. Pelatihan ulang terjadwal atau berbasis pemicu agar model tetap relevan.
d. Pencatatan (logging) dan kemampuan telusur agar setiap keputusan model dapat diaudit.

F. TATA KELOLA, ETIKA, DAN KEAMANAN DATA DALAM IMPLEMENTASI

Saat beroperasi nyata, model menyentuh data pribadi dan memengaruhi keputusan yang berdampak pada manusia. Di Indonesia, hal ini diatur Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi yang menetapkan asas, hak subjek data, serta kewajiban pengendali dan prosesor data (Republik Indonesia, 2022). Tata kelola data sektor publik juga diperkuat kebijakan Satu Data Indonesia yang menuntut data akurat, mutakhir, terpadu, dan dapat dibagipakaikan (Republik Indonesia, 2019). Prinsip implementasi yang bertanggung jawab meliputi:

a. Transparansi dan akuntabilitas, yaitu kemampuan menjelaskan dan mempertanggungjawabkan keputusan modet.
b. Persetujuan dan minimalisasi data, hanya memproses data relevan atas dasar persetujuan yang sah.
c. Keadilan dan mitigasi bias agar modet tidak melanggengkan diskriminasi terhadap kelompok tertentu.
d. Keamanan informasi, mencakup pengendalian akses, enkripsi, dan perlindungan modet dari penyalahgunaan.

Tata kelola yang baik menjadikan implementasi sah secara

hukum dan etis sekaligus membangun kepercayaan pengguna sebagai modal keberlanjutan solusi berbasis data.

G. STUDI KASUS DAN PENERAPAN DI BERBAGAI SEKTOR

Sebagai ilustrasi, perhatikan penerapan pada sektor pendidikan yang banyak dijadikan contoh di bab sebelumnya. Seb uah perguruan tinggi ingin menekan angka mahasiswa yang berisiko gagal atau berhenti studi. Tahapan implementasinya:

a. Pipeline harian menarik nilai tugas, kehadiran, dan aktivitas pada sistem pembelajaran daring, lalu mengo lahnya menjadi fitur seperti rata-rata nilai dan tingkat kehadiran.
b. Model klasifikasi memprediksi tingkat risiko tiap mahasiswa dan dilatih ulang setiap akhir semester.
c. Hasil disajikan lewat dasbor sebagai sistem peringatan dini sehingga dosen pembimbing dapat mendampingi lebih awal.
d. Sistem dipantau agar tetap akurat dan adil serta mematuhi ketentuan perlindungan data pribadi mahasiswa.

Pola serupa dijumpai di sektor lain. E-commerce menerapkan rekomendasi produk daring untuk meningkatkan keterlibatan pelanggan, sedangkan kesehatan mengintegrasikan model pen dukung diagnosis citra medis dengan tenaga medis sebagai penentu akhir. Nilai Data Science terwujud bukan saat model dilatih, melainkan ketika ia bekerja berkelanjutan dalam proses nyata.

H. TANTANGAN DAN PRAKTIK TERBAIK

Implementasi menghadapi tantangan teknis sekaligus organisasional: kesenjangan antara lingkungan eksperimen dan produksi, ketergantungan pada kualitas data operasio na1, keterbatasan keterampilan lintas peran, serta risiko penurunan kinerja model. Praktik terbaik untuk menghadapinya:

a. Mulai dari masalah bis nis yang jelas dengan indikator keberhasilan sejak awal.
b. Bangun pipeline yang otomatis dan dapat direproduksi agar transisi riset ke produksi mulus.
c. Terapkan MLOps secara bertahap sesuai tingkat kematangan organisasi.
d. Siapkan pemantauan dan rencana pelatihan ulang sebagai bagian tak terpisahkan dari sistem.
e. lntegrasikan aspek etika, privasi, dan keamanan sejak tahap perancangan.

BAB 13
MANAJEMEN PROYEK DATA SCIENCE

Pesatnya perkembangan teknologi digital telah menyebabkan peningkatan jumlah data yang dihasilkan dari berbagai aktivitas manusia dan organisasi. Data tersebut berasal dari beragam sumber, seperti media sosial, transaksi bisnis, perangkat elektronik, serta berbagai sistem informasi yang digunakan dalam kehidupan sehari-hari. Keberadaan data dalam jumlah besar menjadi aset yang bernilai karena dapat diolah menjadi informasi dan pengetahuan yang mendukung proses pengambilan keputusan secara lebih akurat. Untuk memanfaatkan potensi tersebut, Data Science hadir sebagai bidang ilmu yang menggabungkan konsep statistika, ilmu komputer, dan kecerdasan buatan dalam proses pengolahan dan analisis data.

Dalam pelaksanaannya, proyek Data Science tidak hanya berorientasi pada pengolahan dan analisis data semata, tetapi juga melibatkan berbagai aktivitas manajerial, seperti perencanaan proyek, pengelolaan sumber daya, koordinasi tim, serta penerapan hasil analisis ke dalam lingkungan kerja organisasi. Oleh karena itu, diperlukan manajemen proyek yang mampu mengarahkan seluruh tahapan pekerjaan agar berjalan sesuai target yang telah ditetapkan. Penerapan manajemen proyek yang efektif dapat membantu meningkatkan kualitas hasil, mengoptimalkan penggunaan sumber daya, serta mengurangi berbagai risiko yang berpotensi menghambat keberhasilan proyek.

Proyek Data Science memiliki karakteristik yang berbeda dibandingkan proyek pengembangan perangkat lunak tradisional. Jika proyek perangkat lunak umumnya dimulai dengan kebutuhan

yang telah dirumuskan secara jelas, proyek Data Science sering kali berkembang secara bertahap melalui proses eksplorasi dan pengujian data yang berulang. Kondisi ini membuat proyek Data Science bersifat iteratif dan fleksibel, karena keputusan yang diambil sangat dipengaruhi oleh temuan selama proses analisis. Dengan dukungan manajemen proyek yang terstruktur, organisasi dapat mengelola kompleksitas tersebut dan menghasilkan solusi berbasis data yang memberikan manfaat nyata bagi peningkatan kinerja dan kualitas pengambilan keputusan.

A. MANAJEMEN PROYEK DATA SCIENCE

Manajemen proyek merupakan serangkaian aktivitas yang meliputi perencanaan, pengorganisasian, pelaksanaan, pemantauan, dan pengendalian sumber daya untuk mencapai tujuan tertentu sesuai dengan batasan waktu, biaya, dan kualitas yang telah ditetapkan. Penerapan manajemen proyek memungkinkan organisasi menjalankan pekerjaan secara lebih terstruktur sehingga target yang telah direncanakan dapat tercapai secara efektif dan efisien. Menurut Project Management Institute (PMI), manajemen proyek adalah penerapan pengetahuan, keterampilan, metode, dan berbagai alat yang diperlukan untuk memenuhi kebutuhan suatu proyek. Oleh karena itu, keberhasilan proyek sangat dipengaruhi oleh kemampuan dalam mengelola aktivitas, sumber daya, dan pihak-pihak yang terlibat agar mampu menghasilkan nilai bagi organisasi maupun pemangku kepentingan.

Dalam konteks Data Science, manajemen proyek merupakan penerapan prinsip-prinsip manajemen proyek pada kegiatan yang berfokus pada pemanfaatan data untuk menghasilkan informasi, wawasan, model prediksi, maupun solusi berbasis data. Proyek Data Science bertujuan membantu organisasi dalam memecahkan

masalah dan mendukung pengambilan keputusan melalui analisis data yang sistematis. Namun demikian, proyek Data Science memiliki karakteristik yang berbeda dibandingkan proyek pada umumnya karena tingkat ketidakpastiannya relatif tinggi. Keberhasilan proyek sangat dipengaruhi oleh kualitas data yang tersedia, metode analisis yang digunakan, serta kemampuan model dalam menghasilkan prediksi atau rekomendasi yang akurat. Oleh sebab itu, pengelolaan proyek Data Science memerlukan pendekatan yang adaptif dan bersifat iteratif untuk menyesuaikan diri dengan perubahan kebutuhan maupun temuan yang muncul selama proses analisis.

Tujuan utama manajemen proyek Data Science adalah memastikan bahwa solusi berbasis data yang dihasilkan mampu menjawab kebutuhan organisasi dan memberikan manfaat yang nyata. Pengelolaan proyek yang baik membantu memastikan bahwa pekerjaan dapat diselesaikan sesuai jadwal, anggaran, dan standar kualitas yang telah ditentukan. Selain itu, manajemen proyek Data Science juga memberikan berbagai manfaat, seperti meningkatkan efisiensi penggunaan sumber daya, mengurangi risiko kegagalan proyek, memperjelas ruang lingkup pekerjaan, serta meningkatkan kualitas hasil analisis yang dihasilkan. Keberadaan manajemen proyek juga berperan penting dalam menjembatani komunikasi antara tim teknis dan pemangku kepentingan sehingga kebutuhan organisasi dapat diterjemahkan menjadi solusi yang tepat sasaran.

Dibandingkan dengan proyek pengembangan perangkat lunak konvensional, proyek Data Science memiliki karakteristik yang berbeda. Pada proyek perangkat lunak, kebutuhan dan spesifikasi sistem umumnya telah ditetapkan secara jelas sejak tahap awal pengembangan. Sebaliknya, proyek Data Science lebih berorientasi pada eksplorasi data untuk menemukan pola, informasi, atau model yang dapat mendukung pengambilan

keputusan. Prosesnya sering kali bersifat eksperimental dan berulang karena hasil yang diperoleh tidak selalu dapat diperkirakan sejak awal. Kondisi tersebut membuat proyek Data Science lebih fleksibel namun juga memiliki tingkat ketidakpastian yang lebih tinggi dibandingkan proyek perangkat lunak tradisional.

Peran manajemen proyek menjadi sangat penting dalam memastikan bahwa setiap tahapan proyek Data Science dapat berjalan secara terarah dan memberikan hasil yang sesuai dengan tujuan organisasi. Tanpa pengelolaan yang baik, proyek berpotensi mengalami berbagai permasalahan, seperti keterlambatan penyelesaian, pembengkakan biaya, atau kegagalan dalam menghasilkan solusi yang dapat diterapkan. Selain itu, proyek Data Science umumnya melibatkan berbagai profesi dengan latar belakang yang berbeda, seperti data scientist, data engineer, business analyst, dan manajer proyek. Oleh karena itu, manajemen proyek diperlukan untuk mengoordinasikan kolaborasi antaranggota tim, mengelola risiko, serta memastikan bahwa hasil analisis dapat diterapkan secara efektif dalam lingkungan organisasi.

