# Pengantar
# Data Science

## Mengambil Keputusan Berdasarkan Data

**Phie Chyan, Zelvi Gustiana, Sitti Arni, Amru Yasir, Hartina Husain, Budi Arif Dermawan, Ade Oktarino, I Putu Tedy Indrayana, Amril Mutoi Siregar, Alfredo Gormantara, Indah Dwi Mumpuni, Medy Wisnu Prihatmono, I Putu Gd Sukenada Andisana, Lenny Maryam AB. Possumah, Ibnu Atho'illah, Siti Aisyah, Santi, Nuk Ghurroh Setyoningrum, Salman Farizy, Vivi Afifah**

# Pengantar Data Science

## Mengambil Keputusan Berdasarkan Data

**Phie Chyan, Zelvi Gustiana, Sitti Arni, Amru Yasir, Hartina Husain, Budi Arif Dermawan, Ade Oktarino, I Putu Tedy Indrayana, Amril Mutoi Siregar, Alfredo Gormantara, Indah Dwi Mumpuni, Medy Wisnu Prihatmono, I Putu Gd Sukenada Andisana, Lenny Maryam AB. Possumah, Ibnu Atho'illah, Siti Aisyah, Santi, Nuk Ghurroh Setyoningrum, Salman Farizy, Vivi Afifah**

**PT. MIFANDI MANDIRI DIGITAL**

**Undang-Undang No. 28 Tahun 2014 Tentang Hak Cipta:**

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam pasal 9 ayat (1) huruf i untuk penggunaan secara komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/ atau pidana denda paling banyak Rp. 100.000.000,- (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/ atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/ atau huruf h untuk penggunaan secara komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/ atau pidana denda paling banyak Rp. 500.000.000,- (lima ratus juta rupiah).
3. Setiap Orang yang dengan tanpa hak dan/ atau tanpa izin pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/ atau huruf g untuk penggunaan secara komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/ atau pidana denda paling banyak Rp. 1.000.000.000,- (satu miliar rupiah).
4. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/ atau pidana denda paling banyak Rp. 4.000.000.000,- (empat miliar rupiah).

# Pengantar Data Science
## Mengambil Keputusan Berdasarkan Data

Phie Chyan, Zelvi Gustiana, Sitti Arni, Amru Yasir, Hartina Husain, Budi Arif Dermawan, Ade Oktarino, I Putu Tedy Indrayana, Amril Mutoi Siregar, Alfredo Gormantara, Indah Dwi Mumpuni, Medy Wisnu Prihatmono, I Putu Gd Sukenada Andisana, Lenny Maryam AB. Possumah, Ibnu Atho'illah, Siti Aisyah, Santi, Nuk Ghurroh Setyoningrum, Salman Farizy, Vivi Afifah

**ISBN: 978-623-8558-39-1**

**Editor** : Sarwandi, M.Pd.T
**Layout** : Miftahul Jannah, M.Kom
**Desain sampul** : Rifki Ramadan

**Penerbit**
PT. Mifandi Mandiri Digital

**Redaksi**
Komplek Senda Residence Jl. Payanibung Ujung D Dalu Sepuluh-B Tanjung Morawa Kab. Deli Serdang Sumatera Utara

**Distributor Tunggal**
PT. Mifandi Mandiri Digital
Komplek Senda Residence Jl. Payanibung Ujung D Dalu Sepuluh-B Tanjung Morawa Kab. Deli Serdang Sumatera Utara

Cetakan Pertama, Agustus 2024

# Kata Pengantar

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa, atas berkat dan rahmat-Nya buku ini yang berjudul "Pengantar Data Science: Mengambil Keputusan Berdasarkan Data" dapat diselesaikan. Buku ini ditulis dengan tujuan memberikan pemahaman yang komprehensif tentang dunia data science, disiplin ilmu yang semakin menjadi tulang punggung dalam pengambilan keputusan di berbagai bidang.

Era digital saat ini ditandai dengan ledakan data yang luar biasa. Setiap detik, sejumlah besar data dihasilkan dari berbagai sumber seperti media sosial, transaksi bisnis, sensor IoT, dan banyak lagi. Data tersebut memiliki potensi besar untuk diolah menjadi informasi yang berharga dan wawasan yang mendalam. Namun, untuk bisa memanfaatkannya secara optimal, diperlukan pengetahuan dan keterampilan khusus dalam bidang data science.

Buku ini dirancang untuk menjadi panduan praktis bagi para mahasiswa, profesional, dan siapa saja yang ingin memahami dasar-dasar data science. Kami membahas berbagai konsep kunci dan teknik yang digunakan dalam data science, mulai dari pengumpulan dan pembersihan data, analisis eksploratif, model pembelajaran mesin, hingga visualisasi data. Setiap bab dilengkapi dengan contoh-contoh nyata dan studi kasus yang relevan untuk membantu pembaca mengaitkan teori dengan praktik.

Selain itu, kami juga menyoroti pentingnya etika dalam data science. Dalam era di mana data pribadi menjadi sangat

berharga, penting bagi kita untuk memahami dan menerapkan prinsip-prinsip etika dalam pengelolaan dan analisis data.

Kami menyadari bahwa buku ini tidak akan terwujud tanpa dukungan dan kontribusi dari berbagai pihak. Oleh karena itu, ucapan terima kasih yang tulus kami sampaikan kepada semua penulis, editor, dan rekan-rekan yang telah memberikan masukan dan dukungan berharga selama proses penulisan buku ini.

Akhir kata, kami berharap buku ini dapat memberikan wawasan yang bermanfaat dan inspirasi bagi pembaca dalam menjelajahi dunia data science. Semoga buku ini dapat menjadi referensi yang berguna dalam pengambilan keputusan berbasis data dan membantu Anda dalam perjalanan karier di bidang yang menarik ini. Selamat membaca dan semoga sukses selalu!

Medan,  Juli 2024

Penulis

# Daftar Isi

# BAB 1 PENGANTAR DATA SCIENCE

## Pendahuluan

Data science adalah bidang yang sedang berkembang pesat dan telah menjadi salah satu disiplin ilmu yang paling diminati dalam beberapa tahun terakhir. Dengan kemajuan teknologi dan ledakan data, kebutuhan akan ahli data semakin meningkat di berbagai industri (Davenport.T & Patil. D, 2012). Artikel ini akan memberikan pengantar lengkap mengenai data science, termasuk definisi, sejarah, komponen utama, proses, alat dan teknik, serta aplikasinya dalam dunia nyata.

## Apa itu Data Science

Data science adalah disiplin ilmu yang menggabungkan keterampilan statistik, ilmu komputer, dan pengetahuan domain untuk mengekstraksi wawasan dan pengetahuan dari data. Tujuan utama dari data science adalah untuk membuat keputusan yang lebih baik dan lebih tepat berdasarkan analisis data yang mendalam. Data science mencakup berbagai teknik seperti pemodelan statistik, *Machine Learning*, data mining, dan visualisasi data (Provost. F & Fawcett. T, 2013)

Data science tidak hanya tentang mengumpulkan dan menganalisis data, tetapi juga tentang memahami konteks di mana data tersebut digunakan dan bagaimana hasil analisis dapat diaplikasikan dalam situasi nyata. Dengan kata lain, data science adalah jembatan antara data mentah dan informasi yang

dapat ditindaklanjuti (McKinney.W, 2013)

## Sejarah Data Science

Sejarah data science dapat ditelusuri kembali ke awal perkembangan statistik dan komputer. Berikut adalah beberapa tonggak penting dalam evolusi data science:

1. Era Awal Statistik
   Ilmu statistik berkembang pada abad ke-17 dengan kontribusi dari para ilmuwan seperti John Graunt dan William Petty, yang mengembangkan metode untuk menganalisis data populasi. Pekerjaan awal mereka dalam pengumpulan dan analisis data populasi dianggap sebagai fondasi dari statistik modern.
2. Komputasi Elektronik
   Pada pertengahan abad ke-20, penemuan komputer memungkinkan analisis data dalam skala yang jauh lebih besar dan lebih cepat. Komputer pertama kali digunakan untuk keperluan ilmiah dan militer, tetapi dengan cepat meluas ke aplikasi bisnis dan komersial.
3. Revolusi Data
   Pada tahun 2000-an, dengan munculnya internet dan teknologi digital, jumlah data yang dihasilkan meningkat secara eksponensial, dikenal sebagai *Big Data*. Data ini berasal dari berbagai sumber seperti media sosial, transaksi online, dan sensor IoT, yang menciptakan tantangan baru dalam pengelolaan dan analisis data.
4. Munculnya *Machine Learning*
   Teknik *Machine Learning*, yang menggunakan algoritma untuk belajar dari data dan membuat prediksi, menjadi pusat perhatian dalam data science modern. Dengan kemampuan untuk mengotomatisasi analisis data dan

membuat prediksi yang akurat, *Machine Learning* telah membuka peluang baru dalam berbagai bidang, dari pengenalan suara hingga deteksi penipuan (Goodfellow et al, 2017).

5. Perkembangan Alat dan Platform Data
   Alat seperti Hadoop, Spark, dan berbagai bahasa pemrograman seperti Python dan R telah membuat analisis data lebih mudah diakses dan lebih efisien. Alat-alat ini memungkinkan para data scientist untuk mengelola dan menganalisis data dalam jumlah besar dengan cara yang lebih terstruktur dan efisien.

## Komponen Utama Data Science

Data science terdiri dari beberapa komponen utama yang saling terkait. Setiap komponen memiliki peran penting dalam keseluruhan proses analisis data. Berikut adalah beberapa komponen utama data science:

1. Pengumpulan Data
   Langkah pertama dalam proses data science adalah pengumpulan data. Data dapat diperoleh dari berbagai sumber seperti *database* perusahaan, sensor, log web, media sosial, dan banyak lagi. Pengumpulan data yang efektif memerlukan pemahaman tentang sumber data dan teknik yang digunakan untuk mengakses dan menyimpan data tersebut.
2. Pembersihan dan Pengolahan Data
   Data mentah seringkali tidak siap untuk dianalisis karena mungkin mengandung kesalahan, duplikasi, atau nilai yang hilang. Pembersihan dan pengolahan data melibatkan kegiatan seperti penanganan *missing values*, normalisasi, dan transformasi data. Proses ini penting

untuk memastikan bahwa data yang akan dianalisis adalah akurat dan konsisten.

3. Analisis Eksplorasi Data (EDA)
   EDA adalah proses untuk memahami karakteristik dan struktur data melalui visualisasi dan statistik deskriptif. EDA membantu dalam mengidentifikasi pola, tren, dan anomali dalam data. Dengan menggunakan alat visualisasi seperti matplotlib dan seaborn di Python, data scientist dapat menggali wawasan awal yang berguna dari dataset (Van der Aalst. W, 2016).
4. Pemodelan Data
   Pemodelan data melibatkan penggunaan algoritma statistik dan *Machine Learning* untuk membangun model yang dapat membuat prediksi atau klasifikasi berdasarkan data. Proses ini termasuk pemilihan fitur, pelatihan model, dan evaluasi model. Pemodelan yang baik memerlukan pemahaman mendalam tentang algoritma yang digunakan dan bagaimana mereka dapat dioptimalkan untuk menghasilkan hasil yang terbaik.
5. Evaluasi dan Validasi Model
   Setelah model dibangun, langkah selanjutnya adalah mengevaluasi kinerjanya menggunakan metrik seperti akurasi, presisi, *recall*, dan F1-score. Validasi model dilakukan untuk memastikan bahwa model tidak *overfitting* atau *underfitting*. Teknik validasi yang umum digunakan termasuk *cross-validation* dan *split train-test*.

## Proses Data Science

Proses data science umumnya mengikuti alur kerja yang terstruktur, meskipun dapat bervariasi tergantung pada proyek dan organisasi (Chollet.F, 2018). Berikut adalah langkah-

langkah umum dalam proses data science:

1. Memahami Masalah Bisnis
   Langkah pertama adalah memahami masalah bisnis yang ingin dipecahkan dan tujuan dari analisis data. Ini melibatkan diskusi dengan para pemangku kepentingan untuk mengidentifikasi kebutuhan dan ekspektasi mereka. Pemahaman yang baik tentang masalah bisnis membantu dalam merancang solusi yang efektif dan tepat sasaran.
2. Pengumpulan Data
   Data yang relevan dikumpulkan dari berbagai sumber. Ini bisa berupa data internal perusahaan atau data eksternal dari pihak ketiga. Proses pengumpulan data memerlukan pemahaman tentang berbagai teknik pengumpulan data dan alat yang digunakan untuk mengakses dan menyimpan data tersebut.
3. Pembersihan dan Pengolahan Data
   Data yang dikumpulkan kemudian dibersihkan dan diolah untuk memastikan bahwa data siap untuk dianalisis. Proses ini termasuk penanganan data yang hilang, penghapusan duplikasi, dan transformasi data. Pembersihan data adalah langkah penting untuk memastikan bahwa analisis yang dilakukan adalah akurat dan dapat diandalkan.
4. Analisis Eksplorasi Data
   EDA dilakukan untuk memahami struktur data dan mengidentifikasi pola yang mungkin ada. Ini melibatkan penggunaan visualisasi data dan statistik deskriptif. EDA membantu dalam mengidentifikasi variabel yang paling penting dan bagaimana mereka saling berinteraksi (Gandomi. A & Haider. M, 2015).

5. Pemodelan Data
   Model *Machine Learning* dibangun dan dilatih menggunakan data yang telah diproses. Pemodelan melibatkan pemilihan algoritma, pemilihan fitur, dan pelatihan model. Pemilihan algoritma yang tepat dan pengoptimalan model adalah langkah penting untuk memastikan bahwa model yang dibangun dapat memberikan hasil yang akurat dan dapat diandalkan.
6. Evaluasi Model
   Model dievaluasi untuk mengukur kinerjanya dan memastikan bahwa model tersebut dapat memberikan hasil yang akurat dan dapat diandalkan. Metrik evaluasi yang umum digunakan termasuk akurasi, presisi, *recall*, dan F1-score. Evaluasi model membantu dalam mengidentifikasi kelemahan model dan bagaimana mereka dapat diperbaiki.
7. Implementasi Model
   Setelah model dievaluasi dan divalidasi, model tersebut diimplementasikan dalam lingkungan produksi untuk digunakan oleh bisnis. Implementasi model melibatkan integrasi model dengan sistem bisnis yang ada dan memastikan bahwa model dapat digunakan secara efektif oleh pengguna akhir.
8. Pemantauan dan Pemeliharaan
   Model yang diimplementasikan harus dipantau secara terus-menerus untuk memastikan bahwa model tersebut tetap akurat dan relevan. Pemeliharaan model melibatkan pembaruan model dengan data baru dan penyesuaian parameter. Pemantauan model juga membantu dalam mengidentifikasi perubahan dalam data yang mungkin mempengaruhi kinerja model.

## Alat dan Teknik dalam Data Science

Data science melibatkan berbagai alat dan teknik yang digunakan untuk analisis data dan pemodelan. Beberapa alat dan teknik yang paling populer dalam data science meliputi:

### Alat

1. Python
   Python adalah bahasa pemrograman yang paling populer dalam data science karena mudah dipelajari dan memiliki banyak pustaka yang kuat untuk analisis data seperti Pandas, NumPy, dan Scikit-learn. Selain itu, Python memiliki komunitas yang besar dan aktif yang terus mengembangkan pustaka dan alat baru untuk data science.
2. R
   R adalah bahasa pemrograman lain yang banyak digunakan dalam data science, terutama untuk analisis statistik dan visualisasi data. R memiliki berbagai paket yang kuat untuk analisis data seperti ggplot2 untuk visualisasi dan dplyr untuk manipulasi data.
3. SQL
   SQL digunakan untuk mengelola dan mengakses data dalam *database* relasional. Kemampuan untuk menulis kueri SQL adalah keterampilan penting bagi data scientist. SQL memungkinkan data scientist untuk mengambil dan mengelola data dengan efisien dari *database* relasional.
4. Apache Hadoop
   Hadoop adalah kerangka kerja *open-source* yang digunakan untuk penyimpanan dan pemrosesan data

besar dalam skala besar. Hadoop memungkinkan penyimpanan data yang terdistribusi dan pemrosesan data secara paralel, yang sangat berguna untuk menangani *Big Data*.

5. Apache Spark
   Spark adalah platform komputasi terdistribusi yang cepat dan dapat digunakan untuk pemrosesan data besar secara *real-time*. Spark menyediakan API yang mudah digunakan untuk pemrosesan data batch dan stream, serta mendukung berbagai bahasa pemrograman seperti Python, Java, dan Scala.
6. Tableau
   Tableau adalah alat visualisasi data yang memungkinkan pengguna untuk membuat dashboard interaktif dan laporan visual. Tableau menyediakan berbagai alat visualisasi yang kuat dan mudah digunakan untuk mengubah data mentah menjadi wawasan yang dapat ditindaklanjuti.
7. Jupyter Notebook
   Jupyter Notebook adalah lingkungan pengembangan interaktif yang banyak digunakan oleh data scientist untuk menulis dan menjalankan kode Python, serta mendokumentasikan proses analisis. Jupyter Notebook memungkinkan integrasi kode, visualisasi, dan dokumentasi dalam satu antarmuka yang mudah digunakan.

## Teknik

1. Regresi Linier
   Regresi linier adalah teknik statistik yang digunakan untuk memodelkan hubungan antara variabel

independen dan variabel dependen dengan cara yang linier. Regresi linier sering digunakan untuk prediksi dan analisis hubungan antara variabel.

2. Klasifikasi
   Klasifikasi adalah teknik *Machine Learning* yang digunakan untuk mengklasifikasikan data ke dalam kategori yang telah ditentukan. Contoh algoritma klasifikasi termasuk *Decision Tree*, *Random Forest*, dan *logistic regression*. Klasifikasi sering digunakan dalam aplikasi seperti deteksi penipuan, diagnosis medis, dan segmentasi pelanggan.
3. *Clustering*
   *Clustering* adalah teknik unsupervised learning yang digunakan untuk mengelompokkan data ke dalam cluster berdasarkan kesamaan antara data. Contoh algoritma *Clustering* adalah *K-means* dan *hierarchical clustering*. *Clustering* digunakan dalam berbagai aplikasi seperti segmentasi pasar, analisis pola perilaku, dan analisis jaringan sosial.
4. *Principal Component Analysis* (PCA)
   PCA adalah teknik reduksi dimensi yang digunakan untuk mengurangi jumlah variabel dalam dataset sambil mempertahankan sebanyak mungkin variasi dalam data. PCA sering digunakan untuk visualisasi data, prapemrosesan data, dan kompresi data.
5. *Natural Language Processing* (NLP)
   NLP adalah cabang data science yang berfokus pada analisis dan pemrosesan teks. Teknik NLP digunakan dalam aplikasi seperti analisis sentimen, penerjemahan mesin, dan chatbot. NLP melibatkan berbagai teknik seperti tokenisasi, stemming, lemmatization, dan analisis

sintaksis.

## Aplikasi Data Science dalam Dunia Nyata

Data science memiliki berbagai aplikasi dalam berbagai industri. Beberapa contoh aplikasi data science dalam dunia nyata misalnya Data science digunakan dalam pemasaran dan periklanan untuk memahami perilaku konsumen, segmentasi pasar, dan personalisasi iklan. Misalnya, perusahaan *e-commerce* menggunakan data science untuk merekomendasikan produk kepada pelanggan berdasarkan riwayat pembelian dan penelusuran mereka (Silver. N, 2012). Dengan analisis data yang mendalam, perusahaan dapat membuat kampanye pemasaran yang lebih efektif dan efisien. Berikutnya dalam bidang Kesehatan, Dalam industri kesehatan, data science digunakan untuk analisis data medis, prediksi penyakit, dan pengembangan perawatan yang dipersonalisasi. Misalnya, *Machine Learning* dapat digunakan untuk menganalisis data gambar medis seperti MRI atau CT scan untuk mendeteksi kelainan atau penyakit. Data science juga digunakan untuk analisis data genetik dan penelitian klinis untuk mengembangkan terapi yang lebih efektif (Koetsier. J, 2020). Kemudian untuk bagian Keuangan, Industri keuangan menggunakan data science untuk analisis risiko, deteksi penipuan, dan pengembangan model kredit. Misalnya, bank menggunakan algoritma *Machine Learning* untuk menilai kelayakan kredit pemohon berdasarkan data historis mereka. Data science juga digunakan untuk analisis pasar saham, manajemen portofolio, dan perencanaan keuangan. Pada bidang Transportasi, Data science digunakan dalam transportasi untuk optimisasi rute, manajemen armada, dan analisis pola lalu lintas. Misalnya, perusahaan *ride-sharing* seperti Uber menggunakan data science untuk memperkirakan permintaan penumpang dan

mengoptimalkan alokasi kendaraan. Data science juga digunakan dalam pengembangan kendaraan otonom dan analisis data transportasi untuk meningkatkan efisiensi sistem transportasi, Bidang Ritel juga tidak luput dari dukungan data science, Dalam industri ritel, data science digunakan untuk manajemen persediaan, analisis penjualan, dan personalisasi pengalaman pelanggan. Misalnya, pengecer besar seperti Walmart menggunakan data science untuk mengelola inventaris dan memprediksi permintaan produk (Marr. B, 2018). Data science juga digunakan untuk analisis perilaku pelanggan dan pengembangan strategi penjualan yang lebih efektif. Selanjutnya dalam bidang Energi, Industri energi menggunakan data science untuk analisis konsumsi energi, optimisasi produksi, dan prediksi kegagalan peralatan. Misalnya, perusahaan utilitas menggunakan algoritma *Machine Learning* untuk memprediksi permintaan energi dan mengoptimalkan distribusi energi. Data science juga digunakan dalam analisis data sensor untuk mendeteksi anomali dan memprediksi pemeliharaan peralatan, dan terakhir pada bidang Pemerintahan, Data science digunakan oleh pemerintah untuk analisis data publik, perencanaan kebijakan, dan pengambilan keputusan. Misalnya, data science digunakan untuk analisis data sensus, prediksi hasil pemilu, dan evaluasi kebijakan publik. Data science juga digunakan dalam analisis data kriminal untuk mendeteksi pola dan mengembangkan strategi penegakan hukum yang lebih efektif.

# BAB 2 SEJARAH DAN EVOLUSI DATA SCIENCE

## Pendahuluan

Data Science, sebagai disiplin ilmu, memiliki akar yang panjang dan luas yang dapat ditelusuri kembali ke awal peradaban manusia. Pada dasarnya, Data Science adalah tentang pengumpulan, analisis, dan interpretasi data untuk memperoleh wawasan yang dapat digunakan dalam pengambilan keputusan. Sejak zaman kuno, manusia telah menggunakan berbagai metode untuk mengumpulkan dan menganalisis data. Misalnya, di Mesir kuno, catatan statistik digunakan untuk mengelola hasil panen dan sumber daya lainnya. Pada abad ke-17, John Graunt dan William Petty memperkenalkan statistik modern, dengan Graunt menerapkan analisis statistik pada data kelahiran dan kematian di London, yang menjadi dasar dari demografi. Abad ke-20 menandai perkembangan signifikan dalam bidang statistika dan komputasi. Pengenalan komputer pada pertengahan abad ke-20 membawa perubahan revolusioner dalam cara data dikumpulkan, disimpan, dan dianalisis. Konsep "data" mulai dipandang sebagai aset yang penting, terutama dalam konteks bisnis dan penelitian ilmiah. Pada tahun 1960-an dan 1970-an, munculnya basis data relasional dan pengembangan bahasa pemrograman seperti SQL memungkinkan pengolahan data dalam jumlah besar dengan lebih efisien.

Memasuki abad ke-21, kita menyaksikan ledakan jumlah data yang dihasilkan oleh berbagai sumber, termasuk media sosial, perangkat seluler, sensor IoT, dan transaksi digital. Era ini

sering disebut sebagai era *Big Data*. Pada tahun 2001, Doug Laney dari META Group (sekarang Gartner) memperkenalkan konsep "3V" - Volume, Velocity, dan Variety - untuk menggambarkan tantangan yang dihadapi dalam mengelola data besar. Volume mengacu pada jumlah data yang besar, Velocity pada kecepatan data dihasilkan dan diproses, dan Variety pada berbagai tipe data yang ada. Data Science sebagai disiplin yang terintegrasi mulai mendapatkan pengakuan pada awal 2000-an. Istilah "Data Scientist" pertama kali diperkenalkan oleh D.J. Patil dan Jeff Hammerbacher, yang bekerja di LinkedIn dan Facebook. Mereka menggambarkan Data Scientist sebagai individu yang mampu memadukan keahlian dalam pemrograman, statistika, dan domain bisnis untuk menganalisis data dan memberikan wawasan yang dapat ditindaklanjuti.

Salah satu perkembangan paling signifikan dalam Data Science adalah kemajuan dalam pembelajaran mesin (*Machine Learning*) dan kecerdasan buatan (*Artificial Intelligence*). Algoritma pembelajaran mesin memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Ini telah membuka berbagai aplikasi baru dalam berbagai bidang, seperti prediksi pasar, diagnosa medis, pengenalan gambar, dan lain-lain. Peningkatan kemampuan komputasi dan ketersediaan data dalam jumlah besar telah mendorong inovasi yang cepat dalam area ini. Meskipun telah banyak kemajuan, Data Science juga menghadapi berbagai tantangan. Isu privasi dan keamanan data, kualitas data, dan interpretabilitas model pembelajaran mesin adalah beberapa tantangan yang terus diperhatikan. Selain itu, ada kebutuhan yang terus berkembang untuk memastikan bahwa data dan algoritma yang digunakan bebas dari bias dan dapat diandalkan.

Melihat ke masa depan, Data Science diharapkan akan terus berkembang dengan integrasi teknologi baru seperti komputasi kuantum dan edge computing. Peran Data Scientist juga diperkirakan akan menjadi semakin penting dalam mengatasi masalah-masalah kompleks di berbagai sektor, mulai dari kesehatan, keuangan, hingga pemerintahan. Perjalanan Data Science dari metode statistik sederhana hingga teknologi canggih saat ini menunjukkan evolusi yang dinamis dan berkelanjutan. Dengan fondasi yang kuat dalam statistika dan komputasi, serta penerapan luas dalam pembelajaran mesin dan AI, Data Science terus menjadi bidang yang menarik dan sangat relevan di era digital ini. Dengan menghadapi tantangan dan memanfaatkan peluang baru, Data Science akan terus memainkan peran kunci dalam inovasi dan pengambilan keputusan berbasis data di masa depan.

## Akar Data Science

Data science memiliki akar yang kuat dalam statistik dan analisis data, yang dapat ditelusuri kembali ke akhir abad ke-19. Tokoh-tokoh seperti Francis Galton dan Karl Pearson memainkan peran penting dalam pengembangan metode statistik awal seperti korelasi dan regresi linear (Galton, 1889; Pearson, 1901). Galton dikenal karena kontribusinya dalam pengembangan konsep korelasi yang mendasari banyak analisis statistik modern, sementara Pearson terkenal dengan koefisien korelasinya yang digunakan luas hingga saat ini.

### Kontribusi Francis Galton dan Karl Pearson

1. Francis Galton
   Mengembangkan konsep korelasi yang mendasari banyak

analisis statistik modern. Misalnya, Galton memperkenalkan konsep regresi ke rata-rata yang menjelaskan bagaimana keturunan dari orang tua tinggi kemungkinan besar akan lebih pendek dari orang tua mereka dan lebih dekat ke rata-rata populasi. Ini membuka jalan bagi pengembangan lebih lanjut dalam teori regresi dan statistik.

2. Karl Pearson
   Mengembangkan koefisien korelasi yang masih digunakan luas hingga saat ini. Pearson juga memperkenalkan metode chi-square, yang merupakan teknik penting dalam statistik inferensial untuk menguji hipotesis tentang distribusi data. Metode ini memungkinkan para ilmuwan untuk menganalisis dan menarik kesimpulan dari data sampel dengan cara yang lebih sistematis dan terstruktur.

Pada pertengahan abad ke-20, perkembangan teknologi komputer membawa revolusi dalam analisis data. Komputer memungkinkan pengolahan data dalam jumlah besar dengan kecepatan yang jauh lebih tinggi dibandingkan metode manual.

### Teori Informasi oleh Claude Shannon

Claude Shannon, dalam karyanya tentang teori informasi yang diterbitkan pada tahun 1948, memberikan kontribusi signifikan pada dasar teori data science (Shannon, 1948). Shannon memperkenalkan konsep entropi informasi yang menjadi landasan bagi banyak perkembangan dalam komunikasi data dan kompresi data. Teori ini menjelaskan bagaimana data dapat dikodekan dan dikompresi secara efisien, yang sangat penting dalam pengembangan teknologi komunikasi dan penyimpanan data. Misalnya, kompresi data menggunakan

algoritma seperti Huffman coding dan teknik lainnya didasarkan pada prinsip-prinsip yang diperkenalkan oleh Shannon.

### Pengembangan *Database* Relasional oleh Edgar F. Codd

Kemajuan lebih lanjut terjadi pada tahun 1960-an dan 1970-an dengan pengembangan *database* relasional oleh Edgar F. Codd. Codd memperkenalkan model data relasional yang memungkinkan pengguna untuk membuat, membaca, memperbarui, dan menghapus data dengan efisien, yang menjadi dasar bagi banyak sistem manajemen basis data modern (Codd, 1970). Model relasional menggunakan tabel untuk menyimpan data dan memungkinkan penggunaan bahasa query, seperti SQL, untuk mengelola data. Ini merevolusi cara data disimpan dan diakses, membuatnya lebih mudah untuk memanipulasi data dalam skala besar. Sistem manajemen basis data relasional (RDBMS) seperti Oracle, MySQL, dan SQL Server didasarkan pada model yang dikembangkan oleh Codd.

### Munculnya Big Data

Pada era 1990-an, konsep *Big Data* mulai muncul seiring dengan ledakan data di era internet. Dengan meningkatnya penggunaan internet dan teknologi digital, jumlah data yang dihasilkan meningkat secara eksponensial. *Big Data* tidak hanya tentang volume, tetapi juga variasi dan kecepatan data yang dihasilkan. Ini memerlukan teknik dan alat baru untuk mengelolanya. Artikel Gil Press di Forbes menyatakan bahwa ledakan data ini memerlukan pendekatan baru untuk mengelola dan menganalisis data (Press, 2013). Data besar ini mencakup berbagai jenis data, seperti teks, gambar, dan video, serta data yang dihasilkan secara *real-time* dari berbagai sumber, termasuk media sosial, sensor, dan perangkat IoT. Misalnya, platform

media sosial seperti Facebook dan Twitter menghasilkan jutaan data setiap detik, menciptakan tantangan dan peluang baru bagi para ilmuwan data untuk mengelola dan menganalisis informasi ini secara efektif.

## Evolusi Data Science

Istilah "data science" diperkenalkan oleh William S. Cleveland pada tahun 2001, di mana ia mengusulkan penggabungan statistik dengan ilmu komputer untuk membentuk disiplin baru yang berfokus pada analisis data besar (Cleveland, 2001). Cleveland berpendapat bahwa data science harus mencakup pemrograman komputer, keterampilan statistik, dan pemahaman domain spesifik untuk memberikan wawasan yang berharga dari data. Ini mengubah cara pandang tradisional terhadap analisis data, yang sebelumnya didominasi oleh metode statistik murni.

### Penggunaan dalam Berbagai Bidang

Sejak itu, data science berkembang pesat dengan aplikasi dalam berbagai bidang seperti kesehatan, keuangan, dan pemasaran (Provost & Fawcett, 2013).

1. Kesehatan
   Data science digunakan untuk analisis data pasien, prediksi penyakit, dan personalisasi perawatan. Misalnya, dengan menggunakan teknik pembelajaran mesin, data science dapat membantu dalam mendeteksi pola yang menunjukkan awal mula penyakit tertentu, seperti kanker atau diabetes, sebelum gejala klinis muncul. Ini memungkinkan intervensi medis yang lebih dini dan lebih efektif. Selain itu, data science digunakan dalam analisis

genetik untuk memahami penyakit yang diwariskan dan mengembangkan terapi gen.

2. Keuangan
   Data science digunakan untuk deteksi penipuan, analisis risiko, dan pengambilan keputusan investasi. Dengan menganalisis pola transaksi, algoritma pembelajaran mesin dapat mendeteksi aktivitas yang mencurigakan yang mungkin menandakan penipuan. Hal ini membantu lembaga keuangan mengurangi kerugian akibat penipuan dan membuat keputusan investasi yang lebih baik. Data science juga digunakan dalam analisis pasar untuk memprediksi tren harga saham dan mengidentifikasi peluang investasi.
3. Pemasaran
   Data science membantu memahami perilaku konsumen, segmentasi pasar, dan optimalisasi kampanye pemasaran. Perusahaan dapat menggunakan data dari interaksi pelanggan, pembelian, dan media sosial untuk menargetkan iklan secara lebih efektif dan meningkatkan kepuasan pelanggan. Dengan menganalisis data pelanggan, perusahaan dapat mengidentifikasi tren dan preferensi yang membantu mereka merancang produk dan layanan yang lebih sesuai dengan kebutuhan pasar. Data science juga memungkinkan personalisasi pengalaman pelanggan dengan memberikan rekomendasi produk yang relevan berdasarkan riwayat pembelian dan preferensi.

