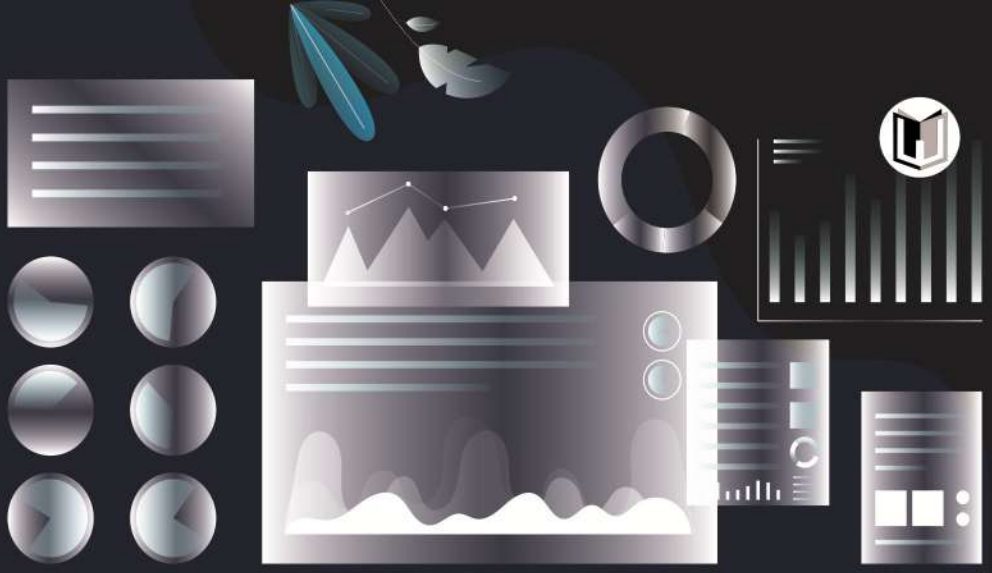

Pengantar

DATA SCIENCE

Memahami Konsep dan Teknik Analisis Data

Dr. Dian Ferriswara, S.E., M.M.

Pengantar

DATA SCIENCE

Memahami Konsep dan Teknik Analisis Data

Dr. Dian Ferriswara, S.E., M.M.

Penerbit Widina

PENGANTAR *DATA SCIENCE*:
Memahami Konsep dan Teknik Analisis Data

Penulis:
Dr. Dian Ferriswara, S.E., M.M.

Desain Cover:
Septian Maulana

Sumber Ilustrasi:
www.freepik.com

Tata Letak:
Handarini Rohana

Editor:
Dr. Iskandar Ahmaddien, S.ST., S.E., S.H., M.M.

ISBN:
978-634-246-454-0

Cetakan Pertama:
November, 2025

PENERBIT:
WIDINA MEDIA UTAMA
Komplek Puri Melia Asri Blok C3 No. 17 Desa Bojong Emas
Kec. Solokan Jeruk Kabupaten Bandung, Provinsi Jawa Barat

Anggota IKAPI No. 519/JBA/2025
Website: www.penerbitwidina.com
Instagram: @penerbitwidina
Telepon (022) 87355370

KATA PENGANTAR

Rasa syukur yang teramat dalam dan tiada kata lain yang patut kami ucapkan selain mengucap rasa syukur. Karena berkat rahmat dan karunia Tuhan Yang Maha Esa, buku yang berjudul "Pengantar *Data Science*: Memahami Konsep dan Teknik Analisis Data" telah selesai disusun dan berhasil diterbitkan, semoga buku ini dapat memberikan sumbangsih keilmuan dan penambah wawasan bagi siapa saja yang memiliki minat terhadap pembahasan tentang Pengantar *Data Science*: Memahami Konsep dan Teknik Analisis Data.

Akan tetapi pada akhirnya kami mengakui bahwa tulisan ini terdapat beberapa kekurangan dan jauh dari kata sempurna, sebagaimana pepatah menyebutkan "*tiada gading yang tidak retak*" dan sejatinya kesempurnaan hanyalah milik Tuhan semata. Maka dari itu, kami dengan senang hati secara terbuka untuk menerima berbagai kritik dan saran dari para pembaca sekalian, hal tersebut tentu sangat diperlukan sebagai bagian dari upaya kami untuk terus melakukan perbaikan dan penyempurnaan karya selanjutnya di masa yang akan datang.

Terakhir, ucapan terima kasih kami sampaikan kepada seluruh pihak yang telah mendukung dan turut andil dalam seluruh rangkaian proses penyusunan dan penerbitan buku ini, sehingga buku ini bisa hadir di hadapan sidang pembaca. Semoga buku ini bermanfaat bagi semua pihak dan dapat memberikan kontribusi bagi pembangunan ilmu pengetahuan di Indonesia.

November, 2025

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENGENALAN *DATA SCIENCE*

A. DEFINISI DAN RUANG LINGKUP *DATA SCIENCE*

Data science adalah disiplin ilmu interdisipliner yang fokus pada ekstraksi wawasan dan pengetahuan dari data dalam berbagai bentuk, baik terstruktur maupun tidak terstruktur. Disiplin ini menggabungkan pendekatan dari statistik, ilmu komputer, dan pemahaman konteks domain untuk mengolah, menganalisis, dan menginterpretasi data secara efektif. Provost dan Fawcett (2013) mendefinisikan *data science* sebagai ilmu yang mempelajari bagaimana memperoleh pengetahuan dari data melalui proses dan teknik ilmiah yang sistematis. Oleh karena itu, *data science* bukan sekadar analisis statistik, melainkan juga tentang bagaimana data digunakan untuk mendukung pengambilan keputusan yang cerdas.

Berikut adalah lima definisi *data science* dari para ahli, lengkap dengan nama ahli dan sumber referensinya:

1. Provost & Fawcett (2013) "*Data science is the science of extracting useful knowledge from data to solve real-world problems*." → Dalam bukunya *Data Science for Business*, mereka menekankan bahwa *data science* menggabungkan prinsip statistik, ilmu komputer, dan bisnis untuk menghasilkan wawasan yang dapat ditindaklanjuti.
2. Cao (2017) "*Data science is a new interdisciplinary field that combines techniques from computer science, statistics, and domain knowledge to extract meaningful insights from large-scale data*." → Dalam artikelnya *Data Science: A Comprehensive Overview*, Cao menekankan sifat multidisipliner *data science* dan fokusnya pada pengambilan wawasan yang berguna.
3. Dhar (2013) "*Data science is the study of the generalizable extraction of knowledge from data*." → Dalam jurnal *Communications of the ACM*, Dhar menggarisbawahi bahwa *data science* berusaha menemukan pola atau pengetahuan yang dapat digeneralisasi dari kumpulan data yang besar.
4. Berkeley School of Information (2018) "*Data science involves the principles, processes, and techniques for understanding phenomena via the (automated) analysis of data*."→ Dalam kurikulum resminya, Berkeley mendefinisikan *data science* sebagai ilmu yang mencakup pendekatan teknis dan analitis untuk memahami fenomena melalui data.