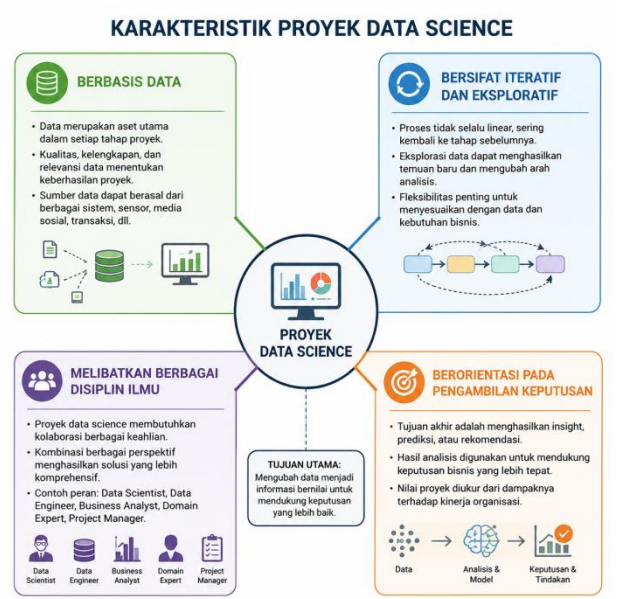
Meskipun menawarkan banyak manfaat, proyek Data Science juga menghadapi berbagai tantangan. Salah satu tantangan utama adalah kualitas data yang sering kali belum memenuhi standar untuk dianalisis secara langsung. Data yang tidak lengkap, tidak konsisten, atau mengandung kesalahan dapat memengaruhi validitas hasil analisis dan menurunkan performa model yang dikembangkan. Akibatnya, proses pembersihan dan persiapan data sering menjadi tahapan yang paling banyak menyita waktu dalam proyek Data Science. Selain itu, perubahan kebutuhan bisnis, keterbatasan sumber daya, serta kesulitan dalam mengimplementasikan model ke lingkungan operasional juga menjadi tantangan yang perlu dihadapi. Dari sisi nonteknis, komunikasi antara tim teknis dan pemangku kepentingan sering

kali menjadi faktor penentu keberhasilan proyek karena hasil analisis harus dapat dipahami dan dimanfaatkan secara optimal dalam proses pengambilan keputusan.

B. KARAKTERISTIK PROYEK DATA SCIENCE

Proyek Data Science memiliki sejumlah karakteristik khusus yang membedakannya dari proyek teknologi informasi maupun pengembangan perangkat lunak tradisional. Perbedaan tersebut muncul karena fokus utama proyek Data Science adalah mengolah dan menganalisis data untuk menghasilkan wawasan, pengetahuan, prediksi, atau rekomendasi yang dapat memberikan dukungan dalam proses pengambilan keputusan. Dengan demikian, keberhasilan proyek tidak hanya diukur dari keberhasilan pengembangan sistem atau aplikasi, tetapi juga dari sejauh mana hasil yang diperoleh mampu memberikan manfaat dan nilai tambah bagi organisasi.

Di samping itu, proyek Data Science umumnya melibatkan proses eksplorasi data yang mendalam, penerapan metode statistika dan machine learning, serta kerja sama antara berbagai bidang keahlian. Hasil yang diperoleh sering kali tidak dapat diprediksi secara pasti sejak awal karena dipengaruhi oleh kualitas data, teknik analisis yang digunakan, serta pola-pola yang ditemukan selama proses pengolahan data. Oleh sebab itu, proyek Data Science memiliki beberapa karakteristik utama, yaitu berfokus pada data sebagai sumber utama informasi, bersifat iteratif dan eksploratif, melibatkan kolaborasi multidisiplin, serta berorientasi pada penyediaan informasi yang mendukung pengambilan keputusan secara efektif.



Gambar 2. Karakteristik Proyek Data Science

1. Berbasis Data

Proyek Data Science menempatkan data sebagai sumber daya utama dalam proses menghasilkan informasi dan pengetahuan yang bernilai. Melalui pengolahan dan analisis data, berbagai pola, tren, serta hubungan antarvariabel dapat diidentifikasi untuk menghasilkan wawasan, membangun model prediktif, dan mendukung proses pengambilan keputusan. Data yang digunakan dalam proyek dapat berasal dari berbagai sumber, seperti sistem basis data organisasi, media sosial, perangkat Internet of Things (IoT), transaksi bisnis, maupun sumber data publik. Oleh karena itu, kualitas, kelengkapan, konsistensi, dan relevansi data menjadi aspek yang sangat penting karena berpengaruh langsung terhadap akurasi hasil analisis dan keberhasilan proyek secara keseluruhan.

2. Bersifat Iteratif dan Eksploratif

Proyek Data Science memiliki karakteristik yang bersifat iteratif dan eksploratif, sehingga setiap tahapan analisis dapat dilakukan berulang kali untuk memperoleh hasil yang lebih akurat dan sesuai dengan tujuan yang ditetapkan. Dalam prosesnya, tim sering menemukan pola, hubungan, anomali, maupun informasi baru yang sebelumnya belum diketahui. Temuan tersebut dapat mendorong dilakukannya penyesuaian terhadap data yang digunakan, metode analisis yang diterapkan, maupun pendekatan yang diambil. Oleh karena itu, pelaksanaan proyek Data Science membutuhkan tingkat fleksibilitas yang tinggi agar mampu beradaptasi dengan perubahan dan perkembangan yang muncul selama proses analisis berlangsung.

3. Melibatkan Berbagai Disiplin Ilmu

Proyek Data Science merupakan bidang yang mengintegrasikan berbagai disiplin ilmu, seperti statistika, matematika, ilmu komputer, kecerdasan buatan, serta pengetahuan khusus yang berkaitan dengan domain permasalahan yang dihadapi. Perpaduan berbagai bidang keilmuan tersebut diperlukan untuk mengolah data menjadi informasi dan solusi yang dapat memberikan nilai bagi organisasi. Dalam pelaksanaannya, proyek Data Science melibatkan berbagai profesional, seperti data scientist, data engineer, business analyst, machine learning engineer, dan pakar domain, yang bekerja sama secara terpadu untuk menghasilkan solusi berbasis data yang efektif, akurat, dan selaras dengan kebutuhan organisasi.

4. Berorientasi pada Pengambilan Keputusan

Proyek Data Science dirancang untuk mendukung proses pengambilan keputusan melalui pemanfaatan analisis data yang

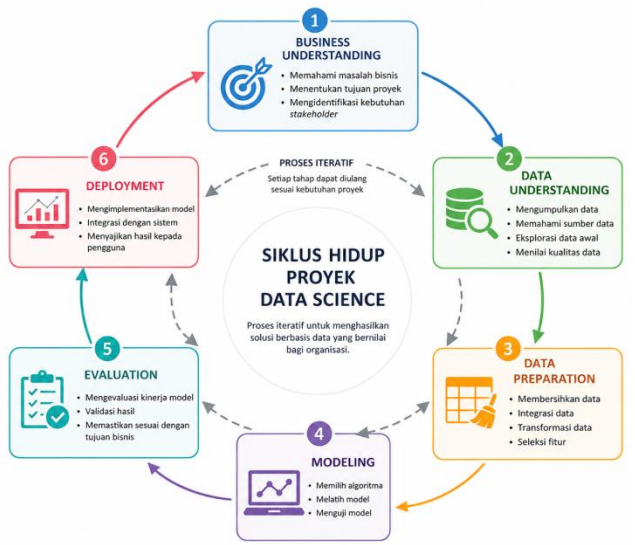
sistematis, objektif, dan terukur. Hasil yang dihasilkan dari proses tersebut dapat berupa wawasan, prediksi, rekomendasi, maupun sistem pendukung keputusan yang membantu organisasi dalam menentukan kebijakan dan strategi yang lebih tepat. Dengan memanfaatkan data sebagai dasar pertimbangan, organisasi dapat mengambil keputusan secara lebih cepat, akurat, dan berdasarkan bukti yang dapat dipertanggung jawabkan.

C. SIKLUS HIDUP PROYEK DATA SCIENCE

Siklus hidup proyek Data Science merupakan rangkaian tahapan yang dilaksanakan secara terstruktur untuk menghasilkan solusi yang memanfaatkan data sebagai dasar pengambilan keputusan. Salah satu framework yang banyak digunakan dalam pengelolaan proyek Data Science adalah CRISP-DM (Cross Industry Standard Process for Data Mining). Framework ini dikembangkan sebagai pedoman standar dalam pelaksanaan proyek data mining dan telah diterapkan secara luas di berbagai bidang industri. CRISP-DM menyediakan pendekatan yang sistematis, namun tetap fleksibel dan iteratif, sehingga memungkinkan tim proyek melakukan penyesuaian pada setiap tahap sesuai dengan karakteristik data dan kebutuhan organisasi.

CRISP-DM terdiri atas enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Melalui tahapan-tahapan tersebut, framework ini membantu memastikan bahwa proses analisis data tidak hanya menghasilkan model yang memiliki kinerja yang baik secara teknis, tetapi juga mampu memberikan manfaat yang nyata bagi organisasi. Dengan pendekatan yang terstruktur dan mudah diterapkan, CRISP-DM hingga saat ini masih menjadi salah satu framework yang paling banyak digunakan dalam pengelolaan

proyek Data Science dan data mining.



Gambar 1. Siklus Hidupp Proyek Data Science

1. Business Understanding (Pemahaman Bisnis)

Tahap pertama dalam CRISP-DM berfokus pada pemahaman terhadap permasalahan yang dihadapi organisasi, kebutuhan bisnis, serta tujuan yang ingin dicapai melalui proyek Data Science. Pada tahap ini, tim proyek berkolaborasi dengan para pemangku kepentingan (stakeholder) untuk memahami konteks permasalahan, menetapkan sasaran proyek, dan menentukan kriteria atau indikator yang akan digunakan untuk mengukur keberhasilan proyek.

2. Data Understanding (Pemahaman Data)

Tahap Data Understanding bertujuan untuk memperoleh pemahaman yang mendalam mengenai data yang akan digunakan dalam proyek. Kegiatan pada tahap ini mencakup pengumpulan data, identifikasi sumber data, eksplorasi awal terhadap karakteristik data, serta penilaian kualitas data. Selain

itu, dilakukan pula pemeriksaan untuk mengidentifikasi berbagai permasalahan yang mungkin terdapat dalam data, seperti nilai yang hilang (missing values), inkonsistensi data, duplikasi, maupun kesalahan pencatatan yang dapat memengaruhi hasil analisis.

3. Data Preparation (Persiapan Data)

Tahap Data Preparation dilakukan untuk mempersiapkan data agar siap digunakan dalam proses pemodelan. Pada tahap ini, data yang telah dikumpulkan akan dibersihkan dan diolah melalui berbagai aktivitas, seperti menghapus data duplikat, menangani nilai yang hilang (missing value), melakukan transformasi data, menggabungkan data dari berbagai sumber, serta memilih atribut atau fitur yang paling relevan. Tujuan dari tahap ini adalah menghasilkan dataset yang berkualitas dan sesuai untuk mendukung proses analisis serta pembangunan model.