## Kemajuan dalam Pembelajaran Mesin dan Kecerdasan Buatan

Kemajuan dalam pembelajaran mesin (*Machine*

*Learning*) dan kecerdasan buatan (*Artificial Intelligence*) dalam dekade terakhir telah menjadi pendorong utama evolusi data science. Teknik pembelajaran mendalam (*Deep Learning*) memungkinkan komputer untuk belajar dari data dengan cara yang meniru cara kerja otak manusia, membuka peluang baru untuk analisis prediktif dan pengambilan keputusan otomatis (Goodfellow, Bengio, & Courville, 2016). *Deep Learning* menggunakan jaringan saraf tiruan yang terdiri dari banyak lapisan (layer) untuk mempelajari representasi data yang kompleks dan non-linier.

### Contoh Aplikasi

1. Pengenalan Wajah dan Suara
   Algoritma *Deep Learning* digunakan dalam aplikasi pengenalan wajah dan suara untuk keamanan dan kenyamanan pengguna. Misalnya, teknologi pengenalan wajah digunakan dalam sistem keamanan untuk mengenali individu, sedangkan pengenalan suara digunakan dalam asisten virtual seperti Siri dan Alexa. Algoritma ini mampu mengenali pola wajah dan suara dengan akurasi tinggi, bahkan dalam kondisi yang sulit seperti pencahayaan yang buruk atau kebisingan latar belakang.
2. Kendaraan Otonom
   Teknologi ini digunakan dalam kendaraan otonom untuk memproses data dari berbagai sensor dan membuat keputusan mengemudi secara *real-time*. Kendaraan otonom dapat menganalisis lingkungan sekitarnya dan mengambil tindakan yang sesuai untuk mengemudi dengan aman tanpa intervensi manusia. Misalnya, kendaraan otonom menggunakan sensor LIDAR, kamera,

dan radar untuk mendeteksi objek di sekitarnya dan menghindari tabrakan. Algoritma pembelajaran mendalam memungkinkan kendaraan untuk belajar dari pengalaman mengemudi sebelumnya dan meningkatkan kinerja mereka dari waktu ke waktu.

## Tantangan dan Masa Depan Data Science

Meskipun data science telah mencapai banyak kemajuan, tantangan seperti privasi data, bias dalam algoritma, dan kebutuhan untuk keterampilan teknis yang lebih tinggi masih perlu diatasi.

### Tantangan Privasi Data dan Bias Algoritma

1. Privasi Data
   Masalah privasi data menjadi semakin penting dengan meningkatnya jumlah data pribadi yang dikumpulkan dan dianalisis. Organisasi perlu memastikan bahwa mereka mengumpulkan dan menggunakan data dengan cara yang transparan dan sesuai dengan peraturan privasi, seperti GDPR di Eropa. Ini termasuk memberi tahu pengguna tentang bagaimana data mereka digunakan dan memberikan opsi untuk mengontrol penggunaan data pribadi mereka. Selain itu, organisasi perlu mengimplementasikan langkah-langkah keamanan yang kuat untuk melindungi data dari akses tidak sah dan kebocoran data.
2. Bias Algoritma
   Algoritma pembelajaran mesin dapat memperkuat bias yang ada dalam data, yang dapat menyebabkan keputusan yang tidak adil. Misalnya, algoritma perekrutan otomatis

yang dilatih dengan data dari perusahaan yang memiliki bias gender dalam sejarah perekrutan mereka mungkin cenderung mengabaikan kandidat dari kelompok tertentu. Penting untuk mengembangkan metode untuk mendeteksi dan mengurangi bias dalam algoritma untuk memastikan keadilan dan akurasi. Salah satu pendekatan untuk mengurangi bias adalah dengan menggunakan teknik *fairness-aware machine learning* yang dirancang untuk mengidentifikasi dan memperbaiki bias dalam data dan model.

### Integrasi Teknologi Baru

Di masa depan, data science diharapkan akan terus berkembang dengan integrasi teknologi baru seperti komputasi kuantum dan *Internet of Things* (IoT) (Marr, 2018).

1. Komputasi Kuantum
   Menjanjikan kemampuan untuk memproses data dalam jumlah besar dengan kecepatan yang jauh lebih tinggi daripada komputer klasik, yang dapat membuka peluang baru untuk analisis data yang kompleks. Komputasi kuantum menggunakan prinsip-prinsip mekanika kuantum untuk memproses informasi dengan cara yang lebih efisien daripada komputer tradisional. Misalnya, algoritma kuantum seperti Shor's algorithm dapat memecahkan masalah faktorisasi bilangan bulat dengan lebih cepat daripada algoritma klasik, yang memiliki implikasi besar untuk keamanan kriptografi.
2. IoT
   Menghasilkan data dalam jumlah besar dari perangkat yang saling terhubung, menciptakan peluang baru untuk analisis data. Misalnya, data dari sensor IoT dalam *smart*

*city* dapat digunakan untuk mengoptimalkan manajemen lalu lintas dan mengurangi kemacetan. Sensor-sensor ini dapat mengumpulkan data tentang berbagai aspek lingkungan, seperti kualitas udara, suhu, dan kelembaban, yang dapat digunakan untuk meningkatkan kualitas hidup di kota-kota pintar. Selain itu, IoT digunakan dalam berbagai aplikasi industri untuk pemantauan dan pengendalian proses secara *real-time*, meningkatkan efisiensi operasional dan mengurangi biaya.

### Pendidikan dan Pelatihan

Pendidikan dan pelatihan dalam data science menjadi semakin penting seiring dengan meningkatnya permintaan akan keterampilan data science di pasar kerja. Banyak universitas dan institusi pendidikan menawarkan program studi dalam data science, mulai dari tingkat sarjana hingga doktor. Program-program ini sering kali mencakup topik seperti pemrograman, statistik, pembelajaran mesin, dan visualisasi data. Selain itu, terdapat banyak kursus online dan bootcamp yang menawarkan pelatihan intensif dalam data science. Program-program ini dirancang untuk memberikan keterampilan praktis yang dapat langsung diterapkan dalam industri. Misalnya, kursus online di platform seperti Coursera dan edX menawarkan pelatihan dalam berbagai topik data science yang diajarkan oleh pakar industri dan akademisi terkemuka.

### Etika dalam Data Science

Etika menjadi aspek yang semakin penting dalam data science. Dengan meningkatnya penggunaan data pribadi dan keputusan otomatis yang dibuat oleh algoritma, penting untuk

memastikan bahwa data digunakan dengan cara yang etis dan bertanggung jawab. Isu-isu seperti privasi data, keamanan data, dan bias dalam algoritma perlu diperhatikan dengan serius. Organisasi perlu memastikan bahwa mereka mengumpulkan dan menggunakan data dengan cara yang transparan dan sesuai dengan peraturan privasi. Selain itu, penting untuk mengembangkan dan menerapkan algoritma yang adil dan tidak bias untuk memastikan bahwa keputusan yang dibuat berdasarkan data adalah adil dan tidak diskriminatif. Contoh pentingnya etika dalam data science adalah kasus Cambridge Analytica, di mana data pribadi jutaan pengguna Facebook digunakan tanpa izin untuk mempengaruhi pemilihan politik. Kasus ini menyoroti perlunya regulasi yang lebih ketat dan praktik yang lebih etis dalam pengumpulan dan penggunaan data.

Masa depan data science penuh dengan peluang dan tantangan. Dengan terus berkembangnya teknologi, data science akan terus menjadi bidang yang dinamis dan penting. Integrasi dengan teknologi baru seperti komputasi kuantum dan IoT akan membuka peluang baru untuk analisis data. Namun, penting juga untuk mengatasi tantangan yang ada, seperti privasi data dan bias algoritma, untuk memastikan bahwa data science digunakan dengan cara yang etis dan bertanggung jawab. Selain itu, kolaborasi antara akademisi, industri, dan pemerintah akan menjadi kunci untuk mengembangkan solusi inovatif yang dapat mengatasi tantangan ini dan memanfaatkan potensi penuh dari data science.

# BAB 3 PROSES DALAM DATA SCIENCE

## Pendahuluan

Proses data science adalah tahapan yang diikuti untuk mengolah dan menganalisis *Big Data* untuk menghasilkan informasi bermanfaat, membuat prediksi atau membantu pengambilan keputusan. Proses ini mencakup beberapa tahapan penting, mulai dari memahami masalah yang dihadapi, mengumpulkan dan membersihkan data, menganalisis dan memodelkan data, dan kemudian mengkomunikasikan hasil dan menerapkan solusi (Rismayani et al., 2024).

Proses data science melibatkan sejumlah langkah atau tahapan sebagai berikut (Gede Aditra Pradnyana, n.d.).

1. Identifikasi masalah dan tentukan tujuan
   Tahap identifikasi masalah dan penetapan tujuan adalah langkah awal yang krusial dalam data science. Pada tahap ini, kita perlu memahami permasalahan bisnis yang ingin dipecahkan dengan data science (Mirqotussa'adah, et all., 2019).
2. Akuisisi Data
   Tahap akuisisi data adalah mengumpulkan data yang akan digunakan untuk Analisa. Data bisa berasal dari berbagai sumber seperti basis data, sensor, web scraping dan eksperimen.
3. Pembersihan dan Pemrosesan Awal Data
   Tahap pembersihan dan pemrosesan awal data menyiapkan data agar bisa dianalisa lebih lanjut. Data yang tidak bersih dan terproses dengan baik bisa

menghasilkan kesimpulan yang salah dan model yang tidak akurat (Joseph Santoso, 2023). Hal ini mungkin melibatkan penghapusan kesalahan, mengisi nilai yang hilang, dan memformat data secara konsisten.

4. Analisis Data Eksplorasi (EDA)
   Tahap Analisis Data Eksplorasi (*Exploratory Data Analysis* - EDA) yang mengamati dan memahami data untuk menemukan pola dan tren. EDA membantu data scientist menjadi familiar dengan data sebelum melakukan analisis yang lebih kompleks (Joseph Santoso, 2023).
5. Pemodelan
   Tahap pemodelan (*modeling*) adalah tahap krusial dalam data science dimana dibuat representasi matematis dari hubungan yang ditemukan dalam data. Model ini pada dasarnya adalah rumus atau program komputer yang mempelajari data untuk membuat prediksi atau klasifikasi baru.
6. Interpretasi
   Tahap interpretasi data dalam proses data science menjelaskan arti dari pola dan tren yang ditemukan dalam data. Pada tahap ini, data scientist tidak hanya melihat angka-angka saja, tetapi juga berupaya untuk menceritakan histori yang disampaikan oleh data tersebut.
7. Evaluasi
   Tahap evaluasi menilai performa model yang telah dibangun. Evaluasi memastikan model dapat digunakan dengan benar dan menghasilkan prediksi akurat untuk data yang tidak terlihat.
8. *Deployment*

Tahap deployment adalah tahap menggunakan model yang sudah selesai dibangun untuk menghasilkan sesuatu yang bermanfaat.

## Identifikasi Masalah dan Tentukan Tujuan Bisnis

Identifikasi masalah adalah langkah awal yang krusial dalam proses data science. Pada tahap ini perlu memahami permasalahan bisnis yang akan dipecahkan dengan data science. Data scientist mengidentifikasi kebutuhan organisasi dan menggunakannya untuk menemukan jawaban atau Solusi dari masalah yang dihadapi. Selain itu, data scientist harus mampu merumuskan insights spesifik yang akan digali dari data. (Program Data Science, 2020). Dengan identifikasi masalah yang baik dapat menentukan arah yang tepat untuk proses data science selanjutnya.

Beberapa hal yang dilakukan pada tahap identifikasi masalah dapat dilihat pada tabel 3.1 berikut.

*Tabel 3.1 Kegiatan Identifikasi MasalahSilahkan tambahkan pendahuluan sesuai dengan tema bab-nya, paling sedikit 3 paragrap*

| Kegiatan | Keterangan |
|---|---|
| Mengerti tujuan bisnis | Diskusi dengan stakeholders untuk memahami tujuan dan target yang ingin dicapai |
| Mendefinisikan pertanyaan penelitian | Menerjemahkan tujuan bisnis menjadi pertanyaan terstruktur yang bisa dijawab menggunakan data |
| Evaluasi kelayakan | Mempertimbangkan ketersediaan data, sumber daya, dan keahlian yang dibutuhkan untuk menyelesaikan masalah |

Identifikasi masalah dan penetapan tujuan yang tepat, kita dapat mengarahkan analisa data ke arah yang benar dan menghasilkan solusi yang efektif melalui data science.

## Akuisisi Data

Data scientist harus memiliki data yang dibutuhkan berdasarkan insights yang akan digali. Data scientist membutuhkan bantuan dari praktisi lain, khususnya data engineer yang tugasnya lebih berfokus pada infrastruktur dan sistem pengelolaan data organisasi. Tugas pengumpulan data yang kompleks memerlukan akses ke berbagai sumber data dalam sistem yang besar. Jika beberapa data belum terekam di sistem organisasi tetapi tersedia di luar organisasi, data scientist (mungkin dengan bantuan data engineer) perlu mengambil data tersebut. Pada tahap ini, dilakukan pengumpulan data yang relevan untuk menjawab permasalahan yang ingin dipecahkan.

Beberapa hal yang dikukan pada tahap akuisisi data dapat dilihat pada tabel 3.2 berikut.

*Tabel 3.2 Kegiatan akuisisi data*

| Kegiatan | Keterangan |
|---|---|
| Pendefinisian sumber data | Mencari dan mengidentifikasi sumber data yang sesuai dengan kebutuhan analisis Sumber data bisa berasal dari internal perusahaan, sumber publik, atau pembelian data pihak ketiga |
| Pengumpulan data | Mengumpulkan data dari sumber yang sudah diidentifikasi. Teknik pengumpulan data bisa berbeda-beda, misalnya scraping web, menggunakan API atau mengimpor data dari file. |
| Transformasi data | Mengubah data ke dalam format yang bisa diterima dan diolah pada tahap selanjutnya. Proses ini bisa meliputi pembersihan data dan formatting data. |

Dengan akuisisi data yang baik, pastikan memiliki data yang relevan dan cukup untuk proses selanjutnya dalam data science (Santoso et al., 2020).

## Pembersihan dan Pemrosesan Data Awal

Salah satu tahapan dalam proses data science yang menentukan kualitas dan keandalan hasil analisis adalah pembersihan dan pemrosesan data awal. Proses pembersihan data adalah langkah penting dalam persiapan data sebelum memasuki tahap analisis data karena kualitas input menentukan kualitas hasil pengolahan. Data yang diperoleh dari sumbernya biasanya masih "kotor" dan tidak siap untuk digunakan langsung dalam model. Istilah "GIGO" (*Garbage In, Garbage Out*) digunakan untuk menggambarkan situasi tersebut. Tahapan ini melibatkan berbagai metode dan teknik untuk menangani masalah dalam data mentah seperti duplikasi, kesalahan input, outlier dan nilai yang hilang. Pembersihan dan pemrosesan data awal bertujuan untuk meningkatkan kualitas data sehingga hasil analisis menjadi lebih akurat dan bisa diandalkan.

Beberapa kegiatan dalam pembersihan dan pemrosesan data awal dapat dilihat pada tabel 3.3 berikut.

*Tabel 3.3 Kegiatan pembersihan dan pemrosesan data awal*

| Kegiatan | Keterangan |
|---|---|
| Pembersihan Data (Data Cleaning) | – Menangani data yang hilang. Data hilang bisa diisi dengan nilai estimasi atau dihapuskan tergantung kasusnya.<br>– Penanganan outlier. Outlier adalah data yang nilainya jauh berbeda dari data lainnya. Outlier bisa diinvestigasi lebih lanjut atau dihapus tergantung kasusnya<br>– Format data. Memastikan data memiliki format yang konsisten, misalnya format tanggal atau format huruf |
| Transformasi Data (Data Transformation) | – Skala data. Beberapa algoritma *Machine Learning* mengharuskan data memiliki skala yang sama.<br>– Pembuatan fitur baru. Data baru bisa dibuat dari kombinasi fitur yang ada untuk meningkatkan performa model |

Pembersihan dan pemrosesan awal data yang baik, dapat meningkatkan kualitas data dan mendapatkan hasil analisis yang lebih akurat.

## Analisis Data Eksplorasi

Analisis Data Eksplorasi (EDA) adalah tahap awal yang penting untuk memahami karakteristik data sebelum dilakukan analisis lebih lanjut. Pada tahap ini, secara ringkas mengeksplorasi data untuk menemukan pola tersembunyi, anomali, dan untuk mendapatkan gambaran umum tentang data.

Jika data sudah tersedia, maka analisis data lebih mudah dilakukan. Data scientist sudah memahami metode, algoritma, teknologi atau alat yang akan digunakan. Metode atau algoritma yang paling sesuai dipilih berdasarkan insights yang akan digali yang dapat berasal dari algoritma pembelajaran mesin, yang merupakan bagian dari kecerdasan buatan atau kecerdasan buatan. Agar dapat memilih algoritma yang tepat, data scientist harus memahami data yang ditangani, perilaku, prinsip kerja dan kelebihan dan kekurangan berbagai algoritma. Jika tujuannya adalah untuk membuat model, algoritma digunakan untuk mengolah data yang telah disiapkan untuk menghasilkan model, misalnya model klasifikasi atau prediksi. Selanjutnya model dievaluasi untuk memastikan memenuhi persyaratan tertentu. Untuk menguji model, seperti menguji keakuratan model prediksi, data scientist harus memahami metode pengukuran model dan memilih metode yang tepat. Selanjutnya hasil uji dinilai. Model mungkin tidak dapat digunakan jika kualitasnya buruk, jadi perlu dibuat lagi. Jika ada kemungkinan bahwa data masukan berbeda, tahap pertama harus diulangi dan dilanjutkan sampai hasil analisis data memadai (Program Data

Science, 2020).

Beberapa kegiatan dalam Analisis Data Eksplorasi, dapat dilihat pada tabel 3.4 berikut.

*Tabel 3.4 Kegiatan analisis data eksplorasi*

| Kegiatan | Keterangan |
|---|---|
| Mendeskripsikan data | EDA membantu kita untuk memahami struktur data, seperti jumlah data, jenis variabel, dan nilai statistik dasar |
| Membersihkan data | EDA dapat membantu mengidentifikasi data yang hilang, outlier, dan kesalahan pada data |
| Membangun hipotesis | Melalui EDA kita bisa menemukan hubungan antar variabel yang mungkin bisa dijadikan hipotesis untuk diuji selanjutnya |
| Mempersiapkan data untuk analisis lanjut | EDA membantu kita memilih fitur yang relevan dan melakukan transformasi data yang diperlukan untuk tahap modeling |

Dengan melakukan EDA yang baik, kita bisa meningkatkan kualitas analisis data dan mendapatkan hasil yang lebih akurat.

## Pemodelan

Pemodelan data adalah langkah penting dalam data science untuk menyederhanakan dan mengatur data. Proses ini membuat data menjadi lebih terstruktur dan mudah dipahami. Model data akan digunakan untuk membangun basis penyimpanan data. Model data dibuat untuk mencapai tujuan yang diinginkan. Metode regresi dan prediksi digunakan untuk memperkirakan nilai di waktu mendatang, serta untuk mengklasifikasikan dan mengelompokkan nilai yang ada dalam kumpulan data.

Berikut adalah beberapa manfaat dari pemodelan data, yaitu:

1. Memudahkan pengelolaan data: Model data yang baik membuat pengelolaan data menjadi lebih efisien.

2. Meminimalkan redundansi: Pemodelan data dapat membantu mengurangi duplikasi data.
3. Meningkatkan komunikasi: Model data menjadi bahasa yang mudah dipahami antara tim teknis dan non-teknis.

Beberapa kegiatan pada pemodelan data dapat dilihat pada tabel 3.5 berikut.

*Tabel 3.5 Kegiatan pemodelan data*

| Kegiatan | Keterangan |
|---|---|
| Pendefinisian Kebutuhan | Mendefinisikan tujuan dan kebutuhan analisis data |
| Perancangan Model | Membuat model yang menggambarkan entitas (data) dan hubungan antar entitas |
| Dokumentasi | Mendokumentasikan model data yang telah dibuat |

Pemodelan yang efektif membutuhkan pemilihan algoritma yang tepat, pelatihan model dengan data yang baik, dan evaluasi kinerja model untuk memastikan keakuratan dan generalisasinya.

## Interpretasi

Interpretasi data merupakan proses dimana model dan data diinterpretasikan. Orang awam yang tidak memahami istilah teknis harus dapat memahami hasil interpretasi data. Data yang diperoleh digunakan dalam presentasi untuk menjawab pertanyaan bisnis. Selain itu, penting untuk memiliki kemampuan komunikasi yang baik pada tahap interpretasi data agar poin penting dapat disampaikan secara efektif kepada semua pihak yang berkepentingan.

Seorang data scientist harus dapat mengkomunikasikan proses dan hasil analisis data dengan cara yang sistematis, menarik, tidak ambigu dan mudah dipahami bagi yang berkepentingan dengan proses dan hasil. Komunikasi dapat

dilakukan secara tatap muka pada rapat atau seminar atau secara tertulis dalam bentuk laporan, bergantung pada kebutuhan organisasi tempat data scientist bekerja. Hasil yang dikomunikasikan melalui pemaparan harus dipresentasikan dalam bentuk visual yang tepat agar menarik dan mudah dipahami. Data scientist harus dapat membuat laporan yang sistematis, jelas dan berkualitas tinggi serta menguasai teknik presentasi yang efektif.

Beberapa kegiatan yang dilakukan pada tahap interpretasi, dapat dilihat pada tabel 3.6 berikut.

*Tabel 3.6 Kegiatan Interpretasi Data*

| Kegiatan | Keterangan |
|---|---|
| Menjelaskan hasil model | Menjelaskan bagaimana model yang dihasilkan bekerja dan kemungkinan bias yang ada. |
| Visualisasi data | Menggunakan visualisasi data untuk mengungkapkan pola dan hubungan tersembunyi dalam data |
| Komunikasi hasil | Kita harus bisa menyampaikan hasil analisis data dengan jelas kepada orang yang awam dengan ilmu data |

Interpretasi data adalah tahap terakhir yang penting dalam proses data science. Pada tahap ini dijelaskan arti dari hasil analisis data. Interpretasi yang baik akan membantu memperoleh wawasan yang berguna dari data.

Dengan interpretasi yang tepat, penggunaan hasil data science untuk membuat keputusan dan pemecahan masalah yang tepat sasaran.

## Evaluasi

Evaluasi model adalah tahap penting untuk memastikan kinerja model yang telah dibangun. Dengan evaluasi, kita dapat mengetahui apakah model bisa memberikan prediksi dan hasil yang akurat sebelum digunakan untuk hal-hal yang lebih

penting.

Pada tahap ini, kita menggunakan berbagai metrik evaluasi untuk mengukur kelebihan dan kelemahan model. Beberapa contoh metrik evaluasi adalah akurasi, presisi, dan *recall*.

Beberapa hal yang dilakukan pada tahap evaluasi, dapat dilihat pada tabel 3.7 berikut.

*Tabel 3.7 Kegiatan Evaluasi*

| Kegiatan | Keterangan |
|---|---|
| Membandingkan model dengan data uji | Membandingkan hasil prediksi model dengan data yang belum pernah dilihat model sebelumnya (data uji). Dengan cara ini bisa dilihat kemampuan generalisasi model. |
| Menganalisis kesalahan mode | Kesalahan yang dibuat oleh model untuk mengetahui kelemahannya perlu dianalisis. Dengan begitu, kita bisa memperbaiki model agar performanya menjadi lebih baik. |

Evaluasi yang baik akan menghasilkan model yang:

1. Akurat
   Model bisa memberikan prediksi yang mendekati kenyataan.
2. Andal
   Model bisa diandalkan untuk menghasilkan prediksi yang konsisten.
3. Generalizable
   Model bisa memberikan performa yang baik pada data yang belum pernah dilihat sebelumnya.

Dengan evaluasi yang baik, kita bisa mengetahui apakah model perlu diperbaiki atau bahkan dibangun kembali. Dengan begitu, kita bisa membuat keputusan dan pemecahan masalah yang lebih tepat menggunakan data science.

## *Deployment* (Penerapan Model)

Tahap penggunaan model dalam data science dikenal dengan istilah deployment. Setelah model selesai dibangun dan dievaluasi, tahap ini berfokus pada penggunaan model untuk menghasilkan prediksi atau keputusan bisnis.

Beberapa kegiatan yang dilakukan pada tahap penggunaan model dapat dilihat pada tabel 3.8 berikut.

*Tabel 3.8 Kegiatan penggunaan model (deployment)*

| Kegiatan | Keterangan |
|---|---|
| Memintegrasikan model ke dalam sistem | Model perlu diintegrasikan dengan sistem lain yang akan memanfaatkan hasil prediksi. Misalnya, model prediksi harga saham bisa diintegrasikan ke dalam platform trading saham |
| Monitoring performa model | Model yang sudah diterapkan perlu dimonitor secara berkala untuk memastikan kinerjanya tetap baik. Jika performa menurun, model perlu diperbaiki atau bahkan dibangun ulang dengan data baru |
| Memberikan interpretasi hasil | Tidak cukup hanya dengan hasil prediksi, pengguna perlu memahami alasan di balik prediksi tersebut. Hal ini bisa membantu pengguna mengambil keputusan yang lebih baik |

Deployment merupakan tahap krusial dalam data science karena dampak bisnis biasanya baru terlihat pada tahap ini. Keberhasilan deployment bergantung pada kerjasama antara data scientist dan tim IT.

# BAB 4 PERAN DATA SCIENTIST DI ERA DIGITAL

## Pendahuluan

Di era digital saat ini, data memainkan peran yang sangat penting dalam pengambilan keputusan bisnis dan pengembangan teknologi. Volume data yang dihasilkan setiap hari oleh berbagai sumber seperti media sosial, transaksi bisnis, dan perangkat *Internet of Things* (IoT) terus meningkat secara eksponensial (Press, 2013). Data ini memiliki potensi besar untuk memberikan wawasan berharga yang dapat digunakan untuk meningkatkan efisiensi operasional, mengidentifikasi peluang bisnis, dan menciptakan produk serta layanan baru yang lebih baik.

Data scientist adalah para profesional yang dilatih untuk mengumpulkan, menganalisis, dan menginterpretasikan data ini. Mereka menggunakan berbagai teknik dan alat untuk mengekstrak informasi berharga dari data mentah dan mengubahnya menjadi wawasan yang dapat digunakan untuk pengambilan keputusan (Provost & Fawcett, 2013). Peran data scientist menjadi semakin penting seiring dengan meningkatnya ketergantungan bisnis pada data untuk mendukung strategi dan operasi mereka.

Namun, dengan berkembangnya teknologi dan volume data yang terus meningkat, data scientist juga menghadapi berbagai tantangan. Tantangan ini mencakup masalah kualitas data, privasi dan keamanan data, serta kebutuhan untuk terus mengembangkan keterampilan mereka agar tetap relevan dengan perkembangan teknologi terbaru (Marr, 2018). Dalam

sub bab ini, akan mengeksplorasi peran data scientist di era digital, keterampilan yang diperlukan, serta tantangan dan peluang yang dihadapi.

## Definisi dan Tanggung Jawab Data Scientist

Data scientist bertanggung jawab untuk mengumpulkan, menganalisis, dan menginterpretasikan data dalam jumlah besar. Mereka menggunakan berbagai teknik dan alat untuk mengekstrak informasi berharga yang dapat digunakan untuk membuat keputusan bisnis yang lebih baik (Davenport & Patil, 2012).

1. Pengumpulan Data
   Data scientist mengumpulkan data dari berbagai sumber, termasuk *database* internal, aplikasi web, dan sumber data eksternal. Mereka harus memahami bagaimana data dikumpulkan dan disimpan untuk memastikan integritas dan kualitas data (Provost & Fawcett, 2013). Misalnya, data dapat dikumpulkan dari transaksi penjualan, interaksi pelanggan, sensor IoT, media sosial, dan berbagai sumber lainnya. Tantangan utama dalam pengumpulan data adalah memastikan bahwa data yang diperoleh adalah representatif, akurat, dan relevan untuk analisis yang akan dilakukan.
   Contoh nyata dari pengumpulan data adalah dalam industri *e-commerce*, di mana data tentang perilaku pelanggan di situs web dikumpulkan untuk menganalisis pola pembelian dan preferensi produk. Data ini dapat mencakup informasi tentang halaman yang dikunjungi, produk yang dilihat, dan transaksi yang dilakukan. Analisis data ini memungkinkan perusahaan untuk mengoptimalkan pengalaman pelanggan dan

meningkatkan penjualan.

2. Pembersihan dan Pengolahan Data
   Setelah data dikumpulkan, langkah berikutnya adalah pembersihan data. Ini termasuk mengidentifikasi dan memperbaiki kesalahan dalam data, menangani nilai yang hilang, dan mengubah data ke format yang sesuai untuk analisis lebih lanjut (Goodfellow, Bengio, & Courville, 2016). Proses ini dikenal sebagai data wrangling atau data munging. Pembersihan data adalah langkah penting karena data yang kotor atau tidak akurat dapat menghasilkan analisis yang menyesatkan. Misalnya, nilai yang hilang dapat diimputasi menggunakan metode statistik, dan outlier dapat diidentifikasi dan ditangani untuk mencegah distorsi dalam analisis.
   Dalam konteks analisis kesehatan, data dari pasien sering kali tidak lengkap atau tidak konsisten. Pembersihan data dalam kasus ini mungkin melibatkan pengisian data yang hilang berdasarkan catatan medis lainnya atau menggunakan metode statistik untuk mengestimasi nilai yang hilang. Data yang bersih dan konsisten sangat penting untuk membuat model prediktif yang akurat dalam diagnosis dan perawatan pasien.
3. Analisis Data
   Data scientist menggunakan berbagai teknik analisis data, termasuk statistik, *Machine Learning*, dan pembelajaran mendalam, untuk menemukan pola dan wawasan dalam data. Mereka juga menggunakan alat visualisasi data untuk membantu dalam interpretasi hasil analisis (Provost & Fawcett, 2013). Teknik-teknik ini memungkinkan data scientist untuk mengidentifikasi tren, pola, dan anomali dalam data yang mungkin tidak

terlihat dengan metode analisis sederhana. Contohnya, analisis regresi dapat digunakan untuk memprediksi nilai masa depan berdasarkan tren historis, sementara algoritma *Clustering* dapat mengelompokkan data menjadi segmen-segmen yang berbeda berdasarkan karakteristik yang serupa.

Misalnya, dalam industri keuangan, data scientist dapat menggunakan analisis regresi untuk memprediksi harga saham berdasarkan data historis. Selain itu, algoritma *Clustering* dapat digunakan untuk segmentasi pelanggan, mengidentifikasi kelompok pelanggan dengan karakteristik serupa untuk mengembangkan strategi pemasaran yang lebih efektif. Pembelajaran mendalam juga diterapkan dalam analisis sentimen di media sosial untuk memahami persepsi publik terhadap merek atau produk tertentu.

4. Interpretasi dan Presentasi Hasil

   Setelah analisis selesai, data scientist harus menginterpretasikan hasil dan menyampaikan temuan mereka kepada pemangku kepentingan non-teknis. Ini sering kali melibatkan pembuatan laporan, dashboard, dan presentasi yang mudah dipahami (Davenport & Patil, 2012). Visualisasi data adalah alat penting dalam tahap ini, karena grafik dan diagram dapat membantu menjelaskan temuan dengan lebih jelas dibandingkan dengan tabel angka saja. Data scientist harus dapat menjelaskan implikasi bisnis dari temuan mereka dan memberikan rekomendasi yang dapat diimplementasi-kan untuk meningkatkan kinerja organisasi.

Contoh penerapan interpretasi hasil dapat dilihat dalam presentasi temuan analisis pasar kepada tim manajemen

perusahaan. Data scientist dapat menggunakan dashboard interaktif untuk menunjukkan tren penjualan, segmentasi pelanggan, dan prediksi permintaan produk. Dengan menggunakan alat seperti Tableau atau Power BI, mereka dapat membuat visualisasi yang dinamis dan mudah dipahami yang membantu manajemen dalam mengambil keputusan strategis.

## Keterampilan dan Teknologi yang Diperlukan

Untuk berhasil dalam peran ini, data scientist harus memiliki berbagai keterampilan teknis dan non-teknis.

### Keterampilan Teknis

1. Statistik dan Matematika
   Data scientist harus memiliki pemahaman yang kuat tentang konsep statistik dan matematika untuk menganalisis data secara efektif. Mereka harus memahami konsep seperti distribusi probabilitas, pengujian hipotesis, analisis regresi, dan teknik statistik lainnya (Provost & Fawcett, 2013).
2. Pemrograman
   Keterampilan pemrograman dalam bahasa seperti Python, R, dan SQL sangat penting untuk mengolah dan menganalisis data. Python dan R adalah bahasa pemrograman yang populer di kalangan data scientist karena memiliki banyak pustaka dan alat untuk analisis data, seperti NumPy, pandas, dan Scikit-learn untuk Python, serta dplyr dan ggplot2 untuk R. SQL digunakan untuk mengelola dan mengakses data dalam sistem manajemen basis data (Goodfellow, Bengio, & Courville, 2016).

