



PEMAHAMAN TENTANG DATA SCIENCE

Konsep, Proses, Manfaat, dan Tantangan

Oleh: Rudy C Tarumingkeng



1 Apa itu Data Science?

“Data Science adalah bidang interdisipliner yang menggabungkan statistika, matematika, komputasi, dan pengetahuan domain untuk mengubah data menjadi informasi, wawasan, prediksi, dan keputusan.”

2 Komponen Inti



Kelima unsur ini bekerja bersama untuk menghasilkan keputusan berbasis bukti.

3 Siklus Kerja Data Science



4 Jenis Analitik



5 Metode & Teknologi



6 Manfaat Utama

- Mendukung pengambilan keputusan
- Memahami perilaku pelanggan
- Meningkatkan efisiensi operasi
- Memprediksi risiko & permintaan
- Mendorong inovasi dan strategi
- Menciptakan nilai bisnis dan sosial

7 Tantangan

- Kualitas data
- Bias algoritmik
- Privasi & keamanan
- Kurangnya literasi data
- Kesulitan implementasi
- Kebutuhan tata kelola dan etika



8 Prinsip Penting



Data Science bukan sekadar teknologi, tetapi proses ilmiah untuk menghasilkan keputusan yang lebih cerdas, adaptif, dan bertanggung jawab.



*Rudy C Tarumingkeng: DATA SCIENCE - Konsep, Metodologi, Teknologi,
Analisis, Tata Kelola, dan Implikasinya bagi Pengambilan Keputusan*

Oleh:

[Prof Ir Rudy C Tarumingkeng, PhD](#)

Professor of Management NUP: 9903252922

Professor Emeritus, IPB-University

Anggota MPR-RI, 1982-1987

Dekan Fak. Kehutanan IPB University. 1967-1968, 1968-1971, 1977-1978

Rektor, Universitas Cenderawasih, Papua (1978-1988, dan

Rektor, Kampus AGRO Manokwari sekarang Universitas Papua Manokwari

Coordinator, CIDA/DIKTI SFU Burnaby BC Canada 1988-1991

Rektor, Universitas Kristen Krida Wacana, Jakarta (1991-2000)

Ketua, Dewan Guru Besar, IPB-University, Bogor (2005-2006)

AI - Data Analyst, dan Ketua, Senat Akademik, IBM-ASMI, Jakarta 2024-

© RudyCT Academic Series

rudyclt75@gmail.com

14 Juli 2026

DATA SCIENCE

Konsep, Metodologi, Teknologi, Analisis, Tata Kelola, dan Implikasinya bagi Pengambilan Keputusan

Abstrak

Data Science merupakan bidang interdisipliner yang mengintegrasikan statistika, matematika, ilmu komputer, teknologi informasi, pengetahuan tentang bidang tertentu, serta kemampuan komunikasi untuk mengubah data menjadi pengetahuan, prediksi, rekomendasi, dan keputusan yang bernilai. Kehadiran Data Science semakin penting karena organisasi modern menghasilkan data dalam jumlah besar melalui transaksi digital, media sosial, perangkat bergerak, sensor, sistem perusahaan, layanan publik, serta interaksi manusia dengan teknologi. Namun, ketersediaan data tidak otomatis menghasilkan pengetahuan. Data harus dikumpulkan, dibersihkan, diintegrasikan, dianalisis, ditafsirkan, serta dikelola secara etis agar dapat memberikan manfaat.

Pembahasan ini menjelaskan Data Science sebagai proses ilmiah dan manajerial, bukan sekadar penggunaan perangkat lunak atau algoritma. Data Science mencakup perumusan masalah, pengumpulan data, eksplorasi, visualisasi, rekayasa fitur, pembangunan model, evaluasi, implementasi, pemantauan, serta penerjemahan hasil analisis ke dalam keputusan organisasi. Di dalamnya terdapat analisis deskriptif, diagnostik, prediktif, preskriptif, dan kausal. Teknik yang digunakan meliputi statistika

klasik, machine learning, deep learning, natural language processing, analisis jaringan, analisis deret waktu, sistem rekomendasi, serta kecerdasan buatan generatif.

Dalam konteks manajemen, Data Science dapat digunakan untuk memahami pelanggan, meningkatkan produktivitas, memprediksi risiko, mengoptimalkan rantai pasok, mengelola sumber daya manusia, mendukung keputusan investasi, memperbaiki pelayanan kesehatan, serta merancang kebijakan publik berbasis bukti. Meskipun demikian, penggunaan Data Science juga menimbulkan tantangan berupa kualitas data, bias algoritmik, privasi, keamanan, transparansi, kesenjangan kompetensi, serta ketergantungan organisasi pada sistem otomatis. Oleh sebab itu, Data Science perlu ditempatkan dalam kerangka tata kelola yang bertanggung jawab, berorientasi manusia, dapat dijelaskan, adil, aman, dan selaras dengan tujuan strategis organisasi.

Tulisan ini menekankan bahwa keberhasilan Data Science tidak terutama ditentukan oleh kecanggihan algoritma. Keberhasilannya bergantung pada ketepatan masalah, kualitas data, kemampuan organisasi, keterlibatan pemangku kepentingan, integritas analisis, serta keberanian manajemen mengubah hasil analisis menjadi tindakan. Data Science pada akhirnya merupakan disiplin untuk belajar secara sistematis dari data demi menghasilkan keputusan yang lebih rasional, adaptif, dan bertanggung jawab.

Kata Kunci: *Data Science, analitik data, big data, statistika, machine learning, kecerdasan buatan, pengambilan keputusan, tata kelola data, visualisasi data, transformasi digital.*

1. Pendahuluan

Perkembangan teknologi digital telah mengubah cara manusia menghasilkan, menyimpan, mengolah, serta memanfaatkan informasi. Hampir setiap aktivitas modern meninggalkan jejak data. Ketika seseorang melakukan pembelian melalui platform perdagangan elektronik, data mengenai produk, harga, waktu pembelian, lokasi, metode pembayaran, dan pola pencarian akan tercatat. Ketika seorang pasien datang ke rumah sakit, berbagai data klinis dan administratif dihasilkan. Ketika sebuah perusahaan mengoperasikan mesin produksi, sensor dapat merekam temperatur, getaran, tekanan, konsumsi energi, kecepatan, dan tingkat kerusakan. Ketika masyarakat menggunakan media sosial, jutaan teks, gambar, video, komentar, serta pola hubungan terbentuk setiap saat.

Data tersebut memiliki potensi besar. Namun, data mentah belum tentu memberikan arti. Sebuah organisasi dapat menyimpan jutaan catatan transaksi, tetapi tetap tidak memahami mengapa pelanggan meninggalkan perusahaan. Sebuah rumah sakit dapat memiliki ribuan data pasien, tetapi belum tentu mampu memprediksi risiko pasien kembali dirawat. Sebuah pemerintah dapat memiliki data kemiskinan, pendidikan, kesehatan, dan kependudukan, tetapi belum tentu mampu mengintegrasikan data tersebut untuk merancang kebijakan yang tepat sasaran.

Dalam kondisi inilah Data Science menjadi penting. Data Science berusaha menjembatani jarak antara data dan keputusan. Data Science tidak hanya bertanya, "Data apa yang tersedia?" tetapi juga, "Masalah apa yang perlu diselesaikan?", "Pengetahuan apa yang dapat diperoleh dari data?", "Seberapa dapat dipercaya hasil analisis tersebut?", serta "Bagaimana hasil analisis dapat diterapkan dalam keputusan nyata?"

Data Science sering dipahami secara terlalu sempit sebagai kemampuan membuat program komputer, menjalankan algoritma machine learning, atau menghasilkan visualisasi yang menarik. Pemahaman tersebut belum lengkap. Pemrograman hanyalah salah satu alat. Algoritma hanyalah salah satu komponen. Visualisasi hanyalah salah satu cara menyampaikan hasil.

Data Science merupakan keseluruhan proses membangun pengetahuan dari data melalui pendekatan ilmiah.

Dalam pendekatan ilmiah, seorang data scientist harus mampu merumuskan pertanyaan, mengembangkan hipotesis, mengumpulkan bukti, memeriksa kualitas bukti, memilih metode analisis, mengevaluasi ketidakpastian, menguji kesimpulan, serta mengomunikasikan hasil secara jujur. Oleh karena itu, Data Science memiliki hubungan erat dengan metode penelitian.

Namun, Data Science juga memiliki orientasi praktis. Hasil analisis tidak berhenti pada publikasi atau penjelasan teoritis. Dalam organisasi, Data Science diharapkan menghasilkan perbaikan keputusan, efisiensi, inovasi, pengurangan risiko, peningkatan pelayanan, serta penciptaan nilai ekonomi dan sosial.

Data Science berkembang karena beberapa perubahan besar. Pertama, volume data meningkat secara drastis. Kedua, biaya penyimpanan data semakin rendah. Ketiga, kemampuan komputasi semakin tinggi. Keempat, algoritma statistik dan machine learning semakin berkembang. Kelima, muncul berbagai platform komputasi awan yang memungkinkan organisasi mengakses sumber daya komputasi dalam skala besar. Keenam, kebutuhan organisasi terhadap keputusan berbasis bukti semakin meningkat.

Walaupun demikian, meningkatnya penggunaan Data Science tidak selalu menghasilkan keputusan yang lebih baik. Organisasi dapat terjebak dalam apa yang disebut *data-rich but insight-poor*, yaitu memiliki banyak data tetapi sedikit wawasan. Organisasi juga dapat mengalami *automation bias*, yakni kecenderungan mempercayai hasil sistem otomatis tanpa pemeriksaan kritis. Selain itu, model yang dibangun dari data masa lalu dapat mempertahankan ketidakadilan masa lalu apabila bias dalam data tidak dikenali.

Karena itu, pemahaman Data Science harus mencakup dimensi teknis, metodologis, organisasi, strategis, sosial, dan etis.

2. Pengertian Data Science

Secara umum, Data Science dapat didefinisikan sebagai bidang interdisipliner yang menggunakan metode ilmiah, statistika, matematika, komputasi, algoritma, serta pengetahuan bidang tertentu untuk memperoleh wawasan dan menghasilkan keputusan dari data.

Definisi tersebut mengandung beberapa unsur utama.

Pertama, Data Science bersifat interdisipliner. Tidak ada satu disiplin ilmu yang sepenuhnya mencakup Data Science. Seorang ahli statistika mungkin sangat kuat dalam inferensi dan pengujian hipotesis, tetapi belum tentu memiliki kemampuan membangun sistem perangkat lunak berskala besar. Seorang ahli komputer mungkin mampu membangun algoritma yang efisien, tetapi belum tentu memahami karakteristik bisnis atau konteks sosial tempat model diterapkan. Seorang manajer mungkin memahami masalah organisasi, tetapi belum tentu memahami keterbatasan analisis statistik. Data Science mempertemukan kompetensi-kompetensi tersebut.

Kedua, Data Science menggunakan data sebagai bahan utama pembelajaran. Data dapat berbentuk angka, kategori, teks, suara, gambar, video, sinyal, lokasi, jaringan hubungan, atau kombinasi berbagai bentuk tersebut.

Ketiga, Data Science berorientasi pada penemuan pola. Pola dapat berupa hubungan, kecenderungan, segmentasi, anomali, urutan, atau struktur tersembunyi dalam data.

Keempat, Data Science memiliki tujuan prediktif. Model dapat digunakan untuk memperkirakan kejadian di masa depan, seperti permintaan pasar, tingkat kegagalan mesin, kemungkinan kredit macet, risiko penyakit, atau kemungkinan pelanggan berhenti menggunakan layanan.

Kelima, Data Science memiliki tujuan preskriptif. Data Science tidak hanya menjelaskan apa yang mungkin terjadi, tetapi juga membantu menentukan tindakan terbaik.

Keenam, Data Science harus mempertimbangkan konteks. Sebuah pola statistik belum tentu memiliki makna substantif. Korelasi antara dua variabel tidak selalu menunjukkan hubungan sebab-akibat. Karena itu, pengetahuan bidang atau *domain knowledge* sangat diperlukan.