4. Modeling (Pemodelan)

Tahap Modeling merupakan proses pembangunan model dengan memilih dan menerapkan algoritma analisis data atau machine learning yang sesuai dengan tujuan proyek. Pada tahap ini, model dikembangkan menggunakan data yang telah dipersiapkan sebelumnya, kemudian dilakukan proses pelatihan (training) dan pengujian (testing) untuk mengevaluasi kinerjanya. Hasil evaluasi tersebut digunakan untuk menentukan apakah model yang dibangun telah mampu menghasilkan prediksi atau klasifikasi yang sesuai dengan kebutuhan proyek.

5. Evaluation (Evaluasi)

Tahap Evaluation bertujuan untuk menilai kinerja model yang telah dibangun dan memastikan bahwa hasil yang diperoleh

sesuai dengan tujuan bisnis yang telah ditetapkan. Pada tahap ini, model dievaluasi menggunakan berbagai ukuran kinerja yang relevan, seperti accuracy, precision, recall, F1-score, atau Root Mean Square Error (RMSE). Hasil evaluasi tersebut menjadi dasar untuk menentukan apakah model sudah layak digunakan atau masih memerlukan perbaikan dan pengembangan lebih lanjut.

6. Deployment (Implementasi)

Tahap Deployment merupakan tahap akhir dalam siklus proyek Data Science yang berfokus pada penerapan model atau hasil analisis ke dalam lingkungan operasional organisasi. Tujuannya adalah agar hasil yang telah dikembangkan dapat dimanfaatkan secara langsung oleh pengguna dalam mendukung aktivitas dan pengambilan keputusan. Implementasi dapat dilakukan dalam berbagai bentuk, seperti dashboard analitik, aplikasi berbasis web, sistem pendukung keputusan, maupun layanan yang memanfaatkan teknologi kecerdasan buatan. Secara umum, setiap tahapan dalam proyek Data Science memiliki keterkaitan yang saling mendukung, mulai dari perencanaan hingga implementasi, dengan menghasilkan keluaran (output) yang menjadi dasar bagi tahapan berikutnya seperti terlihat pada Tabel. 1 dibawah ini :

Tabel 1. Deskripsi dan Aktivitas Utama dalam Proyek Data Science

Tahap	Deskripsi	Aktivitas Utama	Output
Business Understanding	Memahami permasalahan bisnis, kebutuhan organisasi, dan tujuan proyek yang ingin dicapai.	Identifikasi masalah, penentuan tujuan, analisis kebutuhan stakeholder, penyusunan rencana proyek.	Rumusan masalah, tujuan proyek, dan kriteria keberhasilan.
Data Understanding	Mengumpulkan dan memahami data yang akan	Pengumpulan data, eksplorasi data, identifikasi kualitas	Dataset awal dan laporan

	digunakan dalam proyek.	data, analisis karakteristik data.	pemahaman data.
Data Preparation	Mempersiapkan data agar siap digunakan dalam proses pemodelan.	Pembersihan data, integrasi data, transformasi data, seleksi fitur, penanganan missing value.	Dataset yang bersih dan siap untuk analisis atau pemodelan.
Modeling	Membangun model analitik atau machine learning untuk menyelesaikan permasalahan yang telah ditentukan.	Pemilihan algoritma, pelatihan model, pengujian model, tuning parameter.	Model yang telah dilatih dan hasil prediksi awal.
Evaluation	Mengevaluasi performa model dan kesesuaiannya dengan tujuan bisnis.	Pengukuran akurasi model, validasi hasil, analisis kesesuaian dengan kebutuhan bisnis.	Hasil evaluasi dan rekomendasi perbaikan model.
Deployment	Mengimplementasikan model atau hasil analisis agar dapat digunakan oleh organisasi.	Implementasi model, pembuatan dashboard, integrasi sistem, monitoring dan pemeliharaan.	Solusi yang siap digunakan dan memberikan nilai bagi organisasi.

D. STUDI KASUS SEDERHANA : IMPLEMENTASI METODE NAÏVE BAYES UNTUK PREDIKSI DAERAH TERJANGKIT DEMAM BERDARAH DENGUE

Implementasi Metode Naïve Bayes untuk Prediksi Daerah Terjangkit Demam Berdarah Dengue (DBD) digunakan untuk mengidentifikasi dan memprediksi wilayah yang berpotensi mengalami penyebaran DBD berdasarkan data historis dan faktor-faktor yang memengaruhinya. Metode ini menganalisis berbagai variabel, seperti curah hujan, suhu udara, kepadatan penduduk,

kondisi sanitasi, serta jumlah kasus DBD pada periode sebelumnya untuk menghitung probabilitas suatu daerah termasuk kategori terjangkit atau tidak terjangkit DBD. Hasil prediksi tersebut dapat dimanfaatkan oleh dinas kesehatan dan pemerintah daerah sebagai dasar dalam menentukan wilayah prioritas untuk pelaksanaan fogging, pemberantasan sarang nyamuk, penyuluhan kesehatan, distribusi sumber daya kesehatan, serta langkah-langkah pencegahan lainnya. Dengan demikian, penggunaan metode Naïve Bayes dapat membantu meningkatkan efektivitas pengendalian DBD melalui pengambilan keputusan yang lebih cepat, tepat, dan berbasis data.Data historis yang digunakan ditunjukkan pada tabel berikut. Sebagai awalan dibutuhkan data latih sebagai berikut :

Tabel 2. Data latih Prediksi Daerah Terjangkit DBD Dengue

No	Curah Hujan	Kepadatan Penduduk	Sanitasi	Status DBD
1	Tinggi	Tinggi	Buruk	Terjangkit
2	Tinggi	Tinggi	Buruk	Terjangkit
3	Sedang	Tinggi	Buruk	Terjangkit
4	Rendah	Rendah	Baik	Tidak Terjangkit
5	Rendah	Sedang	Baik	Tidak Terjangkit
6	Sedang	Sedang	Baik	Tidak Terjangkit
7	Tinggi	Sedang	Buruk	Terjangkit
8	Sedang	Rendah	Baik	Tidak Terjangkit
9	Tinggi	Tinggi	Buruk	Terjangkit
10	Sedang	Sedang	Baik	Tidak Terjangkit

Dari data tersebut diatas terdapat:

- 5 daerah terjangkit DBD
- 5 daerah tidak terjangkit DBD

Selanjutnya akan dilakukan pengujian terhadap data baru (data uji). Misalkan akan diprediksi suatu daerah dengan karakteristik seperti tampak pada Tabel 3 dibawah ini dengan status daerah yang belum diketahui.

Tabel 3. Data latih (data uji)

Variabel	Nilai
Curah Hujan	Tinggi
Kepadatan Penduduk	Tinggi
Sanitasi	Buruk

Adapun langkah-langkah penyelesaian yang dapat dilakukan adalah :

1. **Menghitung Probabilitas Awal (Prior)**
   Dengan menggunakan P(A|B) = {P(B|A)P(A)}{P(B)} dimana :
   Jumlah data = 10
   Kelas Terjangkit = 5 data
   Kelas Tidak Terjangkit = 5 data
   Maka:
   Probabilitas (Terjangkit) = 5/10 = 0,5
   Probabilitas (Tidak Terjangkit) = 5/10 = 0,5

2. **Menghitung Probabilitas Kondisional**
   Kelas Terjangkit yang di ambil dari 5 data terjangkit :

Tabel 4. Kelas Terjangkit

Variabel	Perhitungan	Nilai
Curah Hujan Tinggi	4/5	0,8
Kepadatan Tinggi	4/5	0,8
Sanitasi Buruk	5/5	1

Untuk Kelas Tidak Terjangkit juga terdiri dari 5 data seperti tampak di table dibawah ini :

Tabel 5. Kelas Tidak Terjangkit

Variabel	Perhitungan	Nilai
Curah Hujan Tinggi	0/5	0
Kepadatan Tinggi	0/5	0
Sanitasi Buruk	0/5	0

3. Menghitung Nilai Naïve Bayes

Untuk kelas Terjangkit:

P(Terjangkit|X)=0,5×0,8×0,8×1 = 0,32

Untuk kelas Tidak Terjangkit:

P(Tidak|X)=0,5×0×0×0 = 0

4. Menentukan Hasil Prediksi

Tabel 6. Hasil Prediksi

Kelas	Nilai Probabilitas
Terjangkit	0,32
Tidak Terjangkit	0

Karena:

0,32>00,32 > 00,32>0

maka daerah tersebut diprediksi: Terjangkit Demam Berdarah Dengue (DBD)

Darai hasil prediksi menunjukkan bahwa daerah dengan curah hujan tinggi, kepadatan penduduk tinggi, dan kondisi sanitasi yang buruk memiliki kecenderungan lebih besar untuk mengalami kasus Demam Berdarah Dengue. Oleh karena itu, wilayah tersebut dapat dijadikan prioritas dalam pelaksanaan program pencegahan, seperti fogging, pemberantasan sarang nyamuk, dan penyuluhan kesehatan kepada masyarakat.

BAB 14
ETIKA DAN KEAMANAN DATA

Data telah menjadi bagian penting dalam kehidupan digital. Hampir setiap aktivitas manusia menghasilkan data, mulai dari media sosial, transaksi elektronik, pembelajaran daring, layanan kesehatan, transportasi digital, perbankan, hingga aplikasi berbasis kecerdasan buatan. Data tidak lagi dipahami sebagai catatan administratif, tetapi sebagai aset digital yang dapat digunakan untuk membaca pola, memperbaiki layanan, menyusun strategi, memprediksi perilaku pengguna, dan mendukung pengambilan keputusan organisasi.

Pada bidang data science, data memiliki nilai besar karena dapat diolah menjadi informasi, pengetahuan, dan dasar pengambilan keputusan. Namun, nilai tersebut harus dikelola secara bertanggung jawab. Data yang dikumpulkan tanpa tujuan jelas, disimpan tanpa perlindungan, atau digunakan tanpa persetujuan dapat menimbulkan pelanggaran privasi, pencurian identitas, manipulasi informasi, diskriminasi berbasis algoritma, kebocoran data, penipuan digital, serta hilangnya kepercayaan publik.

Etika data dan keamanan data perlu dipahami sebagai satu kesatuan. Etika data mengatur penggunaan data agar sah, adil, transparan, dan dapat dipertanggungjawabkan, sedangkan keamanan data melindungi data dari akses tidak sah, perubahan, kehilangan, kerusakan, dan penyalahgunaan. Regulasi pelindungan data pribadi, seperti Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi dan General Data Protection Regulation, menegaskan bahwa pemrosesan data harus

memperhatikan hak subjek data, dasar pemrosesan yang sah, tujuan yang jelas, serta kewajiban pengendali data (European Parliament and Council, 2016; Republik Indonesia, 2022).