3. Pembelajaran Mesin dan AI
   Pengetahuan tentang algoritma *Machine Learning* dan kecerdasan buatan adalah kunci untuk membuat model prediktif dan analitik yang canggih. Data scientist harus menguasai algoritma seperti regresi linear, pohon keputusan, jaringan saraf tiruan, dan algoritma *Clustering*. Mereka juga harus memahami bagaimana menerapkan teknik pembelajaran mendalam (*Deep Learning*) dengan menggunakan pustaka seperti TensorFlow dan PyTorch (Goodfellow, Bengio, & Courville, 2016).
4. *Database* dan *Big Data*
   Kemampuan untuk bekerja dengan sistem manajemen basis data dan teknologi *Big Data* seperti Hadoop dan Spark. Data scientist harus mampu mengelola dan menganalisis data dalam skala besar, menggunakan alat-alat ini untuk pemrosesan data yang terdistribusi dan analisis data dalam jumlah besar yang tidak dapat ditangani oleh alat konvensional (Provost & Fawcett, 2013).

### Keterampilan Non-Teknis

1. Pemecahan Masalah
   Data scientist harus mampu memecahkan masalah kompleks dengan menggunakan pendekatan analitis. Mereka harus dapat merancang eksperimen, menganalisis hasil, dan mengidentifikasi solusi yang efektif berdasarkan data (Davenport & Patil, 2012).
2. Komunikasi
   Keterampilan komunikasi yang baik diperlukan untuk menyampaikan temuan kepada pemangku kepentingan

non-teknis. Data scientist harus dapat menjelaskan konsep teknis dan hasil analisis dalam bahasa yang dapat dimengerti oleh audiens yang tidak memiliki latar belakang teknis (Davenport & Patil, 2012).

3. Kolaborasi
   Data scientist sering bekerja dalam tim yang terdiri dari anggota dengan latar belakang yang berbeda, sehingga kemampuan untuk berkolaborasi sangat penting. Mereka harus dapat bekerja sama dengan profesional lain, seperti analis bisnis, pengembang perangkat lunak, dan eksekutif, untuk mencapai tujuan bersama (Provost & Fawcett, 2013).

## Teknologi dan Alat

1. Alat Visualisasi Data
   Alat seperti Tableau, Power BI, dan Matplotlib digunakan untuk membuat visualisasi data yang menarik dan mudah dipahami. Visualisasi data membantu dalam menyajikan informasi kompleks dengan cara yang lebih sederhana dan intuitif (Provost & Fawcett, 2013).
2. Alat Pembelajaran Mesin
   Libraries seperti TensorFlow, Scikit-Learn, dan PyTorch digunakan untuk mengembangkan dan menerapkan model pembelajaran mesin. Alat-alat ini menyediakan berbagai algoritma dan fungsi yang memudahkan data scientist dalam membangun, melatih, dan menguji model *Machine Learning* (Goodfellow, Bengio, & Courville, 2016).
3. Manajemen Data
   Alat seperti SQL, NoSQL *database*s, Hadoop, dan Spark digunakan untuk mengelola dan memproses data dalam

skala besar. Data scientist harus mampu bekerja dengan berbagai jenis *database*, termasuk relational dan non-relational, serta memahami bagaimana memproses data dalam lingkungan komputasi terdistribusi (Provost & Fawcett, 2013).

4. Tantangan dan Peluang di Era Digital
   Meskipun data scientist memiliki banyak peluang di era digital, mereka juga menghadapi sejumlah tantangan yang harus diatasi.

**Tantangan**

1. Volume dan Kompleksitas Data
   Data yang dihasilkan oleh organisasi modern sangat besar dan kompleks, membuat proses pengumpulan, pembersihan, dan analisis menjadi tugas yang menantang (Marr, 2018). Data scientist harus mampu menangani data dari berbagai sumber dengan format yang berbeda-beda dan mengintegrasikannya untuk analisis yang komprehensif.
2. Kualitas Data
   Data scientist sering menghadapi masalah dengan data yang tidak lengkap, tidak akurat, atau tidak konsisten, yang dapat mempengaruhi hasil analisis (Provost & Fawcett, 2013). Mereka harus mampu mengembangkan teknik untuk memastikan kualitas data dan menangani masalah yang timbul selama proses analisis.
3. Privasi dan Keamanan Data
   Dengan meningkatnya perhatian terhadap privasi data, data scientist harus memastikan bahwa mereka mematuhi regulasi dan standar privasi yang ketat. Mereka juga harus melindungi data dari akses tidak sah dan

kebocoran. Ini termasuk menerapkan enkripsi data, kontrol akses, dan audit keamanan untuk melindungi informasi sensitif (Marr, 2018).

4. Perubahan Teknologi
   Teknologi data dan alat analitik terus berkembang, sehingga data scientist harus terus belajar dan mengadaptasi keterampilan mereka untuk tetap relevan. Mereka harus mengikuti perkembangan terbaru dalam teknologi dan alat analitik untuk memastikan bahwa mereka menggunakan metode yang paling efektif dan efisien (Provost & Fawcett, 2013).

## Peluang

1. Permintaan yang Tinggi
   Permintaan untuk data scientist terus meningkat karena organisasi dari berbagai sektor mencari cara untuk memanfaatkan data mereka dengan lebih baik (Davenport & Patil, 2012). Data scientist memiliki peluang untuk bekerja di berbagai industri, termasuk teknologi, keuangan, kesehatan, pemasaran, dan banyak lagi.
2. Inovasi dan Pengembangan Teknologi
   Dengan kemajuan dalam teknologi seperti AI, *Machine Learning*, dan *Big Data*, data scientist memiliki peluang untuk bekerja pada proyek-proyek inovatif yang dapat mengubah cara kita memahami dan menggunakan data. Mereka dapat berkontribusi dalam pengembangan teknologi baru yang dapat meningkatkan efisiensi, produktivitas, dan kualitas hidup (Goodfellow, Bengio, & Courville, 2016).
3. Pengaruh pada Pengambilan Keputusan

Data scientist memainkan peran kunci dalam membantu organisasi membuat keputusan yang lebih baik dan lebih cepat berdasarkan wawasan data. Ini mencakup berbagai bidang seperti kesehatan, keuangan, pemasaran, dan banyak lagi. Data scientist dapat membantu organisasi mengidentifikasi tren pasar, meningkatkan operasi, dan mengembangkan strategi bisnis yang lebih efektif (Provost & Fawcett, 2013).

Contoh lain dari peluang yang ada bagi data scientist adalah dalam pengembangan teknologi kendaraan otonom. Data scientist bekerja dengan tim teknik untuk menganalisis data dari sensor kendaraan, mengembangkan algoritma pembelajaran mesin untuk deteksi objek dan pengambilan keputusan, serta mengoptimalkan kinerja sistem secara keseluruhan. Inovasi dalam kendaraan otonom memiliki potensi untuk mengurangi kecelakaan lalu lintas, meningkatkan efisiensi transportasi, dan memberikan manfaat lingkungan melalui pengurangan emisi.

# BAB 5 DATA DAN SKALA PENGUKURAN

## Pendahuluan

Pada era serba digital ini, data menjadi salah satu aset berharga baik bagi individu maupun organisasi. Kehadiran data tidak hanya membantu pengambilan kesimpulan yang lebih baik, melainkan juga memberikan pengetahuan yang mendalam terkait berbagai aspek dalam kehidupan ini. Data memudahkan kita dalam memahami pola dan hubungan yang kompleks dari suatu fenomena, serta menyediakan landasan kuat untuk pengambilan keputusan. Dengan demikian, pemanfaatan data mampu menciptakan solusi tepat sasaran dan efektif dalam berbagai bidang kehidupan.

Seiring dengan meningkatnya volume dan variasi data yang dihasilkan setiap harinya, penting bagi kita untuk memiliki kemampuan dalam mengelola dan menganalisis data tersebut. Data ini dapat bersumber dari transaksi online, media sosial, hingga aktivitas kesehariaan yang terekam melalui perangkat digital. Semakin beragamnya sumber data ini membuat analisis juga menjadi lebih kompleks yang disertai dengan informasi yang semakin detail.

Untuk dapat memanfaatkan data secara optimal, pemahaman tentang skala pengukuran data menjadi sangat krusial. Skala pengukuran data mencakup berbagai dimensi yang masing-masing memiliki karakteristik dan metode analisis yang berbeda. Melalui pemahaman skala pengukuran data, kita dapat mengelola dan menganalisis data dengan lebih tepat, memastikan bahwa interpretasi yang dihasilkan akurat dan

sesuai dengan tujuan yang ingin dicapai. Disisi lain, penggunaaan skala pengukuran yang tepat juga memungkinkan kita untuk mengaplikasikan teknik analisis yang sesuai, sehingga dapat menghasilkan wawasan yang mendalam dan mendukung pengambilan keputusan yang lebih baik.

Dalam pengolahan data, keakuratan dan keandalan data juga perlu diperhatikan. Kecepatan dalam mengakses dan menganalisis data juga menjadi faktor kunci, mengingat semakin cepatnya laju perkembangan informasi. Dengan pengelolaan dan analisis data yang baik, kita dapat mengeksplorasi potensi penuh dari data yang ada, menciptakan solusi inovatif, dan merespons tantangan dengan lebih efektif. Dengan demikian, pemahaman yang mendalam mengenai data dan skala pengukurannya menjadi landasan penting dalam era digital yang terus berkembang ini.

## Pengertian dan Syarat Data

Menurut Kamus Besar Bahasa Indonesia, data merupakan sekumpulan fakta atau informasi yang diperoleh melalui metode seperti pengukuran, penelitian, atau pengamatan. Data ini dapat berupa teks, angka, atau gambar, dan dapat diolah serta dianalisis untuk memperoleh pemahaman atau pengetahuan lebih lanjut. Data dapat pula diartikan sebagai bentuk jamak dari datum, yaitu keterangan atau informasi yang diperoleh dari suatu pengamatan (Nuryadi et al., 2017).

Data berperan penting dalam berbagai sektor, seperti bisnis dan kesehatan, dengan menyediakan wawasan mendalam yang mendasari pengambilan keputusan strategis. Misalnya, dalam bisnis, data penjualan dan perilaku konsumen dapat mengidentifikasi tren pasar dan meningkatkan strategi pemasaran. Sementara itu, data klinis dalam bidang kesehatan

membantu merancang perawatan yang efektif. Beberapa syarat yang harus dipenuhi oleh data antara lain (Prasetyo & Si, 2018):

1. Data harus akurat, mencerminkan kenyataan yang terjadi.
2. Data harus mampu mewakili parameter/variabel yang diukur dengan ukuran variasi yang minimal.
3. Data harus relevan sehingga dapat menjawab permasalahan yang menjadi fokus bahasan.
4. Data harus tersedia tepat waktu, sesuai dengan kebutuhan analisis.

## Kategori Data

Data dapat dikategorikan menjadi beberapa jenis yang terlihat pada Tabel 5.1 berikut ini.

*Tabel 5.1 Kategori Data*

| No | Bentuk Kategori | Data |
|---|---|---|
| 1. | Berdasarkan cara Memperoleh | Primer |
| | | Sekunder |
| 2. | Berdasarkan Jenis Data | Kuantitatif |
| | | Kualitatif |
| 3. | Berdasarkan Sifat Data | Diskrit |
| | | Kontinu |
| 4. | Berdasarkan Waktu Pengumpulan | *Cross-Section* |
| | | *Time Series* |
| 5. | Berdasarkan Sumber Data | Internal |
| | | Eksternal |

Penjelasan dari Tabel 5.1 yaitu antara lain:

1. Kategori Data Berdasarkan Cara Memperolehnya
    a. Data Primer
    Data primer adalah data yang didapatkan langsung dari sumber pertama, tanpa melalui proses atau interpretasi yang lain. Data ini bersifat orisinil dan biasanya dikumpulkan untuk tujuan tertentu (Heryana, 2020). Contohnya data hasil survei

langsung oleh peneliti di lapangan, data dari eksperimen yang dilakukan sendiri, atau data observasi langsung.

b. Data Sekunder

Data sekunder adalah data yang sebelumnya telah dikumpulkan oleh pihak lain dengan tujuan berbeda, namun dapat digunakan kembali untuk tujuan penelitian yang baru. Data ini dapat diperoleh sumber seperti publikasi ilmiah, laporan pemerintah, basis data, dan sumber lainnya (Heryana, 2020). Contohnya data statistik dari Badan Pusat Statistik (BPS), data dari jurnal ilmiah, atau data dari studi sebelumnya yang diambil dari literatur.

2. Kategori Data Berdasarkan Jenis Datanya

a. Data Kuantitatif

Data kuantitatif adalah data yang dinyatakan dalam bentuk kuantitas atau angka. Data ini dapat dihitung, diukur, dan dianalisis secara matematis. Contohnya data seperti tinggi badan, berat badan, pendapatan, atau hasil tes numerik.

b. Data Kualitatif

Data kualitatif adalah data yang dinyatakan dalam bentuk teks atau deskriptif. Data ini menggambarkan kualitas atau karakteristik suatu fenomena tanpa menggunakan angka. Contohnya data seperti pendapat, sikap, preferensi, atau deskripsi verbal dari suatu situasi.

3. Kategori Data Berdasarkan Sifat Datanya

a. Data Diskrit

Data diskrit adalah data kuantitatif yang hanya

mengambil nilai tertentu, biasanya berupa bilangan bulat, dan tidak dapat dibagi menjadi beberapa bagian yang lebih kecil. Data ini biasanya diperoleh dari proses perhitungan. Contohnya jumlah anak dalam suatu keluarga, banyaknya mobil yang dimiliki seseorang, atau jumlah penjualan produk dalam sehari.

b. Data Kontinu

Data kontinu adalah data kuantitatif yang dapat memuat nilai apapun dalam suatu rentang tertentu, termasuk pecahan atau desimal. Data ini biasanya diperoleh dari proses pengukuran. Contohnya tinggi badan, berat badan, waktu yang diperlukan menyelesaikan sebuah tugas, atau suhu udara.

4. Kategori Data Berdasarkan Waktu Pengumpulannya

a. *Data Cross-Section*

*Data Cross-Section* adalah data yang dikumpulkan pada satu periode atau titik waktu tertentu dari beberapa individu atau unit pengamatan yang berbeda. Data ini memberikan gambaran tentang keadaan atau situasi pada waktu tertentu, tanpa memperhatikan perubahan dari waktu ke waktu (Prasetyo & Si, 2018). Contohnya survei pendapatan rumah tangga pada tahun 2024, data jumlah siswa di berbagai sekolah di Provinsi Sulawesi Selatan pada awal semester, atau data curah hujan kota Parepare tahun 2023.

b. *Data Time Series*

*Data time series* adalah data yang terdiri dari pengukuran atau pengamatan yang dilakukan

secara berurutan dan teratur dalam interval waktu yang sama. Data ini menunjukkan bagaimana suatu variabel berubah dari waktu ke waktu, memungkinkan analisis tren, pola, dan peramalan. Contohnya data inflasi bulanan Indonesia 2010-2023, data harga saham harian PT. X selama setahun, atau data penjualan mingguan sebuah toko retail selama setahun.

5. Kategori Data Berdasarkan Sumber Data
    a. Data Internal
    Data internal adalah data yang didapatkan dari dalam perusahaan atau organisasi itu sendiri. Data ini biasanya dikumpulkan melalui proses operasional sehari-hari dan sistem informasi yang ada di dalam organisasi. Contohnya data keuangan perusahaan, data inventaris barang, atau data karyawan.
    b. Data Eksternal
    Data eksternal adalah data yang didapatkan dari luar organisasi atau perusahaan. Data ini dikumpulkan dari sumber-sumber eksternal seperti laporan pemerintah, penelitian pasar, dan data dari pemasok atau pelanggan. Contohnya data pasar dari laporan riset pasar, data ekonomi makro dari BPS, atau data demografis dari lembaga statistik.

## Skala Pengukuran Data

Skala pengukuran data adalah sistem klasifikasi yang digunakan untuk mengukur dan mengkategorikan variabel atau atribut pada penelitian (Dr. Vladimir, 2018). Skala ini penting

untuk menentukan jenis analisis statistik yang sesuai. Terdapat empat skala pengumuran data, yaitu nominal, ordinal, interval, dan rasio.

1. Skala Nominal
   Skala nominal adalah skala pengukuran yang mengklasifikasikan data ke dalam kelompok atau kategori yang berbeda tanpa adanya tingkatan atau urutan. Data nominal hanya berfungsi sebagai label atau identifikasi untuk membedakan satu kategori dari kategori lainnya. Data nominal tidak memiliki nilai numerik atau urutan yang bermakna, sehingga tidak memungkinkan untuk dilakukan operasi matematika (Suparyanto, 2020). Contonya jenis kelamin (Pria, wanita), warna favorit (ungu, abu-abu, hijau), jenis kendaraan (mobil, motor, sepeda), atau status pernikahan (belum menikah, menikah, bercerai). Dalam penggunaannya, skala nomial sering kali dilakukan pengkodean menggunakan angka untuk memudahkan analisis, misalnya 1 kode untuk Pria dan 2 kode untuk Wanita. Angka tersebut tidak memiliki makna numerik melainkan hanya berfungsi sebagai label. Adapun ciri-ciri skala nominal yaitu:
   a. Tidak berurutan, yaitu kategori dalam skala nominal tidak memiliki urutan atau tingkatan. Setiap kategori berdiri sendiri dan tidak lebih besar atau lebih kecil dari kategori lainnya.
   b. Bersifat diskrit, artinya data hanya termuat dalam satu kategori tertentu dan tidak dapat berada di lebih dari satu kategori.
   c. Data kualitatif, yaitu menggambarkan atribut atau karakteristik yang tidak dapat diukur secara numerik.

d. Tidak memiliki nilai numerik yang bermakna.
e. Modus sebagai ukuran pusat.

2. Skala Ordinal

Skala ordinal adalah jenis skala pengukuran yang mengkategorikan data dalam tingkatan atau urutan yang jelas, tetapi jarak antar kategori tidak dapat diukur secara tepat (Zahriyah, 2023). Skala ini digunkan untuk menunjukkan peringkat atau urutan, dan meskipun diketahui bahwa satu nilai lebih tinggi atau lebih rendah dari yang lain, kita tidak bisa menentukan seberapa besar perbedaan antara nilai-nilai tersebut. Contohnya tingkat pendidikan orang tua (SD, SMP, SMA, Diploma, Sarjana, Magister), tingkat kepuasan seorang pelanggan (sangat puas, puas, tidak puas), atau kelas sosial ekonomi (Rendah, Menengah, Tinggi). Adapun ciri-ciri skala ordinal yaitu:

a. Urutan yang jelas, dimana satu kategori bisa dibandingkan dengan kategori lainnya dalam hal mana yang lebih rendah atau lebih kecil.
b. Tidak memiliki interval yang konsisten. Hal ini karena meskipun data diurutkan, perbedaan antara satu kategori dengan kategori lainnya tidak diketahui secara pasti dan tidak harus sama.
c. Bersifat diskrit, dimana setiap nilai hanya dapat menempati satu peringkar atau urutan.
d. Modus dan median sebagai ukuran pusat.
e. Tidak dapat dilakukan operasi matematika.

3. Skala Interval

Skala interval adalah jenis skala pengukuran yang tidak hanya menyediakan urutan kategori atau nilai, tetapi juga memiliki jarak atau interval yang konsisten antara setiap

nilai. Hal ini berarti perbedaan antara dua nilai adalah sama di sepanjang skala. Skala interval memungkinkan untuk melakukan operasi matematika dasar seperti penjumlahan dan pengurangan, tetapi tidak memiliki titik nol mutlak, sehingga rasio tidak bisa dihitung (Tarigan & Frintiana Silaban, 2023). Contohnya suhu dalam derajat Celcius, *IQ Scores*, skala penilaian gejala deperesi (misalnya 0-59 pada Inventaris Depresi Beck). Adapun ciri-ciri skala interval yaitu sebagai berikut.

a. Urutan yang konsisten, dimana data pada skala interval memiliki urutan yang jelas dan teratur.
b. Interval yang sama, dimana perbedaan antara nilai-nilai pada skala interval adalah sama di seluruh skala. Misalnya, jarak antara suhu 10℃ dan 20℃ sama dengan jarak antara 20℃ dan 30℃.
c. Tidak memiliki nol mutlak, artinya tidak memiliki titik nol mutlak yang menunjukkan ketiadaan nilai. Nol pada skala interval bukan berarti "tidak ada" atau "nihil" (Nilda, 2020). Contohnya nilai 0 ℃ tidak berarti tidak ada suhu, melainkan nilai tersebut titik beku air.

4. Skala Rasio

Skala rasio adalah jenis skala pengukuran yang memiliki karakteristik semua skala pengukuran lainnya (nominal, ordinal, dan interval) ditambah dengan adanya nilai nol absolut yang menunjukkan ketiadaan total atribut yang diukur. Skala ini memungkinkan perbandingan rasio antara dua nilai yang berarti, seperti mengatakan bahwa satu nilai adalah dua kali lebih besar dari nilai lainnya (Firmansyah & Data, n.d.). Contonya berat badan, tinggi badan, pendapatan, waktu tempuh, jumlah barang, atau

volume. Adapun ciri-ciri skala rasio antara lain:

a. Memiliki nilai nol mutlak/absolut, yaitu nilai nol pada skala rasio menunjukkan ketiadaan total dari variabel yang diukur.
b. Jarak yang konsisten, yaitu jarak atau interval antara setiap nilai adalah sama dan tetap. Misalnya jarak antara 10 kg dan 30 kg sama saja dengan jarak antara 40 kg dan 60 kg.
c. Perbandingan rasio, yaitu dapat dibuat perbandingan yang bermakna antara dua nilai. Misalnya kita bisa mengatakan bahwa 20 kg adalah dua kali lipat dari 10 kg.
d. Memungkinkan semua operasi matematika, termasuk pengurangan, penjumlahan, pembagian, dan perkalian.

# BAB 6 EKSPLORASI DATA

## Pendahuluan

Eksplorasi data adalah tahap awal dalam analisis data di mana seorang data analyst atau ilmuan data mencoba memahami struktur, pola, dan karakteristik data sebelum melakukan pemodelan lebih lanjut. Tahap ini sangat penting karena dapat memberikan pengetahuan awal yang penting untuk memahami kumpulan data (Joseph M. Tandiallo, 2024) dan membangun model pembelajaran mesin yang lebih baik.

Eksplorasi data yang dikenal juga *dengan exploratory data analysis* (EDA) merupakan tahapan penting yang harus dilakukan dalam siklus hidup data sicence. Dalam prosesnya, EDA dilakukan dengan tujuan melakukan eksplorasi terhadap karakteristik dari suatu kumpulan data dengan menggunakan berbagai metode. Salah satu metode yang digunakan adalah metode plotting secara visual. Plotting dalam EDA dapat dibuat ke dalam berbagai macam di antaranya Histogram, *Box plot*, *Scatter Plot*, dan dalam bentuk lainnya (Imoore, 2020).

Pembelajaran mesin dengan tujuan besarnya yang dapat melakukan prediksi dari data tidak terlepas dari permasalahan yang dialami pada data. Hal ini terlihat ketika model yang dibuat sulit untuk dilakukan peningkatan akurasinya. Teknik eksplorasi data dapat membantu dalam meningkatkan akurasi pada model yang dibuat (Sunil Ray, 2016).

## Analisis Eksplorasi Data Secara Komprehensif

Sebelum mempelajari fase pemodelan, EDA membantu dalam memahami struktur data, ciri-ciri utama dalam data, mengidentifikasi pola, menemukan anomali, memahami hubungan antar variabel, dan memahami variable-variabel yang relevan dengan permasalahan yang dihadapi. EDA juga dapat membantu dalam mengidentifikasi dan menangani nilai yang hilang atau terjadi duplikasi, outlier, dan kesalahan dalam entri data (Joseph M. Tandiallo, 2024). EDA dapat membantu dalam menganalisis kualitas data masukan sehingga dapat dilakukan penanganan yang dapat menjamin kualitas keluaran yang baik (Sunil Ray, 2016). EDA dapat dilakukan secara efektif dengan menggunakan python. Python memiliki ekosistem perpustakaan yang kaya dan menyediakan lingkungan yang kuat dalam melakukan EDA (Grus, 2019).

Dalam melakukan analisis eksplorasi data secara komprehensif diperlukan tahapan sebagai berikut (Sunil Ray, 2016):

1. Pemasangan Perpustakaan Esensial
2. Identifikasi Variabel
3. Analisis Univariat
4. Analisis Bivariat
5. Penanganan Terhadap Missing Value
6. Penanganan Terhadap Outlier
7. Transformasi Variabel
8. Kreasi Variabel

Pada implementasinya, dalam menyempurnakan model, tahap 5 sampai 8 perlu dilakukan secara iteratif guna mendapatkan model yang sempurna.

## Pemasangan Perpustakaan Esensial

Perpustakaan atau disebut juga dengan istilah Library yang dibutuhkan secara umum diantaranya Pandas, Numpy, Matplotlib, Seaborn, dan SciPy. Pandas digunakan untuk melakukan analisis dan manipulasi data. Numpy digunakan untuk melakukan komputasi numerik. Matplotlib dan Seaborn digunakan untuk melakukan vislualisasi data. Sedangkan SciPy digunakan untuk komputasi ilmiah. Beberapa library yang telah disebutkan dapat digunakan tanpa perlu dilakukan instalasi sebelumnya jika seorang data analyst atau ilmuan data menggunakan Google Collaboratory sebagai *Integrated Development Environment* (IDE) dalam menuliskan kode program.

## Identifikasi Variabel

Identifikasi variabel yang dimaksud dimulai dari memahami antara variabel prediktor (*input*) dan variabel target (*output*). Setelah itu identifikasi dilakukan terhadap tipe data dan kategori variabel.

Studi Kasus

Misalkan seorang pelatih ingin melakukan prediksi apakah mahasiswa akan bermain kriket atau tidak (Tabel 6.1).

*Tabel 6.1 Prediksi mahasiswa bermain kriket*

| ID | Jenis Kelamin | Nilai Ujian | Tinggi Badan (cm) | Berat Badan (kg) | Bermain? |
|---|---|---|---|---|---|
| S001 | M | 65 | 178 | 61 | 1 |
| S002 | F | 75 | 174 | 56 | 0 |
| S003 | M | 45 | 163 | 62 | 1 |
| S004 | M | 57 | 175 | 70 | 0 |
| S005 | F | 59 | 162 | 67 | 0 |

Berikut ini definisi berbeda terhadap variabel pada data

Tabel 1:

1. Tipe Variabel: variabel prediktor diantaranya adalah Jenis Kelamin, Nilai Ujian, Tinggi Badan, dan Berat Badan. Sedangkan variabel target adalah Bermain?
2. Tipe Data: tipe data karakter diantaranya adalah ID dan Jenis Kelamin. Sedangkan tipe data numerik diantaranya adalah Bermain? Nilai Ujian, Tinggi Badan dan Berat Badan.
3. Kategori Variabel: kategori variabel kategorikal diantaranya adalah Jenis Kelamin dan Bermain? Sedangkan kategori variabel kontinu adalah Nilai Ujian, Tinggi Badan, dan Berat Badan.

## Analisis Univariat

Pada tahap ini, eksplorasi dilakukan terhadap setiap variabel. Metode yang digunakan bergantung pada jenis variabel yang akan dianalisis apakah kategorikal atau kontinu.

Studi Kasus

Variabel kontinu: pada variabel kontinu perlu dipahami tendensi sentral dan penyebaran variabel. Pengukuran dilakukan dengan menggunakan berbagai metode visualisasi metrik statistik seperti yang ditunjukan Tabel 6.2.

*Tabel 6.2 Metode visualisasi metrik statistik*

| Tendensi Sentral | Ukuran Dispersi | Metode Visualisasi |
|---|---|---|
| Mean | Range | Histogram |
| Median | Quartile | Box Plot |
| Mode | IQR | |
| Min | Variance | |
| Max | Standard Deviation | |
| | Skewness dan Kurtosis | |

Catatan: Analisis Univariat juga digunakan untuk analisis

missing value dan outlier.

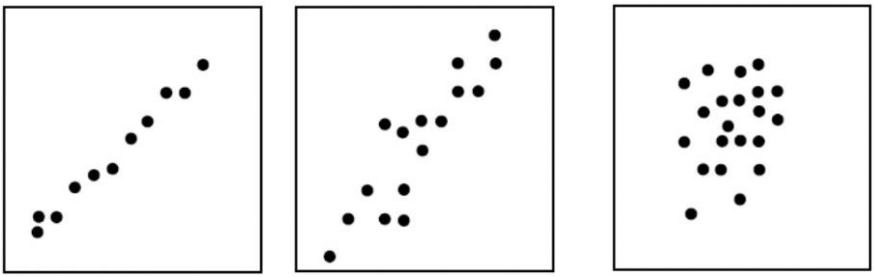
Variabel kategorikal: pada variabel ini, tabel frekuensi dapat digunakan untuk memahami distribusi setiap kategori. Persentase nilai di setiap kategori juga dapat dipahami. Metrik pengukuran yang dapat digunakan yaitu count dan count%. Visualisasi pada variabel ini dapat menggunakan diagram batang.

### Analisis Bivariat

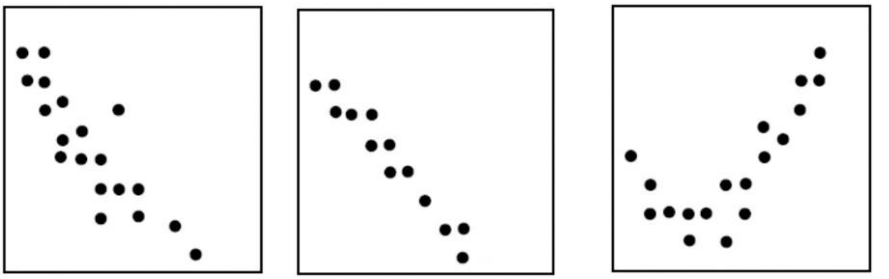
Analisis Bivariat digunakan untuk mengetahui hubungan antara dua variabel. Analisis dapat dilakukan untuk kombinasi variabel kategorikal dan kontinu. Metode berbeda dapat digunakan sesuai dengan tipe kombinasi yang dilakukan pada proses analisis.

Studi Kasus

Kontinu & Kontinu: saat menganalisis bivariat terhadap dua variabel kontinu, hal yang perlu diperhatikan adalah scatter plot. Pola scatter plot menunjukan hubungan antar dua variabel (linier atau non linier) seperti Gambar 6.1.



Korelasi positif yang kuat | Korelasi positif yang sedang | Tidak ada korelasi



Korelasi negatif yang sedang | Korelasi negatif yang kuat | Hubungan lengkung

*Gambar 6.1 Hubungan antar dua variabel*

*Scatter plot* menunjukan hubungan antara dua variabel, namun tidak menunjukan kekuatan hubungan di antara keduanya. Kekuatan hubungan antara dua variabel dapat diketahui menggunakan korelasi yang nilainya di antara -1 dan +1.

Kategorikal & Kategorikal: hubungan antara dua variabel kategorikal dapat diketahui dengan menggunakan metode Two-way Table dan Stacked Column Chart.

Kategorikal & Kontinu: saat melakukan eksplorasi hubungan antara variabel kategorikal dan kontinu, *Box Plot* dapat digunakan untuk setiap tingkat variabel kategorikal. Jika nilainya kecil, maka tidak menunjukkan signifikansi statistik. Signifikansi statistik dapat dilihat dengan melakukan uji Z, uji T, atau ANOVA.

### Penanganan Terhadap *Missing Value*

Data yang hilang dalam kumpulan data pelatihan dapat menyebabkan model menjadi bias karena analisis perilaku dan hubungan belum dilakukan dengan benar. Hal tersebut dapat menyebabkan prediksi atau klasifikasi yang salah.

Beberapa alasan yang menyebabkan terjadinya missing

value yang dapat terjadi dalam dua tahap:

1. Ekstraksi Data
   Proses ekstraksi data memungkinkan terjadinya permasalahan. Pemeriksaan perlu dilakukan kembali terhadap kebenaran data. Beberapa prosedur hashing dapat digunakan untuk memastikan ekstraksi data sudah benar. Kesalahan yang terjadi pada tahap ekstraksi data biasanya mudah ditemukan dan diperbaiki.
2. Koleksi Data
   Pada proses ini, kesalahan terjadi pada saat pengumpulan data dan lebih sulit untuk diperbaiki. Berikut adalah empat jenis kesalahan yang umum terjadi:
   a. Hilang sepenuhnya secara acak
      Ini adalah kasus ketika probabilitas variabel yang hilang sama untuk semua observasinya. Misalnya responden dalam proses pengumpulan data memutuskan bahwa mereka akan menyatakan penghasilannya setelah melempar koin secara adil. Jika muncul kepala, responden menyatakan penghasilannya dan sebaliknya. Di sini setiap observasi memiliki peluang yang sama untuk kehilangan nilai.
   b. Hilang secara acak
      Ini adalah kasus ketika variabel hilang secara acak dan rasio yang hilang bervariasi untuk nilai/tingkat variabel masukan lainnya yang berbeda. Misalnya sekelompok mahasiswa mengumpulkan data yang menunjukkan usia dan perempuan memiliki missing value yang lebih tinggi dibandingkan laki-laki.
   c. Hilangnya bergantung pada prediktor yang tidak

teramati

Ini adalah kasus ketika nilai yang hilang tidak acak dan terkait dengan variabel masukan yang tidak teramati. Misalnya dalam suatu penelitian kedokteran, jika diagnosis tertentu menyebabkan ketidaknyamanan, maka kemungkinan untuk keluar dari penelitian tersebut lebih besar. Nilai yang hilang ini tidak terjadi secara acak kecuali telah dimasukkan "ketidaknyamanan" sebagai variabel masukan untuk semua pasien.

d. Hilangnya tergantung pada nilai yang hilang itu sendiri

Ini adalah kasus ketika kemungkinan nilai yang hilang berkorelasi langsung dengan nilai yang hilang itu sendiri. Misalnya orang dengan pendapatan lebih tinggi atau lebih rendah cenderung tidak memberikan respon terhadap penghasilan mereka.