BAB 2
PENGUMPULAN DAN PENGOLAHAN DATA

A. SUMBER DATA: DATA TERSTRUKTUR DAN TIDAK TERSTRUKTUR

Dalam dunia *data science*, pemahaman tentang jenis-jenis sumber data sangat penting karena kualitas dan bentuk data akan menentukan pendekatan analisis yang digunakan. Secara umum, sumber data dibagi menjadi dua kategori besar: data terstruktur (*structured data*) dan data tidak terstruktur (*unstructured data*). Data terstruktur merujuk pada data yang tersimpan dalam format yang dapat dengan mudah diatur ke dalam baris dan kolom, seperti pada *database* relasional atau *spreadsheet*. Sebaliknya, data tidak terstruktur tidak mengikuti format yang tetap dan sering kali berupa teks bebas, gambar, video, atau audio. Menurut Gandomi dan Haider (2015), lebih dari 80% data yang tersedia di dunia saat ini adalah data tidak terstruktur, sehingga menuntut pendekatan analisis yang lebih kompleks dan inovatif.

Data Terstruktur vs. Tidak Terstruktur

Karakteristik	Data Terstruktur	Data Tidak Terstruktur
Format	Skema yang terdefinisi dengan baik	Tidak ada skema yang terdefinisi
Organisasi	Terorganisir dalam tabel	Tidak terorganisir, format bebas
Akses	Akses mudah dengan SQL	Memerlukan teknik khusus
Contoh	Transaksi, pelanggan, inventaris	Email, media sosial, gambar
Ukuran	Biasanya lebih kecil	Biasanya lebih besar
Alat Analisis	SQL, alat BI	NLP, Pembelajaran Mesin

Gambar 2.1 Perbandingan antara Data Terstruktur dan Data Tidak Terstruktur

Gambar di atas menyajikan perbandingan antara data terstruktur dan data tidak terstruktur berdasarkan berbagai aspek penting. Data terstruktur memiliki format yang jelas dengan skema yang sudah didefinisikan, sehingga mudah diorganisir dalam bentuk tabel seperti pada *database* relasional.

BAB 3
EKSPLORASI DATA
DAN ANALISIS DESKRIPTIF

A. PENGANTAR ANALISIS DESKRIPTIF

Analisis deskriptif merupakan tahap awal dalam eksplorasi data yang berfokus pada penggambaran karakteristik utama dari sekumpulan data. Tujuan utamanya adalah menyederhanakan data kompleks menjadi bentuk yang dapat dipahami dengan lebih mudah, seperti tabel frekuensi, ukuran pemusatan (mean, median, modus), dan ukuran penyebaran (simpangan baku, varians, jangkauan). Langkah ini tidak hanya membantu dalam memahami distribusi data, tetapi juga berfungsi sebagai fondasi untuk analisis statistik lanjutan (Geetha & Sujatha, 2024).

Salah satu metode yang digunakan dalam analisis deskriptif adalah visualisasi data. Grafik batang, diagram lingkaran, histogram, dan *boxplot* adalah beberapa alat yang umum digunakan untuk menggambarkan sebaran data. Visualisasi ini memungkinkan pengguna mengidentifikasi pola, *outlier*, atau tren awal dengan cepat, yang sulit dideteksi hanya dari tabel angka (Fraser, 2019).

Selain visualisasi, analisis deskriptif mencakup pemahaman atas bentuk distribusi data. Misalnya, apakah data terdistribusi normal (*bell-shaped*), simetris, atau miring ke kiri/kanan. Pemahaman distribusi ini sangat penting karena akan memengaruhi teknik statistik apa yang cocok digunakan dalam tahap analisis lebih lanjut (Cooksey, 2020).

Analisis deskriptif juga membantu mengidentifikasi kesalahan pengumpulan data atau *outlier* yang mungkin terjadi. *Outlier* dapat menyebabkan kesimpulan statistik yang bias, sehingga penting untuk dikenali sejak awal. Dalam banyak kasus, pembersihan data akan dilakukan setelah proses analisis deskriptif dilakukan untuk meningkatkan kualitas data (Rajagopalan, 2020).

Dalam konteks bisnis dan industri, analisis deskriptif memberikan gambaran ringkas namun informatif tentang performa atau tren historis. Misalnya, laporan keuangan bulanan atau data penjualan tahunan sering dianalisis menggunakan statistik deskriptif untuk memudahkan pengambilan keputusan strategis oleh manajemen (Geetha & Sujatha, 2024).

BAB 4
PENGANTAR STATISTIKA
UNTUK *DATA SCIENCE*

A. KONSEP DASAR STATISTIKA: POPULASI, SAMPEL, DAN VARIABEL

Berikut adalah 5 definisi konsep dasar statistika: populasi, sampel, dan variabel menurut para ahli, lengkap dan mudah dipahami:

1. Populasi
   Menurut *Sudjana (2005)*, populasi adalah keseluruhan objek penelitian, baik berupa orang, benda, atau kejadian, yang memiliki karakteristik tertentu dan menjadi perhatian seorang peneliti. Populasi bisa bersifat terbatas (*finite*) atau tak terbatas (*infinite*), tergantung pada ruang lingkup studi. Misalnya, seluruh siswa SMA di Indonesia bisa menjadi populasi dalam penelitian pendidikan.
2. Sampel
   Menurut *Sugiyono (2017)*, sampel adalah bagian dari jumlah dan karakteristik yang dimiliki oleh populasi, dan harus dipilih sedemikian rupa agar representatif. Sampel yang baik akan mencerminkan karakteristik populasi, sehingga hasil penelitian yang didapatkan dari sampel dapat digeneralisasi ke seluruh populasi.
3. Variabel
   Kerlinger (2006) menjelaskan bahwa variabel adalah suatu konstruk atau sifat yang dapat diubah-ubah dan diukur secara kuantitatif atau kualitatif. Variabel bisa bersifat independen (bebas), dependen (terikat), kontrol, atau moderator, dan digunakan untuk memahami hubungan antar fenomena yang sedang diteliti.
4. Statistik Deskriptif dan Inferensial. Menurut *Walpole (1995)*, statistik terdiri dari dua cabang utama, yaitu statistik deskriptif yang menggambarkan data melalui grafik, tabel, dan ukuran ringkasan; serta statistik inferensial yang menggunakan sampel untuk menarik kesimpulan tentang populasi. Konsep populasi dan sampel sangat krusial dalam pengambilan keputusan berbasis data.
5. Data dan Unit Analisis. Menurut *Ferdinand (2002)*, data adalah informasi dalam bentuk angka atau simbol yang merepresentasikan variabel dari suatu unit analisis. Unit analisis dapat berupa individu, organisasi, atau peristiwa yang diteliti. Oleh karena itu, pemilihan variabel yang tepat sangat penting dalam mendesain penelitian kuantitatif yang valid.