Ketujuh, Data Science berorientasi pada komunikasi. Temuan analitis harus diterjemahkan ke dalam bahasa yang dapat dipahami oleh pengambil keputusan. Model yang sangat akurat tetapi tidak dipahami, tidak dipercaya, atau tidak digunakan oleh organisasi memiliki nilai praktis yang terbatas.

Dengan demikian, Data Science dapat dirumuskan sebagai:

Proses sistematis untuk memperoleh, mengelola, menganalisis, memodelkan, menafsirkan, dan mengomunikasikan data agar menghasilkan pengetahuan serta tindakan yang bernilai.

3. Hubungan Data Science dengan Data, Informasi, Pengetahuan, dan Kebijakan

Untuk memahami Data Science secara mendalam, penting membedakan data, informasi, pengetahuan, dan kebijakan.

3.1 Data

Data merupakan representasi fakta, kejadian, objek, atau pengamatan. Data dapat berupa angka penjualan, usia pelanggan, temperatur mesin, hasil pemeriksaan laboratorium, komentar pengguna, atau koordinat lokasi.

Data belum tentu memiliki makna apabila dilepaskan dari konteks. Angka "75", misalnya, dapat berarti usia, nilai ujian, berat badan, jumlah barang, atau suhu.

3.2 Informasi

Informasi merupakan data yang telah diorganisasi dan diberi konteks. Apabila angka "75" dijelaskan sebagai nilai rata-rata ujian sebuah kelas, angka tersebut menjadi lebih bermakna.

Informasi menjawab pertanyaan dasar seperti siapa, apa, kapan, di mana, dan berapa.

3.3 Pengetahuan

Pengetahuan muncul ketika informasi diinterpretasikan untuk memahami hubungan, pola, sebab, atau implikasi. Misalnya, analisis menunjukkan bahwa mahasiswa yang memiliki tingkat kehadiran rendah cenderung memperoleh nilai lebih rendah.

Pengetahuan menjawab pertanyaan bagaimana dan mengapa.

3.4 Kebijakan

Kebijakan berhubungan dengan kemampuan menggunakan pengetahuan secara tepat, etis, dan sesuai konteks. Mengetahui bahwa kehadiran berkorelasi dengan prestasi belum otomatis menentukan tindakan. Perguruan tinggi perlu memahami penyebab ketidakhadiran, kondisi mahasiswa, desain pembelajaran, serta dampak kebijakan yang akan diterapkan.

Data Science membantu proses transformasi dari data menjadi informasi dan pengetahuan. Namun, transformasi menjadi kebijakan membutuhkan pertimbangan manusia, nilai, pengalaman, dan etika.

4. Sejarah Perkembangan Data Science

Akar Data Science berasal dari berbagai disiplin yang telah berkembang selama ratusan tahun.

Statistika berkembang sebagai ilmu untuk mengumpulkan, merangkum, menganalisis, serta menarik kesimpulan dari data. Pada masa awal, statistika banyak digunakan untuk administrasi negara, sensus penduduk, pertanian, kesehatan, dan penelitian ilmiah.

Ilmu komputer kemudian menyediakan kemampuan untuk memproses data secara otomatis. Perkembangan basis data memungkinkan organisasi menyimpan dan mengakses informasi dalam jumlah besar. Sistem informasi manajemen membantu perusahaan mengintegrasikan data transaksi ke dalam laporan manajerial.

Pada dekade-dekade berikutnya, berkembang bidang *data mining*, yaitu proses menemukan pola dalam basis data besar. Data mining menggabungkan statistika, kecerdasan buatan, pengenalan pola, dan manajemen basis data.

Perkembangan internet menghasilkan sumber data baru. Data tidak hanya berasal dari sistem internal organisasi, tetapi juga dari situs web, media sosial, perangkat bergerak, serta interaksi daring.

Selanjutnya muncul istilah *big data* yang menggambarkan data dengan volume, kecepatan, dan keragaman yang sulit ditangani menggunakan metode tradisional. Teknologi komputasi terdistribusi, penyimpanan awan, serta pemrosesan paralel dikembangkan untuk menangani tantangan tersebut.

Machine learning menjadi semakin penting karena memberikan kemampuan kepada komputer untuk mempelajari pola dari data tanpa harus diprogram secara eksplisit untuk setiap situasi. Kemajuan ini diperkuat oleh peningkatan kemampuan komputasi, ketersediaan data, dan perkembangan algoritma.

Deep learning kemudian meningkatkan kemampuan mesin dalam mengolah data kompleks, seperti gambar, suara, dan bahasa alami. Sistem pengenalan wajah, penerjemahan otomatis, asisten virtual, kendaraan otonom, serta model bahasa berskala besar merupakan bagian dari perkembangan tersebut.

Data Science modern merupakan hasil konvergensi berbagai perkembangan tersebut. Ia tidak menggantikan statistika, ilmu komputer, atau penelitian operasional. Sebaliknya, Data Science mengintegrasikan berbagai disiplin untuk menjawab masalah yang semakin kompleks.

5. Komponen Utama Data Science

Data Science dapat dipahami sebagai pertemuan beberapa komponen utama.

5.1 Statistika dan Matematika

Statistika memberikan landasan untuk memahami variasi, ketidakpastian, pengambilan sampel, pengujian hipotesis, estimasi parameter, serta evaluasi model.

Matematika mendukung pemahaman terhadap optimasi, aljabar linear, kalkulus, probabilitas, dan struktur algoritma.

Tanpa pemahaman statistik, seorang analis berisiko menghasilkan kesimpulan yang menyesatkan. Model dapat tampak akurat, tetapi sebenarnya mengalami *overfitting*. Korelasi dapat dianggap sebagai sebab-akibat. Sampel yang tidak representatif dapat digunakan untuk membuat generalisasi yang tidak sah.

5.2 Ilmu Komputer

Ilmu komputer menyediakan kemampuan untuk mengembangkan algoritma, mengelola data, membangun aplikasi, serta mengoperasikan sistem komputasi.

Data Science membutuhkan pemrograman karena data nyata biasanya terlalu besar, rumit, dan tidak terstruktur untuk diproses secara manual. Bahasa seperti Python, R, SQL, Julia, Java, atau Scala dapat digunakan sesuai kebutuhan.

5.3 Manajemen Data

Manajemen data mencakup proses pengumpulan, penyimpanan, pengamanan, integrasi, dokumentasi, dan pemeliharaan data.

Kualitas analisis tidak dapat melebihi kualitas data. Prinsip *garbage in, garbage out* menjelaskan bahwa data buruk akan menghasilkan analisis buruk, sekalipun algoritma yang digunakan sangat canggih.

5.4 Pengetahuan Bidang

Pengetahuan bidang diperlukan untuk memahami makna variabel, proses bisnis, karakteristik industri, serta konsekuensi keputusan.

Dalam bidang kesehatan, misalnya, model harus mempertimbangkan pengetahuan klinis. Dalam keuangan, model harus memahami risiko, regulasi, dan perilaku pasar. Dalam manajemen sumber daya manusia, analisis harus mempertimbangkan budaya organisasi, dinamika tenaga kerja, serta perlindungan hak karyawan.

5.5 Komunikasi

Data Science harus dapat menjelaskan hasil kepada pemangku kepentingan. Komunikasi mencakup visualisasi, penulisan laporan, presentasi, serta kemampuan menyederhanakan konsep tanpa menghilangkan ketepatan.

5.6 Etika dan Tata Kelola

Penggunaan data menyangkut privasi, keamanan, keadilan, transparansi, dan tanggung jawab. Karena itu, kemampuan etis bukan tambahan, melainkan bagian inti Data Science.

6. Jenis-Jenis Data

Data dapat diklasifikasikan berdasarkan berbagai sudut pandang.

6.1 Data Kuantitatif

Data kuantitatif berbentuk angka dan dapat digunakan dalam perhitungan matematis.

Data kuantitatif dapat dibedakan menjadi:

a. Data diskrit

Data diskrit biasanya berupa bilangan bulat dan berasal dari proses penghitungan.

Contohnya:

- Jumlah pelanggan.
- Jumlah produk cacat.
- Jumlah kunjungan pasien.
- Jumlah keluhan.
- Jumlah karyawan.

b. Data kontinu

Data kontinu dapat memiliki nilai dalam suatu rentang.

Contohnya:

- Waktu pelayanan.
- Berat badan.
- Tinggi badan.
- Temperatur.
- Pendapatan.
- Jarak perjalanan.

6.2 Data Kualitatif atau Kategorikal

Data kategorikal menggambarkan kelompok atau karakteristik tertentu.

a. Data nominal

Kategori tidak memiliki urutan.

Contohnya:

- Jenis produk.
- Wilayah pemasaran.
- Departemen.
- Metode pembayaran.

b. Data ordinal

Kategori memiliki urutan.

Contohnya:

- Tingkat kepuasan.
- Tingkat pendidikan.
- Kategori risiko.
- Tingkat prioritas.

6.3 Data Terstruktur

Data terstruktur tersusun dalam format yang jelas, seperti tabel dengan baris dan kolom.

Contohnya adalah basis data transaksi, data akuntansi, atau data inventaris.

6.4 Data Semi-Terstruktur

Data semi-terstruktur tidak sepenuhnya berbentuk tabel, tetapi memiliki elemen organisasi tertentu.

Contohnya adalah dokumen JSON, XML, log sistem, dan surat elektronik.

6.5 Data Tidak Terstruktur

Data tidak terstruktur tidak memiliki format tabular yang tetap.

Contohnya:

- Teks bebas.
- Dokumen.
- Gambar.
- Video.
- Rekaman suara.
- Unggahan media sosial.

6.6 Data Primer dan Sekunder

Data primer dikumpulkan langsung untuk tujuan tertentu, misalnya melalui survei, wawancara, eksperimen, atau observasi.

Data sekunder telah dikumpulkan sebelumnya oleh pihak lain, misalnya data sensus, laporan pemerintah, laporan perusahaan, atau basis data penelitian.

6.7 Data Potong Lintang dan Deret Waktu

Data potong lintang menggambarkan banyak unit pada satu periode tertentu.

Data deret waktu menggambarkan satu atau beberapa variabel sepanjang waktu.

Data panel menggabungkan keduanya, yaitu banyak unit yang diamati dalam beberapa periode.

7. Big Data dan Karakteristiknya

Big data bukan hanya data dalam jumlah besar. Big data menggambarkan kondisi ketika karakteristik data menuntut pendekatan teknologi dan analitis yang berbeda dari sistem konvensional.

Karakteristik big data sering dijelaskan melalui konsep beberapa "V".

7.1 Volume

Volume menunjukkan jumlah data. Organisasi dapat menghasilkan terabyte, petabyte, atau lebih besar melalui transaksi, sensor, video, sistem digital, dan interaksi pelanggan.

7.2 Velocity

Velocity menunjukkan kecepatan data dihasilkan dan harus diproses. Dalam perdagangan saham, sistem pembayaran, pemantauan keamanan, atau kendaraan otonom, keputusan perlu dibuat dalam waktu sangat singkat.

7.3 Variety

Variety menunjukkan keragaman format data, mulai dari tabel, teks, gambar, suara, video, hingga data sensor.

7.4 Veracity

Veracity menunjukkan tingkat keandalan data. Data dapat mengandung kesalahan, ketidaklengkapan, bias, duplikasi, atau ketidakkonsistenan.

7.5 Value

Value menunjukkan nilai yang dapat dihasilkan dari data. Data besar yang tidak mendukung tujuan organisasi hanya menjadi beban penyimpanan dan pengelolaan.

7.6 Variability

Variability menggambarkan perubahan pola, makna, dan distribusi data sepanjang waktu.

Sebagai contoh, perilaku konsumen dapat berubah akibat musim, krisis ekonomi, kebijakan pemerintah, tren sosial, atau inovasi teknologi. Model yang efektif pada masa lalu belum tentu tetap akurat pada masa depan.

Big data menimbulkan peluang sekaligus tantangan. Peluangnya adalah kemampuan melihat pola dalam skala yang sebelumnya tidak mungkin. Tantangannya adalah meningkatnya kompleksitas, biaya, risiko keamanan, dan potensi penyalahgunaan.