Berbagai metode yang dapat digunakan untuk menangani *missing value* di antaranya *deletion, mean/mode/median imputation,* dan *prediction model*. Setelah selesai menangani *missing value*, tugas selanjutnya adalah menangani *outlier*. Seringkali outlier diabaikan saat membuat model. *Outlier* cenderung membuat akurasi menurun.

## Penanganan Terhadap *Outlier*

*Outlier* merupakan pengamatan yang tampak jauh dan menyimpang dari pola keseluruhan dalam suatu sampel. Contohnya adalah seseorang yang sedang membuat profil pelanggan dan mengetahui bahwa pendapatan tahunan rata-rata pelanggan adalah $0,8 juta. Namun, terdapat dua pelanggan

yang memiliki pendapatan tahunan sebesar $4 juta dan $4,2 juta. Pendapatan tahunan kedua pelanggan ini jauh lebih tinggi dibandingkan populasi lainnya. Kedua observasi tersebut dianggap sebagai *outlier*.

*Outlier* dapat terdiri dari dua jenis yaitu univariat dan multivariat. Perbedaannya, outlier multivariat dapat ditemukan dalam ruang berdimensi n. Untuk menemukannya diperlukan untuk melihat distribusi dalam multi dimensi.

Setiap kali ditemukannya *outlier*, cara ideal untuk mengatasinya adalah dengan mencari tau alasan kemunculannya. Metode untuk mengatasinya akan bergantung pada alasan kemunculannya. Penyebab *outlier* diantaranya adalah *Data Entry Errors*, *Measurement Error, Experimental Error, Intentional Outlier, Data Processing Error, Sampling Error,* dan *Natural Outlier*.

*Outlier* dapat ditemukan dengan metode yang paling umum digunakan yaitu visualisasi. Beberapa metode visualisasi tersebut diantaranya adalah *Box Plot*, Histogram, dan *Scatter Plot*. Sedangkan penangan terhadap outlier mirip dengan metode penanganan terhadap *missing value* seperti *deleting observation, transformation, binning,* dan lainnya.

## Seni Rekayasa Fitur

Rekayasa fitur adalah ilmu dalam mengekstraksi lebih banyak informasi dari data yang ada. Tidak ada penambahan data dalam hal ini, proses yang dilakukan adalah membuat data yang sudah ada menjadi lebih berguna. Misalnya akan dilakukan prediksi jumlah pengunjung di pusat perbelanjaan berdasarkan tanggal. Jika menggunakan tanggal secara langsung, maka wawasan yang berarti dari ekstraksi data tidak akan didapatkan. Hal ini karena jatuhnya kaki tidak terlalu dipengaruhi oleh hari

dalam sebulan dibandingkan dengan hari dalam seminggu. Informasi tentang hari dalam seminggu ini tersirat dalam data. Hal yang perlu dilakukan adalah mengeluarkannya untuk membuat model yang lebih baik. Proses yang dilakukan untuk mengeluarkan informasi dari data dikenal sebagai rekayasa fitur.

Rekayasa fitur dapat dilakukan setelah melewati proses sebelumnya di mana rekayasa fitur dapat dilakukan dalam 2 langkah yaitu Variable Transformation dan *Variable/Feature Creation*. Kedua Teknik tersebut memiliki dampak luar biasa pada kekuatan prediksi.

1. *Variable Transformation*
   Dalam pemodelan data, transformasi mengacu pada penggantian variabel dengan fungsi. Misalnya, mengganti variabel x dengan akar kuadrat atau algoritma x merupakan suatu transformasi. Dengan kata lain, transformasi adalah proses yang mengubah sebaran atau hubungan suatu variabel dengan variabel lainnya. Dalam melakukan transformasi variabel terdapat banyak metode yang dapat digunakan. Beberapa metode diantaranya akar kuadrat, akar pangkat tiga, logaritma, binning, reciprocal, dan sebagainya.
2. *Feature/Variable Creation*
   Pembuatan fitur/variabel merupakan proses untuk menghasilkan variabel/fitur berdasarkan variabel yang sudah ada. Misalnya dimiliki variabel tanggal (dd/mm/yy) sebagai variabel input dalam kumpulan data. Variabel baru dapat dihasilkan seperti hari, bulan, tahun, minggu, dan hari kerja yang mungkin memiliki hubungan lebih baik dengan variabel target. Langkah tersebut digunakan untuk menyorot hubungan tersembunyi dalam suatu variabel. Terdapat berbagai macam Teknik untuk

membuat fitur baru yang salah satu contohnya telah disampaikan sebelumnya.

# BAB 7 PREPROCESSING DATA

## Pendahuluan

Di era *Big Data* dan *Machine Learning* saat ini, data merupakan aset yang sangat berharga. Data digunakan untuk membuat keputusan bisnis yang kritis, mengembangkan model prediktif, dan menghasilkan wawasan baru yang dapat memandu strategi masa depan. Namun, sebelum data dapat digunakan untuk tujuan ini, data tersebut harus diproses dan dipersiapkan dengan baik. Proses ini dikenal sebagai *preprocessing data*.

*Preprocessing data* adalah langkah awal yang sangat penting dalam analisis data dan *Machine Learning*. Tanpa preprocessing yang tepat, data mentah sering kali mengandung banyak *noise*, ketidakkonsistenan, dan nilai yang hilang, yang dapat mengurangi akurasi dan keandalan model yang dihasilkan. Oleh karena itu, *preprocessing data* adalah langkah yang tidak boleh diabaikan.

Tujuan utama dari *preprocessing data* adalah untuk mengubah data mentah menjadi bentuk yang lebih bersih dan lebih sesuai untuk analisis lebih lanjut atau pelatihan model *Machine Learning*. Berikut adalah beberapa tujuan utama dari *preprocessing data*:

1. Membersihkan Data
   Menghapus atau memperbaiki nilai yang hilang, menangani data yang duplikat, dan mengoreksi kesalahan dalam data.
2. Mengurangi *Noise*

Menghilangkan data yang tidak relevan atau mengganggu yang dapat mempengaruhi hasil analisis.

3. Mengubah Format Data
   Mengubah data ke format yang lebih sesuai untuk analisis atau pemodelan, seperti normalisasi atau standardisasi.
4. Menyederhanakan Data
   Mengurangi dimensi data untuk mempercepat proses analisis dan mengurangi kompleksitas model.
5. Meningkatkan Kualitas Data
   Memastikan bahwa data yang digunakan berkualitas tinggi dan dapat diandalkan untuk analisis dan pemodelan.

## *Data Cleaning*

*Data cleaning* atau pembersihan data adalah proses penting dalam tahap preprocessing data yang bertujuan untuk memperbaiki atau menghapus data yang tidak benar, tidak lengkap, tidak akurat, atau tidak relevan. Proses ini memastikan bahwa data yang akan digunakan dalam analisis atau pelatihan model *Machine Learning* berkualitas tinggi, sehingga menghasilkan hasil yang lebih akurat dan dapat diandalkan.

Pentingnya proses data cleaning pada preprocessing data sebagai berikut.

1. Meningkatkan Kualitas Analisis
   Data yang bersih menghasilkan analisis yang lebih akurat dan wawasan yang lebih dapat diandalkan.
2. Mengurangi Kesalahan Model
   Model *Machine Learning* yang dilatih dengan data yang bersih lebih mungkin menghasilkan prediksi yang akurat dan menghindari kesalahan.
3. Efisiensi Proses

Data yang bersih mempercepat proses analisis dan pemodelan karena mengurangi kebutuhan untuk penanganan error atau inkonsistensi selama proses tersebut.

## Transformasi Data

Transformasi data adalah proses penting dalam *preprocessing data* yang bertujuan untuk mengubah data mentah menjadi format yang lebih sesuai untuk analisis atau pelatihan model *Machine Learning*. Transformasi data melibatkan berbagai teknik untuk mengubah struktur dan nilai data agar lebih konsisten dan relevan. Ini membantu meningkatkan kualitas data dan kinerja model *Machine Learning*.

Adapun jenis-jenis transformasi data yang secara umum di gunakan sebagai berikut.

1. Normalisasi dan Standardisasi
   Normalisasi dan standardisasi adalah dua teknik penting dalam preprocessing data yang digunakan untuk mengubah skala data sehingga lebih sesuai untuk analisis atau pelatihan model *Machine Learning*. Kedua teknik ini bertujuan untuk membuat data lebih konsisten dan mengurangi efek skala yang berbeda pada fitur yang berbeda.
   a. Normalisasi
      Normalisasi adalah teknik transformasi data yang mengubah nilai fitur menjadi rentang [0, 1] atau [-1, 1]. Normalisasi berguna ketika Anda ingin membatasi skala data tanpa mempengaruhi hubungan antar nilai.
      Penggunaan Normalisasi digunakan Ketika ada

kondisi sebagai berikut.

1) Data memiliki skala yang berbeda dan Anda ingin menyamakan skala tersebut.
2) Model yang digunakan sensitif terhadap skala data, seperti K-Nearest Neighbors (KNN) atau *Neural Networks*.

Metode Nonrmalisasi

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

b. Standarisasi

Standardisasi adalah teknik transformasi data yang mengubah data sehingga memiliki mean 0 dan standar deviasi 1. Ini berarti data akan memiliki distribusi yang mengikuti distribusi normal standar.

Penggunaan Standardisasi digunakan Ketika ada kondisi sebagai berikut.

1) Data memiliki distribusi normal atau mendekati normal.
2) Model yang digunakan mengasumsikan bahwa data memiliki distribusi normal, seperti Linear Regression atau *Support Vector Machine*s (SVM).

Metode Standarisasi

$$Xstand = \frac{\mathrm{X} - \mu}{\sigma}$$

di mana $\mu$ adalah mean dan $\sigma$ adalah standar deviasi.

2. Pengkodean Data Kategorikal

Data kategorikal adalah jenis data yang merepresentasikan kategori atau kelompok. Dalam

*Machine Learning*, data kategorikal harus diubah menjadi format numerik sebelum digunakan dalam algoritma yang membutuhkan input numerik. Proses ini disebut pengkodean data kategorikal. Pengkodean yang tepat sangat penting karena dapat mempengaruhi kinerja model dan hasil analisis.

3. Transformasi Log dan Square Root
   Transformasi Log dan Square Root adalah teknik yang digunakan untuk mengubah distribusi data agar lebih mendekati distribusi normal atau untuk mengurangi skewness (kemiringan). Kedua teknik ini sangat berguna ketika data memiliki rentang yang sangat besar atau distribusi yang tidak simetris, yang dapat mempengaruhi kinerja model *Machine Learning*.
   a. Transformasi Log
      Penggunaan Transformasi Log digunkaan Ketika ada kondisi sebagai berikut.
      1) Data memiliki skewness positif yang tinggi.
      2) Rentang nilai data sangat besar.
      3) Anda ingin mengurangi dampak outliers.

      Rumus Transformasi Log
      Untuk nilai $x$ transformasi log diberikan oleh:
      $$xlog = \log(x)$$
      Namun, karena logaritma tidak didefinisikan untuk nilai nol atau negatif, biasanya kita menambahkan konstanta kecil $c$ untuk menangani nilai nol:
      $$xlog = \log(x + c)$$
   b. Transformasi Square
      Transformasi Square Root menggunakan akar kuadrat untuk mengubah data. Teknik ini juga

digunakan untuk mengurangi skewness, terutama pada data yang memiliki distribusi Poisson atau ketika varians meningkat seiring dengan mean.

Penggunaan Transformasi Log digunakan Ketika ada kondisi sebagai berikut.

1) Data memiliki skewness positif.
2) Data memiliki distribusi Poisson.
3) Anda ingin mengurangi varians yang meningkat seiring dengan mean.

Rumus Transformasi Square Root

Untuk nilai $x$ transformasi square root diberikan oleh:

$$Xsqrt = \sqrt{x}$$

4. Scaling

Scaling adalah teknik penting dalam preprocessing data yang bertujuan untuk mengubah skala fitur sehingga lebih sesuai untuk analisis atau pelatihan model *Machine Learning*. Beberapa algoritma *Machine Learning*, seperti *Support Vector Machine*s (SVM), *K-Nearest Neighbors* (KNN), dan *Gradient Descent-based models*, sangat sensitif terhadap skala data. Oleh karena itu, scaling dapat membantu meningkatkan kinerja model dengan memastikan bahwa semua fitur berkontribusi secara proporsional.

Ada beberapa jenis metode scaling yang umum digunakan sebagai berikut.

a. *Standard Scaling* (Z-Score Scaling)

Standard Scaling mengubah data sehingga memiliki mean 0 dan standar deviasi 1. Ini berarti data akan memiliki distribusi normal standar.

Rumus Standard Scaling

$$Xstand = \frac{\mathrm{X} - \mu}{\sigma}$$

di mana $\mu$ adalah mean dan $\sigma$ adalah standar deviasi.

b. *Min-Max Scaling*

*Min-Max Scaling* adalah teknik transformasi data yang digunakan untuk mengubah nilai fitur ke dalam rentang tertentu, biasanya [0, 1]. Teknik ini sangat berguna ketika Anda ingin memastikan bahwa semua fitur memiliki skala yang sama tanpa mengubah distribusi data secara signifikan. *Min-Max Scaling* membantu model *Machine Learning* yang sensitif terhadap skala data untuk berkinerja lebih baik. Rumus untuk *Min-Max Scaling* adalah sebagai berikut.

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

c. *Robust Scaling*

*Robust Scaling* adalah teknik transformasi data yang digunakan untuk mengubah skala fitur sehingga lebih tahan terhadap outliers. Teknik ini menggunakan median dan *Interquartile Range* (IQR) untuk mengubah skala data, berbeda dengan *Standard Scaling* yang menggunakan mean dan standar deviasi. Robust Scaling sangat berguna ketika data mengandung *outliers* yang dapat mempengaruhi skala dan distribusi data.

$$Xroubst = \frac{X - median}{IQR}$$

di mana:

$X$ adalah nilai asli.

median adalah median dari fitur.

$IQR$ adalah Interquartile Range, dihitung sebagai perbedaan antara kuartil ketiga (Q3) dan kuartil pertama (Q1):

$$IQR = Q3 - Q1$$

X robust adalah nilai yang telah diskalakan.

Penggunaan Robust Scaling digunakan Ketika ada kondisi sebagai berikut.

1) Data mengandung *outliers*: Ketika data memiliki outliers yang signifikan, Robust Scaling dapat mengurangi pengaruh outliers tersebut.
2) Distribusi data yang tidak normal: Ketika data tidak mengikuti distribusi normal dan memiliki distribusi yang miring.
3) Fitur dengan skala yang berbeda: Seperti teknik scaling lainnya, *Robust Scaling* membantu menyamakan skala fitur yang memiliki rentang nilai yang sangat berbeda.

d. *Max Abs Scaling*

Max Abs Scaling adalah teknik transformasi data yang digunakan untuk mengubah skala fitur sehingga semua nilai berada dalam rentang [-1, 1] berdasarkan nilai absolut maksimum dari setiap fitur. Teknik ini mempertahankan distribusi data asli dan terutama digunakan untuk data yang bersifat sparse (memiliki banyak nilai nol), seperti data teks yang diubah menjadi fitur frekuensi kata (*bag of words*) atau data berbasis matriks.

5. *Binning*

*Binning* adalah teknik dalam preprocessing data yang

digunakan untuk mengelompokkan data numerik kontinu ke dalam interval atau "*bins*" yang diskrit. Teknik ini membantu dalam mengurangi noise dalam data dan membuat pola dalam data lebih terlihat. *Binning* sering digunakan dalam analisis data eksploratif dan dalam persiapan data untuk model *Machine Learning*, terutama ketika kita ingin mengubah data kontinu menjadi kategori diskrit. Adapun manfaat dalam proses *Binning* sebagai berikut.

a. Mengurangi *Noise*: Dengan mengelompokkan data ke dalam bins, variasi kecil dalam data dapat diabaikan, sehingga mengurangi noise.
b. Mempermudah Analisis: Data yang dibinned lebih mudah dianalisis dan diinterpretasikan, terutama dalam visualisasi data.
c. Mengatasi *Skewness*: *Binning* dapat membantu dalam mengatasi skewness dengan mengelompokkan nilai ekstrem ke dalam satu bin.
d. Membantu dalam Model: Dalam beberapa algoritma *Machine Learning*, terutama yang berbasis pohon keputusan, binning dapat meningkatkan kinerja model.

Adapun jenis-jenis *Binning* sebagai berikut.

a. *Equal-Width Binning* (Binning Lebar Sama)
   *Equal-Width Binning* membagi rentang data ke dalam beberapa bins yang memiliki lebar yang sama. Misalnya, jika data berkisar dari 0 hingga 100 dan kita ingin membuat 5 bins, maka setiap bin akan memiliki lebar 20
b. *Equal-Frequency Binning* (Binning Frekuensi Sama)

*Custom Binning* memungkinkan pengguna untuk menentukan batas-batas bins sesuai dengan kebutuhan spesifik mereka. Ini memberikan fleksibilitas untuk mengelompokkan data berdasarkan kriteria yang lebih relevan.

c. *Custom Binning* (Binning Kustom)
   *Custom Binning* memungkinkan pengguna untuk menentukan batas-batas bins sesuai dengan kebutuhan spesifik mereka. Ini memberikan fleksibilitas untuk mengelompokkan data berdasarkan kriteria yang lebih relevan.

6. Transformasi untuk Data Waktu
   Data waktu (*time series*) memiliki karakteristik khusus yang memerlukan teknik preprocessing yang berbeda dari data biasa. Transformasi data waktu bertujuan untuk mengekstrak informasi penting dari data dan membuatnya lebih siap untuk analisis atau pemodelan. Data waktu dapat mencakup data finansial, data sensor, log aktivitas, dan banyak lagi. Adapun Teknik-Teknik dalam proses transformasi untuk data waktu sebagai berikut.
   a. Ekstraksi Komponen Waktu
   b. Resampling
   c. Rolling Statistics
   d. Differencing
   e. Lag Features
   f. Handling Missing Values
   g. Normalization and Standardization
   h. Pembuatan Fitur

## Reduksi Dimensi

Reduksi dimensi adalah teknik dalam *preprocessing data* yang bertujuan untuk mengurangi jumlah fitur (dimensi) dalam dataset tanpa kehilangan informasi penting. Teknik ini sangat berguna ketika Anda memiliki dataset dengan jumlah fitur yang sangat besar, yang dapat menyebabkan masalah seperti *overfitting*, peningkatan waktu komputasi, dan kesulitan dalam visualisasi data.

Berikut manfaat dari proses Reduksi dimensi dalam teknik *preprocessing*:

1. Mengurangi *Overfitting*
   Dengan mengurangi jumlah fitur, risiko *overfitting* dapat diminimalkan karena model menjadi lebih sederhana dan tidak terlalu menyesuaikan diri dengan data pelatihan.
2. Mengurangi Waktu Komputasi
   Mengurangi jumlah fitur berarti mengurangi jumlah perhitungan yang diperlukan, sehingga proses pelatihan dan prediksi menjadi lebih cepat.
3. Memudahkan Visualisasi
   Data dengan dimensi rendah lebih mudah untuk divisualisasikan dan dianalisis.
4. Menghilangkan Redundansi
   Fitur yang sangat berkorelasi dapat dihapus, menghilangkan redundansi dan meningkatkan interpretabilitas model.

## Pembagian Data

Pembagian data adalah langkah penting dalam proses *Machine Learning* yang melibatkan pemisahan dataset asli menjadi beberapa subset untuk melatih, memvalidasi, dan menguji model. Langkah ini memastikan bahwa model *Machine Learning* dapat digeneralisasi dengan baik pada data yang tidak

terlihat sebelumnya dan membantu dalam mengukur kinerja model secara objektif.

Adapun tujuan dalam proses Pembagian Data sebagai berikut.

1. Melatih Model
   Dataset pelatihan digunakan untuk melatih model.
2. Validasi Model
   Dataset validasi digunakan untuk menyetel hyperparameter model dan mencegah overfitting.
3. Mengukur Kinerja Model
   Dataset pengujian digunakan untuk mengukur kinerja akhir model pada data yang benar-benar baru.

Berikut adalah Teknik-Teknik dalam proses Pembagian Data:

1. *Train-Test Split*
2. *Cross-Validation*
3. *Stratified Sampling*
4. *Time Series Split*

# BAB 8 STATISTIK DESKRIPTIF DAN INFERENSIAL

## Pendahuluan

Perkembangan ilmu pengetahuan dan teknologi dewasa ini mendorong perkembangan kehidupan manusia dalam berbagai sektor kehidupan. Berbagai sektor tersebut berkembang semakin kompleks dibarengi dengan arus informasi dan data yang semakin kompleks pula. Dengan demikian, informasi dalam bentuk data menjadi bagian vital dalam perkembangan kehidupan manusia dewasa ini. Data dapat dianalogikan sebagai sebuah sumber energi baru untuk kehidupan manusia. Bagi sebuah lembaga, perusahaan, maupun negara, data menjadi aset berharga yang dapat digunakan untuk melakukan perencanaan program-program pengembangan lembaga bahkan dasar perumusan strategi untuk mencapai keuntungan secara finansial bagi perusahaan. Oleh karena itu, pemahaman terhadap data itu sendiri menjadi sangat penting. Memahami data dapat dilakukan melalui ilmu data (sains data).

Sains data mencakup berbagai upaya untuk mengekstrak pengetahuan yang berguna dan valid dari data. Sains data mencakup berbagai aktivitas seperti pengembangan infrastruktur data, pengumpulan dan pengelolaan data, pengembangan dan penerapan model dan algoritma untuk analisis data, serta visualisasi dan komunikasi data secara kuantitatif. Sains data adalah ilmu tentang proses mengekstraksi pengetahuan dari data dalam berbagai bentuk. Proses ini

melibatkan pembersihan data, integrasi, visualisasi, dan analisis kumpulan data untuk mengungkap pola dan tren.

Sains data adalah campuran dari berbagai algoritme, alat, prinsip, dan bahasa untuk mengidentifikasi pola yang tersembunyi dalam variabel dalam kumpulan data. Sains data digunakan untuk membuat keputusan berdasarkan prediksi yang dibuat menggunakan kumpulan data yang ada. Dengan demikian, sains data adalah proses penggunaan data dalam jumlah besar untuk memperoleh wawasan bermanfaat dan membantu pengambilan keputusan. Salah satu aspek penting dalam sains data adalah statistik yang merupakan tulang punggung interpretasi dan prediksi data. Ilmuwan data menggunakan teknik statistik untuk mendapatkan interpretasi dari kumpulan data yang besar dan kompleks. Mereka menggunakan metode statistik untuk mengidentifikasi pola, hubungan, dan tren data.

Statistik adalah cabang matematika. Statistik berkaitan dengan pengumpulan, analisis, interpretasi, penyajian, dan pengorganisasian data numerik. Ini memberikan metode dan teknik untuk merangkum dan mendeskripsikan data, membuat kesimpulan, dan menarik kesimpulan tentang populasi berdasarkan data sampel. Statistik memainkan peran penting dalam sains data. Menyusun pertanyaan secara statistik memungkinkan pemanfaatan sumberdaya data untuk mengekstraksi pengetahuan dan mendapatkan jawaban yang lebih baik. Dengan demikian, sains data adalah bagian terapan dari statistik yang menggunakan metode statistik untuk menganalisis data dalam jumlah besar dan memahami hasilnya dengan lebih baik.

Berdasarkan metodologinya, statistik dibedakan menjadi dua yaitu statistik deskriptif dan statistik inferensial. Statistik

deskriptif memberikan ringkasan singkat untuk kumpulan data tertentu. Ringkasan ini dapat berlaku untuk keseluruhan kumpulan data atau sampel tertentu dalam kumpulan data. Statistik deskriptif tidak digunakan untuk perumusan kesimpulan terhadap kumpulan data. Sementara itu, statistik inferensial adalah statistik yang digunakan untuk mengidentifikasi dan membuat berbagai kesimpulan penting tentang populasi. Statistik inferensial berguna ketika data yang diberikan tidak layak atau praktis untuk menganalisis setiap kelompok dari keseluruhan populasi. Statistik inferensial terdiri dari dua tipe, yaitu statistik parametrik dan statistik non-parametrik. Statistik parametrik digunakan untuk menganalisis data rasio atau interval dari sebuah populasi yang berdistribusi normal. Sementara itu, statistik non-parametrik digunakan untuk menganalisis data ordinal dan ordinal dari populasi yang bebas distribusi.

Pada chapter ini akan dibahas secara detail penggunaan statistik dalam sains data, dimulai dengan ulasan persamaan sains data dan statistik, kemudian peran statistik dalam sains data, dan terakhir ulasan tentang statistik deskriptif dan inferensial beserta contoh-contohnya. Dengan demikian, melalui pemahaman materi dalam chapter ini maka pembaca diharapkan mampu mengkonstruksi pengetahuannya secara komprehensif tentang konsep dasar statistik deskriptif dan inferensial sebagai perangkat analisis dalam sains data.

## Persamaan Sains Data dan Statistik

Sains data dan statistik memiliki beberapa persamaan, diantaranya dalam konteks analisis data, pengumpulan data dan pra-pemrosesan data, pembuatan model, penalaran deduktif-induktif, ukuran ketidakpastian, dan presentasi hasil.

## Analisis Data

Analisis data berfungsi sebagai landasan bersama antara sains data dan statistik. Baik pada sains data dan statistik, tujuan analisis data adalah untuk mengekstrak informasi dan menarik kesimpulan yang bermakna. Ilmuwan data dan ahli statistik sama-sama terlibat dalam proses pengumpulan data, pengorganisasian, dan penggunaan metode kuantitatif untuk memahami fenomena, dan membuat prediksi yang akurat.

## Pengumpulan dan Pra-Pemrosesan Data

Baik dalam sains data maupun statistik, langkah awal proses analisis adalah pengumpulan data. Proses pengumpulan data memiliki beberapa langkah penting untuk memperoleh data yang relevan dan andal. Pada kedua bidang tersebut, para ilmuwan data maupun ahli statistik mengidentifikasi sumber data yang sesuai berisi informasi yang diperlukan untuk proses analisis. Mereka mengekstraksi data dari berbagai sumber dan menggabungkannya ke dalam format yang sesuai untuk analisis. Pengumpulan data dilakukan melalui web scraping, data mining, atau perekaman data melalui sensor dan perangkat. Proses pengumpulan data tidak lengkap tanpa menjamin kualitas data. Baik sains data maupun statistik menekankan pada validasi dan verifikasi data. Ini mencakup pemeriksaan keakuratan, konsistensi, kelengkapan, dan keandalan data yang dikumpulkan. Proses penjaminan kualitas dilakukan untuk menyelesaikan masalah apa pun terkait integritas data, nilai yang hilang, duplikasi, atau outlier.

Setelah tahap pengumpulan data dan penjaminan kualitas, baik ilmuwan data maupun ahli statistik menyadari pentingnya pra-pemrosesan data. Pra-pemrosesan data melibatkan pembersihan data dengan menghilangkan segala

ketidakkonsistenan, kesalahan, atau gangguan yang dapat mempengaruhi keakuratan analisis selanjutnya. Penanganan nilai-nilai yang hilang dan outlier diperlukan untuk memastikan integritas dan keandalan data.

Dengan melakukan pengumpulan data secara menyeluruh dan prosedur pra-pemrosesan data, ilmuwan data dan ahli statistik dapat membangun landasan yang kuat untuk analisis lebih lanjut. Data yang bersih dan andal ini memungkinkan mereka melakukan analisis statistik yang bermakna, mengembangkan model yang akurat, dan menarik kesimpulan yang andal yang mendorong wawasan dan pengambilan keputusan.

### Pengembangan Model Statistik

Kesamaan signifikan lainnya antara sains data dan statistik terletak pada pengembangan model statistik atau matematika. Pembuatan dan pemanfaatan model diperlukan dalam menganalisis data dan mengekstraksi informasi. Model statistik dan matematika berfungsi sebagai alat ampuh yang memungkinkan para profesional untuk mewakili situasi dunia nyata dengan cara yang terstruktur dan terukur. Model memiliki jenis yang berbeda yaitu model regresi, model deret waktu, algoritma pengelompokan, atau model pembelajaran mesin. Model tersebut dirancang untuk menangkap dan mewakili hubungan, atau ketergantungan yang ada dalam data.

Ilmuwan data dan ahli statistik menggunakan model ini untuk membuat prediksi, memperkirakan hasil di masa depan, mengidentifikasi tren, atau mengeksplorasi hubungan antar variabel. Dengan menyesuaikan data ke model yang dipilih, praktisi dapat memperkirakan parameter, mengevaluasi signifikansi variabel, dan menghasilkan wawasan yang

mendorong pengambilan keputusan.

### Penalaran Deduktif - Induktif

Metode utama untuk memahami dan menilai data baik dalam sains data maupun statistik bergantung pada penalaran deduktif dan induktif. Para ahli statistik sering kali menggunakan penalaran deduktif dalam pekerjaannya. Mereka memanfaatkan teori, prinsip, dan kerangka statistik yang ada untuk merumuskan hipotesis dan mengembangkan teori tentang hubungan antar variabel atau struktur data. Mengawali dengan pengetahuan yang sudah ada dan membuat kesimpulan logis, ahli statistik dapat menghasilkan prediksi yang dapat diuji dan merancang eksperimen untuk mengumpulkan data yang dapat mendukung atau menyangkal hipotesis mereka. Penalaran deduktif memungkinkan ahli statistik untuk menerapkan prinsip dan teori yang diketahui pada konteks tertentu dan membuat prediksi yang tepat tentang data.

Sebaliknya, ilmuwan data sering kali mengandalkan penalaran induktif saat mereka mengeksplorasi kumpulan data yang besar dan kompleks. Mereka menganalisis pola, tren, dan hubungan yang diamati dalam data untuk menarik kesimpulan umum dan mengembangkan algoritma pembelajaran mesin. Penalaran induktif membantu ilmuwan data untuk mengamati kesamaan dan keteraturan dalam data dan menggeneralisasi-kannya ke dalam model atau algoritma prediktif. Mereka menggunakan pola yang diamati sebagai dasar untuk menyimpulkan prinsip-prinsip yang lebih luas yang dapat diterapkan pada data baru.

### Ukuran Ketidakpastian

Ilmu data dan statistik memiliki konsep yang sama dalam

mengukur ketidakpastian. Oleh karena itu, untuk membuat keputusan yang lebih baik, kedua bidang tersebut menyadari pentingnya memahami dan mengukur ketidakpastian ini. Para ahli statistik menggunakan berbagai teori dan metode untuk mengukur dan mengelola ketidakpastian. Interval kepercayaan biasanya digunakan untuk menghitung kisaran nilai yang kemungkinan besar termasuk dalam suatu parameter atau statistik. Interval kepercayaan memberikan ukuran ketidakpastian yang terkait dengan estimasi dan membantu ahli statistik menilai ketepatan dan keandalan temuan mereka. Pengujian hipotesis adalah alat statistik lain yang digunakan untuk mengukur ketidakpastian dengan menentukan kemungkinannya.

Demikian pula, ilmuwan data menggunakan metode serupa untuk mengatasi ketidakpastian dalam model pembelajaran mesin mereka. Ilmuwan data mengevaluasi efektivitas dan ketidakpastian model mereka selama fase pelatihan menggunakan metode seperti validasi silang atau bootstrapping. Mereka dapat memperkirakan keakuratan prediksi model dan mengevaluasi konsistensi hasil di beberapa subkumpulan data menggunakan metode ini. Model probabilistik, teknik Bayesian, dan metode ansambel semuanya berguna untuk memperoleh perkiraan ketidakpastian model pembelajaran mesin.

### Presentasi Hasil

Baik dalam ilmu data maupun statistik diperlukan penyajian hasil yang efektif dengan cara yang jelas dan mudah dipahami. Ilmuwan data maupun ahli statistik mengetahui pentingnya menyajikan informasi dan temuan kepada pengambil keputusan dan pemangku kepentingan.

Ilmuwan data dan ahli statistik menyajikan informasi rumit dengan cara yang menarik secara visual. Mereka menggunakan grafik, bagan, dan infografis untuk membuat presentasi. Pengambil keputusan dan pemangku kepentingan dapat dengan cepat memahami temuan visual dari ilmuwan data.

Selain itu, interpretasi hasil merupakan langkah umum dalam ilmu data dan statistik. Ahli statistik dan ilmuwan data melakukan lebih dari sekadar menunjukkan angka mentah dan hasil statistik. Mereka menggambarkan apa arti hasil dan bagaimana kaitannya dengan masalah atau pertanyaan yang ada. Penjelasan ini membantu masyarakat memahami betapa pentingnya temuan tersebut dan menjelaskan kepada mereka apa yang harus dilakukan selanjutnya berdasarkan hasil tersebut.