BAB 5
DATA *WRANGLING* DAN MANIPULASI DATA

A. TEKNIK DATA *WRANGLING*: MENGUBAH DAN MENYARING DATA

Data *wrangling*, atau data munging, adalah proses penting dalam siklus kerja data *science* yang mencakup pengumpulan, pembersihan, transformasi, dan penyaringan data mentah menjadi format yang dapat digunakan untuk analisis lebih lanjut. Proses ini bertujuan untuk memastikan bahwa data yang digunakan dalam analisis memiliki kualitas tinggi, lengkap, relevan, dan bebas dari kesalahan yang dapat mengganggu hasil. Menurut Kandel *et al*. (2011), sekitar 50-80% waktu kerja *data analyst* dihabiskan untuk data *wrangling*, menandakan betapa krusialnya proses ini dalam *pipeline* data *science*.

Salah satu langkah utama dalam data *wrangling* adalah data *transformation*, yaitu mengubah struktur atau tipe data agar sesuai dengan kebutuhan analisis. Contohnya adalah mengonversi kolom tanggal dari format *string* ke *datetime*, atau menggabungkan beberapa kolom menjadi satu. Transformasi juga dapat mencakup normalisasi skala data, *encoding* data kategorikal, dan membuat fitur-fitur baru dari data yang ada. Menurut Dasu dan Johnson (2003), transformasi ini meningkatkan kegunaan data dalam model statistik dan *machine learning*.

Langkah penting lainnya adalah data *filtering*, yaitu proses menyaring data untuk memilih subset yang relevan atau membuang entri yang tidak memenuhi kriteria tertentu. Misalnya, menyaring data pelanggan yang hanya melakukan pembelian di atas Rp500.000 atau menghapus baris data yang memiliki nilai kosong pada variabel kunci. *Filtering* membantu mengurangi *noise* dalam data dan fokus pada observasi yang penting untuk tujuan analisis (Han, Kamber & Pei, 2012).

Selain transformasi dan *filtering, handling outliers* dan data duplikat juga merupakan bagian dari data *wrangling*. *Outlier* adalah nilai yang jauh berbeda dari sebagian besar data dan dapat mengganggu model prediksi. Teknik umum meliputi *winsorization, trimming*, atau menghapus *outlier* berdasarkan batas statistik (misal Z-*score* > 3). Data duplikat harus diidentifikasi dan dihapus untuk menghindari bias, apalagi dalam analisis frekuensi atau regresi (Aggarwal, 2015).

Dalam praktiknya, proses data *wrangling* biasanya dilakukan menggunakan *tools* seperti *Python* (pandas, numpy), R (dplyr, tidyr), atau bahkan SQL. *Tools* ini menyediakan fungsi-fungsi seperti dropna(), fillna(), groupby(), dan merge() untuk membantu manipulasi data secara efisien.

BAB 6
ALGORITMA DAN TEKNIK PEMODELAN DATA

A. PEMODELAN PREDIKTIF DAN ALGORITMA YANG DIGUNAKAN

Pemodelan prediktif adalah teknik dalam *data science* yang digunakan untuk memperkirakan nilai masa depan berdasarkan pola dari data historis. Tujuannya adalah membangun model statistik atau *machine learning* yang dapat mengenali hubungan antar variabel dan menghasilkan prediksi akurat untuk data baru. Pemodelan prediktif banyak digunakan dalam berbagai industri seperti keuangan (deteksi *fraud*), pemasaran (prediksi *churn* pelanggan), kesehatan (diagnosis penyakit), dan ritel (peramalan permintaan produk) [(James *et al*., 2013)].

Langkah awal dalam pemodelan prediktif adalah pemilihan variabel target dan fitur prediktor. Variabel target adalah apa yang ingin diprediksi (misalnya, apakah pelanggan akan membeli atau tidak), sedangkan fitur prediktor adalah informasi relevan yang digunakan untuk membuat prediksi (misalnya, umur, pendapatan, histori pembelian). Proses ini membutuhkan eksplorasi data yang menyeluruh untuk memahami korelasi dan signifikansi masing-masing fitur terhadap target [(Kuhn & Johnson, 2013)].

Setelah itu, dilakukan pemisahan data menjadi dua subset: data pelatihan (*training* data) dan data pengujian (testing data). Model dilatih menggunakan data pelatihan dan dievaluasi kinerjanya menggunakan data pengujian. Tujuannya adalah menghindari *overfitting*, yaitu kondisi di mana model terlalu cocok dengan data pelatihan dan tidak mampu generalisasi ke data baru. Teknik validasi silang (*cross-validation*) sering digunakan untuk meningkatkan keandalan model [(Hastie, Tibshirani & Friedman, 2009)].

Berbagai algoritma dapat digunakan dalam pemodelan prediktif tergantung pada jenis data dan masalah yang dihadapi. Untuk klasifikasi, algoritma populer meliputi *Decision Tree*, *Random Forest*, *Logistic Regression*, *Support Vector Machines* (SVM), dan Naive Bayes. Untuk regresi (prediksi nilai numerik), digunakan *Linear Regression, Ridge/Lasso Regression*, dan *Gradient Boosting*. Sedangkan untuk prediksi *time series*, model seperti ARIMA dan LSTM sangat populer [(Geron, 2019)].

Machine learning modern juga banyak mengandalkan algoritma *ensemble* seperti *Random Forest* dan *Gradient Boosting Machines* (GBM) karena kemampuannya untuk mengurangi varians dan meningkatkan akurasi prediksi.

BAB 7
PENGANTAR PEMBELAJARAN MESIN (*MACHINE LEARNING*)

A. KONSEP PEMBELAJARAN MESIN: *SUPERVISED* VS *UNSUPERVISED LEARNING*

Pembelajaran mesin (*machine learning*) terdiri dari dua pendekatan utama: *supervised learning* dan *unsupervised learning*. *Supervised learning* melibatkan pelatihan model dengan *dataset* yang telah dilabeli, di mana *input* dan *output* sudah diketahui, sedangkan *unsupervised learning* bekerja dengan data yang belum dilabeli dan mencoba menemukan pola tersembunyi di dalamnya. Kedua pendekatan ini memiliki kekuatan dan kelemahan masing-masing serta digunakan dalam konteks yang berbeda sesuai dengan tujuan analisis data.