8. Siklus Hidup Proyek Data Science

Proyek Data Science sebaiknya tidak dimulai dengan memilih algoritma. Proyek dimulai dengan memahami masalah.

8.1 Perumusan Masalah

Tahap pertama adalah mendefinisikan masalah secara jelas.

Pertanyaan yang perlu dijawab meliputi:

- Apa masalah organisasi yang hendak diselesaikan?
- Siapa pemangku kepentingan?
- Keputusan apa yang akan dipengaruhi?
- Apa indikator keberhasilan?
- Data apa yang tersedia?
- Apa batasan hukum dan etika?
- Apa konsekuensi apabila model salah?

Contoh masalah yang terlalu umum adalah:

“Perusahaan ingin menggunakan AI untuk meningkatkan penjualan.”

Masalah tersebut perlu dirumuskan lebih spesifik, misalnya:

"Perusahaan ingin memprediksi pelanggan yang memiliki kemungkinan tinggi berhenti berlangganan dalam 60 hari mendatang agar tim pelayanan dapat melakukan intervensi."

Perumusan yang jelas menentukan jenis data, metode, dan ukuran evaluasi.

8.2 Pemahaman Konteks Organisasi

Data scientist harus memahami bagaimana organisasi bekerja. Proses bisnis perlu dipetakan. Peran manusia, sistem, aturan, insentif, dan hambatan perlu dikenali.

Model prediksi pelanggan berhenti berlangganan tidak berguna apabila perusahaan tidak memiliki program retensi atau sumber daya untuk menindaklanjuti hasil model.

8.3 Pengumpulan Data

Data dapat berasal dari:

- Basis data internal.
- Sistem transaksi.
- Sistem ERP.
- Sistem CRM.
- Sensor.
- Survei.
- Media sosial.
- Data pemerintah.
- Penyedia data eksternal.
- Eksperimen.
- Observasi.

Pengumpulan data harus memperhatikan legalitas, persetujuan, privasi, dan tujuan penggunaan.

8.4 Pemahaman Data

Pada tahap ini dilakukan pemeriksaan awal:

- Jumlah observasi.
- Jumlah variabel.
- Tipe data.
- Distribusi.
- Nilai hilang.
- Duplikasi.
- Nilai ekstrem.
- Kesalahan pengkodean.
- Hubungan awal antarvariabel.

Pemahaman data sering mengungkap masalah yang sebelumnya tidak terlihat.

8.5 Pembersihan Data

Pembersihan data mencakup:

- Menghapus duplikasi.
- Memperbaiki kesalahan penulisan.
- Menangani nilai hilang.
- Menyesuaikan format tanggal.
- Menyamakan satuan.
- Memeriksa nilai tidak masuk akal.
- Menyelaraskan kategori.

- Mengidentifikasi pencilan.

Pembersihan data dapat menyita sebagian besar waktu proyek karena data organisasi jarang tersedia dalam kondisi sempurna.

8.6 Integrasi Data

Data sering tersebar di berbagai sistem. Data pelanggan dapat berada di sistem penjualan, pelayanan, keuangan, aplikasi, dan media sosial.

Integrasi membutuhkan identitas penghubung yang konsisten. Kesalahan integrasi dapat menyebabkan catatan milik orang berbeda dianggap sebagai satu individu atau sebaliknya.

8.7 Eksplorasi Data

Exploratory Data Analysis bertujuan memahami struktur, pola, dan anomali.

Teknik eksplorasi meliputi:

- Statistik deskriptif.
- Distribusi frekuensi.
- Histogram.
- Boxplot.
- Diagram pencar.
- Korelasi.
- Tabulasi silang.
- Segmentasi awal.
- Analisis tren.

Eksplorasi bukan sekadar membuat grafik. Eksplorasi merupakan proses bertanya kepada data.

8.8 Rekayasa Fitur

Fitur adalah variabel yang digunakan model untuk belajar.

Rekayasa fitur dapat mencakup:

- Mengubah tanggal menjadi hari, bulan, atau musim.
- Menghitung lama pelanggan.
- Menghitung frekuensi transaksi.
- Membuat rasio.
- Menggabungkan beberapa variabel.
- Mengubah teks menjadi representasi numerik.
- Menghasilkan indikator perilaku.
- Melakukan normalisasi atau standarisasi.

Rekayasa fitur sering lebih menentukan keberhasilan model dibandingkan perbedaan kecil antaralgoritma.

8.9 Pemodelan

Model dipilih berdasarkan jenis masalah.

Masalah klasifikasi memprediksi kategori, misalnya pelanggan akan berhenti atau tidak.

Masalah regresi memprediksi nilai numerik, misalnya jumlah permintaan.

Masalah clustering mengelompokkan objek tanpa label sebelumnya.

Masalah deret waktu memprediksi pola berdasarkan urutan waktu.

Masalah rekomendasi menyarankan produk atau tindakan.

8.10 Evaluasi

Model harus dievaluasi menggunakan ukuran yang sesuai.

Evaluasi tidak hanya mempertimbangkan akurasi statistik, tetapi juga:

- Nilai bisnis.
- Stabilitas.

- Keadilan.
- Kemampuan dijelaskan.
- Biaya kesalahan.
- Kemudahan implementasi.
- Kecepatan prediksi.
- Kepatuhan terhadap regulasi.

8.11 Implementasi

Model yang telah dikembangkan perlu diterapkan ke dalam proses organisasi.

Implementasi dapat berbentuk:

- Dashboard.
- Sistem peringatan.
- Aplikasi.
- API.
- Laporan berkala.
- Integrasi dengan sistem operasional.
- Rekomendasi keputusan.

8.12 Pemantauan

Model perlu dipantau karena lingkungan berubah. Distribusi data dapat bergeser. Perilaku pelanggan berubah. Kebijakan berubah. Kompetitor berubah.

Fenomena ketika performa model menurun akibat perubahan data disebut *model drift* atau *concept drift*.

8.13 Perbaikan Berkelanjutan

Data Science merupakan siklus pembelajaran. Hasil implementasi menghasilkan data baru. Data baru digunakan untuk memperbarui model dan memperbaiki keputusan.

9. Analisis Deskriptif, Diagnostik, Prediktif, Preskriptif, dan Kausal

9.1 Analisis Deskriptif

Analisis deskriptif menjawab pertanyaan:

Apa yang telah terjadi?

Contohnya:

- Berapa penjualan bulan lalu?
- Berapa rata-rata waktu pelayanan?
- Berapa tingkat absensi?
- Produk apa yang paling banyak dibeli?

Analisis deskriptif menggunakan tabel, grafik, persentase, rata-rata, median, dan ukuran penyebaran.

9.2 Analisis Diagnostik

Analisis diagnostik menjawab:

Mengapa hal tersebut terjadi?

Contohnya, penjualan menurun karena keterlambatan pengiriman, perubahan harga, promosi kompetitor, atau penurunan daya beli.

Analisis diagnostik dapat menggunakan korelasi, segmentasi, analisis akar masalah, atau perbandingan kelompok.

9.3 Analisis Prediktif

Analisis prediktif menjawab:

Apa yang kemungkinan akan terjadi?

Contohnya:

- Berapa permintaan bulan depan?
- Pelanggan mana yang mungkin berhenti?
- Mesin mana yang berisiko rusak?
- Kredit mana yang berpotensi macet?

Prediksi tidak berarti kepastian. Prediksi merupakan estimasi berdasarkan pola data dan mengandung probabilitas kesalahan.

9.4 Analisis Preskriptif

Analisis preskriptif menjawab:

Tindakan apa yang sebaiknya dilakukan?

Contohnya:

- Berapa persediaan optimal?
- Pelanggan mana yang perlu diberikan promosi?
- Bagaimana jadwal tenaga kerja disusun?
- Rute distribusi mana yang paling efisien?

Analisis preskriptif dapat menggabungkan prediksi, optimasi, simulasi, serta aturan keputusan.

9.5 Analisis Kausal

Analisis kausal menjawab:

Apakah suatu tindakan menyebabkan perubahan tertentu?

Misalnya, apakah program pelatihan benar-benar meningkatkan produktivitas? Apakah promosi meningkatkan penjualan, atau pelanggan yang menerima promosi memang sejak awal memiliki kemungkinan membeli lebih tinggi?

Prediksi dan kausalitas berbeda. Model prediksi dapat sangat akurat tanpa menjelaskan sebab. Sebaliknya, penelitian kausal berusaha mengisolasi pengaruh suatu tindakan.

Metode kausal dapat mencakup eksperimen acak, eksperimen lapangan, *difference-in-differences*, *regression discontinuity*, *instrumental variables*, *matching*, dan pendekatan kausal lainnya.

10. Statistika sebagai Fondasi Data Science

Statistika memberikan kerangka untuk memahami data dan ketidakpastian.

10.1 Populasi dan Sampel

Populasi adalah keseluruhan objek yang menjadi perhatian. Sampel adalah sebagian populasi yang diamati.

Sampel harus dipilih dengan benar agar kesimpulan dapat digeneralisasikan. Sampel yang bias dapat menghasilkan keputusan keliru.

Sebagai contoh, survei kepuasan pelanggan yang hanya dikirim kepada pelanggan aktif mungkin mengabaikan pelanggan yang telah kecewa dan meninggalkan perusahaan.

10.2 Ukuran Pemusatan

Ukuran pemusatan meliputi:

- Mean.
- Median.
- Modus.

Mean sensitif terhadap nilai ekstrem. Median lebih tahan terhadap pencilan.

Dalam data pendapatan yang sangat tidak merata, median sering lebih representatif dibandingkan mean.

10.3 Ukuran Penyebaran

Ukuran penyebaran meliputi:

- Rentang.
- Varians.
- Simpangan baku.
- Interquartile range.
- Koefisien variasi.

Dua kelompok dapat memiliki rata-rata sama tetapi tingkat variasi berbeda. Variasi penting dalam manajemen risiko dan pengendalian kualitas.

10.4 Probabilitas

Probabilitas mengukur kemungkinan suatu kejadian.

Dalam Data Science, probabilitas digunakan untuk:

- Mengukur ketidakpastian.
- Mengembangkan model klasifikasi.
- Mengestimasi risiko.
- Menentukan ambang keputusan.
- Membandingkan alternatif.

10.5 Distribusi

Distribusi menggambarkan pola penyebaran nilai.

Distribusi normal sering digunakan sebagai pendekatan dalam statistika, tetapi data nyata tidak selalu normal. Pendapatan, ukuran transaksi, jumlah pengikut, atau waktu tunggu dapat memiliki distribusi miring.

10.6 Pengujian Hipotesis

Pengujian hipotesis membantu menilai apakah perbedaan atau hubungan dalam sampel cukup kuat untuk dianggap tidak sekadar kebetulan.

Namun, nilai statistik signifikan tidak selalu memiliki makna praktis. Pada sampel sangat besar, perbedaan kecil dapat menjadi signifikan secara statistik tetapi tidak relevan secara manajerial.

10.7 Interval Kepercayaan

Interval kepercayaan memberikan rentang estimasi yang menggambarkan ketidakpastian.

Penyajian interval sering lebih informatif dibandingkan hanya menampilkan satu nilai prediksi.

10.8 Korelasi

Korelasi mengukur hubungan antara dua variabel.

Namun:

Korelasi tidak membuktikan kausalitas.

Hubungan dapat dipengaruhi variabel ketiga, kebetulan, arah sebab yang terbalik, atau kesalahan pengukuran.

11. Machine Learning

Machine learning adalah cabang kecerdasan buatan yang memungkinkan sistem mempelajari pola dari data dan meningkatkan kinerjanya pada tugas tertentu.

11.1 Supervised Learning

Dalam supervised learning, model belajar dari data yang memiliki label.

Contohnya:

- Data kredit dengan label lancar atau macet.

- Data pasien dengan label sembuh atau tidak.
- Data pelanggan dengan label berhenti atau bertahan.
- Data rumah dengan harga jual.

Tujuan model adalah mempelajari hubungan antara fitur dan target.