A. DATA SEBAGAI ASET DIGITAL

Data disebut sebagai aset digital karena memiliki nilai bagi individu, organisasi, pemerintah, dan masyarakat. Lembaga pendidikan dapat menggunakan data akademik untuk memantau perkembangan mahasiswa, perusahaan dapat memakai data pelanggan untuk memperbaiki layanan, dan pemerintah dapat mengolah data kependudukan untuk meningkatkan pelayanan publik. Dengan pengelolaan yang tepat, data membantu proses kerja menjadi lebih cepat, terukur, dan berbasis bukti.

Meski bernilai, data tidak boleh diperlakukan seperti barang bebas. Setiap data memiliki pemilik, tujuan penggunaan, tingkat sensitivitas, dan risiko. Data pribadi seperti nama, alamat, nomor identitas, lokasi, riwayat kesehatan, transaksi, dan pendidikan perlu dikelola secara hati-hati. Organisasi juga harus mengetahui data apa yang dikumpulkan, alasan pengumpulan, pihak yang berwenang mengakses, masa penyimpanan, tujuan pembagian, serta waktu penghapusan agar kepatuhan, keamanan, dan tanggung jawab penggunaan data tetap terjaga.

B. PRINSIP ETIKA DATA

Etika data berfungsi sebagai pedoman agar data digunakan secara sah, adil, transparan, dan menghormati hak pemilik data. Prinsip ini penting agar pemanfaatan data tidak hanya berorientasi pada hasil analisis, tetapi juga memperhatikan tanggung jawab sosial dan perlindungan individu.

a. Legalitas

   Data harus dikumpulkan dan diproses berdasarkan dasar yang sah. Organisasi tidak boleh mengambil data secara tersembunyi atau menggunakan data untuk tujuan yang tidak pernah dijelaskan kepada pemilik data.

b. Transparansi

   Pemilik data berhak mengetahui jenis data yang dikumpulkan, tujuan penggunaan, pihak yang dapat mengakses data, masa penyimpanan, serta mekanisme pengajuan keberatan. Informasi tersebut perlu disampaikan dengan bahasa yang jelas dan mudah dipahami.

c. Keadilan

   Data tidak boleh digunakan untuk merugikan individu atau kelompok tertentu secara tidak wajar. Pada sistem berbasis algoritma, keadilan penting karena data historis dapat mengandung bias yang memengaruhi hasil analisis atau rekomendasi.

d. Minimisasi Data

   Organisasi sebaiknya hanya mengumpulkan data yang benar-benar dibutuhkan. Pengumpulan data secara berlebihan dapat memperbesar risiko kebocoran, penyalahgunaan, dan pelanggaran privasi.

e. Akuntabilitas

   Organisasi harus mampu menjelaskan proses, keputusan, dan tindakan yang dilakukan terhadap data. Akuntabilitas diperlukan agar pengelolaan data dapat diaudit, dipertanggungjawabkan, dan diperbaiki apabila terjadi kesalahan.

C. PRIVASI DAN PERSETUJUAN

Privasi merupakan hak individu untuk mengendalikan informasi tentang dirinya. Dalam layanan digital, privasi dapat terancam ketika aplikasi mengumpulkan lokasi, kamera, mikrofon, kontak, atau riwayat aktivitas tanpa penjelasan yang memadai.

Persetujuan penggunaan data harus diberikan secara sadar, bebas, spesifik, dan dapat ditarik kembali. Organisasi perlu menjelaskan tujuan, manfaat, risiko, pihak yang terlibat, serta hak pemilik data dengan bahasa yang ringkas dan mudah dipahami agar persetujuan menjadi tindakan bermakna, bukan formalitas administratif.

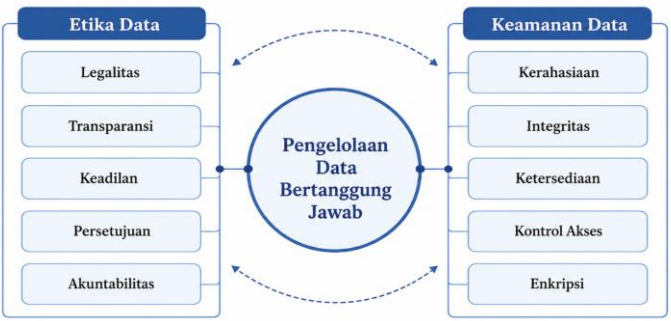
D. KEAMANAN DATA

Keamanan data bertujuan menjaga kerahasiaan, integritas, dan ketersediaan data. Kerahasiaan berarti data hanya dapat diakses pihak berwenang, integritas berarti data tetap akurat dan tidak diubah tanpa izin, sedangkan ketersediaan berarti data dapat digunakan saat dibutuhkan oleh pihak yang berhak (National Institute of Standards and Technology, 2020).

Keamanan data tidak hanya bergantung pada teknologi, tetapi juga pada kebijakan, prosedur, perilaku pengguna, dan pengawasan organisasi. ISO/IEC 27001:2022 menekankan bahwa keamanan informasi perlu dikelola secara berkelanjutan melalui komitmen, pembagian tanggung jawab, pengukuran risiko, dan perbaikan berulang (International Organization for Standardization, 2022).

Etika data dan keamanan data perlu berjalan bersama. Etika data mengarahkan tujuan dan tanggung jawab penggunaan data, sedangkan keamanan data memastikan data terlindungi selama proses pengumpulan, penyimpanan, pemrosesan, dan pembagian.

Hubungan tersebut ditunjukkan pada Gambar 14.1.



Gambar 14.1. Keterpaduan Etika Data dan Keamanan Data

E. ANCAMAN TERHADAP DATA

Ancaman terhadap data dapat berasal dari pihak luar maupun pihak internal organisasi. Ancaman eksternal meliputi phishing, malware, ransomware, dan peretasan sistem. Phishing dilakukan dengan menipu pengguna agar memberikan informasi sensitif, seperti kata sandi, kode verifikasi, atau kredensial akun. Malware dan ransomware dapat merusak sistem, mencuri data, atau mengunci data sehingga tidak dapat diakses. Peretasan sistem dapat menyebabkan data sensitif bocor ke pihak yang tidak berwenang.

Ancaman internal dapat muncul dari pegawai, mitra, atau pihak lain yang memiliki akses sah, tetapi menggunakan data secara keliru. Bentuknya dapat berupa penyalahgunaan akses, pengunduhan data tanpa izin, salah kirim dokumen, atau penyimpanan data sensitif di perangkat pribadi tanpa pengamanan. Ancaman internal sering sulit terdeteksi karena pelaku memiliki akses legal ke sistem. Karena itu, organisasi perlu membatasi akses sesuai kebutuhan kerja, mencatat aktivitas akses, dan melakukan audit berkala.

Ringkasan ancaman terhadap data dan bentuk

pengendaliannya disajikan pada Tabel 14.1.

Tabel 14.1. Ancaman terhadap Data dan Bentuk Pengendaliannya

No	Ancaman Data	Sumber Ancaman	Risiko yang Muncul	Pengendalian yang Disarankan
1	Phishing	Eksternal	Pencurian akun dan kredensial	Edukasi pengguna, autentikasi dua faktor
2	Malware/ Ransomware	Eksternal	Data terkunci, hilang, atau rusak	Antivirus, backup berkala, pembaruan sistem
3	Peretasan sistem	Eksternal	Kebocoran data sensitif	Firewall, enkripsi, audit keamanan
4	Penyalahgunaan akses	Internal	Data digunakan tanpa izin	Pembatasan hak akses dan audit log
5	Kelalaian pengguna	Internal	Salah kirim data atau paparan dokumen	SOP keamanan dan pelatihan rutin

F. TATA KELOLA DATA

Tata kelola data merupakan sistem pengaturan agar data dikelola secara tertib, aman, dan dapat dipertanggungjawabkan. Tata kelola data mencakup kebijakan, struktur tanggung jawab, prosedur operasional, klasifikasi data, manajemen akses, pengawasan, dan respons insiden. Tanpa tata kelola yang baik, organisasi akan sulit mengetahui siapa yang bertanggung jawab

terhadap data, bagaimana data digunakan, serta tindakan apa yang harus dilakukan ketika terjadi insiden.

Langkah awal tata kelola data adalah klasifikasi data. Data perlu dikelompokkan berdasarkan tingkat sensitivitas, misalnya data publik, data internal, data rahasia, dan data sangat sensitif. Setelah data diklasifikasikan, organisasi dapat menentukan perlakuan yang sesuai, termasuk hak akses, metode penyimpanan, enkripsi, dan durasi penyimpanan.

Organisasi juga perlu memiliki prosedur respons insiden. Ketika terjadi kebocoran data atau gangguan keamanan, organisasi harus mampu mendeteksi, melaporkan, menangani, memulihkan, dan mengevaluasi penyebab insiden. Respons yang lambat dapat memperbesar kerugian dan menurunkan kepercayaan pemilik data.

G. ETIKA DATA DALAM KECERDASAN BUATAN

Kecerdasan buatan membutuhkan data dalam jumlah besar untuk melatih model, mengenali pola, dan menghasilkan prediksi. AI bermanfaat bagi pendidikan, bisnis, kesehatan, transportasi, dan pelayanan publik, tetapi juga menimbulkan persoalan etis seperti bias, transparansi, privasi, dan akuntabilitas. OECD menekankan bahwa AI tepercaya perlu menghormati hak asasi manusia, transparansi, keamanan, dan akuntabilitas (OECD, 2016).

Dalam pengembangan AI, data harus sah, relevan, berkualitas, dan bebas dari bias yang merugikan. AI sebaiknya digunakan sebagai alat bantu keputusan, bukan pengganti tanggung jawab profesional, terutama pada seleksi kerja, pinjaman, pendidikan, dan layanan kesehatan.

H. LITERASI KEAMANAN DATA

Literasi keamanan data menjadi kebutuhan bagi individu dan organisasi. Individu perlu memahami cara menjaga akun digital, mengenali tautan mencurigakan, mengatur privasi aplikasi, menggunakan jaringan yang aman, dan tidak membagikan data sensitif secara sembarangan. Kebiasaan sederhana seperti menggunakan kata sandi yang berbeda, mengaktifkan autentikasi dua faktor, dan memeriksa alamat pengirim pesan dapat mengurangi risiko serangan digital.