Dalam *supervised learning*, model belajar dari pasangan *input-output* untuk membuat prediksi pada data baru. Contohnya termasuk klasifikasi (seperti mengidentifikasi email *spam*) dan regresi (seperti memprediksi harga rumah). Kelebihannya adalah akurasi yang tinggi karena menggunakan data yang sudah dianotasi, tetapi proses pelabelan bisa sangat memakan waktu dan mahal (Pothos, Edwards, & Perlman, 2015).

Sebaliknya, *unsupervised learning* digunakan ketika data tidak memiliki label. Algoritma ini mencoba mengidentifikasi pola seperti klasterisasi atau asosiasi dalam data. Misalnya, dalam analisis pasar, *unsupervised learning* dapat digunakan untuk mengelompokkan konsumen berdasarkan perilaku pembelian mereka tanpa mengetahui kategori sebelumnya (Retnoningsih & Pramudita, 2020).

Dalam praktiknya, perbedaan antara kedua pendekatan ini tidak selalu hitam putih. Beberapa peneliti berpendapat bahwa semua bentuk pembelajaran mesin memiliki unsur pengawasan, baik secara eksplisit melalui label, maupun secara implisit melalui struktur data itu sendiri (Odaibo, 2019).

Penggunaan kedua metode ini juga sangat bergantung pada ketersediaan data. Misalnya, *supervised learning* sangat berguna jika tersedia dataset besar yang telah dilabeli, sedangkan *unsupervised* lebih fleksibel dalam lingkungan data yang tidak terstruktur dan sulit dikategorikan sebelumnya (Khodabandelou *et al*., 2014).

Beberapa aplikasi praktis menggabungkan kedua pendekatan ini dalam teknik semi-*supervised learning* atau self-*supervised learning*, yang memanfaatkan sedikit data berlabel bersama data tidak berlabel untuk

BAB 8
PEMBELAJARAN MENDALAM (*DEEP LEARNING*)

A. PENGANTAR *DEEP LEARNING* DAN JARINGAN *NEURAL*

Deep learning merupakan subbidang dari *machine learning* yang menekankan pada penggunaan *artificial neural networks* (ANN) dengan banyak lapisan tersembunyi (*hidden layers*) untuk mengekstraksi representasi data yang kompleks dan berskala besar. Pendekatan ini terinspirasi dari cara kerja otak manusia yang terdiri dari neuron-neuron saling terhubung dan berfungsi secara paralel untuk memproses informasi. Istilah "*deep*" dalam *deep learning* merujuk pada banyaknya lapisan dalam jaringan, yang memungkinkan pembelajaran fitur dari data mentah secara bertingkat dan hierarkis [(LeCun, Bengio, & Hinton, 2015)].

Jaringan *neural* buatan (*Artificial Neural Networks*/ANN) terdiri atas tiga jenis lapisan utama: lapisan *input* (*input layer*), lapisan tersembunyi (*hidden layers*), dan lapisan *output* (*output layer*). Setiap lapisan terdiri dari sejumlah unit atau neuron buatan yang menerima *input*, mengolahnya menggunakan bobot dan fungsi aktivasi, dan meneruskan hasilnya ke lapisan berikutnya. Proses ini dikenal sebagai *forward propagation*. Tujuan utama dari pelatihan jaringan *neural* adalah menemukan bobot optimal yang meminimalkan kesalahan antara prediksi dan nilai sebenarnya melalui metode *backpropagation* dan optimisasi seperti *gradient descent* [(Goodfellow, Bengio, & Courville, 2016)].

Salah satu kekuatan utama *deep learning* terletak pada kemampuannya untuk melakukan *feature learning* secara otomatis. Pada *machine learning* tradisional, proses ekstraksi fitur biasanya dilakukan secara manual oleh ahli domain. Sebaliknya, dalam *deep learning*, model dapat belajar langsung dari data mentah seperti citra, suara, atau teks untuk menemukan representasi terbaik yang relevan bagi tugas tertentu. Ini membuat *deep learning* sangat unggul dalam menangani data tidak terstruktur dan kompleks [(Schmidhuber, 2015)].

Deep learning telah merevolusi berbagai bidang seperti *computer vision*, *natural language processing* (NLP), *speech recognition*, dan robotika. Misalnya, *Convolutional Neural Networks* (CNN) banyak digunakan untuk klasifikasi gambar dan deteksi objek, sedangkan *Recurrent Neural Networks* (RNN) dan *Long Short-Term Memory* (LSTM) digunakan dalam pengolahan data berurutan

BAB 9
ANALISIS TEKS DAN *NATURAL LANGUAGE PROCESSING* (NLP)

A. KONSEP DASAR *NATURAL LANGUAGE PROCESSING* (NLP)

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (AI) yang berfokus pada interaksi antara komputer dan bahasa manusia (*natural language*). Tujuan utama NLP adalah memungkinkan mesin untuk memahami, menafsirkan, menghasilkan, dan merespons bahasa alami manusia dengan cara yang bermakna dan berguna. NLP mencakup berbagai teknik yang menggabungkan linguistik komputasional, pembelajaran mesin (*machine learning*), dan *deep learning* untuk memungkinkan pemrosesan teks atau ucapan. Dalam konteks modern, NLP digunakan dalam berbagai aplikasi mulai dari chatbot, sistem penerjemah otomatis, hingga analisis sen*time*n media sosial [(Jurafsky & Martin, 2021)].

Komponen utama dalam NLP mencakup beberapa tahapan penting, yakni *tokenization*, *morphological analysis*, *syntactic parsing*, *semantic analysis*, dan *pragmatic understanding*. *Tokenization* adalah proses memecah kalimat menjadi unit-unit kata (tokens), sedangkan *morphological analysis* melihat struktur internal kata untuk memahami makna morfem. *Parsing* bertujuan menganalisis struktur sintaksis kalimat berdasarkan aturan tata bahasa, dan *semantic analysis* mencoba memahami arti dari kalimat atau frasa tersebut. Proses ini memungkinkan mesin untuk membangun pemahaman semantik yang dapat digunakan untuk membuat keputusan atau merespons *input* pengguna [(Bird, Klein, & Loper, 2009)].

Salah satu tantangan utama dalam NLP adalah am*big*uitas bahasa alami. Bahasa manusia penuh dengan makna ganda, idiom, metafora, dan variasi konteks. Misalnya, kata “bank” bisa berarti institusi keuangan atau tepi sungai, tergantung pada konteksnya. Untuk mengatasi tantangan ini, NLP menggunakan pendekatan statistik dan pembelajaran mendalam yang dapat belajar dari data dalam skala besar untuk menangkap pola-pola linguistik dan konteks semantik yang kompleks [(Cambria & White, 2014)].