Algoritma supervised learning

- Linear regression.
- Logistic regression.
- Decision tree.
- Random forest.
- Support vector machine.
- K-nearest neighbors.
- Gradient boosting.
- Neural network.

11.2 Unsupervised Learning

Dalam unsupervised learning, data tidak memiliki label target.

Model digunakan untuk menemukan struktur tersembunyi.

Contohnya:

- Segmentasi pelanggan.
- Pengelompokan wilayah.
- Deteksi pola pembelian.
- Pengurangan dimensi.
- Deteksi anomali.

Algoritma unsupervised learning

- K-means clustering.
- Hierarchical clustering.
- DBSCAN.
- Principal Component Analysis.
- Autoencoder.
- Association rule mining.

11.3 Semi-Supervised Learning

Semi-supervised learning menggunakan sejumlah kecil data berlabel dan sejumlah besar data tanpa label.

Pendekatan ini berguna ketika pelabelan mahal, seperti dalam diagnosis medis, analisis dokumen, atau pengenalan gambar.

11.4 Reinforcement Learning

Reinforcement learning melibatkan agen yang belajar melalui interaksi dengan lingkungan. Agen memperoleh penghargaan atau hukuman berdasarkan tindakannya.

Reinforcement learning dapat digunakan dalam:

- Robotika.
- Kendaraan otonom.
- Pengelolaan energi.
- Penjadwalan.
- Strategi permainan.
- Optimasi keputusan berurutan.

11.5 Deep Learning

Deep learning menggunakan jaringan saraf dengan banyak lapisan.

Deep learning unggul pada data kompleks seperti:

- Gambar.
- Suara.
- Video.
- Bahasa alami.
- Data sensor berskala besar.

Namun, deep learning biasanya membutuhkan data besar, komputasi tinggi, serta mekanisme pengendalian yang kuat.

12. Pemodelan dan Evaluasi Model

Model yang baik tidak sekadar cocok dengan data pelatihan. Model harus mampu bekerja pada data baru.

12.1 Data Pelatihan, Validasi, dan Pengujian

Data biasanya dibagi menjadi:

- Training set.
- Validation set.
- Test set.

Training set digunakan untuk melatih model.

Validation set digunakan untuk memilih parameter dan membandingkan alternatif.

Test set digunakan untuk memberikan evaluasi akhir.

Data pengujian tidak boleh digunakan selama proses pelatihan agar evaluasi tetap objektif.

12.2 Overfitting

Overfitting terjadi ketika model terlalu menyesuaikan diri dengan data pelatihan, termasuk gangguan acak.

Model overfit memiliki performa tinggi pada data pelatihan tetapi buruk pada data baru.

12.3 Underfitting

Underfitting terjadi ketika model terlalu sederhana untuk menangkap pola data.

Model memiliki performa buruk pada data pelatihan maupun data baru.

12.4 Bias-Variance Trade-Off

Model sederhana cenderung memiliki bias tinggi tetapi varians rendah.

Model kompleks cenderung memiliki bias rendah tetapi varians tinggi.

Tujuan pemodelan adalah menemukan keseimbangan.

12.5 Cross-Validation

Cross-validation membagi data menjadi beberapa bagian secara bergantian untuk pelatihan dan validasi.

Metode ini membantu memperoleh estimasi performa yang lebih stabil.

12.6 Evaluasi Klasifikasi

Ukuran evaluasi klasifikasi meliputi:

Akurasi

Proporsi prediksi yang benar.

Akurasi dapat menyesatkan pada data tidak seimbang.

Apabila hanya satu persen transaksi merupakan penipuan, model yang selalu memprediksi "bukan penipuan" memiliki akurasi 99 persen tetapi tidak berguna.

Precision

Proporsi prediksi positif yang benar-benar positif.

Precision penting ketika biaya alarm palsu tinggi.

Recall

Proporsi kasus positif yang berhasil ditemukan.

Recall penting ketika kegagalan mendeteksi kasus positif memiliki konsekuensi besar.

F1-score

Rata-rata harmonik precision dan recall.

ROC-AUC

Mengukur kemampuan model membedakan kelas pada berbagai ambang keputusan.

Confusion matrix

Menampilkan:

- True positive.
- True negative.
- False positive.
- False negative.

12.7 Evaluasi Regresi

Ukuran evaluasi regresi meliputi:

- Mean Absolute Error.
- Mean Squared Error.
- Root Mean Squared Error.
- Mean Absolute Percentage Error.
- R-squared.

Pemilihan ukuran harus mempertimbangkan konteks.

12.8 Biaya Kesalahan

Tidak semua kesalahan memiliki biaya sama.

Dalam diagnosis penyakit berbahaya, gagal mendeteksi pasien berisiko dapat lebih serius daripada memberikan peringatan kepada pasien yang ternyata sehat.

Dalam deteksi penipuan, terlalu banyak alarm palsu dapat memperlambat transaksi yang sah dan mengganggu pelanggan.

Karena itu, ambang klasifikasi harus disesuaikan dengan konsekuensi keputusan.

13. Rekayasa Fitur

Rekayasa fitur merupakan proses mengubah data mentah menjadi representasi yang lebih berguna bagi model.

Misalnya, perusahaan memiliki data tanggal pembelian. Tanggal tersebut dapat diubah menjadi:

- Hari dalam minggu.
- Bulan.
- Musim.
- Jarak waktu sejak pembelian terakhir.
- Frekuensi pembelian.
- Pola pembelian akhir pekan.
- Lama hubungan pelanggan.

Dalam analisis pelanggan, pendekatan RFM sering digunakan:

- *Recency*: seberapa baru transaksi terakhir.

- *Frequency*: seberapa sering pelanggan bertransaksi.
- *Monetary*: seberapa besar nilai transaksi.

Fitur yang baik mencerminkan proses nyata. Karena itu, pengetahuan domain sangat penting.

Namun, rekayasa fitur juga menimbulkan risiko *data leakage*. Data leakage terjadi ketika model secara tidak sengaja menggunakan informasi yang tidak tersedia pada saat prediksi dilakukan.

Sebagai contoh, model memprediksi apakah pasien akan kembali dirawat, tetapi menggunakan variabel yang baru tercatat setelah pasien kembali. Model tampak sangat akurat, tetapi tidak dapat digunakan secara nyata.

14. Natural Language Processing

Natural Language Processing atau NLP merupakan bidang yang memungkinkan komputer mengolah bahasa manusia.

Data teks sangat penting karena sebagian besar pengetahuan organisasi tersimpan dalam bentuk tidak terstruktur, seperti:

- Laporan.
- Surat elektronik.
- Keluhan pelanggan.
- Catatan medis.
- Kontrak.
- Artikel.
- Dokumen kebijakan.
- Transkrip percakapan.
- Media sosial.

14.1 Tugas Dasar NLP

Tugas NLP meliputi:

- Tokenisasi.
- Stemming.
- Lemmatization.
- Part-of-speech tagging.
- Named entity recognition.
- Sentiment analysis.
- Topic modeling.
- Text classification.
- Information extraction.
- Machine translation.
- Question answering.
- Text summarization.

14.2 Analisis Sentimen

Analisis sentimen berusaha mengenali apakah teks memiliki nada positif, negatif, atau netral.

Organisasi dapat menggunakannya untuk menganalisis ulasan pelanggan. Namun, analisis sentimen menghadapi tantangan seperti ironi, konteks budaya, bahasa campuran, dan ambiguitas.

14.3 Topic Modeling

Topic modeling membantu menemukan tema-tema utama dalam kumpulan dokumen.

Misalnya, ribuan keluhan pelanggan dapat dikelompokkan menjadi masalah harga, kualitas, pelayanan, pengiriman, dan aplikasi.

14.4 Model Bahasa

Model bahasa mempelajari pola penggunaan kata dan struktur bahasa.

Perkembangan model bahasa besar memungkinkan sistem:

- Menjawab pertanyaan.
- Merangkum dokumen.
- Membantu penulisan.
- Menghasilkan kode.
- Menganalisis teks.
- Mendukung pencarian informasi.

Namun, model bahasa dapat menghasilkan informasi keliru atau *hallucination*. Model juga dapat mereproduksi bias data pelatihan. Karena itu, hasilnya perlu diverifikasi.

15. Computer Vision

Computer vision memungkinkan komputer menganalisis gambar dan video.

Penerapannya meliputi:

- Pengenalan objek.
- Deteksi wajah.
- Pemeriksaan kualitas produk.
- Analisis citra medis.
- Pemantauan lalu lintas.
- Analisis citra satelit.
- Deteksi kerusakan infrastruktur.

- Pengawasan keselamatan kerja.

Dalam manufaktur, kamera dapat mendeteksi cacat produk lebih cepat. Dalam pertanian, citra drone dapat digunakan untuk memantau kesehatan tanaman. Dalam kehutanan, citra satelit dapat membantu mendeteksi perubahan tutupan lahan.

Meskipun demikian, penggunaan computer vision untuk pengawasan individu menimbulkan risiko privasi dan kebebasan sipil. Akurasi sistem juga dapat berbeda antar kelompok apabila data pelatihan tidak representatif.

16. Analisis Deret Waktu

Analisis deret waktu digunakan ketika data memiliki urutan waktu.

Contohnya:

- Penjualan bulanan.
- Harga saham.
- Konsumsi listrik.
- Jumlah pasien.
- Tingkat inflasi.
- Curah hujan.
- Permintaan transportasi.

Komponen deret waktu dapat meliputi:

- Tren.
- Musiman.
- Siklus.
- Gangguan acak.

Model deret waktu mencakup:

- Moving average.
- Exponential smoothing.
- ARIMA.
- Vector autoregression.
- State-space models.
- Prophet.
- Recurrent neural networks.
- Long Short-Term Memory.
- Transformer untuk deret waktu.

Validasi deret waktu berbeda dari validasi data biasa. Data masa depan tidak boleh digunakan untuk memprediksi masa lalu. Pembagian data harus menghormati urutan waktu.

Peramalan selalu mengandung ketidakpastian. Semakin jauh horizon prediksi, biasanya semakin besar ketidakpastian.

17. Sistem Rekomendasi

Sistem rekomendasi membantu memilih item yang relevan bagi pengguna.

Penerapannya terdapat pada:

- E-commerce.
- Platform video.
- Musik digital.
- Berita.
- Pendidikan.

- Pariwisata.
- Layanan keuangan.

17.1 Content-Based Filtering

Sistem merekomendasikan item yang mirip dengan item yang sebelumnya disukai pengguna.

17.2 Collaborative Filtering

Sistem menggunakan pola perilaku banyak pengguna. Apabila dua pengguna memiliki selera serupa, item yang disukai salah satu pengguna dapat direkomendasikan kepada pengguna lain.

17.3 Hybrid Recommendation

Pendekatan hibrida menggabungkan berbagai metode.

Sistem rekomendasi meningkatkan personalisasi, tetapi dapat menciptakan *filter bubble*. Pengguna terus menerima konten serupa dan semakin sedikit terpapar perspektif berbeda.

18. Visualisasi Data

Visualisasi data merupakan proses menyajikan data secara visual agar pola, hubungan, tren, dan anomali lebih mudah dipahami.

Visualisasi yang baik harus:

- Akurat.
- Sederhana.
- Relevan.
- Tidak menyesatkan.
- Sesuai dengan jenis data.
- Mendukung keputusan.

18.1 Jenis Visualisasi

- Bar chart untuk membandingkan kategori.
- Line chart untuk tren waktu.
- Scatter plot untuk hubungan dua variabel.
- Histogram untuk distribusi.
- Boxplot untuk penyebaran dan pencilan.
- Heatmap untuk intensitas atau korelasi.
- Peta untuk pola geografis.
- Network graph untuk hubungan.
- Dashboard untuk pemantauan indikator.

18.2 Risiko Visualisasi Menyesatkan

Visualisasi dapat menyesatkan melalui:

- Sumbu yang dipotong.
- Skala tidak proporsional.
- Penggunaan grafik tiga dimensi yang tidak perlu.
- Pemilihan warna yang membingungkan.
- Terlalu banyak informasi.
- Pemilihan periode tertentu secara selektif.
- Mengabaikan ketidakpastian.