## Peran Statistik Dalam Sains Data

Secara umum, terdapat beberapa peran statistik dalam sains data, seperti:

1. Eksplorasi data
   Statistik menyediakan alat untuk mengeksplorasi dan memahami data dengan menghitung ukuran data seperti mean, median, dan deviasi standar, serta memvisualisasikan data melalui grafik dan bagan.
2. Pemrosesan awal data
   Sebelum melakukan analisis atau pemodelan apa pun, ilmuwan data sering kali perlu membersihkan dan memproses data terlebih dahulu. Statistik membantu mereka dalam menangani data yang hilang, outlier, serta mengubah variabel jika diperlukan.
3. Pengujian hipotesis

Pengujian hipotesis memungkinkan ilmuwan data membentuk asumsi tentang kumpulan data tertentu dan menguji validitasnya secara ketat. Proses ini melibatkan pendefinisian hipotesis nol. Ini mewakili tidak adanya pengaruh atau hubungan yang signifikan. Melalui penghitungan dan analisis statistik, ilmuwan data mengevaluasi bukti yang bertentangan dengan hipotesis nol dan menentukan apakah hipotesis tersebut harus ditolak atau dipertahankan. Ilmuwan data menggunakan pengujian hipotesis untuk membuat kesimpulan tentang suatu populasi berdasarkan data sampel. Hal ini membantu mereka mengevaluasi signifikansi hubungan atau perbedaan antar variabel.

4. Analisis deskriptif
   Hasil analisis statistik deskriptif memberikan informasi tentang karakteristik dan perilaku suatu kumpulan data. Dengan memanfaatkan berbagai ukuran seperti mean, median, mode, dan deviasi standar, ilmuwan data dapat meringkas dan memahami fitur data secara efektif. Analisis statistik deskriptif ini memberikan ringkasan singkat dari kumpulan data bagi ilmuwan data untuk memahami kecenderungan sentral, penyebaran, dan distribusi titik-titik data secara keseluruhan. Tanpa analisis statistik deskriptif ini, ilmuwan data akan menghadapi kesulitan dalam memahami karakteristik mendasar kumpulan data.
5. Pemodelan prediktif
   Statistik memberikan landasan untuk membangun dan memvalidasi model prediktif. Ini membantu ilmuwan data dalam memilih algoritme yang sesuai dan mengevaluasi performa model untuk meramalkan dan

memprediksi hasil berdasarkan informasi yang diambil dari data historis. Dalam pemodelan prediktif, ilmuwan data menggunakan model statistik dan algoritme untuk menemukan hubungan dan ketergantungan dalam data. Dengan menganalisis data masa lalu dan mengidentifikasi variabel yang relevan, para ilmuwan data dapat mengembangkan model yang menangkap pola mendasar dan membuat prediksi tentang kejadian atau perilaku di masa depan.

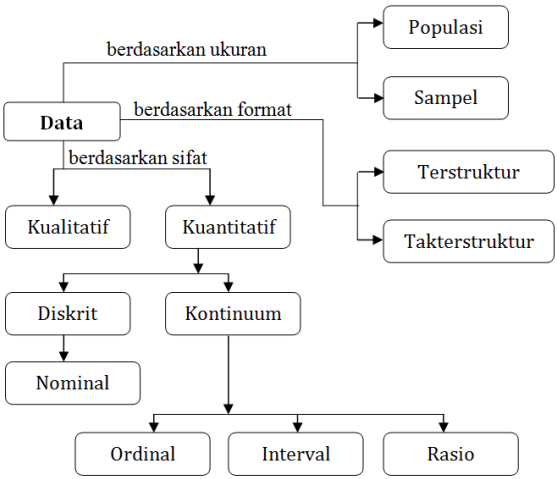
## Data dan Karakteristiknya

Data adalah kumpulan informasi faktual berdasarkan angka, kata, pengamatan, pengukuran yang dapat dimanfaatkan untuk perhitungan, pembahasan dan penalaran. Berdasarkan sifatnya maka terdapat dua tipe data, yaitu data kualitatif dan data kuantatif (Gambar 8.1). Data kualitatif adalah data yang berupa kata, kalimat, serta gambar, contohnya adalah hari ini hujan; besok siang diprediksi akan hujan deras; semua dosen di perguruan tinggi negeri di Indonesia bergelar akademik doktor. Data ini tidak dinyatakan dengan angka-angka sehingga tidak dapat dikuantifikasi atau dilakukan pengukuran secara numerik. Sementara itu, data kuantitatif adalah data yang berbentuk angka sehingga dapat dilakukan kuantifikasi atau penyajian secara numerik. Contoh data kuantitatif adalah jumlah pendudukan usia produktif (15–64 tahun) di Indonesia pada tahun 2022 diperkirakan sebanyak 190,98 juta jiwa atau 69,25%; jumlah penduduk usia non-produktif pada tahun 2022 mencapai 84,8 juta jiwa atau 30,75%.

Data kualitatif dapat dibedakan menjadi tiga tipe data, yaitu data Binomial, data Nominal, dan data Ordinal. Data Binomial adalah set data berupa pasangan seperti “bagus atau

jelek"; "benar atau salah". Data nominal adalah data yang sifatnya tidak berurutan (*unordered data*), contohnya warna "merah, kuning, hijau". Sementara itu, data ordinal adalah data yang memperhatikan level atau urutan data, seperti "pendek, medium, panjang".

Data kuantitatif dapat dibedakan menjadi dua jenis, yaitu data diskrit dan data kontinuum. Data diskrit adalah data yang diperoleh dari hasil menghitung atau membilang (bukan hasil pengukuran), contohnya jumlah siswa di kelas XA SMA Saverius X adalah 30 orang. Data seperti ini juga dapat dinyatakan sebagai contoh data nominal dalam kelompok data kuantitatif. Data kontinuum adalah data yang diperoleh dari hasil pengukuran. Kelompok data ini terdiri dari data ordinal, interval, dan data rasio.




*Gambar 8.1 Klasifikasi data*

Data ordinal dalam kelompok data kuantitatif adalah data yang sifatnya memiliki urutan, jenjang, atau level, contohnya data juara kelas seperti juara I, juara II, dan juara III; data jabatan dalam sebuah instansi pemerintahan seperti eselon I,

eselon II, eselon III; serta data golongan pegawai seperti golongan I, golongan II, golongan III, serta golongan IV. Data interval adalah data yang berupa interval dan memiliki rentangan tertentu, contohnya data temperatur udara, data tekanan udara, data kelembaban, data persebaran pendudukan berdasarkan umur, serta contoh data sejenis lainnya. Data rasio adalah data yang jaraknya sama dan memiliki nilai nol absolut. Data ini dapar dikenai operasi penjumlahan dan perkalian. Contoh data rasio adalah data hasil pengukuran panjang, berat, volume, maupun luas suatu bidang.

Berdasarkan ukurannya data diklasifikasikan menjadi dua kelompok, yaitu data populasi dan data sampel. Data populasi adalah data semesta dari objek yang dikaji (objek yang menjadi perhatikan/pusat pengkajian) dan dinyatakan dengan notasi "N". Contohnya adalah data jumlah pendudukan Indonesia berusia produktif pada tahun 2024. Dalam konteks ini, objek yang dikaji adalah penduduk Indonesia berusia produktif pada tahun 2024. Data semestanya adalah data jumlah penduduk Indonesia yang berusia produktif (15 – 64 tahun) pada tahun 2024. Angka-angka yang kita peroleh saat menggunakan populasi disebut parameter. Sementara itu, data sampel adalah bagian dari populasi yang dilambangkan dengan "n" dan angka-angka yang kita peroleh saat menggunakan sampel disebut statistik.

Berdasarkan formatnya, data diklasifikasikan menjadi dua kelompok, yaitu data terstruktur dan data tak terstruktur. Data terstruktur adalah data yang memiliki format tertentu, data yang teroganisasi, serta data yang mudah dipahami oleh bahasa mesin (machine language). Contohnya adalah data nama, alamat, serta tanggal lahir. Sementara itu, data tak terstruktur adalah data yang tidak terformat, tidak terorganisasi, tidak

mampu dikenali dengan mudah oleh bahasa mesin, serta data yang harus dianalisis menggunakan metode konvensional, contohnya data text, audio, video, serta aktivitas sosial media.

## Statistik Deskriptif

Statistik deskriptif berfungsi untuk merepresentasikan data sampel atau populasi dalam berbagai metode penyajian data tanpa disertai dengan proses analisis untuk memperoleh kesimpulan yang dapat berlaku umum. Dengan demikian, bahasan statistik deskriptif ini meliputi penyajian data, pemusatan data, serta pengukuran variasi data sampel maupun populasi.

### Penyajian Data

Penyajian data dapat dilakukan dengan berbagai bentuk, seperti tabel dan diagram. Pertama, tabel merupakan visualisasi matrik, yaitu perpaduan antara baris dan kolom. Terdapat dua jenis tabel, yaitu tabel baris – kolom dan tabel distribusi frekuensi. Berikut ini adalah contoh tabel baris – kolom yang umum kita jumpai dalam buku-buku bacaan statistika serta matematika.

*Tabel 8.1 Jumlah lulusan SMA Negeri 1 Denpasar pada tahun 2004 – 2008*

| Tahun | Jenis Kelamin | | Jumlah |
|---|---|---|---|
| | Perempuan | Laki-laki | |
| 2004 | 82 | 100 | 182 |
| 2005 | 90 | 80 | 170 |
| 2006 | 95 | 90 | 185 |
| 2007 | 100 | 95 | 195 |
| 2008 | 100 | 110 | 210 |

Tabel distribusi frekuensi sama dengan tabel pada umumnya, hanya saja pada salah satu kolom terdiri dari data berupa data tunggal atau interval dan kolom lainnya adalah frekuensi (jumlah data) yang relevan dengan data tunggal atau data interval tersebut. Berikut ini contoh penyajian tabel distribusi frekuensi data tunggal.

*Tabel 8.2 Distribusi frekuensi data tunggal hasil tes statistik lanjut mahasiswa semester IV Prodi Sains Data Universitas X tahun akademik 2023/2024*

| Skor | *f* | Skor | *f* | Skor | *f* |
|---|---|---|---|---|---|
| 34 | 1 | 56 | 5 | 77 | 2 |
| 35 | 2 | 58 | 1 | 78 | 3 |
| 37 | 1 | 60 | 1 | 79 | 1 |
| 40 | 1 | 65 | 4 | 80 | 2 |
| 42 | 1 | 67 | 4 | 84 | 1 |
| 43 | 1 | 68 | 1 | 85 | 3 |
| 44 | 1 | 70 | 2 | 87 | 6 |
| 45 | 5 | 72 | 1 | 88 | 5 |
| 47 | 1 | 73 | 1 | 89 | 5 |
| 49 | 1 | 74 | 1 | 90 | 2 |
| 54 | 2 | 75 | 1 | 93 | 1 |
| 55 | 1 | 76 | 8 | 95 | 1 |

Tabel distribusi frekuensi data tunggal ini dapat disajikan dalam bentuk tabel distribusi frekuensi data berkelompok yaitu seperti Tabel 8.3. Oleh karena itu, untuk menyajikan data dalam bentuk tabel distribusi data berkelompok yang bersumber dari data tunggal maka dilakukan melalui beberapa tahapan berikut.

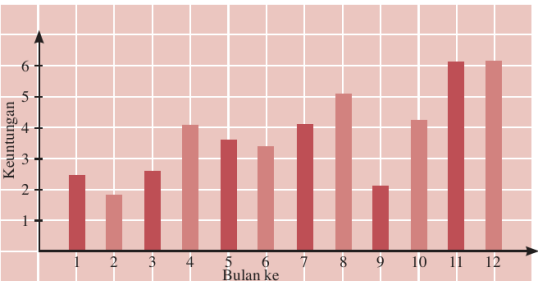
1. Menentukan jangkauan data;
2. Menentukan banyaknya kelas interval;
3. Menentukan panjang kelas interval;
4. Menentukan tepi bawah kelas interval data yang pertama; dan
5. Menyusun tabel distribusi frekuensi.

Data yang tersaji pada Tabel 8.2 dapat disajikan dalam bentuk tabel distribusi frekuensi sebagai berikut.

*Tabel 8.3 Distribusi frekuensi data berkelompok hasil tes statistik lanjut mahasiswa semester IV Prodi Sains Data Universitas X tahun akademik 2023/2024*

| Nilai | Tally | Frekuensi |
|---|---|---|
| 33 – 40 | 𝍸 | 5 |
| 41 – 48 | 𝍸 IIII | 9 |
| 49 – 56 | 𝍸 IIII | 9 |
| 57 – 64 | II | 2 |
| 65 – 72 | 𝍸 𝍸 II | 12 |
| 73 – 80 | 𝍸 𝍸 𝍸 IIII | 19 |
| 81 - 88 | 𝍸 𝍸 𝍸 | 15 |
| 89 – 96 | 𝍸 IIII | 9 |
| | Jumlah | 80 |

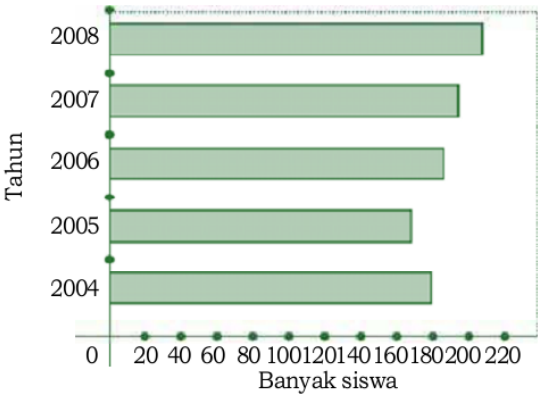
Penyajian data juga dapat dilakukan dalam bentuk diagram. Terdapat beberapa jenis diagram untuk penyajian data, seperti diagram batang (histogram), diagram garis, dan diagram lingkaran yang diilustrasikan berturut-turut oleh Gambar 8.2 – 8.4. Gambar 8.2 mengilustrasikan diagram batang (histogram vertikal) tentang data keuntungan produksi sebuah perusahaan setiap bulannya dalam satu tahun terakhir.




*Gambar 8.2 Keuntungan penjualan roti PT Cipta Rasa Abadi pada tahun 2024 (dalam satuan miliar rupiah)*

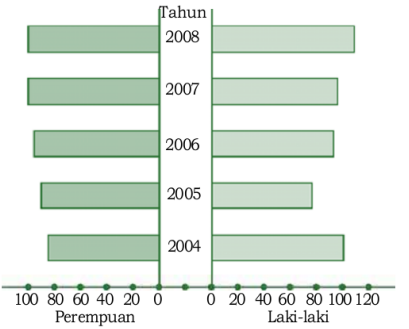
Secara horisontal, diagram batang juga dapat disajikan

sebagai berikut.




*Gambar 8.3 Jumlah lulusan siswa SMA Saverius X pada kurun waktu 2004 – 2008*

Apabila data pada Gambar 8.3 disajikan lebih spesifik lagi, maka kita dapat menggunakan diagram batang yang terkatagorisasi seperti Gambar 8.4.



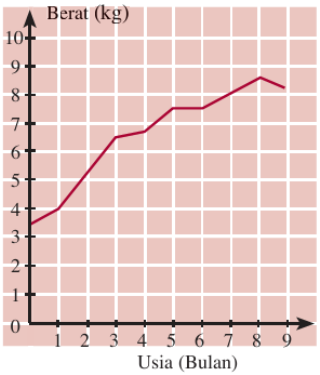

*Gambar 8.4 Jumlah lulusan siswa SMA Saverius X berdasarkan jenis kelamin pada kurun waktu 2004 – 2008*

Cara lain untuk menyajikan data adalah menggunakan diagram garis seperti diilustrasikan oleh Gambar 8.5. Sebagai contoh kita memiliki data perkembangan berat badan bayi selama 9 bulan pertama.

*Tabel 8.4 Berat badan bayi selama 9 bulan pertama*

| Usia (bulan) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Berat Badan (kg) | 3,5 | 4 | 5,2 | 6,4 | 6,8 | 7,5 | 7,5 | 8 | 8,8 | 8,6 |

Data pada Tabel 8.4 dapat dikonversi dalam bentuk diagram garis berikut.




*Gambar 8.5 Berat badan bayi selama 9 bulan pertama*

Ada kalanya sebuah kumpulan data dapat disajikan dengan menggunakan diagram batang daun. Misalnya disediakan data tunggal seperti berikut.

45 10 20 31 48 20 29 27 11 8
25 21 42 24 22 36 33 22 23 13
34 29 25 39 32 38 50 5

Data tersebut dapat disajikan dalam bentuk diagram batang daun seperti diilustrasikan pada Gambar 8.6.

| Batang | Daun |
|---|---|
| 5 | 0 |
| 4 | 2 5 8 |
| 3 | 1 2 3 4 6 8 9 |
| 2 | 0 0 1 2 2 3 4 5 5 7 9 9 |
| 1 | 0 1 |
| 0 | 5 8 |

*Gambar 8.6 Contoh ilustrasi diagram batang daun*

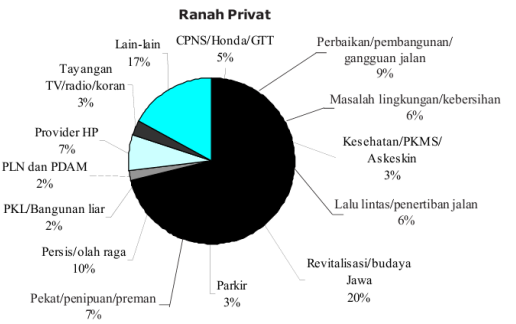
Pada Gambar 8.6 tampak dua kolom yaitu kolom batang dan kolom daun. Kolom batang menyatakan angka puluhan dan kolom daun menyatakan angka satuan. Data tunggal tersebut disajikan secara urut dari data terbesar hingga data terkecil.

Selain menggunakan ketiga jenis diagram tersebut, data dapat juga disajikan dalam bentuk diagram lingkaran. Diagram lingkaran adalah bentuk penyajian data statistik dengan menggunakan takaran berupa lingkaran. Juring-juring dari daerah lingkaran menunjukkan porsi masing-masing kelompok data. Gambar 8.7 mengilustrasikan diagram lingkaran yang disajikan dalam versi persentase dari jumlah pengaduan (ranah privat) masyarakat yang dilansir oleh Koran Xpress pada tanggal 30 Juni 2024 dengan data mentahan sebagai berikut.

*Tabel 8.5 Jumlah pengaduan (ranah privat) masyarakat yang dilansir oleh Koran Xpress pada tanggal 30 Juni 2024*

| No | Ranah Privat | Persentase |
|---|---|---|
| 1. | CPNS/Honda/GTT | 5 % |
| 2. | Perbaikan/pembangunan/gangguan jalan | 9 % |
| 3. | Masalah lingkungan/ kebersihan | 6 % |
| 4. | Kesehatan/PKMS/Askeskin | 3 % |
| 5. | Lalu lintas/penertiban jalan | 6 % |
| 6. | Revitalisasi/budaya Jawa | 20 % |
| 7. | Parkir | 3 % |
| 8. | Pekat/penipuan/preman | 7 % |
| 9. | Persis/olahraga | 10 % |
| 10. | PKL/bangunan liar | 2 % |
| 11. | PLN dan PDAM | 2 % |
| 12. | Provider HP | 7 % |
| 13. | Tayangan TV/radio/koran | 3 % |
| 14. | Lain-lain | 17 % |
| | Jumlah | 100 % |

Diagram lingkaran dari data pada Tabel 8.5 disajikan seperti berikut.




*Gambar 8.7 Jumlah pengaduan (ranah privat) masyarakat yang dilansir oleh Koran Xpress pada tanggal 30 Juni 2024*

## Ukuran Pemusatan Data

Ukuran pemusatan data dalam suatu rangkaian data adalah sebuah nilai yang dapat mewakili rangkaian data tersebut. Suatu rangkaian data umumnya terkonsentrasi atau terpusat pada suatu nilai pemusatan ini. Ukuran statistik yang dapat menjadi pusat dari rangkaian data disebut ukuran pemusatan data. Ukuran pemusatan data dapat digunakan untuk menganalisis data lebih lanjut. Ukuran pemusatan data terdiri dari tiga komponen, yaitu rata-rata (*mean*), nilai tengah (*median*), dan nilai yang paling sering muncul/nilai dengan frekuensi terbesar (*modus*).

1. Rata-rata

   Rata-rata sama dengan jumlah seluruh nilai dalam kumpulan data dibagi dengan jumlah nilai dalam kumpulan data. Oleh karena itu, jika kita memiliki n nilai dalam suatu kumpulan data dan nilai tersebut adalah x1, x¬2, x3, ..., xn, maka rata-rata sampelnya diberikan seperti di bawah ini.

$$\text{Rataan} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \quad \text{atau} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Apabila rangkaian data tersebut disajikan dalam bentuk tabel distribusi frekuensi, maka nilai rata-rata sampelnya dapat dihitung menggunakan persamaan berikut.

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + .... + f_nx_n}{f_1 + f_2 + ... + f_n} \quad \text{atau} \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

2. Median

Median adalah nilai tengah dari rangkaian data yang telah diurut berdasarkan besarnya (kuantitasnya). Sebagai contoh, kita memiliki rangkaian data sebagai berikut:

2, 5, 4, 5, 6, 7, 5, 9, 8, 4, 6, 7, 8

Setelah dilakukan pengurutan, maka diperoleh rangkaian data sebagai berikut.

2, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8, 8, 9

Nilai tengah pada rangkaian data tersebut adalah 6, yaitu:

2, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8, 8, 9

↓

*Me*

Secara umum, rumus untuk menentukan nilai median rangkaian data yang bersifat tunggal adalah

- Untuk *n* ganjil: $Me = x_{\frac{1}{2}(n+1)}$
- Untuk *n* genap: $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

Sementara itu, untuk data berkelompok atau data distribusi frekuensi berkelompok, maka nilai mediannya dapat dihitung menggunakan persamaan berikut.

$$Me = b_2 + c\left(\frac{\frac{1}{2}N - F}{f}\right)$$

dimana:

$b2$ = tepi bawah kelas median
$c$ = lebar kelas
$N$ = banyaknya data
$F$ = frekuensi kumulatif kelas sebelum kelas median
$f$ = frekuensi kelas median

3. Modus

Modus adalah nilai yang paling sering muncul dalam rangkaian data yang diketahui. Perhatikan data yang diketahui berikut.

| Nilai | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frekuensi | 1 | 2 | 1 | 3 | 1 | 1 | 2 | 1 | 2 | 1 |

Data ini merupakan data yang tersebar secara tunggal, sehingga modusnya adalah 4 karena nilai 4 memiliki frekuensi terbesar yaitu 3.

Apabila data yang diketahui memiliki distribusi frekuensi berkelompok dengan beberapa kelas intervalnya seperti berikut maka nilai modusnya terletak pada kelas interval 70 – 74.

| Interval Kelas | Frekuensi |
|---|---|
| 40 – 44 | 2 |
| 45 – 49 | 2 |
| 50 – 54 | 6 |
| 55 – 59 | 8 |
| 60 – 64 | 10 |
| 65 – 69 | 11 |
| 70 – 74 | 15 |
| 75 – 79 | 6 |
| 80 – 84 | 4 |
| 85 – 89 | 4 |
| 90 – 94 | 3 |

Nilai modus sebaran data tersebut dapat dihitung menggunakan persamaan berikut.

$$Mo = b_0 + l \left( \frac{d_1}{d_1 + d_2} \right)$$

dimana:

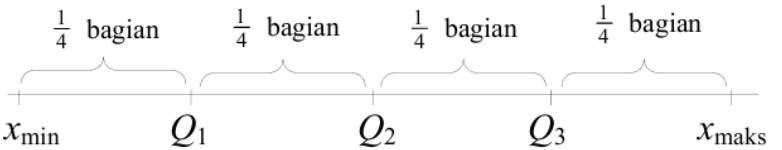
| | | |
|---|---|---|
| $b_o$ | = | tepi bawah kelas modus |
| $l$ | = | lebar kelas modus |
| $d_1$ | = | selisih frekuensi kelas modus dengan kelas sebelumnya |
| $d_2$ | = | selisih frekuensi kelas modus dengan kelas setelahnya |

**Ukuran Letak Data**

Apabila ukuran pemusatan data fokus pada sebuah nilai yang dapat mewakili rangkaian data tersebut, maka ukuran letak data fokus pada penentuan letas data dalam serangkaian data yang diketahui. Ukuran letak data meliputi, kuartil, desil, dan persentil.

1. Kuartil

   Kuartil adalah membagi data yang telah diurutkan menjadi empat bagian yang sama banyak. Oleh karena itu terdapat tiga kuartil, yaitu kuartil 1 (*Q1*), kuartil 2 (*Q2*), dan kuartil 3 (*Q3*). Berikut ilustrasi kuartil tersebut.




   Apabila rangkaian data yang diketahui adalah data tunggal, maka penentuan letak kuartil ke-i dapat dilakukan menggunakan persamaan berikut.

$$\text{Letak } Q_i = \frac{i(n+1)}{4}$$

dimana $i$ = 1, 2, dan 3 serta $n$ adalah jumlah data yang diketahui.

Apabila rangkaian data yang diketahui adalah berupa distribusi frekuensi data berkelompok dalam kelas-kelas interval tertentu, maka penentuan letak kuartil ke-i dapat dilakukan menggunakan persamaan berikut.

$$Q_i = b_i + l\left(\frac{\frac{i}{4}N - F}{f}\right)$$

dimana:

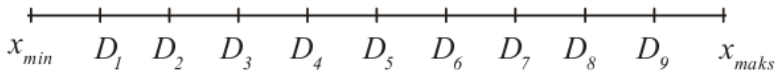
*Qi* = kuartil ke-i (1, 2, atau 3)
*bi* = tepi bawah kelas kuartil ke-i
*N* = banyaknya data
*F* = frekuensi kumulatif kelas sebelum kelas kuartil
*l* = lebar kelas kuartil ke-i
*f* = frekuensi kelas kuartil

Berdasarkan nilai kuartil tersebut, maka dapat ditentukan beberapa nilai besaran berikut.

a. Jangkauan (J), yaitu selisih nilai maksimum dan nilai maksimum data.
b. Jangkauan interkuartil (H) adalah selisih nilai kuartil ketiga dan kuartil kesatu
c. Jangkauan semi interkuartil (Qd) adalah setengah dari jangkauan interkuartil (H)
d. Langkah (L) adalah tiga kali nilai jangkauan semi interkuartil (Qd),

2. Desil

Jika kuartil membagi data menjadi empat bagian, maka desil membagi data menjadi sepuluh bagian, yaitu:




Apabila diketahui rangkaian data tunggal, maka letak desil ke-i dapat ditentukan dengan persamaan berikut.

$$\text{Letak } D_i \text{ di urutan data ke - } \frac{i(n+1)}{10}$$

dimana:

$D_i$ = desil ke-i (1, 2, 3, 4, 5, ..., 9)

$n$ = banyaknya data

Apabila diketahui data memiliki distribusi frekuensi berkelompok, maka letak desil ke-i dapat ditentukan dengan persamaan berikut.

$$D_i = b + l\left(\frac{\frac{i \cdot n}{10} - F}{f}\right)$$

dimana:

$D_i$ = desil ke-i (1, 2, ... 9)

$bi$ = tepi bawah kelas desil ke-i

$n$ = banyaknya data

$F$ = frekuensi kumulatif kelas sebelum kelas desil

$l$ = lebar kelas desil ke-i

$f$ = frekuensi kelas desil

3. Persentil

Persentil membagi data menjadi seratus bagian yang sama. Oleh karena itu, letak persentil ke-i untuk rangkaian data tunggal ditentukan dengan persamaan.

$$\text{Letak } P_i \text{ di urutan data ke - } \frac{i(n+1)}{100}$$

dimana:

$P_i$ = persentel ke-i (1, 2, 3, 4, 5, ..., 99)
$n$ = banyaknya data

Apabila diketahui data memiliki distribusi frekuensi berkelompok, maka letak persentil ke-i dapat ditentukan dengan persamaan berikut.

$$P_i = b + l\left(\frac{\frac{i \cdot n}{100} - F}{f}\right)$$

dimana:

| | | |
|---|---|---|
| $P_i$ | = | persentil ke-i (1, 2, ... 99) |
| $bi$ | = | tepi bawah kelas persentil ke-i |
| $n$ | = | banyaknya data |
| $F$ | = | frekuensi kumulatif kelas sebelum kelas persentil |
| $l$ | = | lebar kelas persentil ke-i |
| $f$ | = | frekuensi kelas persentil |

## Ukuran Penyebaran Data

Ukuran penyebaran data memberikan informasi sebaran data dari titik-titik pemusatan. Ukuran penyebaran data meliputi: jangkauan, simpangan rata-rata, simpangan baku, dan ragam atau variansi.

1. Jangkauan

   Jangkauan atau range adalah selisih data bernilai maksimum dan data bernilai minimum. Perhatikan data berikut: 6, 7, 3, 4, 8, 3, 7, 6, 10, 15, 20. Jangkauan data tersebut adalah J = 20 – 3 = 17. Apabila diketahui data terdistribusi berkelompok seperti berikut.

| Nilai | Frekuensi |
|---|---|
| 3 – 5 | 3 |
| 6 – 8 | 6 |
| 9 – 11 | 16 |
| 12 – 14 | 8 |
| 15 – 17 | 7 |
| 18 – 20 | 10 |

Jangkauannya adalah ditentukan dengan mengetahui terlebih dahulu nilai tengah kelas interval terbawah dan nilai tengah kelas interval teratas. Nilai tengah tersebut masing-masing mewakili data bernilai minimum dan data bernilai maksimum.

Nilai tengah kelas interval terendah, yaitu (3 + 5)/ 2 = 4. Sementara nilai tengah kelas interval teratas, yaitu (20 + 18)/2 = 19. Jadi, jangkauan data tersebut adalah 19 – 4 = 15.

2. Simpangan rata-rata

Simpangan rata-rata dari rangkaian data adalah nilai rata-rata dari selisih setiap data dengan nilai rerata totalnya. Data yang tersebar secara tunggal dapat dihitung nilai simpangan rata-ratanya menggunakan persamaan berikut.

$$SR = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

dimana:

$SR$ = simpangan rata-rata

$x_i$ = nilai data ke-i

$\bar{x}$ = rata-rata total serangkaian data

Apabila data yang diketahui memiliki distribusi frekuensi berkelompok, maka simpangan rata-rata data dapat

dihitung menggunakan persamaan.

$$SR = \frac{\sum_{i=1}^{n} f_i \left| x_i - \overline{x} \right|}{\sum_{i=1}^{n} f_i}$$

dimana:

*SR* = simpangan rata-rata
*xi* = nilai tengah data untuk interval kelas ke-i
$\overline{x}$ = rata-rata total serangkaian data
$f_i$ = frekuensi kelas interval ke-i

3. Simpangan baku

Simpangan baku digunakan untuk mengestimasi akurasi pengukuran data yang dihasilkan. Sebagai contoh ketika seorang mahasiswa melakukan praktikum pengukuran periode osilasi sebuah bandul matematis maka simpangan baku data periode osilasi bandul tersebut dapat digunakan untuk menyatakan tingkat akurasi pengukuran periode bandul itu. Simpangan baku juga diistilahkan dengan standar deviasi. Apabila data yang disajikan berupa data tunggal, maka simpangan baku data tersebut dihitung menggunakan persamaan berikut.

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} \overline{x} \right)^2}{n(n-1)}}$$

Persamaan ini digunakan untuk data yang jumlahnya lebih kecil dari 30 ($n < 30$) atau sering disebut sebagai statistik (data sampel). Sementara itu, apabila jumlah data yang diketahui lebih besar dari 30 ($n > 30$) maka simpangan baku yang dihitung adalah simpangan baku

populasi dengan perhitungan sebagai berikut.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Dimana:

$$n = \sum_{i=1}^{n} f_i$$

Kedua persamaan tersebut juga dapat disajikan dalam bentuk yang berbeda yaitu:

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} \bar{x}_i\right)^2}{n(n-1)}}$$

Untuk data yang jumlahnya lebih kecil dari 30 ($n$ < 30) dan

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} \bar{x}_i\right)^2}{n^2}}$$

Untuk data yang jumlahnya lebih besar dari 30 ($n$ > 30). Ingat bahwa $x_i$ menyatakan nilai data ke-i.

Apabila data yang tersedia memiliki distribusi frekuensi berkelompok, maka standar deviasi untuk data yang jumlahnya lebih kecil dari 30 ($n$ < 30) dapat ditentukan dengan persamaan berikut.

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}}$$

Sementara untuk data yang jumlahnya lebih besar dari 30 ($n$ > 30) dapat dihitung standar deviasinya menggunakan

persamaan berikut.

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n}}$$

Kedua persamaan tersebut juga dapat disajikan dalam bentuk lain yaitu:

$$s = \sqrt{\frac{n\sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i \bar{x}_i\right)^2}{n(n-1)}}$$

Untuk data yang jumlahnya lebih kecil dari 30 ($n$ < 30) dan

$$s = \sqrt{\frac{n\sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i \bar{x}_i\right)^2}{n^2}}$$

Untuk data yang jumlahnya lebih besar dari 30 ($n$ > 30).