Perkembangan NLP sangat dipengaruhi oleh kemajuan dalam representasi teks, terutama melalui teknik seperti *word embedding*. Model seperti Word2Vec, GloVe, dan Fast*Text* mengubah kata-kata menjadi vektor angka dalam ruang berdimensi tinggi, di mana makna semantik kata ditangkap dalam hubungan spasial antar vektor. Dengan representasi ini, kata-kata yang

BAB 10
PENGOLAHAN DATA BESAR (*BIG DATA*)

A. KONSEP *BIG DATA* DAN KARAKTERISTIKNYA

Big Data merupakan istilah yang digunakan untuk menggambarkan kumpulan data yang sangat besar, kompleks, dan terus berkembang sehingga tidak dapat ditangani dengan teknologi pengolahan data konvensional. Data tersebut berasal dari berbagai sumber seperti media sosial, transaksi keuangan, sensor, rekam medis, log aktivitas pengguna, hingga interaksi digital lainnya. *Big Data* bukan hanya soal ukuran, tetapi juga mencakup kecepatan data datang (*velocity*), beragamnya jenis data (*variety*), serta potensi nilai informasi (*value*) yang terkandung di dalamnya. Menurut Mayer-Schönberger dan Cukier (2013), *Big* Data memungkinkan pengambilan keputusan berbasis data (*data-driven decision making*) yang lebih akurat dan *real-time*.

Dalam perkembangannya, *Big* Data telah mengubah lanskap teknologi informasi. Pendekatan tradisional yang bergantung pada sistem relasional tidak mampu menangani data dalam skala *petabyte* atau bahkan *exabyte*. Oleh karena itu, lahirlah berbagai *platform* dan arsitektur baru seperti Hadoop, Spark, dan sistem file terdistribusi lainnya untuk mengelola penyimpanan dan pemrosesan data secara paralel. Menurut Gandomi dan Haider (2015), pendekatan ini memungkinkan analisis data besar dilakukan dengan efisien dalam waktu yang jauh lebih singkat, bahkan untuk data yang tidak terstruktur.

Karakteristik utama dari *Big* Data dikenal dengan istilah 3V: Volume, *Velocity*, dan *Variety*. Volume merujuk pada jumlah data yang sangat besar dan terus bertambah setiap detik. *Velocity* menunjukkan kecepatan data dihasilkan dan diproses, seperti data *stream*ing dari sensor IoT atau media sosial. *Variety* mencerminkan keberagaman bentuk data, termasuk data terstruktur (misalnya *database*), semi-terstruktur (seperti JSON atau XML), dan tidak terstruktur (seperti video, audio, dan teks bebas). Konsep ini pertama kali diperkenalkan oleh Doug Laney pada tahun 2001 dan menjadi fondasi dalam memahami dinamika *Big* Data (Laney, 2001).

Seiring perkembangan, para peneliti menambahkan dua dimensi karakteristik lain, yaitu *Veracity* dan *Value*. *Veracity* merujuk pada keakuratan dan keandalan data, mengingat tidak semua data yang dikumpulkan memiliki kualitas baik atau bebas dari bias. *Value* menggambarkan nilai yang bisa diekstraksi dari data untuk mendukung pengambilan keputusan strategis. Seperti dijelaskan oleh Schroeder, Cowls, dan Taddeo (2016), tantangan

BAB 11
EVALUASI DAN INTERPRETASI MODEL

A. EVALUASI KINERJA MODEL: *CONFUSION MATRIX*, *PRECISION*, *RECALL*, F1-*SCORE*

Evaluasi kinerja model dalam pembelajaran mesin merupakan tahap krusial untuk menentukan seberapa baik suatu model mampu melakukan prediksi terhadap data baru. Salah satu cara paling umum untuk melakukan evaluasi terhadap model klasifikasi adalah dengan menggunakan metrik-metrik seperti *confusion matrix*, *precision*, *recall*, dan F1-*score*. Metrik ini memberikan gambaran menyeluruh mengenai performa model, terutama pada data yang memiliki distribusi kelas tidak seimbang (*imbalanced classes*) (Powers, 2011).

Confusion matrix adalah representasi matriks dari hasil prediksi model yang dibandingkan dengan label sebenarnya. Matriks ini terdiri dari empat komponen utama yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). TP menunjukkan jumlah kasus di mana model memprediksi kelas positif dengan benar, sedangkan TN menunjukkan prediksi benar terhadap kelas negatif. FP merupakan prediksi salah terhadap kelas positif (*false alarm*), dan FN adalah kegagalan model dalam mendeteksi kasus positif (Han *et al*., 2011).

Precision mengukur proporsi prediksi positif yang benar-benar relevan. Formula *precision* adalah TP / (TP + FP). *Precision* sangat penting dalam kasus di mana biaya dari prediksi positif yang salah sangat tinggi, misalnya dalam diagnosis penyakit langka atau sistem deteksi penipuan. Model dengan *precision* tinggi akan mengurangi jumlah *false positive*, sehingga lebih dapat diandalkan dalam konteks di mana kesalahan tipe I (*false positive*) harus dihindari (Sokolova & Lapalme, 2009).

Recall, atau juga dikenal sebagai *sensitivity* atau *true positive rate*, adalah ukuran dari proporsi kasus positif yang berhasil diidentifikasi oleh model. Formula *recall* adalah TP / (TP + FN). *Recall* menjadi sangat penting dalam konteks di mana kegagalan mendeteksi kasus positif dapat menimbulkan dampak besar, seperti dalam deteksi kanker atau sistem peringatan dini. Model dengan *recall* tinggi akan meminimalisir kesalahan tipe II (*false negative*) (Powers, 2011).

F1-*Score* merupakan rata-rata harmonik dari *precision* dan *recall*. F1-*Score* digunakan untuk mendapatkan keseimbangan antara *precision* dan *recall*, terutama saat distribusi kelas tidak seimbang. Formula F1-*score* adalah 2 *

BAB 12
ETIKA DAN TANTANGAN DALAM DATA *SCIENCE*

A. ETIKA PENGUMPULAN DAN PENGGUNAAN DATA

Etika dalam pengumpulan dan penggunaan data merupakan prinsip fundamental dalam dunia analitik data dan kecerdasan buatan. Di era digital saat ini, data menjadi aset strategis yang bernilai tinggi, namun penggunaannya harus tetap memperhatikan aspek moral, hukum, dan sosial. Etika data merujuk pada seperangkat pedoman yang memastikan bahwa proses pengumpulan, penyimpanan, analisis, dan distribusi data dilakukan dengan menghormati hak individu dan kelompok (Zwitter, 2014). Tanpa etika yang kuat, penggunaan data berisiko melanggar privasi, menimbulkan diskriminasi, atau menciptakan ketidakadilan sosial.