Visualisasi bukan dekorasi. Visualisasi adalah bagian dari argumentasi ilmiah.

19. Data Storytelling

Data storytelling menggabungkan data, visualisasi, dan narasi.

Data saja tidak selalu menggerakkan keputusan. Manusia memahami informasi melalui cerita, konteks, konflik, dan konsekuensi.

Struktur data storytelling dapat mencakup:

1. Situasi awal.
2. Masalah.
3. Bukti.
4. Penjelasan.
5. Alternatif.
6. Rekomendasi.
7. Dampak yang diharapkan.

Sebagai contoh, laporan tidak cukup mengatakan bahwa tingkat pelanggan berhenti meningkat dari 5 persen menjadi 8 persen. Laporan perlu menjelaskan segmen mana yang paling terdampak, kapan peningkatan terjadi, faktor apa yang berkaitan, berapa kerugian yang ditimbulkan, serta tindakan apa yang perlu dilakukan.

Data storytelling yang baik tidak memanipulasi emosi. Narasi harus tetap setia kepada bukti.

20. Peran Data Scientist

Data scientist memiliki peran yang beragam.

20.1 Perumus Masalah

Data scientist menerjemahkan masalah organisasi ke dalam pertanyaan analitis.

20.2 Peneliti Data

Data scientist mengeksplorasi data, mengembangkan hipotesis, serta menguji pola.

20.3 Pembangun Model

Data scientist mengembangkan model statistik dan machine learning.

20.4 Komunikator

Data scientist menjelaskan hasil kepada manajer, praktisi, regulator, dan masyarakat.

20.5 Penjaga Integritas Analisis

Data scientist harus melaporkan keterbatasan, ketidakpastian, bias, dan risiko.

20.6 Penghubung Antardisiplin

Data scientist bekerja bersama:

- Data engineer.
- Software engineer.
- Business analyst.
- Domain expert.
- Product manager.
- Legal officer.
- Risk manager.
- Cybersecurity specialist.
- Eksekutif.

Karena kompleksitas tersebut, Data Science lebih tepat dipahami sebagai fungsi tim daripada pekerjaan satu orang.

21. Perbedaan Data Scientist, Data Analyst, Data Engineer, dan Profesi Terkait

21.1 Data Analyst

Data analyst biasanya berfokus pada:

- Laporan.
- Dashboard.
- Statistik deskriptif.
- Analisis bisnis.
- SQL.
- Visualisasi.
- Pemantauan KPI.

21.2 Data Scientist

Data scientist lebih banyak terlibat dalam:

- Pemodelan statistik.
- Machine learning.
- Eksperimen.
- Prediksi.
- Rekayasa fitur.
- Analisis kompleks.

21.3 Data Engineer

Data engineer membangun infrastruktur data:

- Data pipeline.
- Data warehouse.

- Data lake.
- Integrasi sistem.
- Pemrosesan data.
- Reliabilitas dan skalabilitas.

21.4 Machine Learning Engineer

Machine learning engineer berfokus pada implementasi model ke dalam sistem produksi.

21.5 Business Intelligence Specialist

Spesialis BI mengembangkan laporan, dashboard, serta sistem pendukung keputusan.

21.6 Data Architect

Data architect merancang struktur, standar, dan arsitektur data organisasi.

21.7 Data Steward

Data steward bertanggung jawab atas definisi, kualitas, kepemilikan, dan penggunaan data.

Dalam organisasi kecil, beberapa peran dapat dilakukan oleh orang yang sama. Dalam organisasi besar, fungsi-fungsi tersebut biasanya lebih terspesialisasi.

22. Infrastruktur Teknologi Data Science

22.1 Basis Data Relasional

Basis data relasional menyimpan data dalam tabel yang saling berhubungan.

SQL digunakan untuk mengambil, menggabungkan, menyaring, dan mengolah data.

22.2 NoSQL

NoSQL digunakan untuk data yang membutuhkan fleksibilitas struktur atau skala besar.

Jenisnya meliputi:

- Document database.
- Key-value database.
- Column-family database.
- Graph database.

22.3 Data Warehouse

Data warehouse menyimpan data terintegrasi untuk analisis dan pelaporan.

22.4 Data Lake

Data lake menyimpan data mentah dalam berbagai format.

Data lake memberikan fleksibilitas, tetapi tanpa tata kelola dapat berubah menjadi *data swamp*, yaitu kumpulan data yang sulit dipahami dan digunakan.

22.5 Data Lakehouse

Data lakehouse berusaha menggabungkan fleksibilitas data lake dengan keandalan data warehouse.

22.6 Cloud Computing

Komputasi awan memungkinkan organisasi menggunakan:

- Penyimpanan.
- Komputasi.
- Basis data.
- Machine learning platform.

- Sistem analitik.

Layanan awan memberikan skalabilitas, tetapi juga menimbulkan isu biaya, ketergantungan penyedia, keamanan, dan lokasi data.

22.7 Edge Computing

Edge computing memproses data lebih dekat dengan sumbernya.

Pendekatan ini penting ketika:

- Respons harus cepat.
- Koneksi terbatas.
- Privasi tinggi.
- Volume data terlalu besar untuk dikirim seluruhnya.

Contohnya adalah kendaraan, mesin pabrik, kamera, dan perangkat kesehatan.

23. MLOps dan Operasionalisasi Model

MLOps merupakan praktik untuk mengelola siklus hidup machine learning secara sistematis.

MLOps menggabungkan:

- Pengembangan model.
- Rekayasa perangkat lunak.
- Operasi teknologi.
- Tata kelola.
- Pemantauan.

Komponen MLOps meliputi:

- Version control.

- Data versioning.
- Experiment tracking.
- Model registry.
- Automated testing.
- Deployment.
- Monitoring.
- Retraining.
- Audit trail.

Tanpa MLOps, organisasi dapat mengalami kesulitan:

- Tidak mengetahui model mana yang sedang digunakan.
- Tidak dapat mereproduksi hasil.
- Tidak mengetahui data pelatihan.
- Tidak mengetahui kapan performa menurun.
- Sulit memperbarui model.
- Tidak memiliki dokumentasi keputusan.

MLOps mengubah model dari eksperimen menjadi aset organisasi yang dapat dipelihara.

24. Kualitas Data

Kualitas data mencakup beberapa dimensi.

24.1 Accuracy

Data mencerminkan kondisi sebenarnya.

24.2 Completeness

Data memiliki elemen yang diperlukan.

24.3 Consistency

Data tidak bertentangan antar sistem.

24.4 Timeliness

Data tersedia pada waktu yang dibutuhkan.

24.5 Validity

Data mengikuti format dan aturan yang ditetapkan.

24.6 Uniqueness

Tidak terdapat duplikasi yang tidak semestinya.

24.7 Relevance

Data sesuai dengan masalah yang dianalisis.

Kualitas data harus dikelola sejak sumbernya. Membersihkan data pada akhir proses tidak cukup apabila proses organisasi terus menghasilkan kesalahan.

25. Bias dalam Data Science

Bias dapat muncul pada berbagai tahap.

25.1 Bias Pengumpulan

Data tidak mewakili populasi.

25.2 Bias Historis

Data mencerminkan ketidakadilan atau pola diskriminasi masa lalu.

25.3 Bias Pengukuran

Variabel tidak mengukur konsep secara tepat.

25.4 Bias Pelabelan

Label yang digunakan mengandung penilaian subjektif atau kesalahan.

25.5 Bias Seleksi

Kasus tertentu lebih mungkin masuk ke dalam data.

25.6 Bias Algoritmik

Pilihan algoritma, fitur, target, atau ambang dapat menghasilkan dampak berbeda antar kelompok.

25.7 Bias Konfirmasi

Analisis mencari pola yang mendukung keyakinan awal.

25.8 Survivorship Bias

Analisis hanya melihat unit yang berhasil bertahan dan mengabaikan yang gagal.

Sebagai contoh, penelitian tentang karakteristik perusahaan sukses yang hanya menganalisis perusahaan yang masih beroperasi dapat menghasilkan kesimpulan yang keliru karena perusahaan dengan strategi serupa tetapi gagal tidak dimasukkan.

26. Etika Data Science

Etika Data Science membahas pertanyaan mengenai apa yang seharusnya dilakukan, bukan hanya apa yang dapat dilakukan.

26.1 Privasi

Individu berhak mengetahui bagaimana data dikumpulkan, digunakan, disimpan, dan dibagikan.

Pengumpulan data harus mempertimbangkan:

- Persetujuan.
- Tujuan penggunaan.

- Minimalisasi data.
- Waktu penyimpanan.
- Hak mengakses.
- Hak memperbaiki.
- Hak menghapus sesuai ketentuan.

26.2 Keadilan

Sistem tidak boleh memberikan perlakuan tidak adil tanpa alasan yang sah. Namun, definisi keadilan tidak selalu sederhana. Beberapa ukuran keadilan statistik dapat saling bertentangan. Karena itu, penilaian etis dan konteks sosial diperlukan.

26.3 Transparansi

Pemangku kepentingan perlu memahami bahwa keputusan didukung oleh algoritma, data apa yang digunakan, serta bagaimana keputusan dapat dipertanyakan.

26.4 Explainability

Explainability adalah kemampuan menjelaskan alasan di balik prediksi.

Dalam keputusan berisiko tinggi, seperti kredit, kesehatan, pendidikan, atau pekerjaan, penjelasan menjadi sangat penting.

26.5 Akuntabilitas

Harus jelas siapa yang bertanggung jawab ketika sistem menimbulkan kerugian.

Algoritma tidak dapat dijadikan alasan untuk menghindari tanggung jawab manusia.

26.6 Keamanan

Data dan model harus dilindungi dari akses tidak sah, manipulasi, pencurian, serta serangan.

26.7 Human Oversight

Keputusan penting sebaiknya tetap menyediakan mekanisme pengawasan manusia.

Namun, pengawasan manusia harus nyata, bukan sekadar formalitas. Manusia harus memiliki kompetensi, informasi, dan kewenangan untuk menolak rekomendasi sistem.

27. Explainable AI

Explainable AI atau XAI berusaha membuat hasil model dapat dipahami.

Model sederhana seperti regresi atau pohon keputusan relatif mudah dijelaskan. Model kompleks seperti ensemble dan deep neural network lebih sulit diinterpretasikan.

Pendekatan penjelasan dapat dibedakan menjadi:

27.1 Global Explanation

Menjelaskan cara model bekerja secara keseluruhan.

27.2 Local Explanation

Menjelaskan mengapa model memberikan prediksi tertentu untuk satu kasus.

27.3 Feature Importance

Menunjukkan fitur yang paling memengaruhi prediksi.

27.4 Counterfactual Explanation

Menjelaskan perubahan apa yang diperlukan agar hasil prediksi berubah.

Contohnya:

“Permohonan kredit belum memenuhi kriteria. Apabila rasio utang turun dan riwayat pembayaran membaik, hasil penilaian dapat berubah.”

Penjelasan harus disesuaikan dengan audiens. Penjelasan untuk data scientist berbeda dari penjelasan untuk pelanggan atau regulator.

28. Data Governance

Data governance merupakan kerangka untuk memastikan data dikelola secara konsisten, aman, berkualitas, dan bertanggung jawab.

Data governance mencakup:

- Kebijakan.
- Standar.
- Peran.
- Kepemilikan data.
- Kualitas data.
- Metadata.
- Hak akses.
- Keamanan.
- Retensi.
- Audit.
- Kepatuhan.

28.1 Data Ownership

Organisasi harus menentukan siapa yang bertanggung jawab atas domain data tertentu.

28.2 Data Stewardship

Data steward memastikan definisi, kualitas, dan penggunaan data tetap konsisten.

28.3 Metadata Management

Metadata menjelaskan arti, sumber, format, pemilik, serta riwayat data.

28.4 Data Lineage

Data lineage mencatat perjalanan data dari sumber hingga laporan atau model.

Tanpa lineage, organisasi sulit mengetahui asal angka dalam dashboard atau keputusan model.