4. Ragam atau variansi
   Tingkat variansi data terhadap nilai rata-rata sampel atau populasi disebut dengan ragam (variansi). Nilai ragam dapat dihitung dengan mengkuadratkan nilai simpangan baku (s) baik untuk simpangan baku sampel atau simpangan baku populasi. Jadi, Variansi = s2. Koefesien variansi data sampel atau populasi dapat dihitung menggunakan persamaan berikut.

$$KV = \frac{s}{\bar{x}}$$

## Ukuran Asimetri

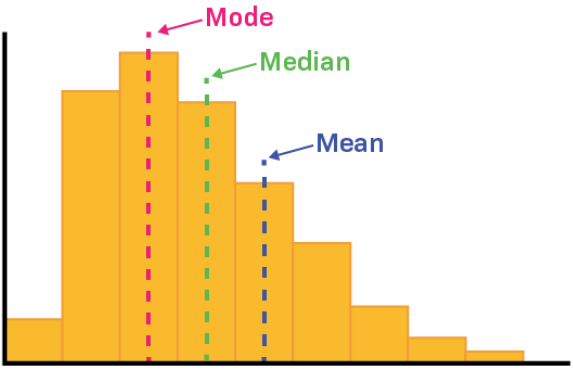
Ukuran asimetri menunjukkan apakah pengamatan dalam suatu kumpulan data terfokus pada satu sisi. Ukuran asimetri ini dinyatakan dengan tingkat kemiringan distribusi

data (Skewness). Nilai skewness dapat dihitung dengan persamaan berikut.

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i-}\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i-}\bar{x})^2}^3}$$

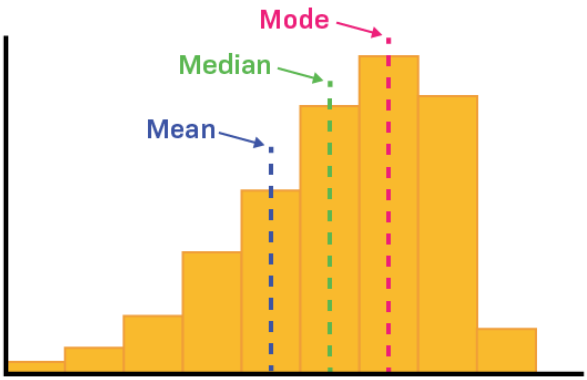
Ada dua jenis skewness, yaitu:

1. Kemiringan Kanan atau Positif (*positive skweness*)
   Kemiringan kanan atau positif berarti outlier (pencilan) berada di sebelah kanan (ekor panjang ke kanan) seperti ditunjukkan oleh Gambar 8.8.



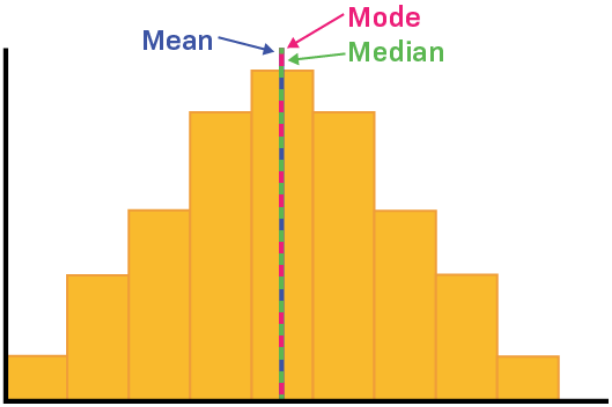

*Gambar 8.8 Ilustrasi data yang memiliki grafik distribusi dengan skweness positif*

2. Kemiringan Kiri atau Negatif
   Kemiringan ke kiri atau negatif berarti outlier (pencilan) berada di sebelah kiri (ekor panjang ke kiri) seperti ditunjukkan oleh Gambar 8.9.




*Gambar 8.9 Ilustrasi data yang memiliki grafik distribusi dengan skweness negatif*

Jika nilai rata-rata (mean) = median = modus maka grafik distribusi data tersebut bersifat simetris, seperti diilustrasikan oleh Gambar 8.10.




*Gambar 8.10 Ilustrasi data yang memiliki grafik distribusi dengan skweness positif*

## Statistik Inferensial

Statistik inferensial adalah kelompok statistik yang digunakan untuk menganalisis data berupa interval atau rasio.

Analisis statistik ini bertujuan untuk melakukan uji hipotesis dan penarikan kesimpulan umum yang dapat berlaku bagi seluruh data populasi. Data-data yang dianalisis menggunakan perangkat statistik inferensial haruslah memenuhi syarat distribusi normal.

Pada umumnya terdapat beberapa jenis distribusi peluang sebaran data, yaitu distribusi normal, distribusi binomial, distribusi T, distribusi homogen (*uniform*), serta distribusi Poisson.

### Distribusi Normal

Distribusi ini menunjukkan distribusi yang diikuti oleh sebagian besar peristiwa alam. Dilambangkan dengan Y ~ (μ, σ2). Ciri-ciri utama distribusi normal adalah:

1. Grafik yang diperoleh dari distribusi normal berbentuk kurva lonceng, simetris dan mempunyai ekor melengking.
2. 68% dari semua nilainya harus berada dalam interval, yaitu (μ – σ, μ + σ)
3. E(*Y*) = μ
4. Var(*Y*) = σ2

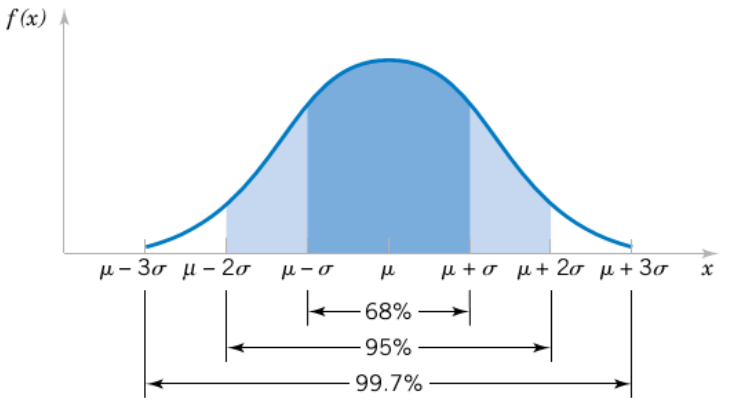
Contoh dan kegunaan distribusi normal adalah:

1. Distribusi normal sebagian besar terlihat pada berbagai fenomena alam maupun sosial.
2. Kita dapat mengubah distribusi normal apa pun menjadi distribusi normal standar. Distribusi normal dapat distandarisasi dengan menggunakan tabel Z melalui persamaan berikut.

$$z = \frac{y - \mu}{\sigma}$$

Dimana, σ menyatakan standar deviasi populasi dan μ

adalah rata-rata populasi. Berikut ini disajikan kurva distribusi normal.




*Gambar 8.11 Ilustrasi grafik data yang berdistribusi normal*

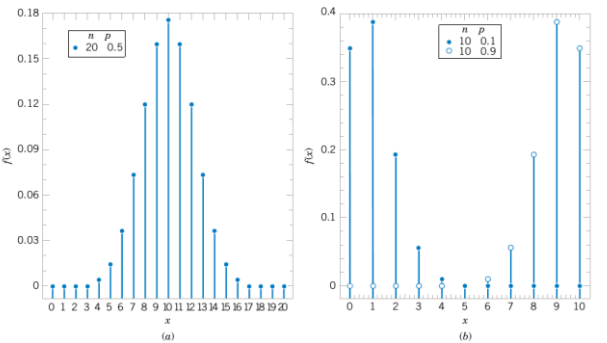
## Distribusi Binomial

Urutan kejadian Bernoulli yang identik disebut Binomial dan mengikuti distribusi Binomial. Dilambangkan dengan Y ~B(n, p). Karakteristik distribusi Binomial yaitu:

1. Selama n percobaan, distribusi mengukur frekuensi kemunculan salah satu hasil yang mungkin.
2. E(Y) = n × p
3. f(y) = Comb.(y, n) × py× (1 – p)n-y
4. Var(Y) = n × p × (1 – p)

Contoh dan kegunaan distribusi Binomial adalah:

1. Menghitung kemungkinan memperoleh gambar jika kita melempar koin sebanyak 10 kali.
2. Distribusi ini sebagian besar digunakan ketika kita mencoba memprediksi seberapa besar kemungkinan suatu peristiwa terjadi melalui serangkaian percobaan.



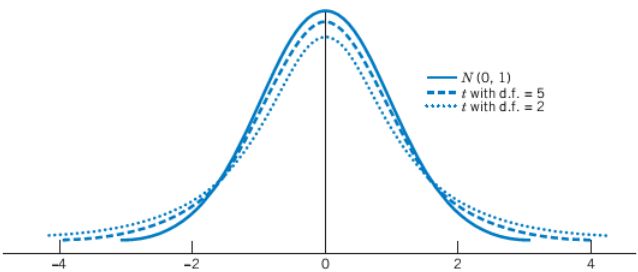

*Gambar 8.12 Ilustrasi grafik data yang berdistribusi Binomial f(y) untuk nilai n dan p yang berbeda*

## Distribusi T (Student-T Distribution)

Distribusi T digunakan untuk memperkirakan batasan populasi ketika ukuran sampel kecil dan varians populasi tidak diketahui. Distribusi ini dilambangkan dengan Y~ t(k). Karakteristik distribusi T adalah:

1. Estimasi ukuran sampel kecil dari distribusi normal
2. Grafiknya simetris dan melengkung berbentuk lonceng, namun memiliki ekor yang besar.
3. Jika $k > 1$ maka E(Y) = μ dan Var(Y) = $s^2 \times \frac{k}{k-2}$

Distribusi T digunakan dalam pemeriksaan data sampel kecil yang biasanya mengikuti distribusi normal.




*Gambar 8.13 Perbandingan grafik distribusi normal N (0,1) dan distribusi T untuk derajat kebebasan 2 dan 5*

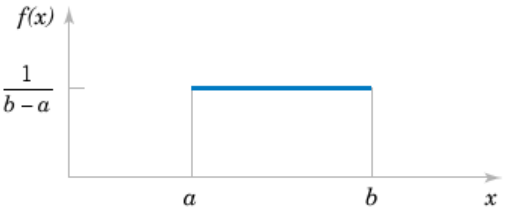
## Distribusi Seragam (Homogen)

Dalam distribusi yang seragam, semua hasil memiliki kemungkinan yang sama. Fungsi distribusi ini dilambangkan dengan Y~U(a,b) atau f(x). Jika nilainya bersifat kategoris, maka kita cukup menunjukkan jumlah kategori, seperti Y~U(a). Ciri-ciri distribusi seragam adalah:

1. Dalam distribusi yang seragam, semua hasil memiliki kemungkinan yang sama.
2. Dalam grafik, semua batang memiliki tinggi yang sama.
3. Nilai dan varians yang diharapkan tidak memiliki kekuatan prediksi.

   Contoh dan kegunaan distribusi seragam adalah

1. Hasil yang diperoleh setelah pelemparan sebuah dadu
2. Karena kesetaraannya, distribusi ini banyak digunakan dalam algoritma pengacakan.




*Gambar 8.14 Grafik distribusi homogen untuk rentang nilai data a dan b*

## Distribusi Poisson

Distribusi Poisson digunakan untuk menentukan seberapa besar kemungkinan suatu peristiwa tertentu terjadi dalam selang waktu atau jarak tertentu. Fungsi distribusi ini dilambangkan dengan Y ~ Po(λ). Ciri-ciri distribusi Poisson adalah

- Ini mengukur frekuensi selama interval waktu atau jarak.

- $E(Y) = \lambda$
- $P(Y = y) = \frac{\lambda^y}{\lambda! e^{-\lambda}}$
- $Var(Y) = \lambda$

Contoh dan kegunaan distribusi Poisson adalah

1. Distribusi ini digunakan untuk menentukan seberapa besar kemungkinan suatu peristiwa tertentu terjadi dalam interval waktu atau jarak tertentu.
2. Banyak digunakan dalam analisis pemasaran untuk mengetahui apakah kunjungan lebih dari rata-rata merupakan hal yang luar biasa atau sebaliknya.

**Teorema Limit Pusat**

Teorema ini menyatakan bahwa distribusi rata-rata sampel mendekati distribusi normal seiring dengan semakin besarnya ukuran sampel (dengan asumsi bahwa semua sampel memiliki ukuran yang sama), terlepas dari bentuk distribusi populasi. Jika ukuran sampel ≥ 30 dianggap cukup untuk memenuhi Teorema Limit Pusat. Aspek utama dari teorema ini adalah bahwa rata-rata mean sampel dan deviasi standar akan sama dengan mean populasi dan deviasi standar.

Selain itu, ukuran sampel yang cukup besar dapat meramalkan karakteristik suatu populasi secara akurat. Teorema Limit Pusat berlaku dengan kriteria:

1. Tidak peduli distribusinya
2. Semakin banyak sampel, semakin mendekati Normal ($k \rightarrow \infty$)
3. Semakin besar sampelnya, semakin mendekati Normal ($n \rightarrow \infty$)

## Estimator dan Estimasi

Estimator adalah adalah fungsi matematika dari sampel yang memberi tahu kita cara menghitung estimasi parameter dari sampel. Semakin kecil variansnya maka estimatornya semakin efisien. Contoh penduga dan parameter ekuivalen diberikan pada tabel di bawah ini.

*Tabel 8.6 Penduga dan parameter ekuivalen*

| Term | Estimator | Parameter |
|---|---|---|
| Mean | $\bar{x}$ | $\mu$ |
| Variance | $s^2$ | $\sigma^2$ |
| Correlation | R | $\rho$ |

Estimasi adalah nilai output yang bisa kita peroleh dari estimator. Ada jenis perkiraan berikut.

1. Perkiraan Poin – nilai tunggal, mis. 1, 6, 12,34, 0,123
2. Perkiraan Interval Keyakinan – interval, mis. (1,4), (43, 45), (3.22, 5.33), (-0.24, 0.26).

Dalam statistik kita kebanyakan menggunakan perkiraan interval kepercayaan saat membuat kesimpulan karena lebih tepat dibandingkan dengan perkiraan titik.

## Interval Keyakinan/*Confidency Interval*

Ini adalah interval dimana kita yakin dengan terhadap kesimpulan yang kita ambil pada tingkat kepercayaan tertentu.

## Margin Kesalahan/*Margin of Error*

Margin kesalahan menjelaskan berapa poin persentase hasil Anda akan berbeda dari nilai populasi sebenarnya. Itu dapat dihitung dengan dua cara berikut.

1. *Margin of error* = Nilai kritis x Standar deviasi

2. *Margin of error* = Nilai kritis x Kesalahan standar statistik

## Hipotesis

Hipotesis adalah asumsi yang didasarkan pada bukti yang tidak memadai yang memerlukan pengujian dan eksperimen lebih lanjut. Setelah pengujian lebih lanjut, suatu hipotesis secara umum dapat dipastikan benar atau salah. Terdapat dua jenis hipotesis, yaitu hipotesis nol dan hipotesis alternatif. Hipotesis nol adalah hipotesis yang perlu diuji. Ini adalah hipotesis yang coba dibuktikan oleh peneliti sebagai hipotesis yang salah. Hipotesis ini bersifat status quo. Sementara itu, hipotesis alternatif adalah kebalikan dari hipotesis nol yang biasanya didasarkan pada pendapat kita sendiri.

Contoh hipotesis nol dan hipotesis alternatif dalam sebuah penelitian pendidikan tentang eksperimen model pembelajaran tertentu dalam perkuliahan Fisika terhadap kemampuan pemecahan masalah fisika para mahasiswa.

Ho: Tidak terhadap perbedaan kemampuan pemecahan masalah antara mahasiswa yang belajar menggunakan model Problem Solving dan model pembelajaran konvensional.

H1: Terdapat perbedaan kemampuan pemecahan masalah antara mahasiswa yang belajar menggunakan model Problem Solving dan model pembelajaran konvensional.

## Pengujian Hipotesis

Pengujian hipotesis adalah sebuah proses pengujian terhadap hipotesis yang telah dirumuskan. Dalam pengujian hipotesis digunakan perangkat statistik sebagai alat pengujiannya seperti uji normalitas menggunakan fungsi

distribusi Z sebagai perangkat statistiknya. Uji korelasi menggunakan fungsi distribusi Chi-Square sebagai perangkat statistiknya. Uji perbandingan dapat menggunakan perangkat statistik berupa fungsi distribusi T maupun fungsi distribusi F.

### Pengambilan Keputusan Hasil Uji Statistik

Setelah dilakukan pengujian, akan ada dua kemungkinan keputusan yaitu menerima hipotesis nol atau menolak hipotesis nol. Menerima hipotesis nol berarti tidak ada cukup data untuk mendukung perubahan atau kebaruan yang dibawa oleh hal-hal yang tidak konvensional. Menolak hipotesis nol berarti terdapat cukup bukti statistik yang menunjukkan hipotesis nol tersebut salah.

### Tingkat Signifikansi (*Significance Level*)

Tingkat signifikansi adalah probabilitas menolak hipotesis nol melalui pengujian hipotesis. Tingkat signifikansi dilambangkan dengan α (Alfa).

### Tingkat kepercayaan diri (*Confidency Level*)

Tingkat kepercayaan adalah kemungkinan suatu parameter berada dalam rentang nilai tertentu. Tingkat kepercayaan dilambangkan dengan *C*. Apabila tingkat signifikansi dihubungkan dengan tingkat kepercayaan maka keduanya dapat dinyatakan dalam sebuah hubungan matematis yaitu: $C = 1 - \alpha$. Tingkat signifikansi umum dan tingkat kepercayaan yang terkait diberikan dalam tabel berikut.

| Tingkat signifikansi ($\alpha$) | Tingkat kepercayaan diri ($C$) |
|---|---|
| 0,10 | 90% |
| 0,05 | 95% |
| 0,01 | 99% |

### Nilai-*p*

Nilai *p* adalah tingkat signifikansi marjinal terkecil di mana hipotesis nol akan ditolak. Nilai *p* yang lebih kecil berarti terdapat bukti yang lebih kuat yang mendukung hipotesis alternatif. Biasanya nilai p ditemukan dengan 3 digit setelah titik (x.xxx).

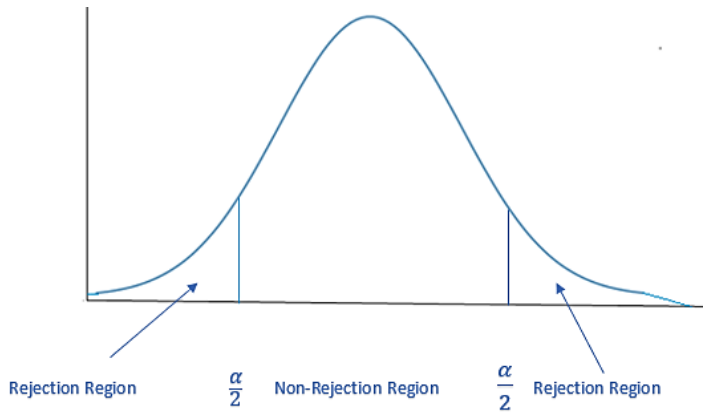
Angka dari nilai p terletak antara 0 – 1, yaitu:

1. Nilai p yang kecil (biasanya ≤ 0,05) memberikan bukti kuat terhadap hipotesis nol, jadi, kami menolak hipotesis nol.
2. Nilai p yang besar (> 0,05) menunjukkan bukti yang lemah terhadap hipotesis nol, sehingga kita gagal menolak hipotesis nol. 0,05 sering kali merupakan "batas batas".

Aturan penolakan hipotesis nol berlaku, yaitu:

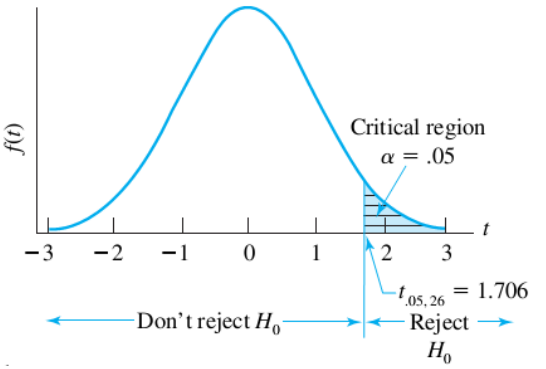
1. Jika nilai-p ≤ tingkat signifikansi, maka tolak hipotesis nol.
2. Jika nilai-p > tingkat signifikansi, maka hipotesis nol tidak ditolak.

Dalam hal penolakan (*rejection*) dan penerimaan (*acceptance*) hipotesis nol, maka terdapat dua wilayah yang masing-masing menyatakan wilayah penolakan hipotesis nol dan wilayah penerimaan berlakunya hipotesis nol, seperti diilustrasikan oleh Gambar 8.15.




*Gambar 8.15 Ilustrasi daerah penerimaan dan penolakan Ho*

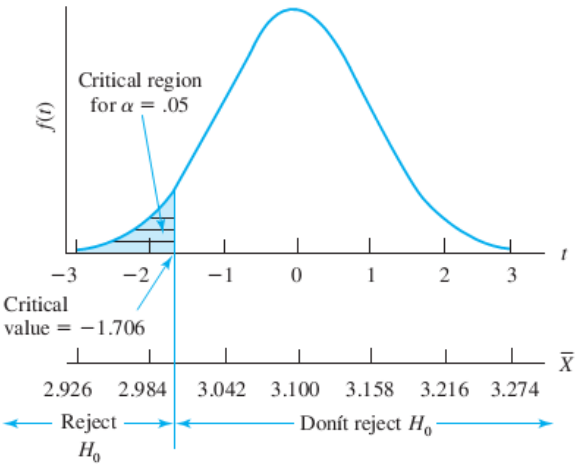
Dalam pengujian hipotesis dikenal adanya metode pengujian satu sisi (*one–tailed testing*) dan pengujian dua sisi (*double-tailed testing*). Uji satu sisi digunakan bila nolnya tidak mengandung tanda persamaan atau pertidaksamaan (<, >, ≤, ≥). Daerah penolakan untuk pengujian satu sisi ditunjukkan pada Gambar 8.16.



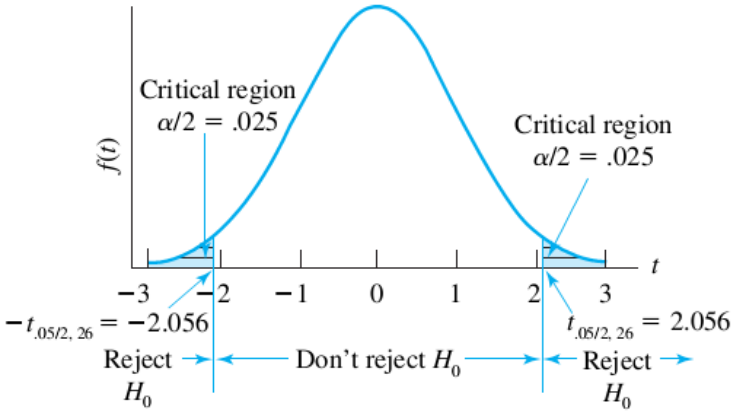

*Gambar 8.16 Ilustrasi daerah pengujian satu sisi kanan dengan tingkat signifikansi 0,05.*

Pada pengujian arah kanan, daerah penolakan diarsir di

sisi kanan (seperti terlihat pada Gambar 8.16). Sementara itu, pada pengujian sisi kiri, daerah penolakan diarsir pada sisi kiri (Gambar 8.17). Khusus pengujian dua sisi (*two-tailed testing*) digunakan bila nol mengandung tanda persamaan (=) atau pertidaksamaan (≠). Daerah penolakan hipotesis Ho untuk pengujian dua sisi ditunjukkan pada Gambar 8.18.




*Gambar 8.17 Ilustrasi daerah pengujian satu sisi kiri dengan tingkat signifikansi 0,05*




*Gambar 8.18 Ilustrasi daerah pengujian dua sisi dengan tingkat signifikansi 0,05*

## Kesalahan Statistik

Ada dua jenis kesalahan statistik dalam menaksir parameter populasi, yaitu:

1. Kesalahan tipe I (Positif Palsu), yaitu kesalahan yang terjadi ketika kita menolak hipotesis nol yang sebenarnya benar. Peluang terjadinya kesalahan tipe I dilambangkan dengan α (*alpha*).
2. Kesalahan tipe II (Negatif Palsu) terjadi ketika kita menerima hipotesis nol yang sebenarnya salah. Probabilitas melakukan kesalahan tipe-II dilambangkan dengan β (Beta).

Berdasarkan kriteria tersebut, maka hubungan antara keputusan penerimaan dan penolakan hipotesis nol terhadap jenis kesalahan itu sendiri dapat dinyatakan dengan matrik berikut.

| **Keputusan** | | **Keadaan sebenarnya** | |
|---|---|---|---|
| | | Positif (Ho benar) | Negatif (Ho salah) |
| **Hasil yang diperoleh** | Positif (Ho diterima) | *True positive* (1- $\beta$) | *False positive* (Kesalahan tipe I) |
| | Negatif (Ho ditolak) | *False Negative* (Kesalahan tipe II) | *True Negative* |

# BAB 9 VISUALISASI DATA

## Pendahuluan

Naskah ini akan mengeksplorasi karakteristik dan fungsi yang paling relevan dan khas untuk beragam perpustakaan sebelum menggambarkan proses memvisualisasikan data menggunakannya. Selanjutnya, ini akan memeriksa berbagai kategori data yang dapat direpresentasikan dalam Python, bersama dengan metodologi visualisasi umum, sumber daya, dan jenis grafik.

Sebelum menggali lebih dalam ke perpustakaan itu sendiri, akan menguntungkan untuk mengembangkan pemahaman tentang bagaimana lingkungan perpustakaan visualisasi Python terstruktur. Intinya, memahami desain dan hubungan antara pustaka Python yang berbeda akan membantu dalam memilih pustaka yang paling cocok untuk proyek visualisasi. Berbagai pustaka visualisasi data dan modul kompatibel dengan Python. Pustaka ini umumnya dapat dikategorikan menjadi empat kelompok berdasarkan asal dan fokusnya, yaitu pustaka berbasis MatplotLib, pustaka JavaScript, perpustakaan JSON, dan pustaka WebGL.

Kelompok perpustakaan terkemuka awal didirikan di Matplotlib. Matplotlib berdiri sebagai salah satu perpustakaan visualisasi data Python paling awal, dan karena fitur-fiturnya yang luas dan antarmuka yang ramah pengguna, ia tetap digunakan secara luas. Pertama kali diperkenalkan pada tahun 2003, Matplotlib telah mengalami peningkatan berkelanjutan sejak awal (Lee, 2019).

Matplotlib mencakup sejumlah besar alat visualisasi, jenis plot, dan format output, terutama menghasilkan visual statis. Sementara beberapa opsi visualisasi 3D terbatas tersedia, mereka kurang luas dibandingkan dengan pustaka lain seperti Plotly dan VisPy. Plot interaktif juga terbatas, tidak seperti Bokeh, yang akan diuraikan di bagian selanjutnya. Mengingat keberhasilan Matplotlib sebagai alat visualisasi, banyak perpustakaan lain telah memperluas fitur intinya dari waktu ke waktu. Perpustakaan ini, dibangun di atas dasar Matplotlib, memanfaatkannya sebagai mesin untuk fungsi visualisasi unik mereka.

Pustaka yang bergantung pada Matplotlib memperkenalkan fitur tambahan dengan mengkhususkan diri dalam merender tipe data atau domain tertentu, memperkenalkan tipe plot baru, atau mengembangkan API tingkat tinggi baru untuk fungsi Matplotlib. Mereka digunakan bersama dengan Matplotlib untuk meningkatkan kemampuan styling dan plotting (Oberoi & Chauhan, 2019). Pustaka berbasis JavaScript yang dirancang untuk Python berfokus pada visualisasi data dan memanfaatkan interaktivitas yang disediakan oleh browser web melalui HTML5. Hal ini memungkinkan pembuatan grafik dan visualisasi interaktif, bergerak melampaui plot 2D statis. Membuat visualisasi yang menarik secara visual difasilitasi dengan menata halaman HTML dengan CSS (Hunt-Isaak et al., 2024).

Pustaka ini merangkum fungsionalitas dan alat JavaScript/HTML5 dalam Python, memungkinkan pengguna untuk membuat plot interaktif. Mereka menawarkan API tingkat tinggi untuk fungsi JavaScript, memungkinkan penyesuaian primitif JavaScript untuk merancang jenis plot baru langsung di dalam Python.

Pustaka berbasis JSON dirancang untuk menafsirkan dan menampilkan data JSON, format pertukaran data terstruktur yang dapat dipahami tidak hanya untuk pustaka JavaScript tetapi juga untuk hampir semua bahasa pemrograman. Data JSON sepenuhnya terkandung dalam file JSON, sehingga memungkinkan untuk mengintegrasikan plot dengan beragam alat dan teknik visualisasi.

Stkitar WebGL adalah alat grafis yang memungkinkan interaktivitas untuk visualisasi 3D. Mirip dengan dampak HTML5 pada plot 2D, munculnya StKitar WebGL mengarah pada pengembangan pustaka plot 3D interaktif (Oberoi & Chauhan, 2019). Python menawarkan berbagai pustaka yang berfokus pada pengembangan plot WebGL. Perpustakaan ini memfasilitasi integrasi dan berbagi yang mudah melalui notebook Jupyter, serta manipulasi jarak jauh melalui web. Banyak pustaka plot Python ada, beberapa di antaranya menyediakan pembungkus Python untuk berbagai bahasa dan platform visualisasi.

## Visualisasi Data

Publikasi ini akan mengeksplorasi pustaka visualisasi data populer di Python, dikategorikan menjadi lima kelompok. Perpustakaan yang ditampilkan termasuk Matplotlib, PKITAS, Seaborn, Bokeh, Plotly, Altair, GGPlot, Geopkitas, dan Vispy. Memahami kemampuan perpustakaan ini sangat penting untuk memilih yang paling cocok untuk persyaratan proyek tertentu. Mari selidiki perpustakaan ini (Lee, 2019)

## Matplotlib

Matplotlib menonjol sebagai perpustakaan visualisasi

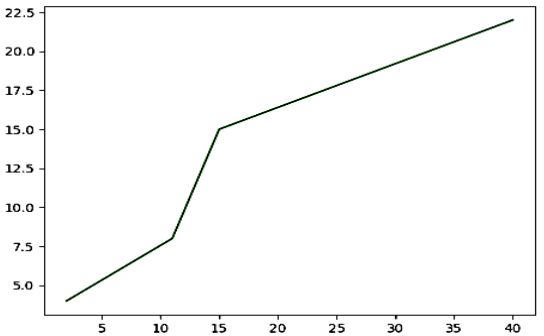
yang banyak digunakan untuk membuat plot 2D statis, dengan dukungan terbatas untuk visualisasi 3D. Strukturnya memungkinkan pengguna untuk menghasilkan dan menyesuaikan beberapa plot dalam satu gambar melalui subplotting. Tujuannya adalah untuk menyederhanakan pembuatan plot dasar dan lanjutan, menawarkan dukungan untuk mode visualisasi statis dan interaktif. Namun, fitur interaktivitasnya agak terbatas (Nagesh et al., 2015).

Pembuatan plot di Matplotlib melibatkan pemanfaatan antarmuka PyPlot, menampilkan perintah yang mengingatkan pada MATLAB. Pengguna dapat menghasilkan visualisasi menggunakan gaya preset atau menyesuaikannya sesuai dengan preferensi mereka. Matplotlib memungkinkan pembuatan gambar dan sumbu untuk mewakili data. Awalnya, kita akan mengeksplorasi metode pembuatan plot sederhana sebelum mempelajari teknik penyesuaian lanjutan.

Melalui PyPlot, pengguna dapat dengan cepat menghasilkan plot profesional dan standar dengan kode minimal. Untuk memulai, modul matplotlib dan pyplot diimpor. Selanjutnya, memanggil berbagai fungsi plot dan menyediakan data yang relevan merampingkan proses pembuatan plot.

Plot dasar dengan nilai acak dapat dihasilkan, di mana set nilai pertama sesuai dengan sumbu X dan set kedua ke sumbu Y. Ini adalah praktik umum untuk hanya memberikan nilai sumbu X, membiarkan Matplotlib menentukan nilai sumbu Y default. Selain itu, warna dapat ditetapkan ke garis, seperti yang ditunjukkan dalam implementasi berikut menggunakan Matplotlib.

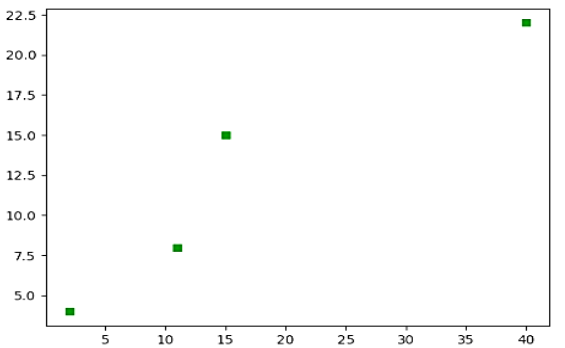
```
1 import matplotlib.pyplot as plt
2 plt.plot([2, 11, 15, 40], [4, 8, 15, 22], color='g')
3 plt.show()
```




Fungsi plot bertanggung jawab untuk membangun plot dengan memanfaatkan elemen-elemennya. Setelah eksekusi kode, fungsi show menampilkan plot. Pyplot, menyerupai gaya plot MATLAB, mengatur plot melalui serangkaian perintah gaya, seperti warna. Pilihan termasuk mengubah simbol seperti lingkaran, kotak, atau segitiga alih-alih garis padat default. Instruksi untuk warna dan simbol dapat diberikan sebagai argumen ketiga dalam panggilan pembuatan plot. Contoh simbol plot termasuk -- untuk garis putus-putus, s untuk kotak, atau ^ untuk segitiga, sedangkan warna seperti r untuk merah, b untuk biru, dan g untuk hijau digunakan. Ilustrasi membuat plot dengan kotak hijau disediakan.