Salah satu isu etis utama dalam pengumpulan data adalah persetujuan yang diinformasikan (*informed consent*). Prinsip ini menekankan bahwa individu harus diberi informasi yang jelas dan memadai mengenai jenis data yang dikumpulkan, tujuan penggunaannya, serta risiko yang mungkin terjadi. Mereka juga harus diberikan kebebasan untuk menyetujui atau menolak partisipasi tanpa tekanan. Dalam konteks digital, persetujuan sering kali dicapai melalui syarat dan ketentuan yang panjang dan rumit, yang pada akhirnya mengaburkan transparansi (Nissenbaum, 2011). Oleh karena itu, perlu adanya desain komunikasi yang lebih etis dan mudah dipahami oleh pengguna.

Privasi data juga menjadi perhatian etis utama. Individu berhak mengetahui siapa yang mengakses data mereka dan untuk tujuan apa. Kebocoran data pribadi dapat berdampak serius terhadap reputasi, keamanan finansial, bahkan keselamatan seseorang. Oleh karena itu, pendekatan *privacy-by-design* dan *data minimization* dianjurkan untuk memastikan bahwa hanya data yang benar-benar diperlukan saja yang dikumpulkan dan diproses (Cavoukian, 2012). Prinsip ini sangat penting dalam peraturan perlindungan data seperti *General Data Protection Regulation* (GDPR) di Uni Eropa, yang memberikan kontrol lebih besar kepada individu atas data pribadi mereka.

Selain itu, terdapat isu etika dalam penggunaan data untuk pengambilan keputusan otomatis, seperti dalam model *machine learning* atau sistem rekomendasi. Keputusan yang dihasilkan dari data bisa saja bias, terutama jika data pelatihan mencerminkan ketimpangan sosial atau diskriminasi historis.

DAFTAR PUSTAKA

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. https://doi.org/10.1002/wics.101

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-47578-3

Agrawal, R., & Srikant, R. (1994). Fast algorithms for *mining* association rules. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, 487–499.

Ali, M., Farid, M., & Rahman, T. (2018). *Applied Statistics for Social and Health Sciences*. Routledge. https://doi.org/10.4324/9781315147853

Altman, N. S. (1992). An Introduction to Kernel and *Nearest*-Neighbor Nonparametric Regression. *The American Statistician*, 46(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Balan, M. (2024). Advances in Visual *Analytics* for *Big* Data: Modern Approaches in Exploratory Data Analysis. *Journal of Data Science & Visualization*, 12(1), 22–38. https://doi.org/10.1016/j.jdsv.2024.02.003

Beam, A. L., & Kohane, I. S. (2018). *Big* data and *machine learning* in health care. *JAMA*, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Berchtold, A., & Raftery, A. E. (2000). *The* mixture transition distribution model for high-order Markov chains and non-Gaussian *time series*. *Statistical Science*, 15(2), 101–124. https://doi.org/10.1214/ss/1009212754

Berkeley School of Information. (2018). *What is Data Science?*. University of California, Berkeley. Diakses dari: https://data*science*.berkeley.edu/about/what-is-data-*science*/

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data *mining* for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Bojarski, M., Del Testa, D., Dworakowski, D., *et al*. (2016). End to End *Learning* for Self-Driving Cars. *arXiv preprint arXiv:1604.07316*. https://arxiv.org/abs/1604.07316

Bojarski, M., *et al*. (2016). End to End *Learning* for Self-Driving Cars. *arXiv preprint arXiv:1604.07316*. https://arxiv.org/abs/1604.07316

Borrohou, F. Z., Bouarfa, H., & El Afia, A. (2023). Anomaly detection and handling *outliers*: A review of methods and applications in *big* data *analytics*. *Big Data Research*, 31, 100336. https://doi.org/10.1016/j.bdr.2022.100336

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

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

Brownlee, J. (2017). *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. *Machine Learning* Mastery.

Bryman, A. (2012). *Social Research Methods* (4th ed.). Oxford University Press.

Burges, C. J. C. (1998). A Tutorial on *Support Vector Machines* for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). New avenues in opinion *mining* and sen*timen*t analysis. *IEEE Intelligent Systems*, 28(2), 15–21. https://doi.org/10.1109/MIS.2013.30

Cao, L. (2017). Data *science*: A comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 43. https://doi.org/10.1145/3076253

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction* (6th ed.). Chapman & Hall/CRC.

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 (KDD)*, 785–794. https://doi.org/10.1145/2939672.2939785

Cooksey, R. W. (2020). *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*. Springer.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). *Nearest* neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. Wiley-Inter*science*.

Date, C. J. (2003). *An Introduction to Database Systems* (8th ed.). Addison Wesley.

Davenport, T. H., & Patil, D. J. (2012). Data *scientist*: *The* sexiest job of *the* 21st century. *Harvard Business Review*, 90(10), 70–76.

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and Statistics* (4th ed.). Pearson Education.

Dhar, V. (2013). Data *science* and prediction. *Communications of the ACM*, 56(12), 64–73. https://doi.org/10.1145/2500499

Domingos, P. (2012). A Few Useful Things to Know About *Machine Learning*. *Communications of the ACM*, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755

Duan, Y., Cao, G., & Edwards, J. S. (2019). *Under*standing *the* impact of business *analytics* on innovation. *Information & Management*, 56(8), 103135. https://doi.org/10.1016/j.im.2018.10.003

El-Nasr, M. S., Drachen, A., & Canossa, A. (2021). *Game Data Science: Introduction to Machine Learning in Games*. Oxford University Press.

Esteva, A., Kuprel, B., Novoa, R. A., *et al*. (2017). Dermatologist-level classification of skin cancer with *deep neural networks*. *Nature*, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Ferdinand, A. (2002). *Metode Penelitian Manajemen: Pedoman Penelitian untuk Penulisan Skripsi, Tesis dan Disertasi Ilmu Manajemen*. Badan Penerbit Universitas Diponegoro.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. *Analytics* Press.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). SAGE Publications.

Fraser, A. (2019). Data visualization in *the* age of *big* data: Charting best practices. *Information Visualization Review*, 35(2), 145–163.