28.5 Access Control

Akses harus diberikan berdasarkan kebutuhan.

Prinsip *least privilege* berarti pengguna hanya memperoleh akses minimum yang diperlukan.

29. Data Science dalam Manajemen Strategis

Data Science membantu manajemen strategis melalui:

- Analisis lingkungan.
- Deteksi tren.
- Pemantauan kompetitor.
- Segmentasi pasar.
- Simulasi skenario.
- Peramalan.
- Analisis risiko.
- Evaluasi strategi.

Namun, strategi tidak dapat sepenuhnya diotomatisasi. Data masa lalu tidak selalu mampu memprediksi disrupsi, perubahan regulasi, inovasi radikal, atau krisis geopolitik.

Data Science sebaiknya digunakan untuk memperkaya penilaian strategis, bukan menggantikan seluruh proses penalaran manajemen.

Contohnya, perusahaan dapat menggunakan data untuk memprediksi pertumbuhan pasar kendaraan listrik. Namun, keputusan investasi juga memerlukan pertimbangan regulasi, infrastruktur, geopolitik mineral, teknologi baterai, serta perilaku konsumen.

30. Data Science dalam Pemasaran

Pemasaran merupakan salah satu bidang yang paling luas menggunakan Data Science.

Penerapannya meliputi:

- Segmentasi pelanggan.
- Prediksi churn.
- Customer lifetime value.
- Personalisasi.
- Analisis kampanye.
- Market basket analysis.
- Pricing.
- Sentiment analysis.
- Marketing attribution.
- Recommendation system.

30.1 Customer Lifetime Value

Customer Lifetime Value memperkirakan nilai ekonomi pelanggan sepanjang hubungan dengan perusahaan.

Perusahaan dapat membedakan strategi untuk pelanggan bernilai tinggi, pelanggan potensial, dan pelanggan berisiko meninggalkan perusahaan.

30.2 Churn Prediction

Model churn mengidentifikasi pelanggan yang kemungkinan berhenti menggunakan layanan.

Intervensi dapat berupa:

- Perbaikan pelayanan.
- Penawaran khusus.
- Edukasi.
- Kontak personal.
- Penyelesaian keluhan.

Namun, perusahaan harus menghindari perlakuan yang tidak adil. Pelanggan loyal tidak seharusnya selalu menerima penawaran lebih buruk dibandingkan pelanggan yang mengancam berhenti.

30.3 Dynamic Pricing

Harga dapat disesuaikan berdasarkan permintaan, waktu, kapasitas, dan kondisi pasar.

Dynamic pricing dapat meningkatkan efisiensi, tetapi dapat dianggap tidak adil apabila tidak transparan atau mengeksploitasi kondisi darurat.

31. Data Science dalam Keuangan

Penerapan Data Science dalam keuangan meliputi:

- Credit scoring.

- Deteksi penipuan.
- Manajemen risiko.
- Peramalan arus kas.
- Optimasi portofolio.
- Analisis pasar.
- Anti-money laundering.
- Audit berbasis data.

31.1 Credit Scoring

Model menilai kemungkinan peminjam memenuhi kewajiban.

Model harus memperhatikan keadilan, transparansi, serta perlindungan data. Variabel yang tampaknya netral dapat menjadi proksi bagi karakteristik sensitif.

31.2 Fraud Detection

Penipuan biasanya merupakan kejadian langka dan terus berubah.

Model harus mampu mendeteksi pola baru dan mengurangi alarm palsu.

31.3 Audit Analytics

Auditor dapat menganalisis seluruh populasi transaksi, bukan hanya sampel kecil.

Teknik deteksi anomali dapat menemukan transaksi tidak biasa, perubahan pola, atau hubungan yang mencurigakan.

32. Data Science dalam Manajemen Sumber Daya Manusia

People analytics menggunakan data untuk memahami tenaga kerja.

Penerapannya meliputi:

- Perencanaan tenaga kerja.
- Prediksi turnover.
- Evaluasi rekrutmen.
- Analisis keterampilan.
- Produktivitas.
- Keterlibatan.
- Absensi.
- Pengembangan karier.

Namun, analisis sumber daya manusia merupakan bidang sensitif.

Organisasi harus menghindari:

- Pengawasan berlebihan.
- Penilaian yang tidak transparan.
- Diskriminasi.
- Penggunaan data di luar tujuan.
- Mengurangi manusia menjadi sekadar skor.

Model turnover, misalnya, seharusnya digunakan untuk memperbaiki lingkungan kerja, bukan menghukum karyawan yang dianggap berisiko keluar.

33. Data Science dalam Operasi dan Rantai Pasok

Data Science dapat meningkatkan:

- Peramalan permintaan.
- Manajemen persediaan.
- Penjadwalan.

- Pemilihan pemasok.
- Optimasi rute.
- Pemeliharaan prediktif.
- Pengendalian kualitas.
- Manajemen kapasitas.

33.1 Demand Forecasting

Peramalan permintaan membantu menentukan produksi dan persediaan.

Kesalahan prediksi dapat menyebabkan:

- Kehabisan stok.
- Persediaan berlebih.
- Biaya penyimpanan.
- Kehilangan penjualan.
- Pemborosan.

33.2 Predictive Maintenance

Sensor digunakan untuk memprediksi kerusakan mesin.

Perawatan dapat dilakukan sebelum kegagalan terjadi, tetapi tidak terlalu dini.

33.3 Supply Chain Risk

Data dapat digunakan untuk memantau:

- Keterlambatan.
- Risiko pemasok.
- Cuaca.
- Konflik.

- Perubahan harga.
- Hambatan logistik.

Ketahanan rantai pasok memerlukan kombinasi efisiensi dan redundansi. Sistem yang sangat efisien tetapi tidak memiliki cadangan dapat rapuh terhadap gangguan.

34. Data Science dalam Pelayanan Kesehatan

Penerapan Data Science dalam kesehatan meliputi:

- Prediksi risiko.
- Diagnosis berbantuan.
- Analisis citra medis.
- Penemuan obat.
- Pengelolaan kapasitas rumah sakit.
- Pemantauan pasien.
- Surveilans penyakit.
- Personalisasi terapi.

Namun, kesehatan merupakan bidang berisiko tinggi. Kesalahan model dapat memengaruhi keselamatan pasien.

Data kesehatan juga sangat sensitif. Penggunaannya harus memperhatikan privasi, persetujuan, keamanan, dan tanggung jawab profesional.

Model tidak seharusnya menggantikan dokter secara mutlak. Model berperan sebagai alat pendukung keputusan.

35. Data Science dalam Pendidikan

Penerapan Data Science dalam pendidikan meliputi:

- Learning analytics.
- Prediksi risiko putus studi.
- Personalisasi pembelajaran.
- Evaluasi kurikulum.
- Analisis keterlibatan.
- Pengelolaan sumber daya.
- Sistem rekomendasi materi.
- Penilaian berbantuan teknologi.

Contohnya, perguruan tinggi dapat mengembangkan sistem peringatan dini bagi mahasiswa berisiko mengalami kesulitan.

Namun, prediksi tidak boleh menjadi stigma. Mahasiswa yang memperoleh skor risiko tinggi harus menerima dukungan, bukan diskriminasi.

Data Science dalam pendidikan harus mempertahankan hubungan manusia, kebebasan akademik, serta pemahaman bahwa keberhasilan belajar tidak sepenuhnya dapat direduksi menjadi data digital.

36. Data Science dalam Pemerintahan dan Kebijakan Publik

Pemerintah dapat menggunakan Data Science untuk:

- Perencanaan pembangunan.
- Penargetan bantuan.
- Pengendalian lalu lintas.
- Kesehatan masyarakat.
- Pengawasan lingkungan.
- Deteksi penyimpangan.

- Manajemen bencana.
- Pelayanan publik.

Namun, penggunaan data pemerintah menimbulkan persoalan serius mengenai pengawasan, hak warga negara, transparansi, dan keadilan.

Sistem penargetan bantuan, misalnya, dapat meningkatkan ketepatan. Tetapi apabila data tidak mutakhir, warga yang berhak dapat terlewat. Karena itu, harus tersedia mekanisme pengaduan dan koreksi.

Kebijakan berbasis data bukan berarti kebijakan ditentukan sepenuhnya oleh data. Kebijakan juga melibatkan nilai, prioritas, distribusi sumber daya, serta pilihan politik.

37. Data Science dalam Pertanian, Kehutanan, dan Lingkungan

Data Science dapat mendukung pengelolaan sumber daya alam.

Penerapannya meliputi:

- Pemantauan tutupan lahan.
- Prediksi hasil pertanian.
- Deteksi kebakaran.
- Analisis kualitas tanah.
- Pemantauan keanekaragaman hayati.
- Optimasi irigasi.
- Pemantauan hama.
- Analisis perubahan iklim.
- Pengawasan deforestasi.

Data satelit, drone, sensor, dan informasi geografis memungkinkan pengamatan dalam skala luas.

Namun, data harus diintegrasikan dengan pengetahuan lokal. Masyarakat lokal memahami kondisi ekologis dan sosial yang tidak selalu terlihat dalam citra satelit.

38. Kasus Ilustratif: Prediksi Pelanggan Berhenti Berlangganan

Sebuah perusahaan telekomunikasi menghadapi peningkatan pelanggan yang berhenti menggunakan layanan.

Tahap pertama adalah mendefinisikan churn. Apakah pelanggan dianggap churn setelah tidak aktif 30 hari, 60 hari, atau setelah secara resmi menutup akun?

Perusahaan kemudian mengumpulkan data:

- Lama menjadi pelanggan.
- Jenis paket.
- Frekuensi keluhan.
- Gangguan layanan.
- Riwayat pembayaran.
- Penggunaan data.
- Interaksi pelayanan.
- Perubahan harga.
- Respons terhadap promosi.

Data dibersihkan dan diintegrasikan. Analisis eksplorasi menunjukkan bahwa churn lebih tinggi pada pelanggan baru, pelanggan yang mengalami gangguan berulang, dan pelanggan yang melakukan beberapa keluhan tanpa penyelesaian.

Model klasifikasi kemudian dibangun. Hasil model memberikan probabilitas churn bagi setiap pelanggan.

Namun, proyek belum selesai. Perusahaan harus menentukan tindakan.

Pelanggan berisiko tinggi dapat menerima:

- Pemeriksaan kualitas layanan.
- Penyelesaian keluhan.
- Penawaran paket lebih sesuai.
- Komunikasi proaktif.

Evaluasi akhir tidak hanya mengukur akurasi model. Evaluasi harus menilai apakah intervensi benar-benar mengurangi churn dan apakah manfaatnya melebihi biaya.

Perusahaan juga harus menghindari penggunaan data yang tidak relevan atau melanggar privasi.

Kasus ini menunjukkan bahwa nilai Data Science berasal dari keseluruhan sistem keputusan, bukan hanya dari model.

39. Kasus Ilustratif: Prediksi Risiko Pasien Kembali Dirawat

Sebuah rumah sakit ingin menurunkan angka pasien kembali dirawat dalam 30 hari.

Data yang digunakan dapat meliputi:

- Diagnosis.
- Usia.
- Riwayat penyakit.
- Lama rawat.
- Hasil laboratorium.
- Obat.
- Kunjungan sebelumnya.

- Kondisi sosial.
- Kepatuhan terhadap kontrol.

Model memprediksi pasien berisiko tinggi.

Namun, prediksi harus dihubungkan dengan intervensi:

- Konsultasi tambahan.
- Edukasi pasien.
- Pengaturan kontrol.
- Pemantauan jarak jauh.
- Koordinasi dengan keluarga.
- Rekonsiliasi obat.

Model tidak boleh digunakan untuk menolak pasien atau membatasi pelayanan. Tujuannya adalah memberikan dukungan tambahan.

40. Kasus Ilustratif: Pemeliharaan Prediktif

Sebuah perusahaan manufaktur menggunakan mesin dengan biaya kerusakan sangat tinggi.

Sensor merekam:

- Getaran.
- Temperatur.
- Tekanan.
- Kecepatan.
- Konsumsi daya.
- Suara.

Data historis kerusakan digunakan untuk membangun model.