Organisasi perlu membangun literasi keamanan melalui pelatihan rutin, panduan penggunaan sistem, simulasi insiden, dan saluran pelaporan yang jelas. Pelatihan tidak cukup hanya menjelaskan teori, tetapi perlu memuat contoh kejadian, langkah pencegahan, dan prosedur pelaporan. Semakin baik literasi pengguna, semakin kecil peluang kesalahan yang membuka celah keamanan.

BAB 15
TREN DAN MASA DEPAN DATA SCIENCE

Data Science telah bertransformasi dari sebuah disiplin ilmu yang khusus dan niche menjadi tulang punggung pengambilan keputusan modern dalam kurun waktu kurang dari dua dekade. Dari awal mula yang hanya berfokus pada statistik dan data mining, evolusi teknologi telah membawa kita ke era di mana kecerdasan buatan (Artificial Intelligence/AI) dan pembelajaran mesin (Machine Learning/ML) terintegrasi ke dalam hampir setiap aspek kehidupan manusia. Pada titik ini, pertanyaan yang muncul bukan lagi "Apa itu Data Science?", melainkan "Ke arah mana Data Science akan melangkah?".

Bab ini, sebagai penutup dari buku Bunga Rampai Pengantar Data Science, dirancang untuk membawa pembaca melihat ke cakrawala masa depan. Kita tidak hanya akan membahas tren teknologi terkini yang sedang hangat diperbincangkan, tetapi juga menganalisis bagaimana tren tersebut akan mengubah lanskap industri, etika, dan keterampilan yang dibutuhkan oleh para praktisi data science.

Masa depan Data Science tidak ditentukan oleh satu teknologi tunggal, melainkan oleh konvergensi berbagai inovasi: dari kebangkitan Generative AI yang revolusioner hingga kebutuhan mendesak akan tata kelola data yang etis. Kita akan membedah bagaimana peran Data Scientist berevolusi, bagaimana hambatan teknologi seperti komputasi kuantum dapat mengubah aturan permainan, dan mengapa "demokratisasi data" menjadi kunci bagi organisasi di masa depan.

A. KECERDASAN BUATAN GENERATIF (GENERATIVE AI) DAN LLM

Salah satu perubahan paling signifikan yang terjadi dalam dua tahun terakhir adalah ledakan Kecerdasan Buatan Generatif (Generative AI) (Jurnal et al., 2024). Jika selama ini model Machine Learning tradisional berfokus pada predictive analytics memprediksi hasil berdasarkan data masa lalu Generative AI berfokus pada creative analytics, yaitu kemampuan mesin untuk menciptakan konten baru.

1. Model Bahasa Besar (Large Language Models/LLM)

Di jantung dari revolusi ini terdapat Large Language Models (LLM) seperti GPT (Generative Pre-trained Transformer), BERT, dan LaMDA. Model-model ini dilatih dengan kumpulan data teks yang sangat besar dari internet, memungkinkan mereka untuk memahami, merangkum, menerjemahkan, dan menghasilkan teks yang koheren dengan kualitas yang hampir menyamai manusia (Haromain et al., 2025).

Dampaknya terhadap Data Science sangat fundamental:

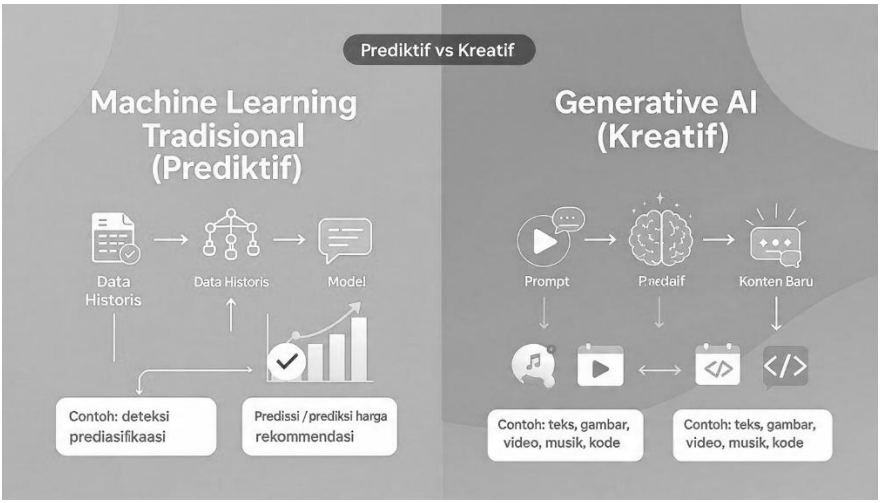
a. Penerjemah Kode: LLM kini berfungsi sebagai asisten pemrograman yang mumpuni. Data Scientist yang sebelumnya terkendala oleh sintaksis pemrograman yang rumit kini dapat menggunakan Natural Language Processing (NLP) untuk menulis kode Python, R, atau SQL (Prasetyo et al., 2021). Hal ini menurunkan hambatan masuk (entry barrier) bagi mereka yang memiliki latar belakang domain knowledge namun lemah dalam coding.

b. Otomatisasi ETL: Proses Extract, Transform, Load (ETL) yang seringkali memakan waktu hingga 70% kerja seorang Data Scientist dapat dipercepat secara drastis dengan bantuan AI yang mampu mendeteksi pola data

yang kotor dan menyarankan fungsi pembersihan (data cleaning) secara otomatis.

2. Generative AI di Luar Teks

Walaupun LLM mendominasi perhatian, tren masa depan juga mencakup generasi media lainnya. Model generatif untuk gambar (seperti Midjourney atau Stable Diffusion) dan video mulai diterapkan dalam konteks bisnis. Di masa depan, Data Science tidak hanya berurusan dengan angka dan teks dalam spreadsheet, tetapi juga dengan analisis pola visual dan generatif aset digital untuk keperluan pemasaran, desain produk, dan simulasi lingkungan.



Gambar 15.1. Perbandingan Machine Learning Tradisional dan Generative AI.

B. DEMOKRATISASI DATA DAN CITIZEN DATA SCIENTIST

Masa depan Data Science bukan lagi tentang dominasi para PhD Statistik atau Software Engineers, tetapi tentang

demokratisasi. Organisasi menyadari bahwa orang-orang yang paling memahami masalah bisnis biasanya bukanlah para ahli data, melainkan para analis bisnis, manajer pemasaran, atau staf operasional (Banerjee, 2018).

1. **Bangkitnya Citizen Data Scientist**

Citizen Data Scientist adalah mereka yang memiliki latar belakang bukan teknis (statistik/komputer) tetapi mampu melakukan analisis data maju menggunakan alat bantu user-friendly (Hidayatullah et al., 2025). Tren ini didorong oleh kemunculan platform AutoML (Automated Machine Learning).

a. AutoML: Alat-alat seperti H2O.ai, DataRobot, atau fitur bawaan di cloud providers memungkinkan pengguna untuk memasukkan data mentah dan mendapatkan model prediktif terbaik hanya dengan beberapa klik tanpa menulis satu baris kode pun. Alat ini secara otomatis melakukan feature engineering, seleksi model, dan hyperparameter tuning.
b. Augmented Analytics: Konsep ini menggunakan AI dan ML untuk meningkatkan cara manusia berinteraksi dengan data. Sistem BI (Business Intelligence) masa depan tidak hanya menampilkan grafik, tetapi akan secara proaktif memberikan wawasan ("Penjualan turun 20% karena faktor X, disarankan untuk melakukan Y").

2. **Tantangan Demokratisasi**

Meskipun menjanjikan, demokratisasi ini membawa risiko baru. Tanpa pemahaman statistik yang mumpuni, Citizen Data Scientist rentan melakukan kesalahan interpretasi, seperti menganggap korelasi sebagai sebab-akibat (spurious correlation) atau mengabaikan bias dalam data.

C. ETIKA DATA, AI YANG DAPAT DIPERCAYA (TRUSTWORTHY AI), DAN REGULASI

Seiring dengan meningkatnya kekuatan Data Science dan AI, meningkat pula kekhawatiran terkait dampak etisnya. Skandal-skandal seperti manipulasi data pemilu, penyalahgunaan data privasi (Cambridge Analytica), dan bias algoritmik dalam perekrutan tenaga kerja telah menjadi peringatan keras.

1. Explainable AI (XAI)

Masa depan Data Science menolak konsep "Kotak Hitam" (Black Box). Regulator dan pengguna menuntut untuk mengetahui mengapa sebuah model mengambil keputusan tertentu.

XAI: Cabang ilmu ini berkembang pesat untuk menciptakan model yang transparan (Sasmita, 2025). Jika sebuah sistem AI menolak pengajuan kredit nasabah, sistem harus bisa menjelaskan faktor apa yang menyebabkan penolakan tersebut (misalnya: pendapatan, riwayat pekerjaan, dll), bukan sekadar memberikan skor probabilitas.

Interpretasi vs Akurasi: Tren ke depan akan menemukan keseimbangan baru. Kadang-kadang model yang sangat akurat seperti (Deep Neural Networks) sulit dijelaskan. Teknik interpretability akan menjadi fitur standar dalam setiap pipeline model Machine Learning, bukan lagi tambahan opsional.

2. Privasi dan Teknik Privasi-Preserving

Regulasi seperti GDPR di Eropa dan UU Pelindungan Data Pribadi (PDP) di Indonesia telah mengubah lanskap Data Science. Mengakses data mentah secara langsung menjadi semakin sulit dan berisiko.

a. Differential Privacy: Teknik ini memungkinkan organisasi

menganalisis data agregat tanpa mengungkapkan informasi tentang individu mana pun. Tambahkan "noise" matematis ke dalam dataset sehingga pola umum tetap terlihat, tetapi data spesifik seseorang tidak dapat dilacak balik.

b. Federated Learning: Ini adalah paradigma di mana model dilatih secara terdesentralisasi di perangkat pengguna (misalnya ponsel pintar) tanpa perlu mengirim data mentah ke server pusat. Google Keyboard menggunakan teknik ini untuk belajar mengetik dari pengguna tanpa pernah melihat pesan pribadi pengguna.

D. KOMPUTASI KUANTUM: ERA PASCA-KLASIK

Mungkin tren yang paling futuristik dan potensialnya paling mengguncang adalah integrasi Data Science dengan Komputasi Kuantum. Komputer klasik berbasis bit (0 atau 1) semakin mendekati batas fisiknya dalam memproses data kompleks.

1. Kecepatan Eksponensial

Komputer kuantum menggunakan qubit yang dapat berada dalam superposisi 0 dan 1 secara bersamaan. Kemampuan ini memungkinkan pemrosesan informasi secara paralel yang jauh melampaui superkomputer saat ini. Bagi Data Science, ini berarti:

a. Optimasi Kompleks: Masalah rute logistik, optimasi portofolio investasi, atau simulasi molekul obat yang membutuhkan waktu bertahun-tahun di komputer klasik dapat diselesaikan dalam hitungan menit atau jam dengan komputer kuantum.

b. Cryptography: Standar keamanan enkripsi saat ini mungkin akan runtuh, memaksa lahirnya "Post-Quantum Cryptography". Data Science akan berperan penting

dalam mengembangkan metode keamanan baru ini (Sinaga & Kasanah, 2023).