Friedman, B. M., & Scharfstein, D. S. (2020). *The* impact of fintech on financial intermediation. *Annual Review of Financial Economics*, 12, 243–266. https://doi.org/10.1146/annurev-financial-112519-115505

Gandomi, A., & Haider, M. (2015). Beyond *the* hype: *Big* data concepts, methods, and *analytics*. *International Journal of Information Management*, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gao, L., Jin, X., Li, Y., & Chen, Y. (2024). DataPrepML: A Transparent and Scalable *Pipeline* for Automating Data *Preparation* in *Machine Learning*. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 18(1), 1–28.

Geetha, S., & Sujatha, R. (2024). Practical Data Analysis: A Business-Centered Approach Using R. *International Journal of Business Analytics*, 11(1), 35–49. https://doi.org/10.4018/IJBA.2024010103

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

Gomez-Uribe, C. A., & Hunt, N. (2015). *The* Netflix Recommender System: Algorithms, business *value*, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4), 13. https://doi.org/10.1145/2843948

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.*deeplearning*book.org/

Goyle, S., Lamba, H., & Thakur, V. (2023). Automating Data Pre*processing* Using *Machine Learning Pipeline*s: *Tools* and Trends. *International Journal of Advanced Computer Science and Applications*, 14(2), 190–199.

Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the Behavioral Sciences* (10th ed.). Cengage *Learning*.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645–1660. https://doi.org/10.1016/j.future.2013.01.010

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.

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

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

Hawkins, D. M. (2004). *The* problem of *overfitting*. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12. https://doi.org/10.1021/ci0342472

Helal, A., & Al-reyashi, H. (2023). Impact of data pre*processing* on *the* performance of *machine learning* models: A review. *Artificial Intelligence Review*, 56, 2063–2089. https://doi.org/10.1007/s10462-022-10226-7

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). O*Text*s. https://o*text*s.com/fpp2/

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2017). Overview of data exploration techniques. *Communications of the ACM*, 60(9), 86–95. https://doi.org/10.1145/3127325

International Journal of Engineering and Advanced Technology. (2019). Data Visualization for Business *Intelligence*: A Case-Based Approach. *IJEAT*, 8(6), 2545–2552.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johnson, R. A., & Kuby, P. (2012). *Elementary Statistics* (11th ed.). Brooks/Cole, Cengage *Learning*.

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

Kalla, A. (2022). *Understanding Statistical Variables in Social Science Research*. *Journal of Educational Measurement and Statistics*, 8(2), 85–97.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3363–3372. https://doi.org/10.1145/1978942.1979444

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Kerlinger, F. N. (2006). *Foundations of Behavioral Research* (4th ed.). Harcourt College Publishers.

Khodabandelou, G., Mili, F., Mendling, J., & Fahland, D. (2014). *Unsupervised* discovery of con*text*-aware activity patterns. *Information Systems*, 39, 1–21. https://doi.org/10.1016/j.is.2013.12.002

Kitchin, R. (2014). *The real-time* city? *Big* data and smart urbanism. *GeoJournal*, 79, 1–14. https://doi.org/10.1007/s10708-013-9516-8

Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.

Kohavi, R. (1995). A Study of *Cross-Validation* and *Bootstrap* for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 2, 1137–1143.

Kohavi, R. (1995). A study of *cross-validation* and *bootstrap* for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143).

Kolmogorov, A. N. (1933). *Foundations of the Theory of Probability*. Chelsea Publishing Company.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with *Deep Convolutional Neural Networks*. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kura, M., & Elmazi, L. (2024). AutoML and *the* Future of Intelligent Model Selection: A Survey. *Journal of Artificial Intelligence Research and Applications*, 6(1), 45–62.

Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Liaw, A., & Wiener, M. (2002). Classification and Regression by *randomForest*. *R News*, 2(3), 18–22.

Lindgren, B. W. (1976). *Statistical Theory* (4th ed.). Macmillan.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Mahalle, P. N., Patil, P. A., & Prasad, N. R. (2021). *Data Science and Analytics: An Introduction to Core Concepts*. Springer.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting: Methods and Applications* (3rd ed.). Wiley.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mark, M. M., & Workman, D. P. (1991). Sampling Methods in Educational Research. *Educational Evaluation and Policy Analysis*, 13(2), 153–164.

Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Wiley.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

Menard, S. (2002). *Applied Logistic Regression Analysis* (2nd ed.). SAGE Publications.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in *Vector* Space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.

Mitchell, R. (2018). *Web Scraping with Python: Collecting Data from the Modern Web* (2nd ed.). O'Reilly Media.

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2015). *Introduction to the Practice of Statistics* (8th ed.). W.H. Freeman.

Mugejo, V. K., & Govere, S. (2024). Handling *missing* data in predictive *analytics*: Comparative performance of imputation techniques. *International Journal of Data Science and Analytics*, 17(1), 115–130. https://doi.org/10.1007/s41060-023-00321-4

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

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, 78. https://doi.org/10.1145/1015330.1015435

Nie, J. (2024). Emerging Trends in Data Visualization: From Static Charts to AI-Enhanced Interactive *Dashboard*s. *Data Science Frontiers*, 3(1), 1–18. https://doi.org/10.1016/dsf.2024.001

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

O'Neil, C., & Schutt, R. (2013). *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media.

O'Brien, F., Remington, R., & Freund, J. (1983). *Statistics: A First Course* (5th ed.). Prentice Hall.

Odaibo, S. G. (2019). On *the* Implicit Supervision in All *Learning* Algorithms. *AI Perspectives*, 1(1), 1–5. https://doi.org/10.1186/s42467-019-0014-2

Olson, D. L. (2003). *Managerial Issues of Enterprise Resource Planning Systems*. McGraw-Hill.

Pothos, E. M., Edwards, D. J., & Perlman, A. (2015). *Supervised* versus *unsupervised learning* in *artificial intelligence*. *Current Directions in Psychological Science*, 24(5), 407–412.

Priestley, M., & McGrath, R. (2019). *The* changing face of knowledge production: From mode 2 to data *science*. *British Educational Research Journal*, 45(3), 501–517. https://doi.org/10.1002/berj.3505

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Quinlan, J. R. (1986). Induction of *Decision Trees*. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Rahm, E., & Do, H. H. (2000). Data *Cleaning*: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

Rajagopalan, B. (2020). Data *Cleaning* and *Validation* in Exploratory Data Analysis. *Journal of Data Analytics*, 8(2), 45–59.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge University Press.

Ramakrishnan, R., & Gehrke, J. (2003). *Database Management Systems* (3rd ed.). McGraw-Hill Education.

Ramaswamy, S., Bizopoulos, P., & Escobar, G. (2021). *Modern Data Architecture: Enabling Business Agility with Data*. O'Reilly Media.