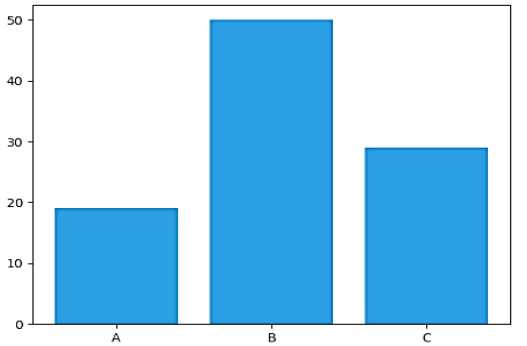
*plt.plot([2, 11, 15, 40], [4, 8, 15, 22], 'gs')*

*2 plt.show()*



Plot yang dibuat sebelumnya melibatkan variabel kontinu; sekarang, fokus bergeser untuk membuat plot dengan variabel kategoris. Variabel kategoris diplot dengan mencantumkan kategori dan nilai yang berbeda, kemudian meneruskannya ke fungsi plot masing-masing. Diagram batang biasanya digunakan untuk nilai kategoris, menunjukkan contoh praktis untuk membuat dan memplot diagram batang:

```
1 names = ['A', 'B', 'C']
2 values = [19, 50, 29]
3 plt.bar(names, values)
4 plt.show()
```

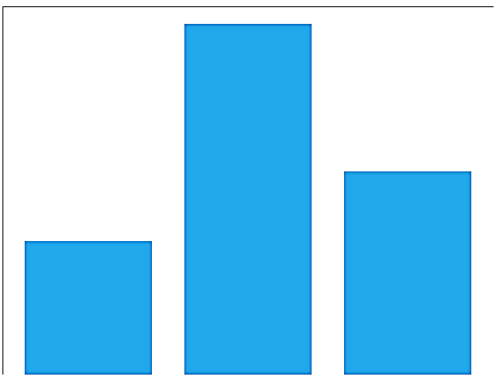



Di Matplotlib, ada metode alternatif untuk membuat plot, menawarkan lebih banyak kontrol atas proses pembuatan plot. Dengan membuat objek Gambar, seseorang dapat menentukan properti plot. Tanpa ini, Matplotlib menghasilkan objek default dengan pengaturan default. Menggunakan fungsi figure () dalam modul pyplot, gambar dibuat dan properti ditetapkan, seperti dimensi gambar yang ditentukan oleh daftar empat nilai antara 0 dan 1 (kiri, bawah, lebar, tinggi). Selain itu, fungsi add_subplot () digunakan untuk penyesuaian lebih lanjut. Sebuah gambar dibuat, disertai dengan informasi mengenai sumbu.

Objek sumbu dalam gambar yang dibuat memberikan

kontrol yang ditingkatkan atas visualisasi data dan elemen plot lainnya. Setiap objek sumbu dapat menyimpan plotnya sendiri di dalam gambar, memberikan kontrol atas tampilan subplot. Contoh Multiple Axes dapat ada dalam satu gambar, berisi berbagai elemen seperti tanda centang, garis, teks, dan poligon. Kustomisasi elemen-elemen ini akan dieksplorasi di bagian Customizing the Plot. Untuk memulai, objek sumbu dibuat dalam gambar:

```
1 fig = plt.figure()
2 ax = fig.add_axes([0, 0, 1, 1])
3 names = ['A', 'B', 'C']
4 values = [19, 50, 29]
5 ax.bar(names, values)
6 plt.show()
```



Fungsi fig add_axis () menghasilkan objek Axes baru yang dienkapsulasi dalam ax. Objek ini memfasilitasi penambahan elemen. Misalnya, ax.bar () dipanggil untuk menggambarkan grafik batang alih-alih plt.bar () sebelumnya.

Hubungan antara kapak dan ara menyiratkan bahwa setiap penambahan yang dibuat pada kapak juga akan mencerminkan gambar.

Fungsi add_axis () mengharuskan penyediaan argumen

[0, 0, 1, 1], menunjuk parameter kiri, bawah, lebar, dan tinggi dari objek axe.

Nilai numerik ini mewakili pecahan gambar yang dicakup oleh objek Axes, sehingga memposisikannya di sudut kiri bawah (0 untuk kiri dan 0 untuk bawah) dengan dimensi yang identik dengan gambar induk (1 untuk lebar dan 1 untuk tinggi).

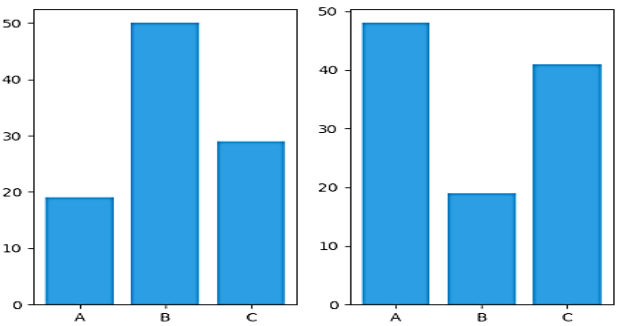
Saat ini, kapak mungkin tidak terlihat, dan plot mungkin kekurangan elemen tertentu dibandingkan dengan pengaturan default. Penghapusan sumbu dapat dicapai dengan menggunakan fungsi delaxis (), seperti yang ditunjukkan di bawah ini.

Setelah membiasakan diri dengan prosedur standar untuk menghasilkan plot di Matplotlib, kami akan mengeksplorasi beragam opsi penyesuaian yang tersedia untuk meningkatkan plot ini.

Matplotlib memungkinkan pembuatan beberapa plot dalam satu gambar. Untuk menggabungkan beberapa plot, seseorang harus membuat "subplot" untuk setiap plot yang diinginkan dalam gambar. Ini dicapai melalui fungsi add_subplot (), yang membutuhkan urutan input numerik. Digit awal menentukan jumlah baris yang diinginkan, digit kedua menunjukkan jumlah kolom yang diinginkan, dan digit ketiga menunjukkan jumlah plot yang akan dimasukkan.

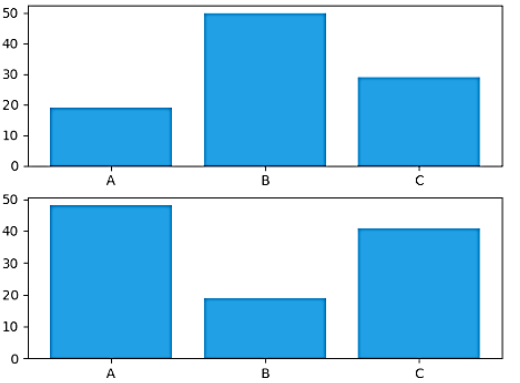
Misalnya, memasukkan angka 111 ke dalam fungsi add_subplot () menghasilkan subplot baru ditambahkan ke gambar. Sebaliknya, memilih angka 221 menghasilkan plot dengan empat sumbu yang diatur dalam dua baris dan dua kolom, dengan subplot yang diinginkan diposisikan terlebih dahulu. Contoh selanjutnya menggambarkan pembuatan dua subplot dalam gambar yang sama, di mana dua objek sumbu telah dihasilkan:

```
1 fig = plt.figure()
2names = ['A', 'B', 'C']
3 values = [19, 50, 29]
4 values_2 = [48, 19, 41]
5 ax = fig.add_subplot(121)
6 ax2 = fig.add_subplot(122)
7 ax.bar(names, values)
8 ax2.bar(names, values_2)
9 plt.show()
```



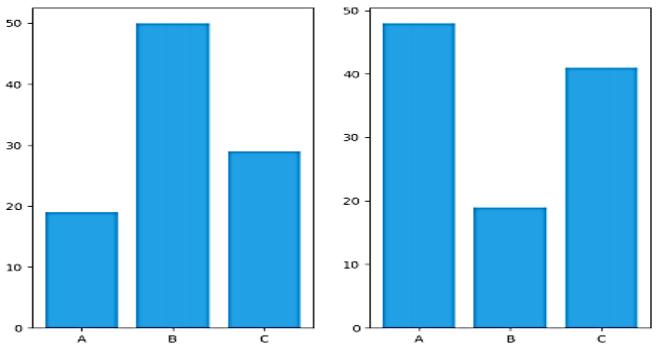

Dua subplot telah dibuat dalam gambar dengan 1 baris dan 2 kolom, diposisikan berdekatan. Konfigurasi gambar serupa akan dicapai dengan menghasilkan gambar dengan 2 baris dan 1 kolom:

```
1 fig = plt.figure()
2 names = ['A', 'B', 'C']
3 values = [19, 50, 29]
4 values_2 = [48, 19, 41]
5 ax = fig.add_subplot(211)
6 ax2 = fig.add_subplot(212)
7 ax.bar(names, values)
8 ax2.bar(names, values_2)
9 plt.show()
```




Memodifikasi Ukuran Gambar, ketika jumlah subplot dan seluk-beluk meningkat, gambar mungkin menjadi terlalu padat dan sulit untuk ditafsirkan. Menyesuaikan ukuran representasi visual kami dapat meningkatkan kejelasan data yang ditampilkan. Dengan memasukkan parameter figsize dalam fungsi figure (), kita dapat mengubah ukuran visualisasi kita sesuai dengan itu. Parameter ini juga dapat digunakan bersama dengan fungsi subplot () untuk menyesuaikan ukuran setiap subplot. Ilustrasi konsep ini ditunjukkan di bawah ini dengan pembuatan plot 8x6:

```
1 names = ['A', 'B', 'C']
2 values = [19, 50, 29]
3 values_2 = [48, 19, 41]
4 fig = plt.figure(figsize=(8.0,6.0))
5 # Adds subplot on position 1
6 ax = fig.add_subplot(121)
7 # Adds subplot on position 2
8 ax2 = fig.add_subplot(122)
9 ax.bar(names, values)
10 ax2.bar(names, values_2)
11 plt.show()
```




Perhatikan bahwa dimensi gambar ditentukan dalam inci, menunjukkan bahwa plot yang dihasilkan memiliki lebar 8 inci dan tinggi 6 inci. Sementara sistem metrik tidak langsung berlaku dalam skenario ini, Kitadapat memperkenalkan fungsi konversi dari sentimeter ke inci:

*1 def cm_to_inch(value):*

*2 return value/2.54*

Dan kemudian sesuaikan ukuran plot seperti ini:

*1 fig = plt.figure(figsize=(cm_to_inch(10),cm_to_inch(15)))*

Customizing A Plot, kami telah menjelaskan proses membangun plot dan mengintegrasikan objek Axes ke dalam Gambar, memfasilitasi penyesuaian lebih lanjut. Selanjutnya, kita dapat memanfaatkan elemen-elemen ini untuk menyempurnakan aspek visual dari plot yang sedang berlangsung. Ini termasuk opsi penyesuaian seperti PenKita, TKitaCheck, lebar garis, gaya garis, legenda, teks, dan anotasi.

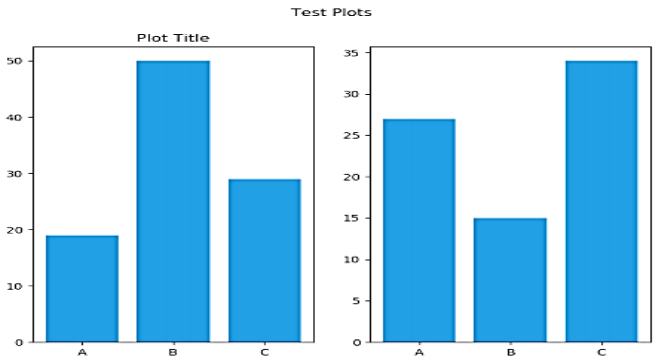
Judul Plot dapat ditunjuk menggunakan fungsi set () dengan parameter title atau menggunakan fungsi set_title (). Untuk objek gambar, fungsi subtitle () disarankan untuk mengatur judul plot secara efektif:

*1 names = ['A', 'B', 'C']*

*2 values = [19, 50, 29]*

*3 values_2 = [27, 15, 34]*

*4 fig = plt.figure(figsize=(8.0,6.0))*
*5 ax = fig.add_subplot(121)*
*6 ax2 = fig.add_subplot(122)*
*7 # Sets the title of the sublot on position 1*
*8 ax.set_title('Plot Title')*
*9 ax.bar(names, values)*
*10 ax2.bar(names, values_2)*
*11 # Sets the title of the entire figure*
*12 plt.suptitle('Test Plots')*
*13 plt.show()*

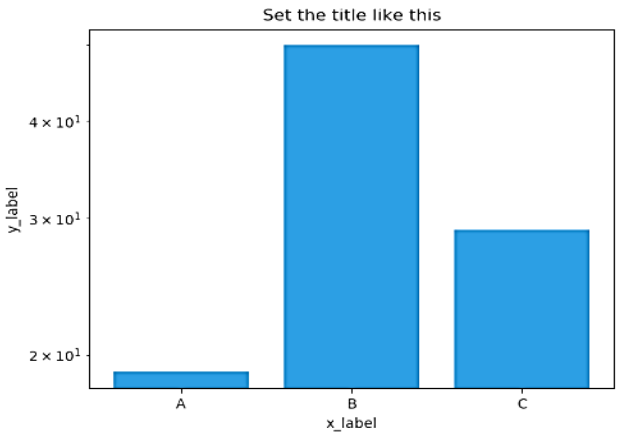



Label dan Legend

Mirip dengan pendekatan Kitadapat dalam memberi judul plot Kita, Kitajuga memungkinkan pelabelan setiap sumbu pada plot Kita. Secara default, pelabelan sumbu dapat dicapai melalui fungsi xlabel () dan ylabel (). Atau, Kitajuga memungkinkan penggunaan fungsi set () pada objek sumbu Kita atau menyetelnya secara individual dengan ax set_xlabel () dan ax set_ylabel ().

*1 names = ['A', 'B', 'C']*
*2 values = [19, 50, 29]*
*3 values_2 = [27, 15, 34]*
*4 plt.xlabel("Label for X")*

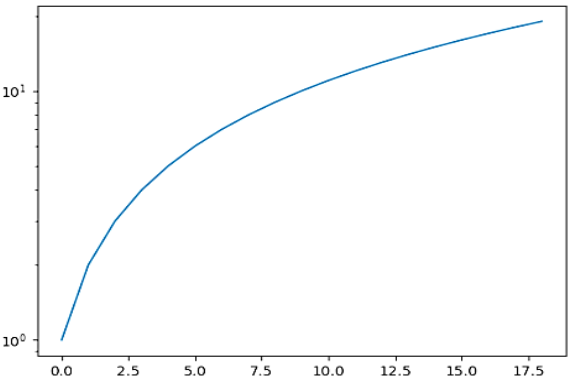
*5 plt.ylabel("Label for Y")*
*6 plt.bar(names, values)*
*7 plt.suptitle('Test Plots')*
*8 plt.show()*
Inilah cara kami mengaturnya menggunakan set_xlabel() dan set_ylabel():
*1 fig = plt.figure()*
*2 ax = fig.add_subplot(111)*
*3 ax.set_title('Set the title like this')*
*4 ax.set_xlabel('x_label')*
*5 ax.set_ylabel('y_label')*
*6 names = ['A', 'B', 'C']*
*7 values = [19, 50, 29]*
*8 values_2 = [27, 15, 34]*
*9 plt.bar(names, values)*
*10 plt.show()*




Saat memplot data pada skala nonlinier, penyesuaian penskalaan sumbu dapat dilakukan menggunakan fungsi xscale () dan yscale () dengan jenis skala tertentu seperti 'log'. Matplotlib menawarkan dukungan untuk berbagai skala

termasuk skala linier, log, log simetris, dan logit:

```
1 values = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
19]
2 plt.plot(values)
3 plt.yscale('log')
4 plt.show()
```




Mengamati perbedaan saat ini dalam skala sumbu x dan y, terlepas dari nilai liniernya, penyesuaian diperlukan. Posisi legenda dapat dikontrol dengan menggunakan parameter loc atau lokasi pada objek sumbu/gambar dan menentukan lokasi yang diinginkan. Dalam hal ini, kami mengarahkan fungsi untuk memposisikan legenda di sudut “kanan atas”. Selain itu, fungsi dapat menerima daftar nilai untuk diwakili; karena grafik berisi tipe data tunggal, elemen tunggal sudah cukup.

# BAB 10 METODE PENGUMPULAN DATA

## Pendahuluan

Pengumpulan data merupakan langkah awal yang krusial dalam setiap proyek data science. Dalam era digital saat ini, data menjadi aset yang sangat berharga bagi organisasi dan peneliti karena dapat memberikan wawasan mendalam yang mendukung pengambilan keputusan berbasis bukti. Proses pengumpulan data tidak hanya tentang mengumpulkan informasi secara acak, tetapi juga tentang mengumpulkan data yang relevan, berkualitas, dan dapat diandalkan yang dapat digunakan untuk analisis lebih lanjut.

Dalam data science, teknik pengumpulan data yang tepat sangat menentukan kualitas hasil analisis dan model yang dikembangkan. Metode pengumpulan data yang buruk dapat mengakibatkan bias, inkonsistensi, dan ketidakakuratan yang pada akhirnya mempengaruhi validitas temuan dan prediksi. Oleh karena itu, pemilihan teknik yang sesuai harus didasarkan pada tujuan penelitian, jenis data yang dibutuhkan, dan sumber daya yang tersedia.

Teknik pengumpulan data dalam data science melibatkan berbagai metode, mulai dari survei dan wawancara tradisional hingga teknik digital canggih seperti web scraping dan penggunaan API. Setiap metode memiliki kelebihan dan kekurangan yang perlu dipertimbangkan. Misalnya, survei dapat memberikan data langsung dari responden tetapi bisa mahal dan memakan waktu, sementara web scraping dapat mengumpulkan data dalam jumlah besar dengan cepat namun menghadapi

tantangan legalitas dan etika.

Dalam bab ini, kita akan membahas berbagai teknik pengumpulan data yang sering digunakan dalam data science seperti survey, wawancara, observasi, pengumpulan data digital serta eksperimen. Pemahaman yang komprehensif mengenai instrumen dan prosedur pengumpulan data akan memberikan landasan yang kuat bagi penelitian dalam mengumpulkan data.

## Teknik Pengumpulan Data

Berikut merupakan teknik pengumpulan data mencakup berbagai metode yang digunakan untuk mengumpulkan informasi dari berbagai sumber, seperti survei, wawancara, observasi, dan penggunaan perangkat digital seperti web scraping dan API.

### Survei & Kuisioner

Survei merupakan salah satu teknik pengumpulan data yang paling umum dan efektif dalam data science, digunakan untuk mengumpulkan data langsung dari responden dalam bentuk kualitatif maupun kuantitatif. Melalui survei, data scientist dapat memperoleh informasi yang mendalam tentang opini, perilaku, dan karakteristik dari populasi yang diteliti. Keunggulan survei terletak pada fleksibilitasnya; survei dapat disesuaikan untuk berbagai macam kebutuhan penelitian, mulai dari survei sederhana dengan beberapa pertanyaan hingga survei kompleks dengan berbagai jenis pertanyaan dan skala. Dengan desain survei yang tepat, data yang dikumpulkan dapat memberikan wawasan yang signifikan untuk analisis lebih lanjut.

Namun, agar survei dapat memberikan hasil yang akurat

dan reliabel, perlu diperhatikan beberapa aspek penting, seperti pemilihan sampel yang representatif, penyusunan pertanyaan yang jelas dan tidak bias, serta metode distribusi survei yang efektif. Teknik pengumpulan data melalui survei juga harus memperhatikan etika penelitian, termasuk mendapatkan persetujuan dari partisipan dan menjaga kerahasiaan informasi mereka. Dengan demikian, survei yang dilakukan secara cermat dan metodis dapat menjadi alat yang sangat berharga dalam mengumpulkan data yang valid dan berguna dalam proyek-proyek data science (Fan & Yan, 2022).

Kuesioner adalah salah satu alat utama dalam teknik pengumpulan data yang sering digunakan dalam data science untuk memperoleh data terstruktur dari sejumlah besar responden. Kuesioner terdiri dari serangkaian pertanyaan yang dirancang secara sistematis untuk mengumpulkan informasi spesifik yang relevan dengan tujuan penelitian. Pertanyaan-pertanyaan dalam kuesioner dapat berbentuk pilihan ganda, skala Likert, pertanyaan terbuka, dan lainnya, yang memungkinkan peneliti untuk mengumpulkan data kuantitatif dan kualitatif.

Dalam konteks data science, kuesioner digunakan untuk mengumpulkan data dari populasi yang luas dengan cara yang efisien dan terstandardisasi. Kuesioner dapat disebarkan melalui berbagai media seperti kertas, email, atau platform online, yang memungkinkan jangkauan yang lebih luas dan biaya yang lebih rendah dibandingkan metode pengumpulan data lainnya seperti wawancara langsung. Dengan teknologi digital, kuesioner online dapat dirancang untuk otomatis mengumpulkan dan menyimpan data, serta memungkinkan analisis data yang lebih cepat dan akurat menggunakan perangkat lunak analitik.

Kelebihan utama dari kuesioner adalah kemampuannya

untuk mengumpulkan data dalam jumlah besar dengan cepat dan biaya yang relatif rendah. Kuesioner juga memberikan responden kebebasan untuk menjawab pertanyaan secara anonim, yang dapat meningkatkan kejujuran dan akurasi jawaban. Namun, kuesioner juga memiliki tantangan tersendiri, seperti risiko interpretasi pertanyaan yang berbeda oleh responden, tingkat respons yang rendah, dan potensi bias dalam pertanyaan atau pemilihan sampel (Dillman et al., 2022).

### Wawancara

Wawancara adalah salah satu teknik pengumpulan data yang digunakan dalam data science untuk mendapatkan wawasan mendalam dan detail dari responden. Dalam wawancara, data scientist atau peneliti berbicara langsung dengan responden, baik secara tatap muka maupun melalui telepon atau video call, untuk menggali informasi yang tidak dapat diperoleh melalui metode lain seperti survei. Wawancara dapat bersifat terstruktur dengan pertanyaan yang sudah ditentukan sebelumnya, semi-terstruktur dengan panduan umum tetapi fleksibel, atau tidak terstruktur yang lebih bebas dan mirip dengan percakapan biasa.

Kelebihan dari teknik wawancara adalah kemampuannya untuk menggali data kualitatif yang mendalam, menangkap nuansa dan konteks yang mungkin terlewatkan dalam metode pengumpulan data lainnya. Wawancara memungkinkan peneliti untuk mengajukan pertanyaan lanjutan berdasarkan jawaban responden, sehingga mendapatkan informasi yang lebih kaya dan rinci. Namun, wawancara juga memerlukan keterampilan khusus dari peneliti dalam hal komunikasi dan analisis, serta membutuhkan waktu dan biaya yang lebih besar dibandingkan dengan metode lain seperti survei tertulis. Selain itu, penting

untuk memastikan bahwa proses wawancara dilakukan dengan mematuhi prinsip-prinsip etika, termasuk menjaga kerahasiaan dan mendapatkan persetujuan dari responden (Smith & Noble, 2022).

## Observasi

Observasi adalah teknik pengumpulan data di mana peneliti mengamati dan mencatat perilaku, kejadian, atau kondisi tanpa interaksi langsung dengan subjek penelitian. Teknik ini digunakan dalam data science untuk memperoleh data yang akurat dan kontekstual yang mungkin tidak terdeteksi melalui metode lain seperti survei atau wawancara. Observasi dapat dilakukan secara partisipatif, di mana peneliti menjadi bagian dari lingkungan yang diamati, atau non-partisipatif, di mana peneliti tetap sebagai pengamat luar.

Keunggulan dari teknik observasi adalah kemampuannya untuk mengumpulkan data yang natural dan *real-time*, yang sangat berharga dalam memahami konteks dan dinamika yang kompleks. Misalnya, dalam studi tentang perilaku konsumen di toko ritel, observasi langsung dapat memberikan wawasan tentang bagaimana konsumen berinteraksi dengan produk dan tata letak toko. Namun, observasi juga memiliki keterbatasan, seperti potensi bias dari peneliti, kesulitan dalam mereplikasi kondisi yang sama, dan tantangan dalam mengamati perilaku yang jarang terjadi. Selain itu, penting untuk mempertimbangkan aspek etika, seperti memastikan bahwa subjek yang diamati memberikan persetujuan jika mereka sadar sedang diamati, dan menjaga kerahasiaan informasi yang diperoleh (Jones & Bradley, 2021).

## Pengumpulan Data Digital

Pengumpulan data digital adalah metode yang memanfaatkan teknologi untuk mengakses, mengumpulkan, dan menganalisis data dari berbagai sumber digital. Dalam era informasi saat ini, data digital berasal dari berbagai platform online, sensor IoT, aplikasi, dan banyak lagi. Penggunaan teknik pengumpulan data digital memungkinkan data scientist untuk mengumpulkan data dalam skala besar dengan efisiensi tinggi, memberikan wawasan yang mendalam dan komprehensif yang tidak mungkin dicapai melalui metode tradisional. Salah satu keuntungan utama dari pengumpulan data digital adalah kemampuannya untuk terus-menerus mengumpulkan data secara *real-time*, yang sangat penting untuk analisis tren dan pengambilan keputusan yang cepat.

Salah satu metode populer dalam pengumpulan data digital adalah web scraping, yang melibatkan ekstraksi data dari situs web menggunakan skrip atau alat otomatis. Web scraping memungkinkan peneliti untuk mengakses data yang tidak tersedia melalui API resmi atau yang sulit dikumpulkan secara manual. Misalnya, peneliti dapat menggunakan web scraping untuk mengumpulkan data harga dari berbagai situs *e-commerce* atau data ulasan produk dari forum online. Namun, teknik ini memerlukan pemahaman yang baik tentang struktur HTML dan kemungkinan menghadapi tantangan hukum terkait hak cipta dan kebijakan situs web yang melarang *scraping* (Mitchell, 2022).

Selain web scraping, penggunaan API (Application Programming Interface) juga merupakan metode penting dalam pengumpulan data digital. API memungkinkan aplikasi untuk berkomunikasi dan bertukar data satu sama lain secara terstruktur. Banyak platform besar seperti Google, Twitter, dan

Facebook menyediakan API yang memungkinkan peneliti mengakses data pengguna, aktivitas, dan metrik lainnya. Penggunaan API menawarkan cara yang lebih resmi dan seringkali lebih aman untuk mengumpulkan data dibandingkan web scraping, serta memastikan data yang diperoleh lebih terstruktur dan mudah diintegrasikan ke dalam sistem analisis (Vallath, 2023).

Sensor IoT (*Internet of Things*) juga memainkan peran penting dalam pengumpulan data digital. Sensor-sensor ini mengumpulkan data *real-time* dari lingkungan fisik, seperti suhu, kelembaban, dan gerakan, yang kemudian dikirim ke cloud untuk analisis lebih lanjut. Data dari sensor IoT sangat berguna dalam berbagai aplikasi, mulai dari smart home dan kota pintar hingga manajemen rantai pasokan dan pemantauan lingkungan. Meskipun menawarkan banyak keuntungan, pengumpulan data digital juga menghadirkan tantangan terkait privasi dan keamanan data. Oleh karena itu, penting bagi data scientist untuk memastikan bahwa data dikumpulkan dan digunakan dengan mematuhi regulasi dan standar etika yang berlaku, seperti General Data Protection Regulation (GDPR) di Eropa (Gubbi et al., 2021).

### Eksperimen

Eksperimen adalah salah satu teknik pengumpulan data yang paling mendalam dan terkontrol dalam data science, di mana peneliti melakukan manipulasi variabel tertentu untuk mengamati dan mengukur efeknya pada variabel lain. Teknik ini memungkinkan peneliti untuk menentukan hubungan sebab-akibat dengan tingkat kepastian yang tinggi. Dalam konteks data science, eksperimen sering digunakan untuk menguji hipotesis, mengoptimalkan sistem, dan memahami dinamika kompleks

dalam berbagai domain seperti pemasaran, kesehatan, dan rekayasa perangkat lunak.

Dalam eksperimen, peneliti membagi subjek atau sampel menjadi dua kelompok utama: kelompok eksperimen dan kelompok kontrol. Kelompok eksperimen mendapatkan perlakuan atau intervensi yang diuji, sementara kelompok kontrol tidak menerima perlakuan tersebut, atau menerima perlakuan standar. Perbandingan hasil antara kedua kelompok ini memungkinkan peneliti untuk menilai efek perlakuan dengan lebih akurat. Misalnya, dalam pengujian efektivitas kampanye iklan, satu kelompok pelanggan mungkin terpapar iklan tertentu sementara kelompok lain tidak, dan hasil penjualan di antara kedua kelompok dibandingkan untuk menilai dampak iklan tersebut (Lewis & Rao, 2021).

Keunggulan utama dari teknik eksperimen adalah kemampuannya untuk mengendalikan variabel luar yang bisa mempengaruhi hasil, sehingga memberikan hasil yang lebih valid dan dapat diandalkan. Namun, eksperimen juga memiliki tantangan tersendiri. Merancang dan melaksanakan eksperimen yang valid dan etis memerlukan perencanaan yang cermat, sumber daya yang cukup, dan kepatuhan terhadap prinsip-prinsip etika penelitian. Selain itu, beberapa eksperimen mungkin memerlukan waktu yang lama untuk memberikan hasil yang signifikan, terutama jika melibatkan perubahan perilaku manusia atau proses bisnis.

## Prosedur Pengumpulan Data

Pengumpulan data dalam data science adalah langkah awal yang krusial dalam seluruh proses analisis dan pengembangan model. Prosedur pengumpulan data harus dilakukan dengan cermat untuk memastikan bahwa data yang

diperoleh akurat, relevan, dan berkualitas tinggi. Berikut adalah tahapan utama dalam prosedur pengumpulan data dalam data science:

1. Pemahaman Masalah atau Tujuan Penelitian
   Tahap pertama adalah memahami masalah atau tujuan penelitian yang ingin diselesaikan atau dicapai. Ini mencakup identifikasi permasalahan yang ingin dipecahkan, peluang yang ingin diteliti, atau tujuan bisnis yang ingin dicapai. Pemahaman yang jelas tentang tujuan penelitian membantu dalam merumuskan kebutuhan data yang spesifik dan relevan (Provost & Fawcett, 2021).
2. Pemilihan Sumber Data
   Tahap awal dalam prosedur pemilihan sumber data adalah mengidentifikasi berbagai sumber data yang potensial yang dapat digunakan untuk mendukung tujuan penelitian atau analisis. Sumber data ini bisa berasal dari berbagai macam sumber, seperti basis data internal perusahaan, data publik yang tersedia secara online, atau data yang diperoleh melalui survei atau wawancara.
   a. Evaluasi Kredibilitas
      Setelah mengidentifikasi sumber data potensial, langkah selanjutnya adalah mengevaluasi kredibilitas masing-masing sumber data. Ini mencakup mempertimbangkan faktor-faktor seperti keandalan sumber data, metode pengumpulan data yang digunakan, dan reputasi penyedia data. Data yang berasal dari sumber yang kredibel cenderung lebih dapat dipercaya dan akurat (Kelleher & Tierney, 2022).
   b. Pertimbangkan Relevansi
      Selain kredibilitas, penting juga untuk

mempertimbangkan relevansi data dengan tujuan penelitian atau analisis. Data yang relevan adalah data yang secara langsung terkait dengan topik atau variabel yang sedang diteliti. Evaluasi ini membantu memastikan bahwa data yang dikumpulkan dapat memberikan wawasan yang berharga dan relevan untuk analisis yang akan dilakukan.

c. Aksesibilitas Data
Tahap terakhir adalah mempertimbangkan aksesibilitas data. Data yang dipilih harus dapat diakses dengan mudah dan legal. Ini termasuk memeriksa apakah data tersedia secara publik atau apakah perlu izin khusus atau pembayaran untuk mengaksesnya. Aksesibilitas data mempengaruhi kemampuan untuk mengumpulkan data yang diperlukan dalam waktu yang efisien dan efektif (Harris, Murphy, & Vaisman, 2021).

3. Pengelolaan dan Penyimpanan Data
Prosedur pengelolaan dan penyimpanan data merupakan tahap penting dalam siklus hidup data dalam data science. Langkah ini bertujuan untuk memastikan keamanan, ketersediaan, dan kualitas data yang telah dikumpulkan, yang menjadi dasar bagi analisis dan pengambilan keputusan yang akurat. Berikut adalah penjelasan mengenai prosedur ini:

a. Penyimpanan Data
Tahap awal adalah menentukan sistem penyimpanan yang sesuai untuk data yang dikumpulkan. Ini bisa berupa *database* relasional, data warehouse, atau penyimpanan cloud. Penting

untuk memilih sistem yang dapat menangani volume data yang besar, memungkinkan akses cepat, dan memiliki fitur keamanan yang memadai.

b. Pengaturan Keamanan
   Setelah data disimpan, langkah selanjutnya adalah mengatur keamanan data. Ini mencakup pengaturan hak akses yang sesuai untuk memastikan bahwa hanya orang yang berwenang yang dapat mengakses data tersebut. Peran dan izin harus ditetapkan secara jelas berdasarkan kebutuhan dan tanggung jawab masing-masing pengguna.
c. Pemeliharaan Kualitas Data
   Proses pengelolaan data juga melibatkan pemeliharaan kualitas data. Ini mencakup pemantauan secara teratur terhadap kualitas data, deteksi dan perbaikan data yang tidak lengkap, tidak akurat, atau tidak konsisten, serta pencegahan terjadinya duplikasi atau data yang tidak valid.
d. Backup Data
   Langkah berikutnya adalah melakukan backup data secara teratur untuk melindungi data dari kehilangan atau kerusakan. Backup data harus disimpan di lokasi yang terpisah dan aman, sehingga data dapat dipulihkan dengan cepat jika terjadi kegagalan sistem atau bencana alam.
e. Penghapusan Data yang tidak Diperlukan
   Data yang sudah tidak relevan atau tidak diperlukan lagi harus dihapus secara teratur untuk

mengurangi beban penyimpanan dan memastikan kebersihan dan keteraturan data. Proses penghapusan data harus dilakukan dengan hati-hati dan sesuai dengan kebijakan privasi dan regulasi yang berlaku.

f. Audit dan Pelaporan
   Terakhir, penting untuk melakukan audit secara berkala terhadap sistem penyimpanan data untuk memastikan kepatuhan terhadap kebijakan dan regulasi yang berlaku. Hasil audit ini dapat digunakan untuk menyusun laporan tentang status dan kesehatan data, serta rekomendasi perbaikan yang diperlukan.