Model mendeteksi pola yang biasanya muncul beberapa hari sebelum kegagalan. Sistem kemudian memberikan peringatan.

Manfaatnya adalah:

- Mengurangi waktu berhenti.
- Mengurangi kerusakan besar.
- Menurunkan biaya.
- Meningkatkan keselamatan.

Namun, terlalu banyak alarm palsu dapat menyebabkan teknisi mengabaikan sistem. Karena itu, sistem harus dievaluasi berdasarkan biaya nyata dan pengalaman pengguna.

41. Faktor Keberhasilan Proyek Data Science

41.1 Masalah yang Jelas

Proyek harus memiliki tujuan dan pengguna keputusan yang jelas.

41.2 Dukungan Pimpinan

Pimpinan harus menyediakan sumber daya, menghilangkan hambatan, dan mendorong penggunaan hasil.

41.3 Kualitas Data

Data harus relevan, cukup, dan dapat dipercaya.

41.4 Tim Interdisipliner

Tim harus mencakup kemampuan teknis, bisnis, hukum, etika, serta komunikasi.

41.5 Integrasi dengan Proses

Model harus terhubung dengan alur kerja.

41.6 Evaluasi Nilai

Keberhasilan harus diukur berdasarkan dampak, bukan hanya metrik teknis.

41.7 Budaya Eksperimen

Organisasi perlu bersedia menguji gagasan, belajar dari kegagalan, dan memperbaiki sistem.

41.8 Tata Kelola

Proyek harus memiliki dokumentasi, tanggung jawab, dan pengendalian.

42. Penyebab Kegagalan Proyek Data Science

Proyek dapat gagal karena:

- Masalah tidak jelas.
- Mengikuti tren teknologi.
- Data tidak tersedia.
- Kualitas data buruk.
- Tidak ada pengguna hasil.
- Model tidak dapat diimplementasikan.
- Tidak ada dukungan pimpinan.
- Tim bekerja terisolasi.
- Tidak ada perubahan proses.
- Hasil tidak dipercaya.
- Risiko etika diabaikan.
- Biaya melebihi manfaat.
- Model tidak dipantau.

Salah satu kegagalan umum adalah *solutionism*, yakni mencari masalah untuk teknologi yang telah dipilih.

Organisasi membeli platform AI, kemudian mencari cara menggunakannya. Pendekatan yang benar adalah memulai dari masalah dan menentukan apakah Data Science memang solusi yang tepat.

43. Kematangan Data Science Organisasi

Tingkat kematangan dapat digambarkan sebagai berikut.

Tahap 1: Pelaporan Dasar

Organisasi menggunakan laporan manual dan data terpisah.

Tahap 2: Business Intelligence

Organisasi memiliki dashboard dan indikator terintegrasi.

Tahap 3: Analitik Diagnostik

Organisasi menganalisis penyebab dan segmentasi.

Tahap 4: Analitik Prediktif

Organisasi menggunakan model untuk memperkirakan kejadian.

Tahap 5: Analitik Preskriptif

Model terintegrasi dengan optimasi dan rekomendasi.

Tahap 6: Organisasi Berbasis Pembelajaran

Data, eksperimen, dan evaluasi menjadi bagian dari keputusan sehari-hari.

Kematangan bukan hanya masalah teknologi. Ia melibatkan budaya, kompetensi, tata kelola, struktur, dan kepemimpinan.

44. Membangun Budaya Data

Budaya data adalah lingkungan ketika keputusan dibuat berdasarkan bukti, analisis dapat dipertanyakan, dan pembelajaran dihargai.

Budaya data memiliki ciri:

- Data mudah diakses sesuai kewenangan.
- Definisi indikator konsisten.
- Pimpinan menggunakan data.
- Karyawan memiliki literasi data.
- Ketidakpastian dikomunikasikan.
- Pendapat dapat diuji.
- Kesalahan digunakan untuk belajar.
- Manipulasi data tidak ditoleransi.

Budaya data bukan berarti semua keputusan harus menunggu analisis sempurna. Dalam kondisi darurat, keputusan tetap harus dibuat. Namun, asumsi dan bukti perlu dinyatakan secara jelas.

45. Literasi Data bagi Manajer

Manajer tidak harus menjadi programmer, tetapi perlu memiliki literasi data.

Manajer perlu mampu:

- Memahami indikator.
- Mengajukan pertanyaan kritis.
- Menilai kualitas data.
- Memahami probabilitas.
- Membedakan korelasi dan kausalitas.
- Membaca visualisasi.
- Memahami keterbatasan model.

- Menilai dampak etis.
- Menghubungkan analisis dengan strategi.

Pertanyaan yang perlu diajukan manajer kepada tim Data Science antara lain:

1. Masalah apa yang sebenarnya diselesaikan?
2. Siapa yang akan menggunakan hasil?
3. Dari mana data diperoleh?
4. Apakah data representatif?
5. Apa definisi target?
6. Bagaimana model dievaluasi?
7. Apa biaya false positive dan false negative?
8. Apakah model adil?
9. Dapatkah hasil dijelaskan?
10. Bagaimana model dipantau?
11. Apa yang terjadi apabila model salah?
12. Siapa yang bertanggung jawab?

46. Kompetensi Data Scientist

Kompetensi Data Scientist dapat dikelompokkan menjadi beberapa bidang.

46.1 Kompetensi Teknis

- Statistika.
- Probabilitas.
- Pemrograman.

- SQL.
- Machine learning.
- Visualisasi.
- Manajemen data.
- Komputasi awan.
- MLOps.

46.2 Kompetensi Analitis

- Pemecahan masalah.
- Pemikiran kritis.
- Perancangan eksperimen.
- Evaluasi ketidakpastian.
- Interpretasi hasil.

46.3 Kompetensi Bisnis

- Memahami proses.
- Memahami tujuan.
- Menilai dampak.
- Menentukan prioritas.
- Mengelola pemangku kepentingan.

46.4 Kompetensi Komunikasi

- Menulis.
- Presentasi.
- Visualisasi.
- Storytelling.

- Negosiasi.

46.5 Kompetensi Etis

- Privasi.
- Keadilan.
- Transparansi.
- Akuntabilitas.
- Kesadaran dampak sosial.

47. Pembelajaran Data Science

Pembelajaran Data Science sebaiknya dilakukan secara bertahap.

Tahap Fondasi

- Matematika dasar.
- Statistika.
- Logika.
- Spreadsheet.
- SQL.
- Pemrograman dasar.

Tahap Analisis

- Pembersihan data.
- Exploratory Data Analysis.
- Visualisasi.
- Regresi.
- Klasifikasi.

- Clustering.

Tahap Lanjutan

- Deep learning.
- NLP.
- Computer vision.
- Time series.
- Causal inference.
- Big data systems.
- MLOps.

Tahap Aplikasi

Mahasiswa perlu menyelesaikan proyek nyata.

Proyek harus mencakup:

- Definisi masalah.
- Data.
- Analisis.
- Model.
- Evaluasi.
- Etika.
- Komunikasi.
- Implementasi.

Pembelajaran Data Science tidak cukup melalui teori. Namun, praktik tanpa teori juga berbahaya. Keduanya perlu diintegrasikan.

48. Hubungan Data Science dengan Kecerdasan Buatan

Data Science dan Artificial Intelligence saling berhubungan, tetapi tidak identik.

Data Science berfokus pada proses memperoleh pengetahuan dan nilai dari data.

AI berfokus pada pengembangan sistem yang melakukan tugas yang memerlukan kemampuan intelektual.

Machine learning merupakan bagian dari AI dan juga alat penting dalam Data Science.

Tidak semua Data Science menggunakan AI. Analisis deskriptif, statistik sederhana, atau eksperimen dapat dilakukan tanpa AI.

Tidak semua AI berbasis Data Science dalam pengertian analisis organisasi. Sistem berbasis aturan dapat digolongkan sebagai AI tetapi tidak belajar dari data.

Hubungannya dapat digambarkan sebagai wilayah yang saling bertumpang tindih.

49. Data Science dan Generative AI

Generative AI menghasilkan konten baru, seperti teks, gambar, suara, video, dan kode.

Dalam Data Science, generative AI dapat membantu:

- Menulis kode.
- Menjelaskan hasil.
- Menghasilkan dokumentasi.
- Membuat data sintetis.
- Membantu pencarian literatur.

- Merangkum laporan.
- Mendukung analisis teks.
- Membuat antarmuka bahasa alami.

Namun, generative AI tidak boleh digunakan tanpa pengawasan.

Risikonya meliputi:

- Hallucination.
- Kebocoran data.
- Bias.
- Pelanggaran hak cipta.
- Kode yang tidak aman.
- Kesimpulan yang tidak dapat diverifikasi.
- Ketergantungan berlebihan.

Data scientist perlu memverifikasi setiap hasil yang digunakan dalam keputusan penting.

50. Data Sintetis

Data sintetis adalah data buatan yang dirancang menyerupai pola data nyata.

Data sintetis dapat digunakan ketika:

- Data nyata sensitif.
- Data langka.
- Perlu mensimulasikan kejadian.
- Perlu menambah data pelatihan.
- Perlu menguji sistem.

Namun, data sintetis bukan solusi sempurna. Data dapat gagal mencerminkan kompleksitas, bias, atau kejadian langka pada dunia nyata.

Data sintetis harus dievaluasi berdasarkan:

- Kemiripan statistik.
- Kegunaan untuk model.
- Perlindungan privasi.
- Representasi kelompok.
- Risiko identifikasi ulang.

51. Tantangan Masa Depan Data Science

51.1 Pertumbuhan Data

Jumlah dan keragaman data akan terus meningkat.

51.2 Data Multimodal

Sistem semakin banyak menggabungkan teks, gambar, suara, video, sensor, dan data terstruktur.

51.3 Real-Time Analytics

Keputusan akan semakin dilakukan secara langsung.

51.4 Regulasi

Penggunaan data dan AI akan menghadapi pengaturan yang semakin ketat.

51.5 Keamanan Model

Model dapat diserang, dimanipulasi, atau dicuri.

51.6 Sustainability

Komputasi berskala besar membutuhkan energi. Organisasi perlu menilai dampak lingkungan.

51.7 Democratization

Perangkat analitik semakin mudah digunakan. Namun, kemudahan penggunaan dapat meningkatkan risiko analisis yang tidak tepat apabila pengguna tidak memahami metodologi.

51.8 Human-AI Collaboration

Masa depan Data Science bukan sekadar menggantikan manusia, tetapi merancang kolaborasi manusia dan mesin.

Manusia unggul dalam:

- Penilaian kontekstual.
- Empati.
- Nilai.
- Kreativitas.
- Tanggung jawab moral.

Mesin unggul dalam:

- Kecepatan.
- Skala.
- Konsistensi.
- Pengenalan pola.
- Pemrosesan data besar.

Kolaborasi terbaik menggabungkan keunggulan keduanya.

52. Paradigma Data-Driven dan Data-Informed

Istilah *data-driven* sering digunakan untuk menggambarkan keputusan yang didorong oleh data.

Namun, konsep *data-informed* sering lebih tepat. Data menjadi masukan penting, tetapi bukan satu-satunya dasar.

Keputusan juga dipengaruhi oleh:

- Nilai.
- Etika.
- Strategi.
- Pengalaman.
- Pengetahuan lokal.
- Konteks.
- Kepentingan pemangku kepentingan.
- Ketidakpastian.

Data tidak pernah sepenuhnya netral. Pemilihan data, indikator, kategori, dan model selalu melibatkan keputusan manusia.

Karena itu, organisasi sebaiknya berorientasi pada keputusan yang diinformasikan data dan dipertanggungjawabkan manusia.

53. Data Science sebagai Ilmu Pengambilan Keputusan

Pada tingkat yang lebih mendalam, Data Science bukan sekadar ilmu tentang data. Data Science adalah ilmu tentang pengambilan keputusan di bawah ketidakpastian.

Data Science membantu menjawab:

- Apa yang kita ketahui?
- Apa yang belum kita ketahui?