Rashid, T., & Waheed, H. (2021). *Missing* data imputation techniques in *machine learning*: A comparative study. *Journal of Big Data*, 8, 1–28. https://doi.org/10.1186/s40537-021-00495-9

Redman, T. C. (1998). *The* impact of poor data quality on *the* typical enterprise. *Communications of the ACM*, 41(2), 79–82. https://doi.org/10.1145/269012.269025

Retnoningsih, M. & Pramudita, A. A. N. (2020). Segmentasi Pelanggan dengan Metode K-Means Clustering untuk Peningkatan Penjualan. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(2), 341–348.

Richards, D. (2017). Self-*supervised learning*: *The* frontier of semi-*supervised* algorithms. *Journal of Machine Learning Research*, 18, 1–12.

Rogers, R. (1998). Descriptive Statistics in Educational Research. *Educational Research Quarterly*, 21(4), 3–15.

Ross, S. M. (2014). *Introduction to Probability and Statistics for Engineers and Scientists* (5th ed.). Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning* representations by back-propagating errors. *Nature*, 323(6088), 533–536. https://doi.org/10.1038/323533a0

Sahu, P. K. (2016). *Research Methodology: A Guide for Researchers in Agricultural Science, Social Science and Other Related Fields*. Springer.

Saito, T., & Rehmsmeier, M. (2015). *The Precision-Recall* Plot Is More Informative than *the ROC* Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Saito, T., & Rehmsmeier, M. (2015). *The precision-recall* plot is more informative than *the ROC* plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432.

Saleem, M., & Asif, R. (2015). Comparative study of *outlier* detection techniques. *Procedia Computer Science*, 46, 689–695. https://doi.org/10.1016/j.p*roc*s.2015.02.147

Schek, H. J., & Scholl, M. H. (1990). *The* relational model with relation-*value*d attributes. *Information Systems*, 15(3), 303–316.

https://doi.org/10.1016/0306-4379(90)90003-J
Schwabish, J. A. (2021). *Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks*. Columbia University Press.
Sebastiani, F. (2002). *Machine learning* in automated *text* categorization. *ACM Computing Surveys*, 34(1), 1–47. https://doi.org/10.1145/505282.505283
Sekaran, U., & Bougie, R. (2016). *Research Methods for Business: A Skill-Building Approach* (7th ed.). Wiley.
Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples* (4th ed.). Springer.
Stonebraker, M., & Hellerstein, J. M. (2005). What goes around comes around. *Readings in Database Systems* (4th ed.), 2–41. MIT Press.
Strasser, E., & Klettke, M. (2024). Enhancing Transparency in *Machine Learning Pipeline*s through Semantic Data Lineage. *Journal of Data and Information Quality*, 16(1), 1–27.
Strasser, E., & Klettke, M. (2024). Enhancing transparency in *machine learning pipeline*s through semantic data lineage. *Journal of Data and Information Quality, 16*(1), 1–27.
Sudjana. (2005). *Metode Statistika* (6th ed.). Tarsito.
Sugiyono. (2017). *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Alfabeta.
Triola, M. F. (2018). *Elementary Statistics* (13th ed.). Pearson.
Ullman, J. D. (2021). *The Role of Algorithms in Data Science*. Stanford University. (lecture notes)
Urs, S. R., & Minhaj, M. (2022). Curriculum analysis of data *science* programs: A study of iSchools. *Education and Information Technologies*, 27, 1765–1792. https://doi.org/10.1007/s10639-021-10636-3
Valentim, R. A. M., *et al*. (2019). Fairness in *Artificial Intelligence*: Bias, Transparency, and Responsibility. *AI & Ethics*, 1(1), 89–101.
VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.
Von Mises, R. (1957). *Probability, Statistics and Truth*. Macmillan.
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
Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1995). *Probability and Statistics for Engineers and Scientists* (6th ed.). Prentice Hall.
Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer. https://doi.org/10.1007/978-0-387-21736-9
Weiss, N. A. (2016). *Introductory Statistics* (10th ed.). Pearson.

Wickham, H. (2014). *Tidy Data*. *Journal of Statistical Software*, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in *deep learning* based *natural language processing*. *IEEE Computational Intelligence Magazine*, 13(3), 55–75.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in *deep learning* based *natural language processing*. *IEEE Computational Intelligence Magazine*, 13(3), 55–75.

Zhou, K., Fu, C., & Yang, S. (2020). *Big* data-driven smart energy management: From *big* data to *big insight*s. *Renewable and Sustainable Energy Reviews*, 72, 107–112. https://doi.org/10.1016/j.rser.2016.10.005

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill.

Pengantar

DATA SCIENCE

Memahami Konsep dan Teknik Analisis Data

Di era informasi yang didorong oleh volume data yang masif, *Data Science* telah menjelma menjadi disiplin ilmu krusial yang menjembatani kesenjangan antara data mentah dan keputusan bisnis yang strategis. Buku "Pengantar *Data Science*: Memahami Konsep dan Teknik Analisis Data" hadir sebagai panduan komprehensif yang dirancang untuk membekali pembaca dengan fondasi teoretis dan keterampilan praktis yang dibutuhkan oleh seorang *Data Scientist* modern.

Buku ini memandu pembaca melalui seluruh siklus hidup *Data Science*, dimulai dari definisi, peranannya dalam berbagai industri, hingga keterampilan esensial yang wajib dikuasai. Pembahasan dimulai dengan teknik fundamental dalam Pengumpulan dan Pengolahan Data, meliputi penanganan data terstruktur dan tidak terstruktur, pembersihan data (mengatasi *missing values* dan *outliers*), serta teknik Data *Wrangling* dan manipulasi data yang canggih.

Buku ini juga mencakup topik-topik kontemporer seperti Analisis Teks dan *Natural Language Processing* (NLP), termasuk teknik modern seperti Transformer dan BERT, serta pembahasan kritis mengenai Pengolahan Data Besar (Big Data) dan teknologi yang mendasarinya (Hadoop dan Spark).

Ditutup dengan bab penting tentang Evaluasi Kinerja Model (menggunakan *Confusion Matrix* dan ROC *Curve*) dan Etika dalam *Data Science*, buku ini memastikan bahwa pembaca tidak hanya mampu membangun model yang akurat, tetapi juga memahami implikasi moral, tantangan privasi, dan bias dalam pengambilan keputusan berbasis data. Buku ini sangat ideal bagi mahasiswa, akademisi, dan profesional yang ingin membangun karier atau memperdalam pemahaman mereka dalam dunia *Data Science* yang dinamis dan transformatif.

SCANME

www.penerbitwidina.com
@penerbitwidina
penerbit widina
penerbitwidina@gmail.com
widina store
widina bookstore

Layanan Pembaca & Penerbitan Buku
0815-7000-699

Teknologi & Komputer

ISBN 978-634-246-454-0

9 786342 464540