2. **Tantangan Implementasi**

Walaupun cukup menjanjikan, komputasi kuantum masih menghadapi masalah stabilitas (dekoherensi) dan suhu operasi yang ekstrem (hampir nol mutlak). Tren masa depan 10-20 tahun ke depan kemungkinan adalah era Quantum Supremacy di mana komputer hibrida (klasik-kuantum) mulai digunakan untuk tugas-tugas spesifik (seperti optimasi Machine Learning) sebelum komputer kuantum penuh tersedia secara komersial.

E. EVOLUSI PERAN DATA SCIENTIST: HUMAN-AI COLLABORATION

Dengan segala otomatisasi yang terjadi, apakah peran Data Scientist akan hilang? Jawabannya adalah tidak, tetapi akan berubah secara radikal.

1. **Dari "Penyusun Kode" menjadi "Arsitek Solusi"**

Di masa depan, Data Scientist tidak akan menghabiskan waktu untuk menulis fungsi regression atau memanipulasi dataframe secara manual. AI akan melakukan itu. Tugas Data Scientist berubah menjadi merancang arsitektur sistem, menentukan masalah bisnis yang ingin diselesaikan, dan memastikan alur data yang benar.

2. **Soft Skills Menjadi Kunci Diferensiasi**

Karena teknis coding menjadi lebih mudah (didemokratisasi), nilai jual utama seorang Data Scientist terletak pada kemampuan non-teknis:

a. Business Acumen: Memahami industri di mana mereka bekerja. Seorang Data Scientist di bidang kesehatan harus

memahami biologi dan regulasi rumah sakit.

b. Storytelling with Data: Kemampuan untuk menterjemahkan hasil kompleks menjadi narasi yang meyakinkan bagi stakeholder bisnis.
c. Critical Thinking: Dalam dunia yang dipenuhi "halusinasi AI" (saat Generative AI memberi jawaban salah namun percaya diri), kemampuan manusia untuk mengkritisi dan memverifikasi hasil menjadi sangat krusial.

F. STUDI KASUS: TREN PENERAPAN DI INDONESIA

Sebagai penutup, mari kita melihat bagaimana tren-tren global ini merambat ke konteks lokal Indonesia:

a. Pertanian Presisi (Precision Agriculture): Dengan potensi agronomi yang besar, Data Science masa depan di Indonesia berfokus pada analisis citra satelit dan sensor tanah IoT untuk memprediksi masa panen, mendeteksi hama, dan mengoptimalkan penggunaan pupuk, mendukung ketahanan pangan nasional.
b. Smart Government: Inisiatif Satu Data Indonesia mendorong pemerintah daerah untuk menerapkan analisis prediktif dalam pengelolaan lalu lintas, bencana alam (menggunakan data sensor cuaca dan historis), dan distribusi bantuan sosial agar lebih tepat sasaran.

Referensi

Agustina, N., Januhari, N. N. U., Jana, P., Suryawan, I. K. D., Tentua, M. N., Lumba, E., Setyoningrum, N. G., Hardyanto, H., Syah, F., & Anwar, N. (2025). Pengantar Data Science.

Batini, C. (2022). Data and information quality. Springer.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.

Abdelaal, M., Hammacher, C., & Schoening, H. (2023). REIN: A Comprehensive Benchmark Framework for Data Cleaning Methods in ML Pipelines.

Agustina, N., Nyoman, N., Januhari, U., Jana, P., Suryawan, I. K. D., Tentua, M. N., Lumba, E., Setyoningrum, N. G., Hardyanto, R. H., Syah, F., Anwar, N., Purwantara, I. M. A., Fairuzabadi, M., & Pengantar, K. (2025). Pengantar Data Science : Teori, Teknik, dan Aplikasinya di Era Digital. Yash Media.

Bakti Siregar, Andi Pujo Rahadi, M. M. M. (2025). Pengantar Statistik untuk Sains Data.

Banerjee, S. (2018). Citizen Data Science for Social Good in Complex Systems. Interdisciplinary Description of Complex Systems, 16(1), 88–91.

Brown, T. B. , et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Bruce, P. C. ., Bruce, Andrew., & Gedeck, Peter. (2020). Practical statistics for data scientists : 50+ essential concepts using R and Python . O'Reilly Media, Inc.

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python (2nd ed.). O'Reilly Media.

Bruce, P., Bruce, A., & Gedeck, P. (2024). Practical statistics for data scientists. O'Reilly Media.

Bruce, P., Bruce, A., & Gedeck, P. (n.d.). Practical Edition Second Statistics for Data Scientists.

Cambell, A. (2025). Data Science for Beginners (Vol. 6, Issue 0).

Cao, L. (2020). Data science: A comprehensive overview. ACM

Computing Surveys, 55(3), 1–42.
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
Chicco, D., Oneto, L., & Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. PLoS Computational Biology, 18(12).
Cielen, D., Meysman, A., & Ali, M. (2021). Introducing data science. Manning Publications.
D, A. N., M, S., Yashaswini, S., & Tilak, G. (2024). Exploratory Data Analysis (EDA) and Data Visualization. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 12(6).
Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business School Press.
Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90(10), 70-76.
Delen, D. (2021). Predictive analytics, data mining and big data. Pearson.
Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73.
Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 226–231. AAAI Press.
Esteva, A. , K. B. , N. R. A. , et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
European Parliament and Council. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with

regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union.
Evans, J. R. (2019). Business analytics: Methods, models, and decisions (3rd ed.). Pearson Education.
Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54.
Febrieta, D., & Fitriani, Y. (n.d.). STATISTIKA DASAR UNTUK PEMULA.
Feng, P., Bi, Z., Wen, Y., Niu, Q., Liu, J., Wang, T., Li, M., Zhang, S., Peng, B., Liu, M., Pan, X., Yin, C. H., Xu, J., Wang, J., Chen, K., Song, X., & Jiang, Z. (2025). Deep Learning and Machine Learning , Advancing Big Data Analytics and Management : Unveiling AI ' s Potential. Arvxiv, Dec 2025.
George, G., Haas, M. R., & Pentland, A. (2021). Big data and management. Academy of Management Journal.
Geron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.
Grus, J. (2019). Data Science from Scratch: First Principles with Python (2nd ed.) (2nd ed.). O'Reilly Media.
Grus, J. (2023). Data science from scratch. O'Reilly Media.
H, I., & Sarker. (2021). Data Science and Analytics An Overview from Data-Driven Smart.pdf. SN, 2, 22.
Han, J., Kamber, M., & Pei, J. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann Publishers.
Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques. Morgan Kaufmann.
Haromain, I., Munir, S., & Rahmah, A. (2025). Analisa Prompt Engineering pada Large Language Model dengan Retrieval. Indonesian Journal Computer Science, 4(2), 144–153.
Hayashi, C. (1998). What is Data Science ? Fundamental Concepts and a Heuristic Example (C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, & Y. Baba (eds.); pp. 40–51). Springer

Japan.
Healy, K. (2018). Data Visualization: A Practical Introduction. Princeton University Press.
Hidayatullah, R. R., Amir, F., & Oriyasmi, F. (2025). Development of Dasawisma Citizen Data Collection Information System Based on Web Framework. 5(January), 292–299.
IBM. (2021, August 17).
International Organization for Standardization. (2022). ISO/IEC 27001:2022: Information security, cybersecurity and privacy protection—Information security management systems—Requirements. ISO.
Isitor, E., & Stanier, C. (2010). Defining big data. Proceedings of the ACM International Conference, 1–5.
Jain, V. K. (2018). Big data science and analytics. In Big Data Science, Analytics, and Machine Learning (pp. 1–32). Khanna Publishers.
James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2022). An introduction to statistical learning : with applications in R. Springer.
Josyula, H. P., Patel, K. P., Bhanushali, A., Landge, S. R., & Mittal, S. (2023). A Review on Data Visualization for Exploratory Data Analysis. Journal of Data Acquisition and Processing, 38(4).
Jurnal, H., Zalencia Surya, A., Fauzi, A., & Prastomo, A. (2024). Jurnal Riset Teknik Komputer (Jurtikom). 2(2), 100–105.
Kadir. (2022). Statistika Terapan (6th ed.).
Kimball, R., & Ross, M. (2021). The data warehouse toolkit. Wiley.
Kitchin, R. (2021). Big data, new epistemologies and paradigm shifts. Big Data & Society.
Kleppmann, M. (2022). Designing data-intensive applications. O'Reilly Media.
Knaflic, C. N. (2019). Storytelling with Data: A Data Visualization Guide for Business Professionals. John Wiley & Sons.
Kotu, V., & Deshpande, B. (2023). Data science: Concepts and practice. Morgan Kaufmann.
Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best

Practices. In AI (Switzerland) (Vol. 6, Number 10). Multidisciplinary Digital Publishing Institute (MDPI).
Kreuzberger, D., Kuhl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. IEEE Access, 11, 31866-31879.
Lalaoui, I. L., El Haji, E., & Kounaidi, M. (2025). The Evolution and Challenges of Real-Time Big Data: A Review. 11.
Laudon, K. C., & Laudon, J. P. (2024). Management information systems: Managing the digital firm. Pearson.
Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), 413–422.
Liu, X., Martin, I. C., & Trovati, M. (2026). Evaluating Supervised Machine Learning Models : Principles , Pitfalls , and Metric Selection. Arvxiv, 15 April 2026, 1–27.
Marr, B. (2021). Big data in practice. Wiley.
Mayernik, M. S. (2023). Data science as an interdiscipline. Data Science Journal, 22.
McKinney, W. (2017). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.
McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter (3rd Editio). O'Reilly Media.
McKinney, Wes. (2022). Python for Data Analysis, 3rd Edition. O'Reilly Media, Inc.
Mumuni, A., & Mumuni, F. (2024). Automated data processing and feature engineering for deep learning and big data applications: a survey.
National Institute of Standards and Technology. (2020). Security and privacy controls for information systems and organizations (NIST Special Publication 800-53 Revision 5). U.S. Department of Commerce.
OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments.
Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., Singhal, A., Tewari, M., & Choi, S. W. (2024). Data

Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care. In JMIR mHealth and uHealth (Vol. 12, p. e59587).
Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1-29.
Plaue, M. (2023). Data science: An introduction to statistics and machine learning. Springer.
Prasetyo, V. R., Benarkah, N., & Chrisintha, V. J. (2021). Implementasi Natural Language Processing Dalam Pembuatan Chatbot. Teknika, 10(2), 114–121.
Provost, F. , & F. T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.
Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.
Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2023). Data science for business. O'Reilly Media.
Putra Hatoguan, I., & Musllim Karo Karo, I. (2025). KAJIAN KONSEPTUAL MATRIKS SEBAGAI STRUKTUR DASAR DALAM ALJABAR LINEAR. JATI (Jurnal Mahasiswa Teknik Informatika), 9(3), 5325–5329.
Qamar, U., & Raza, M. S. (2023). Data science concepts and techniques with applications. Springer.
Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches.