- Seberapa yakin kita?
- Apa yang mungkin terjadi?
- Apa alternatif tindakan?
- Apa risiko setiap tindakan?
- Siapa yang memperoleh manfaat?
- Siapa yang menanggung biaya?
- Bagaimana kita belajar dari hasil?

Model terbaik bukan model yang paling kompleks. Model terbaik adalah model yang cukup akurat, sesuai tujuan, dapat dipahami, dapat diterapkan, aman, adil, dan menghasilkan nilai.

54. Prinsip-Prinsip Data Science yang Bertanggung Jawab

Data Science yang bertanggung jawab dapat dibangun berdasarkan beberapa prinsip.

54.1 Purpose

Penggunaan data harus memiliki tujuan yang sah dan jelas.

54.2 Proportionality

Tingkat pengumpulan dan analisis harus sebanding dengan manfaat.

54.3 Data Minimization

Hanya data yang diperlukan yang dikumpulkan.

54.4 Quality

Data dan model harus diuji.

54.5 Fairness

Dampak terhadap kelompok berbeda harus diperiksa.

54.6 Transparency

Penggunaan sistem harus dapat dijelaskan.

54.7 Accountability

Harus ada penanggung jawab.

54.8 Security

Data dan model harus dilindungi.

54.9 Human Control

Manusia harus tetap memiliki pengawasan.

54.10 Continuous Monitoring

Sistem harus dievaluasi sepanjang siklus hidupnya.

55. Kerangka Implementasi Data Science bagi Organisasi

Organisasi dapat menggunakan kerangka berikut.

Langkah 1: Menentukan Tujuan Strategis

Identifikasi masalah yang bernilai tinggi.

Langkah 2: Menilai Kesiapan

Periksa data, teknologi, kompetensi, budaya, dan tata kelola.

Langkah 3: Memilih Proyek Percontohan

Pilih proyek yang:

- Cukup penting.
- Memiliki data.
- Dapat diselesaikan.
- Memiliki pengguna.

- Memiliki indikator hasil.

Langkah 4: Membentuk Tim

Libatkan fungsi bisnis, teknologi, analitik, risiko, hukum, dan etika.

Langkah 5: Mengembangkan Prototipe

Bangun solusi awal untuk menguji kelayakan.

Langkah 6: Menguji Dampak

Gunakan eksperimen atau evaluasi yang sesuai.

Langkah 7: Mengintegrasikan Sistem

Masukkan model ke dalam proses kerja.

Langkah 8: Mengelola Perubahan

Latih pengguna dan perbaiki prosedur.

Langkah 9: Memantau

Pantau performa, bias, keamanan, dan manfaat.

Langkah 10: Menskalakan

Perluas setelah manfaat dan risiko dipahami.

56. Refleksi Kritis

Optimisme terhadap Data Science perlu disertai sikap kritis.

Data dapat memperjelas kenyataan, tetapi juga dapat menyederhanakan kenyataan secara berlebihan. Tidak semua aspek manusia dapat diukur secara memadai. Motivasi, martabat, kepercayaan, kreativitas, budaya, dan hubungan sosial memiliki dimensi yang tidak selalu tertangkap dalam angka.

Ketika organisasi hanya mengelola apa yang dapat diukur, aspek penting yang sulit diukur dapat terabaikan. Fenomena ini berkaitan dengan prinsip

bahwa ketika suatu indikator menjadi target, indikator tersebut dapat kehilangan kemampuannya sebagai ukuran yang baik.

Sebagai contoh, apabila sekolah hanya dinilai berdasarkan skor ujian, sekolah dapat memusatkan seluruh pembelajaran pada ujian dan mengabaikan kreativitas, karakter, serta pemikiran kritis.

Apabila karyawan hanya dinilai berdasarkan jumlah tugas, mereka dapat mengejar kuantitas dan mengorbankan kualitas.

Karena itu, indikator harus digunakan dengan kebijaksanaan. Data Science harus membantu manusia memahami kenyataan, bukan menggantikan kenyataan dengan angka.

57. Kesimpulan

Data Science merupakan bidang interdisipliner yang mengubah data menjadi pengetahuan, prediksi, rekomendasi, dan tindakan. Bidang ini menggabungkan statistika, matematika, ilmu komputer, manajemen data, machine learning, visualisasi, pengetahuan domain, komunikasi, serta etika.

Proses Data Science dimulai dari pemahaman masalah, bukan dari pemilihan algoritma. Data harus dikumpulkan, diperiksa, dibersihkan, diintegrasikan, dieksplorasi, dan ditransformasikan. Model kemudian dikembangkan, dievaluasi, diimplementasikan, serta dipantau.

Data Science dapat mendukung pemasaran, keuangan, sumber daya manusia, operasi, kesehatan, pendidikan, pemerintahan, pertanian, kehutanan, dan berbagai bidang lainnya. Manfaatnya meliputi peningkatan efisiensi, prediksi risiko, personalisasi, inovasi, serta pengambilan keputusan berbasis bukti.

Namun, Data Science juga mengandung risiko. Data dapat tidak lengkap atau bias. Model dapat melakukan diskriminasi. Sistem dapat sulit dijelaskan. Data pribadi dapat disalahgunakan. Pengguna dapat terlalu

mempercayai hasil algoritma. Karena itu, Data Science membutuhkan tata kelola, transparansi, keamanan, akuntabilitas, serta pengawasan manusia.

Keberhasilan Data Science tidak diukur dari kompleksitas teknologi. Keberhasilan diukur dari kemampuan menghasilkan keputusan yang lebih baik dan dampak yang bermanfaat.

Data Science juga tidak menghilangkan kebutuhan akan manusia. Sebaliknya, semakin kuat kemampuan mesin, semakin penting kemampuan manusia dalam memberikan tujuan, menafsirkan konteks, menilai konsekuensi, menjaga nilai, dan memikul tanggung jawab.

Dengan demikian, Data Science seharusnya dipahami sebagai disiplin pembelajaran organisasi. Organisasi menggunakan data untuk memahami pengalaman masa lalu, memperkirakan masa depan, memilih tindakan, mengevaluasi hasil, serta terus memperbaiki diri.

Dalam dunia yang semakin kompleks, Data Science dapat menjadi instrumen penting bagi rasionalitas, inovasi, dan adaptasi. Namun, manfaat tersebut hanya dapat diwujudkan apabila Data Science digunakan secara ilmiah, strategis, kritis, etis, dan berorientasi pada kemanusiaan.

Glosarium

Algorithm

Urutan aturan atau prosedur komputasi untuk menyelesaikan masalah.

Artificial Intelligence

Bidang yang mengembangkan sistem untuk melakukan tugas yang memerlukan kemampuan intelektual.

Bias

Penyimpangan sistematis dalam data, analisis, atau keputusan.

Big Data

Data dengan volume, kecepatan, keragaman, atau kompleksitas tinggi.

Classification

Proses memprediksi kategori suatu objek.

Cloud Computing

Penggunaan sumber daya komputasi melalui jaringan.

Clustering

Proses mengelompokkan objek berdasarkan kemiripan tanpa label sebelumnya.

Computer Vision

Bidang yang memungkinkan komputer mengolah gambar dan video.

Concept Drift

Perubahan hubungan antara fitur dan target sepanjang waktu.

Confusion Matrix

Tabel yang membandingkan hasil prediksi dengan kondisi sebenarnya.

Cross-Validation

Teknik evaluasi model melalui pembagian data secara bergantian.

Data Analyst

Profesional yang menganalisis dan melaporkan data untuk mendukung keputusan.

Data Architecture

Struktur dan desain pengelolaan data organisasi.

Data Cleaning

Proses memperbaiki kesalahan dan ketidakkonsistenan data.

Data Engineer

Profesional yang membangun infrastruktur dan aliran data.

Data Governance

Kerangka kebijakan, peran, dan kontrol pengelolaan data.

Data Lake

Penyimpanan data mentah dalam berbagai format.

Data Leakage

Masuknya informasi yang tidak semestinya ke dalam proses pelatihan model.

Data Lineage

Catatan perjalanan data dari sumber hingga penggunaan.

Data Mining

Proses menemukan pola dari kumpulan data.

Data Science

Bidang interdisipliner untuk memperoleh pengetahuan dan nilai dari data.

Data Scientist

Profesional yang menggunakan metode statistik, komputasi, dan pengetahuan domain untuk menganalisis data.

Data Steward

Pihak yang menjaga definisi, kualitas, dan penggunaan data.

Data Warehouse

Sistem penyimpanan data terintegrasi untuk analisis.

Deep Learning

Machine learning berbasis jaringan saraf berlapis banyak.

Descriptive Analytics

Analisis mengenai kejadian yang telah terjadi.

Diagnostic Analytics

Analisis mengenai penyebab suatu kejadian.

Edge Computing

Pemrosesan data di dekat sumber data.

Explainable AI

Pendekatan untuk membuat keputusan sistem AI dapat dipahami.

Feature

Variabel masukan yang digunakan model.

Feature Engineering

Proses menciptakan atau mengubah fitur agar lebih berguna.

Generative AI

AI yang dapat menghasilkan teks, gambar, suara, video, atau kode.

Inference

Proses menarik kesimpulan berdasarkan data.

Machine Learning

Metode yang memungkinkan sistem mempelajari pola dari data.

Metadata

Data yang menjelaskan data lain.

MLOps

Praktik mengelola pengembangan, penerapan, dan pemantauan model machine learning.

Model

Representasi matematis atau komputasional dari pola data.

Model Drift

Penurunan performa model akibat perubahan data atau lingkungan.

Natural Language Processing

Bidang yang memungkinkan komputer mengolah bahasa manusia.

Neural Network

Model komputasi yang terinspirasi secara longgar oleh jaringan neuron.

Overfitting

Kondisi ketika model terlalu cocok dengan data pelatihan dan buruk pada data baru.

Predictive Analytics

Analisis untuk memperkirakan kejadian masa depan.

Prescriptive Analytics

Analisis untuk merekomendasikan tindakan.

Privacy

Hak dan perlindungan individu berkaitan dengan data pribadi.

Regression

Metode untuk memprediksi nilai numerik atau menganalisis hubungan variabel.

Reinforcement Learning

Metode pembelajaran melalui interaksi, penghargaan, dan hukuman.

Structured Data

Data dengan format dan skema yang jelas.

Supervised Learning

Pembelajaran mesin menggunakan data berlabel.

Synthetic Data

Data buatan yang meniru karakteristik data nyata.

Time Series

Data yang tersusun berdasarkan urutan waktu.

Training Data

Data yang digunakan untuk melatih model.

Unstructured Data

Data yang tidak memiliki format tabular tetap.

Unsupervised Learning

Pembelajaran mesin menggunakan data tanpa label target.

Visualization

Representasi data dalam bentuk visual.

Daftar Pustaka

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443–448. <https://doi.org/10.1287/isre.2014.0546>

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Davenport, T. H., & Harris, J. G. (2017). *Competing on analytics: Updated, with a new introduction*. Harvard Business Review Press.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
<https://doi.org/10.1007/978-0-387-84858-7>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
<https://doi.org/10.1007/978-1-0716-1418-1>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.
<https://doi.org/10.1126/science.aaa8415>
- Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.
- Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1). <https://doi.org/10.1177/2053951714528481>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
<https://doi.org/10.1007/978-1-4614-6849-3>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).
- Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. <https://doi.org/10.1145/2939672.2939778>

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70, 263–286. <https://doi.org/10.1016/j.jbusres.2016.08.001>

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77–84. <https://doi.org/10.1111/jbl.12010>

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press. <https://doi.org/10.1017/9781009389433>

Naskah ini merupakan bab pengantar buku, modul mata kuliah **Data Science for Management**, atau landasan untuk pengembangan seri lanjutan mengenai *Business Analytics*, *Machine Learning*, dan *Artificial Intelligence for Decision Making*.

Copilot for this article: ChatGPT, Access date: 14 July 2026. Author's account ([Rudy C Tarumingkeng](https://chatgpt.com/c/6a55aedb-5ee0-83ec-8a3c-cf2be8efd52e)). <https://chatgpt.com/c/6a55aedb-5ee0-83ec-8a3c-cf2be8efd52e>