Republik Indonesia. (2019). Peraturan Presiden Nomor 39 Tahun 2019 tentang Satu Data Indonesia. Lembaran Negara Republik Indonesia Tahun 2019 Nomor 112.

Republik Indonesia. (2022). Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi. Lembaran Negara Republik Indonesia Tahun 2022 Nomor 196.

Republik Indonesia. (2022). Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi. Sekretariat Negara Republik Indonesia.

Russell, S. , & N. P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

Saini, P. (2016). Challenges and opportunities with big data. Database Systems Journal, 5(3), 195–201.

Saltz, J., & Stanton, J. (2021). An introduction to data science. SAGE.

Sari, R. T. K., Handajani, E. T. E., Sholihati, I. D., Hayati, N., Ningsih, S., Gunawan, A., Andrianingsih, Fauziah, Sani, A., & Wijaya, Y. F. (n.d.). DATA SCIENCE : KOMSEP, TEKNIK IMPLEMENTASI DALAM ANALISIS MODERN.

Sasmita, T. P. (2025). Analisis Faktor Risiko Penyakit Jantung Menggunakan Explainable Artificial Intelligence (XAI) dengan Metode SHAP. 6(July), 36–46.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503-2511.

Sengupta, S. (2023). The past, the present, and the future of information and data sciences. Data and Information Management, 7(1).

Setiawan, E. (2021). Metode Statistika untuk Ilmu Terapan. Deepublish.

Shahnawaz, M., & Kumar, M. (2025). A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques. In ACM Computing Surveys (Vol. 57, Number 8, pp. 1–33). Association for Computing Machinery.

Sharda, R., Delen, D., & Turban, E. (2024). Business intelligence,

analytics, and data science. Pearson.
Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.
Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. SBC Reviews on Computer Science, 4(1), 28–43.
Sinaga, R., & Kasanah, U. (2023). Penerapan Komputasi Kuantum dalam Optimasi Masalah Kompleks pada Bidang Kriptografi Modern. (November).
Stair, R., & Reynolds, G. (2023). Principles of information systems. Cengage Learning.
Steorts, R. C. (2023). A Primer on the Data Cleaning Pipeline.
Sudaryono. (2021). Statistik I: Statistik Deskriptif untuk Penelitian. Andi.
Sutojo, T., Mulyanto, E., & Suhartono, V. (2011). Kecerdasan Buatan. Andi Offset.
Suyanto. (2018). Machine Learning : Tingkat Dasar dan Lanjut. Informatika.
Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to data mining (2nd ed.). Pearson Education.
Triola, M. F. . (2022a). Elementary statistics. Pearson.
UNESCO. (2023). Guidance for generative AI in education and research. UNESCO Publishing.
United Nations Global Pulse. (2012). Big data for development: Challenges & opportunities (Global Pulse White Paper). United Nations.
Vanderplas, J. T. . (2023). Python data science handbook : essential tools for working with data. O'Reilly Media, Incorporated.
VanderPlas, Jake. (2016). Python Data Science Handbook. O'Reilly Media.
Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3–28
Vaswani, A. , et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
Venkatraman, R., & Venkatraman, S. (2019). Big data

infrastructure, data visualisation and challenges. In Proceedings of the 2019 International Conference on Big Data and Internet of Things (BDIOT 2019) (pp. 13–17). Association for Computing Machinery.
Wang, R., & Yang, Z. (2017). SQL vs NoSQL: A Performance Comparison.
Waru, D., & Astuti, R. W. (2021). Implementasi metode Naive Bayes untuk prediksi daerah terjangkit demam berdarah dengue di Provinsi Jambi. JTIK (Jurnal Teknik Informatika Kaputama), 5(2).
Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science. O'Reilly Media.
Wilke, C. O. (2020). Fundamentals of Data Visualization. O'Reilly Media.
Wilson, A., & Anwar, M. R. (2024). The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing. International Transactions on Artificial Intelligence (ITALIC), 3(1), 97–107.
Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39.
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.
Ye, F., & Ma, F. (2023). Differences and linkages between data science and information science. Data and Information Management, 7(1).
Yukhno, A. (2024). Digital Transformation: Exploring big data Governance in Public Administration. Public Organization Review, 24(1), 335–349.
Zha, D., Pervaiz Bhat, Z., Herng Lai, K., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric Artificial Intelligence: A Survey. Arvxiv, 1(1), 39.
Zhang, S. , Y. L. , S. A. , & T. Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM

Computing Surveys, 52(1), 1–38.
Zhou, Y., Tu, F., Sha, K., Ding, J., & Chen, H. (2024). A Survey on Data Quality Dimensions and Tools for Machine Learning. Arvxiv, Juni 2024, 12.
Zins, C. (2023). Information and data sciences: Context, units of analysis, meaning, and human impact. Data and Information Management, 7(1).

PROFIL PENULIS

Ismail Setiawan, S.Kom., M.Kom., CDS adalah dosen, peneliti, dan praktisi Teknologi Informasi yang memiliki pengalaman lebih dari 10 tahun dalam bidang Artificial Intelligence, Data Science, Software Development, dan Digital Marketing. Saat ini aktif mengajar di Universitas 'Aisyiyah Surakarta serta terlibat dalam berbagai penelitian dan program pengembangan masyarakat berbasis teknologi digital. Beliau juga memiliki sejumlah sertifikasi profesional di bidang Artificial Intelligence, Data Science, Web Development, dan Mobile Programming, serta aktif sebagai Asesor Kompetensi LSP Informatika. Fokus kajian dan penelitiannya meliputi Artificial Intelligence, Machine Learning, Data Mining, Computer Vision, dan Transformasi Digital.

Setiangga Fachrurrozi, S.Kom., M.Kom, akademisi dan praktisi di bidang teknologi informasi dengan bidang peminatan pada data mining, information retrieval, decision support system, e-business, dan networking. Penulis menyelesaikan Program Sarjana di Universitas Semarang dan melanjutkan Program Magister di Universitas Dian Nuswantoro. Saat ini, berperan aktif sebagai IT Staff sekaligus mengajar sebagai dosen tetap di Universitas Maritim AMNI Semarang, dengan kontribusi aktif dalam pengajaran pada program studi Bisnis Digital, Manajemen Pelabuhan dan Logistik Maritim, serta Teknika. Dalam aktivitas akademik, tidak hanya fokus pada pengajaran, tetapi juga pada penelitian dan pengembangan teknologi di era transformasi digital, khususnya teknologi informasi, bisnis digital dan teknologi

kemaritiman. Melalui buku ini, penulis berharap dapat memberikan wawasan, inspirasi, dan kontribusi nyata bagi pengembangan pembelajaran digital serta pemanfaatan teknologi informasi di era modern.

Sigit Setyowibowo, ST., MMSI, Saat ini penulis aktif mengajar di STMIK PPKIA Pradnya Paramita Malang, Prodi Teknologi Informasi. Buku ini adalah salah satu karya dan inshaa allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku Yang ditulis semata-mata untuk berbagi ilmu pengetahuan.

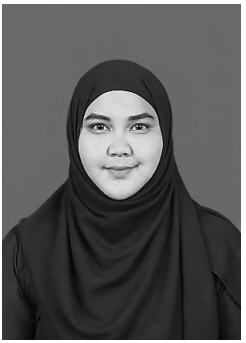
Sukma Puspitorini, S.T., M. Kom. lahir di Blora, 1 April 1982. Beliau merupakan dosen dan peneliti di bidang Teknik Informatika yang telah mengabdikan diri dalam dunia pendidikan tinggi sejak tahun 2006. Saat ini, beliau menjabat sebagai Ketua Program Studi Teknik Informatika Universitas Nurdin Hamzah, serta aktif dalam berbagai kegiatan akademik, penelitian, dan pengembangan keilmuan. Pendidikan Sarjana Teknik Informatika dan Magister Ilmu Komputer beliau selesaikan di Universitas Islam Indonesia, Yogyakarta. Bidang keahlian yang ditekuni adalah Sistem Cerdas (Intelligent Systems) dengan fokus penelitian pada logika fuzzy, Sistem Pendukung Keputusan (Decision Support Systems), serta penerapan Kecerdasan Buatan (Artificial Intelligence) dalam berbagai bidang. Selain menjadi seorang dosen dan peneliti, penulis juga aktif menulis buku, dan serta berperan sebagai reviewer jurnal nasional bereputasi maupun non bereputasi. Beliau dapat dihubungi melalui email: sukmapuspitorini@unh.ac.id.

Fery Purnama, S.Kom., M.Kom. merupakan pendidik/dosen program studi Teknik Informatika Fakultas Ilmu Komputer Universitas Nurdin Hamzah. Lahir di Jambi tanggal 25 September 1989. Memulai pendidikan S1 di STMIK Nurdin Hamzah pada bidang Teknik Informatika pada tahun 2008 kemudian melanjutkan S2 di UPI YTPK Padang, bidang Ilmu Komputer dengan konsentrasi Teknik Informatika tahun 2012. Penulis juga seorang praktisi IT sebagai Web Developer dan saat ini sedang menempuh pendidikan doktoral pada bidang Teknologi Informasi dengan fokus penelitian pada kecerdasan buatan (Artificial Intelligence), pengolahan citra digital, dan pengenalan pola. Minat penelitian penulis meliputi Machine Learning, Deep Learning, Data Science, serta pengembangan sistem berbasis web dan mobile. Selain aktif dalam kegiatan pengajaran, penulis juga terlibat dalam berbagai penelitian dan publikasi ilmiah di bidang teknologi informasi.
Email: ferypurnama@unh.ac.id

Dwi Ismiyana Putri, MMSI. Penulis menyelesaikan pendidikan Sarjana Sistem Informasi di Universitas Sriwijaya pada tahun 2015 dan Magister Manajemen Sistem Informasi dengan konsentrasi Sistem Informasi Bisnis di Universitas Gunadarma pada tahun 2020. Sebelum berkarier di dunia akademik, penulis memiliki pengalaman profesional di industri telekomunikasi hingga menjabat sebagai Floor Manager.
Sejak tahun 2021, penulis aktif sebagai dosen tetap Program Studi Rekayasa Perangkat Lunak, Fakultas Informatika dan Desain, Universitas Bina Insani. Mata kuliah yang diampu antara lain Sistem Informasi Manajemen, Analisis Proses Bisnis, Data Mining, Testing dan Implementasi Sistem, serta Pengantar ICT.
Selain melaksanakan Tri Dharma Perguruan Tinggi, penulis juga

dipercaya sebagai Editor in Chief Jurnal Information System for Educators and Professionals (ISBI) Universitas Bina Insani dan Reviewer Eksternal pada Jurnal Bulletin of Community Service in Information System (BECERIS) Fakultas Ilmu Komputer Universitas Sriwijaya.

M. Fauzan Azhari, S.Stat., M.Kom. adalah seorang dosen di Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Nurdin Hamzah, Jambi. Ia merupakan lulusan Sarjana Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta. Ketertarikan di bidang ilmu sains data membuat ia melanjutkan studi nya di Magister Informatika di Universitas yang sama.
Sejak masa studi hingga saat ini, ia aktif mempelajari berbagai topik yang berkaitan dengan analisis data, sains data, dan computer vision, baik dalam kegiatan akademik maupun penelitian.
Selain itu, ia juga memiliki ketertarikan pada eksplorasi dan visualisasi data. Menurutnya, proses dalam memahami data tidak hanya berkaitan dengan angka dan perhitungan, tetapi juga kemampuan menemukan pola, mengungkap informasi yang tersembunyi, serta menyajikannya dalam bentuk yang mudah dipahami oleh banyak orang.



Cyntia Rivatunisa, S.Kom., M.Kom. merupakan pendidik/dosen program studi Teknik Informatika, Universitas Indonesia Membangun. Lahir di Jambi tanggal 24 Juli 1998. Memulai pendidikan S1 di STMIK Nurdin Hamzah pada bidang Teknik Informatika pada tahun 2016 kemudian melanjutkan S2 di UPI YTPK Padang, bidang Ilmu Komputer dengan konsentrasi Teknik Informatika tahun 2020. Penulis juga seorang Tuton Tutorial Online pada Universitas Terbuka sejak 2025. Minat penelitian penulis meliputi Machine Learning, Deep Learning, Data Science, Selain

aktif dalam kegiatan pengajaran, penulis juga terlibat dalam berbagai penelitian dan publikasi ilmiah di bidang teknologi informasi.
Email: cyntia.rivatunisa@inaba.ac.id

Dr.Sri Handayani, S.T., M.T. Penulis merupakan dosen tetap, dan peneliti dari Universitas Semarang sejak tahun 2007, dengan kompetensi di bidang Jaringan, Hardware komputer, Internet of Thing dan Business Intelligence. Jenjang pendidikan formal yang telah ditempuh penulis menyelesaikan D3 Teknik Telekomunikasi dari Politeknik ITB tahun 1996, menyelesaikan pendidikan Strata-1 Teknik Telekomunikasi dari Sekolah Tinggi Telkom Bandung tahun 2001. Penulis melanjutkan studi Strata-2 di Magister Teknologi Informasi Universitas Gadjah Mada dan lulus tahun 2005. Pada tahun 2025 penulis sudah menyelesaikan studi lanjut Program Doktoral di Doktor Sistem Informasi Sekolah Pasca Sarjana Universitas Diponegoro.

Detin Sofia lahir di Magelang, namun sejak kecil tinggal di Tangerang. Ia tercatat sebagai lulusan Universitas Budi Luhur jurusan Magister Ilmu Komputer. Dosen di bidang Informatika dengan jabatan fungsional saat ini adalah Lektor. Ia aktif menjalankan tridharma melalui pengajaran mata kuliah Pengelolaan Basis Data dan Sistem Pendukung Keputusan, riset di bidang Machine Learning dan Data Mining, serta pengabdian masyarakat yang relevan dengan keahliannya.

Fattachul Huda Aminuddin lahir di Tulungagung, 16 Maret 1993 merupakan akademisi, peneliti dan praktisi di bidang Teknologi Informasi yang saat ini aktif sebagai dosen di Universitas Nurdin Hamzah dan terlibat dalam berbagai kegiatan pengajaran, penelitian, publikasi ilmiah, serta pengembangan solusi berbasis teknologi informasi. Adapun bidang riset yang ditekuni adalah terkait Sistem Multimedia Interaktif, Teknologi Pendidikan berbasis Artificial Intelligence, Computer Vision, Sistem Pendukung Keputusan, dan pengembangan aplikasi berbasis Python dan Streamlit. Penulis juga aktif berperan menjadi narasumber di bidang teknologi AI dan multimedia. Selain menjadi seorang dosen dan peneliti, penulis juga aktif menulis buku, dan serta berperan sebagai reviewer jurnal nasional bereputasi maupun non bereputasi.

Sebagai seorang dosen, penulis memiliki komitmen untuk menghadirkan karya yang tidak hanya berorientasi pada teori, tetapi juga menekankan aspek implementasi yang relevan dengan kebutuhan dunia pendidikan, industri, dan masyarakat. Melalui berbagai publikasi dan karya ilmiah, beliau berupaya menjembatani perkembangan teknologi modern dengan kebutuhan pembelajaran yang mudah dipahami dan aplikatif. Melalui buku ini, penulis berharap dapat memberikan kontribusi dalam meningkatkan literasi Data Science dan Artificial Intelligence di kalangan mahasiswa, dosen, peneliti, maupun praktisi. Dengan pemahaman yang baik terhadap teknologi dan data, diharapkan pembaca mampu mengembangkan solusi inovatif yang bermanfaat serta siap menghadapi tantangan dan peluang di era transformasi digital yang terus berkembang.

Lani Nurlani lahir di Sukabumi pada 3 April 1988. Penulis menempuh pendidikan Diploma III Teknik Komputer di Politeknik Sukabumi, melanjutkan Sarjana Teknik Informatika di STMIK JABAR, dan menyelesaikan Magister Sistem Informasi di STMIK LIKMI. Saat ini penulis aktif sebagai dosen Program Studi D3 Sistem Informasi, Jurusan Teknologi Informasi dan Komputer, Politeknik Negeri Subang. Bidang yang ditekuni adalah sistem informasi, dengan minat riset pada penerapan sistem informasi di bidang kesehatan. Selain melaksanakan Tri Dharma Perguruan Tinggi, penulis juga dipercaya sebagai Editor pada jurnal nasional terakreditasi Sinta 3. Penulis dapat dihubungi melalui surel laninurlani@polteksmi.ac.id.

Reny Wahyuning Astuti, M.Kom. saat ini aktif sebagai dosen program studi Teknik Informatika Fakultas Ilmu Komputer Universitas Nurdin Hamzah. Lahir di Jambi tanggal 16 Mei 1978. Menempuh pendidikan S1 di Universitas DR. Soetomo Suarabaya pada bidang Teknik Informatika pada tahun 1996 kemudian melanjutkan S2 di UPI YTPK Padang pada bidang Ilmu Komputer dengan konsentrasi Teknologi Informasi tahun 2008.
Penulis merupakan dosen yang memiliki minat dan keahlian pada Sistem Basis Data, Sistem Pendukung Keputusan (SPK), dan Data Mining. Dalam kegiatan akademik, Penulis aktif mengampu mata kuliah dan melakukan penelitian serta publikasi ilmiah yang berkaitan dengan pengelolaan data, perancangan basis data, analisis data, serta penerapan metode kecerdasan buatan dan data mining untuk mendukung pengambilan keputusan.

Anita Ratnasari, S.Kom., M.Kom. adalah dosen dan Ketua Program Studi Teknik Informatika, Fakultas Teknik dan Informatika, Universitas Dian Nusantara. Bidang keahliannya meliputi sistem informasi, computer science, data mining, machine learning, text mining, visualisasi data, rekayasa perangkat lunak, dan pengembangan aplikasi berbasis web. Kegiatan penelitian dan publikasinya banyak berfokus pada pengembangan sistem informasi, analisis data, evaluasi layanan pendidikan tinggi, serta penerapan metode komputasi untuk penyelesaian masalah nyata di lingkungan akademik, industri, dan masyarakat.

Miftah Faroq Santoso lahir di Solo pada 28 Desember 1986. Lulus dari SMA Negeri 6 Bekasi pada tahun 2005 dan sempat bekerja beberapa tahun sebelum melanjutkan pendidikan Diploma III Manajemen Informatika di AMIK BSI Bekasi pada tahun 2009 dan lulus pada tahun 2012. Ia kemudian menyelesaikan pendidikan Sarjana Sistem Informasi di Universitas Nusa Mandiri Jakarta pada tahun 2013 serta melanjutkan pendidikan Pascasarjana Program Studi Ilmu Komputer di STMIK Nusa Mandiri Jakarta pada tahun 2017. Saat ini berprofesi sebagai dosen dan praktisi teknologi informasi dengan bidang minat pengembangan web, manajemen proyek, UI/UX, dan digital marketing (SEO). Selain mengajar, ia juga aktif menulis dan membuat konten edukatif seputar teknologi informasi.

PENGANTAR
DATA SCIENCE

KONSEP, METODE, DAN IMPLEMENTASI

Di era transformasi digital, data telah menjadi aset yang sangat berharga. Namun, data hanya akan memberikan manfaat apabila mampu diolah menjadi informasi, pengetahuan, dan solusi yang mendukung pengambilan keputusan. Di sinilah Data Science berperan sebagai disiplin ilmu yang mengintegrasikan statistika, matematika, pemrograman, kecerdasan buatan, dan analisis data untuk menghasilkan wawasan yang bernilai.

Buku Pengantar Data Science: Konsep, Metode, dan Implementasi disusun sebagai panduan komprehensif bagi mahasiswa, dosen, peneliti, maupun praktisi yang ingin memahami dasar-dasar Data Science secara sistematis. Materi disajikan secara bertahap, mulai dari konsep data dan informasi, statistika, pemrograman, basis data, visualisasi data, Machine Learning, Data Mining, Big Data, Deep Learning, hingga penerapan Data Science dalam berbagai bidang serta isu etika dan perkembangan teknologi terkini.

Dengan bahasa yang mudah dipahami, ilustrasi yang informatif, serta contoh implementasi yang relevan, buku ini diharapkan menjadi referensi yang mampu membekali pembaca dengan pemahaman konseptual sekaligus wawasan praktis dalam menghadapi tantangan pengolahan data di era digital.

Mulailah perjalanan Anda memahami Data Science dan jadikan data sebagai dasar dalam menciptakan keputusan yang lebih cerdas, inovatif, dan berdampak.

DITERBITKAN OLEH
PT. SURGA DIGITAL NUSANTARA

Jln. Alam, Bengkalis Kota,
Kec. Bengkalis, Kab. Bengkalis,
Riau

ISBN 978-634-05-4067-3
9 786340 540673