# BAB 11 ANALISIS DATA DALAM DATA SCIENCE

## Pendahuluan

Di era digital saat ini, data telah menjadi salah satu aset paling berharga dalam berbagai disiplin ilmu. Analisis data adalah metode komprehensif untuk memeriksa, membersihkan, mentransformasikan, dan memodelkan data untuk menemukan informasi yang berguna, menarik kesimpulan, dan mendukung pengambilan keputusan. Ini adalah proses multifaset yang melibatkan berbagai teknik dan metodologi untuk menafsirkan data dari berbagai sumber dalam format berbeda, baik terstruktur maupun tidak terstruktur (Matt Crabtree, 2023). Analisis data dalam dalam data science adalah proses yang melibatkan pengumpulan, pembersihan, transformasi, dan pemodelan data dengan tujuan menemukan informasi berguna, mendukung pengambilan keputusan, dan membuat prediksi (Coursera staff, 2024).

Analisis data ilmiah mencakup berbagai metode dan teknik yang digunakan untuk menggali wawasan dari data mentah (Horvitz & Mulligan, 2015). Tujuan utama dari analisis data dalam data science adalah untuk menghasilkan pemahaman yang lebih baik tentang data, mengidentifikasi peluang, mengoptimalkan kinerja, dan membuat keputusan yang lebih tepat dan efektif. Peran analisis data dalam data science semakin signifikan seiring dengan meningkatnya adopsi teknologi *Big Data* dan kecerdasan buatan (AI) (Davenport, 2018). Teknologi ini memungkinkan analisis data yang lebih cepat dan akurat, serta mampu menangani dataset yang sangat besar dan

kompleks. Analisis data yang efektif memerlukan kombinasi dari keterampilan teknis, pemahaman bisnis, dan kemampuan untuk berkomunikasi dengan jelas. Para profesional data science tidak hanya harus menguasai alat dan teknik analisis, tetapi juga harus mampu menginterpretasikan data dalam konteks bisnis dan membuat rekomendasi yang dapat diimplementasikan. Menurut studi oleh McKinsey Global Institute (Davenport, 2018) perusahaan yang memanfaatkan analisis data secara efektif dapat meningkatkan produktivitas dan profitabilitas mereka secara signifikan.

Secara umum, analisis data dalam data science dapat dikelompokkan menjadi beberapa jenis, seperti analisis deskriptif (menjelaskan data secara ringkas), analisis eksploratif (mencari pola dan hubungan dalam data), analisis prediktif (memprediksi hasil berdasarkan data historis), dan analisis preskriptif (memberikan rekomendasi aksi berdasarkan ha5sil analisis). Selain itu, etika penggunaan data, termasuk privasi dan keamanan, menjadi semakin penting seiring dengan peningkatan volume dan kompleksitas data yang dianalisis.

Secara keseluruhan, analisis data dalam data science merupakan bidang yang dinamis dan berkembang pesat, menawarkan peluang besar untuk inovasi dan penemuan baru. Dengan terus berkembangnya teknologi dan metodologi analisis data, potensi untuk pemahaman yang lebih mendalam dan solusi yang lebih efektif untuk masalah ilmiah akan semakin meningkat.

## Tujuan Utama Analisis Data

Tujuan utama dari analisis data dalam data science adalah untuk menghasilkan pemahaman yang lebih baik diantaranya sebagai berikut (Núria Emilio, 2023):

1. Mengidentifikasi Pola dan Tren
   Salah satu fungsi utama analisis data adalah mengidentifikasi pola dan tren dalam dataset. Dengan menggunakan teknik statistik dan visualisasi data, analis dapat mengungkap hubungan tersembunyi, anomali, dan tren temporal yang mungkin tidak terlihat secara kasat mata. Identifikasi pola ini sangat penting untuk memahami perilaku historis dan memprediksi kejadian di masa depan.
2. Mendukung Pengambilan Keputusan
   Analisis data menyediakan landasan empiris bagi pengambilan keputusan yang lebih baik dan lebih cepat. Dengan menganalisis data historis dan *real-time*, perusahaan dapat membuat keputusan yang lebih tepat mengenai strategi bisnis, pemasaran, operasional, dan manajemen risiko. Data yang dianalisis dengan baik memungkinkan organisasi untuk mengurangi ketidakpastian dan meningkatkan efektivitas keputusan yang diambil.
3. Mengukur Kinerja dan Efektivitas
   Melalui analisis data, organisasi dapat mengukur kinerja dan efektivitas berbagai inisiatif dan proses bisnis. Indikator kinerja utama (*Key Performance Indicator*/KPI) dan metrik lainnya dihasilkan dari data yang dianalisis untuk memberikan wawasan tentang seberapa baik suatu organisasi mencapai tujuannya. Analisis ini membantu dalam mengidentifikasi area yang membutuhkan perbaikan dan mengukur dampak dari perubahan yang diterapkan.
4. Memperbaiki dan Mengoptimalkan Proses
   Analisis data memainkan peran penting dalam

memperbaiki dan mengoptimalkan proses bisnis. Dengan menganalisis data operasional, organisasi dapat menemukan bottleneck, inefisiensi, dan area yang dapat dioptimalkan. Pendekatan berbasis data memungkinkan perbaikan berkelanjutan dan inovasi dalam proses, yang pada akhirnya meningkatkan produktivitas dan efisiensi.

5. Mengembangkan Model Prediktif
   Fungsi lainnya dari analisis data adalah mengembangkan model prediktif yang dapat digunakan untuk memprediksi hasil di masa depan. Dengan menggunakan teknik *Machine Learning* dan statistik, model prediktif dapat dibangun untuk berbagai aplikasi seperti prediksi penjualan, analisis risiko, deteksi penipuan, dan personalisasi produk. Model ini membantu organisasi untuk proaktif dalam mengatasi tantangan dan memanfaatkan peluang.
6. Mendukung Inovasi
   Analisis data adalah kunci untuk mendukung inovasi dalam berbagai bidang. Dengan mengumpulkan dan menganalisis data, perusahaan dapat menemukan wawasan baru yang dapat digunakan untuk mengembangkan produk dan layanan baru, meningkatkan pengalaman pelanggan, dan menciptakan solusi yang lebih efisien dan efektif. Inovasi berbasis data memungkinkan organisasi untuk tetap kompetitif dalam lingkungan bisnis yang cepat berubah.

## Tahapan Analisis Data dalam Data Science

Tahapan analisis data dalam data science melibatkan langkah-langkah sistematis yang membantu mengubah data mentah menjadi wawasan yang dapat diandalkan dan digunakan

untuk pengambilan keputusan (Simplelearn, 2024). Berikut adalah tahapan utama dalam analisis data dalam data science:

1. Pengumpulan Data
   Pengumpulan data adalah langkah pertama dalam proses analisis data dan melibatkan pengumpulan data dari berbagai sumber. Data dapat berasal dari sumber internal seperti *database* perusahaan atau sumber eksternal seperti survei dan data publik. Pengumpulan data yang baik memastikan bahwa data yang diperoleh relevan, akurat, dan cukup untuk analisis lebih lanjut.
   a. Sumber Internal dan Eksternal
      Menjelaskan perbedaan antara data yang diperoleh dari dalam organisasi dan dari luar organisasi.
   b. Metode Pengumpulan
      Membahas berbagai metode pengumpulan data seperti survei, wawancara, dan *scraping* data dari web.
   c. Kualitas Data
      Menekankan pentingnya kualitas data dan cara memastikan data yang dikumpulkan bebas dari bias dan kesalahan.
2. Pembersihan Data
   Pembersihan data adalah proses menghilangkan atau memperbaiki data yang kotor atau tidak akurat untuk memastikan kualitas dan konsistensi data. Ini termasuk menghapus duplikasi, menangani data yang hilang, dan memperbaiki kesalahan format.
   *a*. Deteksi dan Penanganan *Outliers*
      Mengidentifikasi dan menangani data yang tidak normal atau ekstrem.
   b. Mengatasi Data Hilang

Strategi untuk mengisi atau mengabaikan data yang hilang.

c. Standardisasi Data
Memastikan data dalam format yang konsisten untuk analisis lebih lanjut.

3. Transformasi Data
Transformasi data melibatkan konversi data mentah menjadi format yang sesuai untuk analisis. Ini dapat mencakup agregasi data, normalisasi, dan pembuatan fitur baru yang relevan untuk model analitik.
   a. Agregasi Data
   Menggabungkan data dari berbagai sumber atau mengelompokkan data berdasarkan kategori tertentu.
   b. Normalisasi dan Skala Data
   Menyelaraskan data dalam rentang tertentu untuk memastikan bahwa semua fitur memberikan kontribusi yang sama dalam analisis.
   c. Pembuatan Fitur
   Membuat fitur baru yang dapat memberikan wawasan lebih mendalam dan relevan.
4. Eksplorasi Data
Eksplorasi data adalah tahap di mana data diperiksa secara mendalam untuk memahami pola, distribusi, dan hubungan antara variabel. Teknik ini termasuk visualisasi data dan analisis statistik deskriptif.
   a. Visualisasi Data
   Menggunakan grafik dan diagram untuk melihat distribusi data dan hubungan antar variabel.
   b. Statistik Deskriptif
   Menghitung metrik seperti mean, median, dan

standar deviasi untuk memahami karakteristik dasar data.

c. Analisis Korelasi
   Menentukan hubungan antara variabel yang berbeda.

5. Pemodelan Data
   Pemodelan data adalah proses membangun model analitik untuk memprediksi atau mengklasifikasikan data. Ini melibatkan pemilihan algoritma, pelatihan model, dan evaluasi kinerja model.
   a. Pemilihan Model
      Memilih algoritma yang paling sesuai untuk masalah yang dihadapi, seperti regresi, klasifikasi, atau *Clustering*.
   b. Pelatihan Model
      Menggunakan data pelatihan untuk mengajarkan model pola dan hubungan dalam data.
   c. Evaluasi Model
      Menggunakan metrik evaluasi seperti akurasi, presisi, dan *recall* untuk menilai kinerja model.
6. Evaluasi dan Validasi
   Evaluasi dan validasi adalah tahap di mana model yang dibangun diuji untuk memastikan bahwa ia bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya. Teknik ini termasuk cross-validation dan penggunaan data uji.
   a. *Cross-Validation*
      Metode untuk mengukur kinerja model dengan membagi data menjadi beberapa subset dan melakukan validasi silang.
   b. Tes Data

Menggunakan subset data yang disimpan khusus untuk menguji kinerja model pada data baru.

c. Penyesuaian Model
Melakukan tuning parameter dan memilih model terbaik berdasarkan hasil evaluasi.

7. Visualisasi dan Interpretasi Hasil
Visualisasi dan interpretasi hasil adalah tahap akhir di mana hasil analisis dan model disajikan dengan cara yang mudah dipahami. Ini melibatkan pembuatan laporan, dashboard, dan interpretasi hasil analitik.
    a. Dashboard dan Laporan
    Menggunakan alat visualisasi seperti Tableau atau PowerBI untuk membuat dashboard interaktif dan laporan yang menggambarkan hasil analisis.
    b. Komunikasi Hasil
    Menjelaskan hasil analisis kepada pemangku kepentingan dengan cara yang jelas dan mudah dipahami.
    c. Interpretasi Bisnis
    Menghubungkan hasil analisis dengan konteks bisnis untuk memberikan rekomendasi yang actionable.

Dengan memahami dan menerapkan setiap langkah dalam proses analisis data ini, organisasi dapat memastikan bahwa mereka memanfaatkan data mereka secara efektif untuk menghasilkan wawasan yang berharga dan mendukung pengambilan keputusan yang lebih baik. Proses yang terstruktur dan sistematis ini merupakan landasan penting dalam praktik data science modern.

## Metode Analisis Data dalam Data Science

Dalam data science, berbagai teknik atau metode analisis data digunakan untuk menggali wawasan dari data mentah. Berikut adalah beberapa metode utama analisis data yang sering digunakan antara lain (Bernardita Calzon, 2023):

1. Metode Eksploratif
   Metode eksploratif digunakan untuk memahami karakteristik dasar dari data dan menemukan pola yang tidak terduga. Teknik ini termasuk eksplorasi visual, analisis statistik dasar, dan analisis multivariat.
   a. Eksplorasi Visual
      Menggunakan alat visualisasi seperti grafik dan diagram untuk melihat distribusi data, mengidentifikasi outliers, dan memahami hubungan antar variabel.
   b. Analisis Statistik Dasar
      Menghitung metrik dasar seperti mean, median, modus, variansi, dan standar deviasi untuk mendapatkan gambaran umum tentang data.
   c. Analisis Multivariat
      Menganalisis lebih dari dua variabel secara bersamaan untuk memahami interaksi dan hubungan kompleks antar variabel.
2. Metode Deskriptif
   Metode deskriptif digunakan untuk merangkum dan mendeskripsikan dataset dengan cara yang sistematis dan informatif. Ini mencakup pembuatan laporan, tabel frekuensi, dan deskriptif statistik.
   a. Laporan Deskriptif
      Membuat laporan yang merangkum data dengan teks, tabel, dan grafik untuk memberikan

gambaran keseluruhan.

b. Tabel Frekuensi
   Menyusun data dalam tabel yang menunjukkan frekuensi kemunculan setiap nilai atau kategori.

c. Deskriptif Statistik
   Menggunakan statistik seperti mean, median, modus, dan rentang untuk memberan ringkasan data yang mudah dipahami.

3. Metode Prediktif
   Metode prediktif digunakan untuk memprediksi nilai masa depan berdasarkan data historis. Teknik ini melibatkan penggunaan algoritma *Machine Learning* seperti regresi linear, *Decision Trees*, dan *Neural Networks*.

   a. Regresi Linear
      Memprediksi nilai masa depan berdasarkan hubungan linear antara variabel.

   b. *Decision Trees*
      Menggunakan model berbentuk pohon untuk membuat prediksi berdasarkan serangkaian keputusan.

   c. *Neural Networks*
      Menggunakan jaringan saraf tiruan untuk memodelkan dan memprediksi hubungan non-linear yang kompleks dalam data.

4. Metode Preskriptif
   Metode preskriptif digunakan untuk memberikan rekomendasi tindakan berdasarkan hasil analisis data. Teknik ini termasuk optimasi dan simulasi.

   a. Optimasi
      Menggunakan model matematika untuk

menemukan solusi terbaik dari serangkaian kemungkinan dengan mempertimbangkan batasan tertentu.

b. Simulasi
Menggunakan model simulasi untuk mengevaluasi dampak dari berbagai skenario dan keputusan dalam lingkungan yang terkontrol.

5. Metode Pembelajaran Mesin (*Machine Learning*)
Metode pembelajaran mesin digunakan untuk mengotomatisasi analisis data dan membuat prediksi dengan menggunakan algoritma yang belajar dari data. Teknik ini melibatkan *supervised learning, unsupervised learning*, dan *reinforcement learning*.
   a. *Supervised Learning*
   Menggunakan data berlabel untuk melatih model yang dapat membuat prediksi berdasarkan input baru.
   b. *Unsupervised Learning*
   Menganalisis data tanpa label untuk menemukan pola dan struktur tersembunyi, seperti *Clustering* dan association.
   c. *Reinforcement Learning*
   Melatih model untuk membuat keputusan melalui percobaan dan kesalahan, dengan mendapatkan umpan balik dari lingkungan.
6. Metode Time Series
Metode time series digunakan untuk menganalisis data yang dikumpulkan dalam urutan waktu tertentu. Teknik ini mencakup analisis musiman, dekomposisi, dan peramalan.
   a. Analisis Musiman

Mengidentifikasi pola berulang dalam data berdasarkan periode waktu tertentu.

b. Dekomposisi
Memisahkan data time series menjadi komponen trend, seasonal, dan residual untuk analisis lebih lanjut.

c. Peramalan
Menggunakan model statistik seperti ARIMA dan exponential smoothing untuk memprediksi nilai masa depan berdasarkan data historis.

7. Metode Text Mining dan Analisis Sentimen
Metode ini digunakan untuk menganalisis data tekstual dan memahami sentimen di balik teks tersebut. Teknik ini melibatkan *text preprocessing, word cloud*, dan *sentiment analysis*.

a. *Text Preprocessing*
Membersihkan dan menyiapkan teks untuk analisis dengan menghapus *stop words*, *stemming*, dan tokenisasi.

b. *Word Cloud*
Visualisasi kata yang paling sering muncul dalam teks untuk mengidentifikasi tema utama.

c. *Sentiment Analysis*
Menggunakan algoritma NLP untuk menentukan sentimen positif, negatif, atau netral dalam teks.

Dengan memahami berbagai metode analisis data ini, kita dapat memilih pendekatan yang paling sesuai dengan kebutuhan dan tujuan spesifik dari setiap proyek data science. Metode-metode ini tidak hanya membantu dalam mengekstrak wawasan yang berharga dari data tetapi juga memungkinkan pengambilan keputusan yang lebih baik dan lebih tepat.

## Studi Kasus dalam Analisis Data dalam Data Science

Berikut beberapa contoh studi kasus dalam analisis data dalam data science merupakan cara yang efektif untuk memahami bagaimana teknik-teknik analisis data diterapkan dalam situasi nyata.

1. Studi Kasus di Industri Kesehatan
   Industri kesehatan adalah salah satu sektor yang sangat diuntungkan dari analisis data. Melalui penggunaan data pasien, riwayat medis, dan data kesehatan populasi, analisis data membantu dalam meningkatkan perawatan pasien dan efisiensi operasional.
   a. Manajemen Rumah Sakit
      Analisis data digunakan untuk mengoptimalkan aliran pasien, manajemen tempat tidur, dan penjadwalan staf. Contoh konkret termasuk prediksi jumlah pasien yang akan datang untuk mengurangi waktu tunggu dan meningkatkan kepuasan pasien.
   b. Diagnosa Penyakit
      Menggunakan algoritma *Machine Learning* untuk menganalisis data medis dan memprediksi kemungkinan penyakit pada tahap awal. Misalnya, analisis citra medis untuk deteksi dini kanker menggunakan *Deep Learning*.
   c. Penelitian Klinis
      Data science membantu dalam menganalisis hasil uji klinis untuk menemukan efek samping obat baru atau menentukan kelompok pasien yang paling mungkin mendapatkan manfaat dari perawatan tertentu.
2. Studi Kasus di Industri Keuangan

Sektor keuangan memanfaatkan analisis data untuk meningkatkan kinerja, manajemen risiko, dan deteksi penipuan.

a. Manajemen Risiko
   Analisis data digunakan untuk mengidentifikasi dan mengukur risiko dalam portofolio investasi. Model prediktif membantu dalam mengantisipasi kerugian potensial dan mengembangkan strategi mitigasi risiko.
b. Deteksi Penipuan
   Algoritma *Machine Learning* diterapkan untuk menganalisis pola transaksi dan mendeteksi aktivitas yang mencurigakan secara *real-time*. Contoh konkret adalah penggunaan *Neural Networks* untuk mendeteksi penipuan kartu kredit.
c. Analisis Kredit
   Penilaian kredit menggunakan data pelanggan untuk menentukan kelayakan kredit dan menetapkan batas kredit. Analisis data membantu dalam membuat keputusan yang lebih akurat mengenai pemberian pinjaman.

3. Studi Kasus di Industri Pemasaran
   Industri pemasaran menggunakan analisis data untuk memahami perilaku konsumen, mengoptimalkan kampanye pemasaran, dan meningkatkan ROI.

   a. Segmentasi Pelanggan
      Analisis cluster digunakan untuk mengelompokkan pelanggan berdasarkan perilaku dan preferensi mereka, memungkinkan pemasaran yang lebih tertarget dan personal.

b. Prediksi Churn
   Model prediktif membantu mengidentifikasi pelanggan yang berisiko meninggalkan layanan, sehingga perusahaan dapat mengambil tindakan proaktif untuk mempertahankan mereka.
c. Optimasi Kampanye
   A/B testing dan analisis data digunakan untuk mengukur efektivitas kampanye pemasaran dan mengoptimalkan strategi pemasaran berdasarkan hasil data.

4. Studi Kasus di Industri Manufaktur
   Analisis data dalam industri manufaktur digunakan untuk meningkatkan efisiensi operasional, kualitas produk, dan mengurangi biaya.
   a. Pemeliharaan Prediktif
      Menggunakan analisis data sensor untuk memprediksi kegagalan mesin sebelum terjadi, sehingga mengurangi waktu henti dan biaya perbaikan.
   b. Optimasi Produksi
      Analisis data untuk mengoptimalkan proses produksi, mengurangi limbah, dan meningkatkan produktivitas. Contoh konkret adalah penggunaan *Machine Learning* untuk mengoptimalkan jadwal produksi dan pengelolaan inventaris.
   c. Kualitas Produk
      Menganalisis data kualitas untuk mengidentifikasi penyebab cacat produk dan mengimplementasikan kontrol kualitas yang lebih baik.
5. Studi Kasus di *E-commerce*
   *E-commerce* memanfaatkan analisis data untuk

meningkatkan pengalaman pelanggan, mengoptimalkan inventaris, dan meningkatkan penjualan.

a. Personalisasi Pengalaman Pengguna
   Menggunakan analisis data untuk memberikan rekomendasi produk yang dipersonalisasi berdasarkan riwayat pembelian dan perilaku browsing pelanggan.
b. Manajemen Inventaris
   Analisis data digunakan untuk mengelola inventaris secara efisien, mengurangi stok berlebih, dan memastikan ketersediaan produk yang populer.
c. Harga Dinamis
   Menggunakan algoritma untuk menyesuaikan harga produk secara *real-time* berdasarkan permintaan, persaingan, dan faktor lainnya untuk memaksimalkan pendapatan.

6. Studi Kasus di Transportasi dan Logistik
   Industri transportasi dan logistik menggunakan analisis data untuk meningkatkan efisiensi operasional, mengurangi biaya, dan meningkatkan layanan pelanggan.
   a. Optimasi Rute
      Menggunakan data GPS dan analisis rute untuk mengoptimalkan pengiriman, mengurangi waktu tempuh, dan menghemat bahan bakar.
   b. Pemantauan *Real-time*
      Menerapkan analisis data untuk pemantauan *real-time* armada kendaraan, membantu dalam pengambilan keputusan yang cepat dan responsif.
   c. Prediksi Permintaan
      Model prediktif digunakan untuk memprediksi

permintaan layanan transportasi dan logistik, memungkinkan perencanaan yang lebih baik dan pengalokasian sumber daya yang efisien.

Melalui studi kasus ini, kita dapat melihat bagaimana analisis data diterapkan dalam berbagai industri untuk meningkatkan efisiensi, kualitas, dan pengambilan keputusan. Setiap studi kasus menunjukkan bagaimana data yang diolah dan dianalisis dapat menghasilkan wawasan yang dapat ditindaklanjuti, mendukung inovasi, dan memberikan keunggulan kompetitif dalam pasar yang semakin kompleks dan kompetitif.

# BAB 12 ETIKA DAN ESTETIKA DARI DATA

## Pendahuluan

Kemajuan teknologi informasi telah mengubah cara kita berinteraksi, bekerja, dan hidup sehari-hari secara signifikan. Selain itu, kemajuan ini telah membuat kehidupan manusia menjadi lebih mudah. Kita dapat bekerja dari mana saja dan kapan saja tanpa batasan ruang dan waktu asalkan kita dapat terhubung ke internet. Jam pintar dan telepon dapat secara *real-time* memprediksi denyut jantung dan jumlah kalori yang dibakar selama latihan. Dengan bersuara, telepon pintar kita dapat mencari apa pun di internet tanpa mengetik. Revolusi industri 4.0 membawa banyak manfaat bagi kehidupan manusia sehingga membuat orang menjadi ketergantungan pada teknologi. Namun, revolusi ini juga membawa dampak negatif, terutama berkaitan dengan etika penggunaan data (Hand, 2018).

Data dalam volume besar, juga dikenal sebagai "*Big Data*", adalah inti dari revolusi industri 4.0 (Shaturaev, 2022). Pada era revolusi industri 4.0, berbagai pihak dapat dengan mudah mengumpulkan data tentang apa pun yang ada di dunia ini, mulai dari data keuangan perusahaan hingga data pergerakan air laut dan pandemi COVID-19. Selanjutnya, data ini dapat digunakan untuk mensimulasikan berbagai skenario dan memprediksi fenomena bisnis yang dapat menguntungkan industri (Gunal, 2019).

Tetapi kemudahan mendapatkan data secara bebas menimbulkan kekhawatiran karena data dapat digunakan untuk berbagai macam alasan, baik itu positif maupun negatif. Dengan

jumlah akun digital yang mengalami kebocoran data 12,74 juta akun pada kuartal III 2022, Indonesia menempati peringkat ketiga negara dengan kasus kebocoran data tertinggi di dunia, menyatakan Data yang dapat diakses dengan mudah oleh siapa saja, baik baik buruk, menimbulkan masalah etika yang memerlukan penyelidikan lebih lanjut (Vaddhano, 2023).

Etika data, juga dikenal sangat penting untuk berbagai jenis penelitian dan pengembangan ilmu pengetahuan, sehingga menjadi salah satu komponen penting. Meskipun memiliki potensi untuk maju, pengambilan dan pengolahannya melibatkan tanggung jawab besar. Meskipun data membantu dalam pekerjaan dan kehidupan sehari-hari, mereka juga menimbulkan masalah yang kompleks. Ketika data mencakup informasi pribadi, penyalahgunaan data, pelanggaran privasi, dan diskriminasi adalah masalah utama. Memahami dan menerapkan data etis sangat penting dalam hal ini. Ketika orang tahu tentang pentingnya menjaga etika data, mereka dapat menghindari banyak masalah yang dapat muncul dari penyalahgunaan data.

## Etika Data

Untuk memulai menerapkan prinsip-prinsip etika, kita harus meminta izin orang sebelum mengumpulkan data; kita harus menghormati dan melindungi identitas dan privasi orang tersebut. Ada kemungkinan bahwa data tidak akan digunakan untuk kepentingan lain jika ada pembatasan yang sesuai dengan tujuan yang telah disetujui sebelumnya. Hasilnya, tercipta proses pengolahan data yang sesuai dengan harapan individu dan bertanggung jawab. Agar data tidak jatuh ke tangan yang salah, keamanan data juga sangat penting. Data harus dilindungi dengan protokol keamanan yang ketat. Teknologi seperti

enkripsi dan akses terbatas hanya untuk orang yang berwenang dapat membantu melindungi data.

Mengolah data dengan tetap menjaga keakuratannya dan menghindari manipulasinya sangat penting. Data tidak boleh diubah untuk mendukung tujuan tertentu, tetapi harus menunjukkan kenyataan yang ada. Hasil data yang benar menjaga kepercayaan publik dan mencegah penyebaran informasi yang salah. Etik data menjamin pengumpulan, penyimpanan, dan pemanfaatan data secara tepat waktu dengan menghormati hak privasi dan mencegah penyalahgunaan. Seseorang dapat menjaga integritas penelitian dan kepercayaan publik terhadap temuannya dengan menerapkan etika ini. Oleh karena itu, penting untuk memahami bagaimana menggunakan data dengan benar dan menerapkan pengetahuan ini dalam kehidupan sehari-hari. Akibatnya, kami dapat menggunakan data dengan jelas, terbuka, dan sesuai tujuan tanpa mengganggu privasi individu.

Open Data Institute mengatakan bahwa data etika adalah bagian dari etika dalam menilai praktik pengelolaan data yang dapat berdampak negatif pada orang dan masyarakat saat mengumpulkan, membagi, dan menggunakan data. Dalam hal etika pengelolaan data, apa yang harus diperhatikan?

1. Dampak terhadap orang. Dalam konteks data pribadi, karakteristik seseorang pasti diwakili dalam data, yang dapat digunakan untuk membuat keputusan yang dapat memengaruhi kehidupan mereka. Misalnya, data kesehatan dan rekam medis. Apakah dampak jika rekam medis bocor? Orang yang tidak bertanggung jawab dapat memanfaatkannya untuk mendapatkan uang, misalnya dengan menjual rekam medis kepada perusahaan yang membutuhkan.

2. Potensi penyalahgunaan. Penyalahgunaan data dapat berdampak buruk pada orang. Sebagai contoh, saat kita memasukkan kartu kredit ke toko di pusat perbelanjaan. Dalam waktu dekat, pasti akan ada iklan, baik dari penyedia kartu kredit lain maupun iklan lainnya. Kami selalu bertanya-tanya dari mana penjual mendapatkan nomor kami. Contoh tambahan adalah insiden daftar pemilih tetap yang bocor, meskipun KPU mengklaim bahwa data tersebut dapat diakses secara publik. Bisakah data ini digunakan? Informasi tersebut pasti dapat digunakan untuk menipu pelaku kriminal selain dapat dijual.
3. Nilai ekonomis data: data yang diolah dengan benar akan memiliki nilai ekonomis. Bagaimana nilai dihasilkan dari data dan siapa yang dapat mengambil nilai ekonominya adalah tanggung jawab pemilik data.

Prinsip-prinsip Etika Data: Etika data berkaitan dengan praktik yang baik dalam pengumpulan, penggunaan, dan penyebaran data. Hal ini jelas sangat penting ketika tindakan yang dilakukan dalam pengelolaan data dapat memengaruhi masyarakat dan individu secara langsung maupun tidak langsung. Sebagai contoh, model otomasi data dapat membuat keputusan tentang apakah seseorang memenuhi syarat untuk mendapatkan asuransi, kredit, atau jenis layanan lain yang ditawarkan oleh bisnis kepada pelanggannya. Aktivitas ini, serta data pengecualian, dapat berdampak pada orang-orang dalam kelompok masyarakat tertentu. Dalam siklus hidup data, setidaknya ada tiga keputusan yang membutuhkan prinsip Etika Data:

1. *Disclose* Data

   Tahap pertama siklus hidup data adalah pembuatan data

sendiri, baik itu dibuat sendiri atau diambil dari sumber lain. Setelah dibuat dan diambil, data akan disimpan. Pada tahap pengambilan data, etika pengungkapan atau pengungkapan data harus dipertimbangkan. Ini termasuk siapa yang memiliki data pribadi, proses pengambilan data, dan sistem yang digunakan untuk membuat dan menyebarkan data.

2. *Manipulate* Data
   Data yang telah dikumpulkan harus diubah atau diproses untuk menjadi informasi bernilai. Dalam proses pengolahan data, tentu saja ada proses perubahan atau manipulasi data untuk menjamin kualitas dan integritas data. Semua aspek manipulasi data, termasuk pemilik data pribadi, proses, dan sistem yang melakukan pengolahan, perubahan, pemindahan, dan analisis data, harus dipertimbangkan dengan hati-hati. Dalam kasus di mana kita mengenal istilah Extract, Transform, and Load (ETL), proses ini terjadi pada tahap pengenalan dan perubahan data.
3. *Consume* Data
   Data yang telah dianalisis pasti akan dikonsumsi, baik oleh individu maupun sistem lain yang menggunakannya, seperti tools BI. Pada tahap ini, diharapkan data ini memberikan insight terbaik sehingga dapat digunakan sebagai acuan untuk proses pengambilan keputusan organisasi. Dalam hal kepercayaan data, sangat penting untuk mempertahankan keterlibatan manusia dalam proses pengambilan keputusan dan tidak bergantung pada sistem sepenuhnya. Tidak jelas pengetahuan dan pengalaman manusia masih diperlukan, tidak peduli seberapa canggih sistem AI saat ini. Karena keduanya

dibuat dan dijalankan oleh manusia, Data Privasi dan Perlindungan Data menjadi pilihan yang baik untuk kontrol data etika.

Menerapkan Data Ethics: Menerapkan etis data pada suatu organisasi sama dengan menerapkan inisiatif apapun, termasuk Program Privasi Data, karena manajemen organisasi dan stakeholder mendukungnya. Struktur data etika yang sederhana terdiri dari enam bagian:

1. *Vision*
   Visi organisasi sangat menentukan jalan dan tujuan organisasi. Organisasi harus menentukan cara mereka menggunakan data secara etis dalam konteks ini. Strategi etis data dapat dipilih oleh manajemen.
2. Strategi
   Untuk mencapai visi, strategi harus dibuat. Dalam hal ini, organisasi harus membuat strategi untuk memasukkan data etika ke dalam kulturnya secara teratur.
3. *Governance*
   Organisasi harus menyusun kebijakan dan prosedur yang kuat untuk "memaksa" pihak terkait untuk menerapkan praktik-praktik data etika. Ini juga harus memastikan bahwa masing-masing pihak terkait memiliki tanggung jawab yang jelas.
4. *Infrastructure & Architecture*
   Mengendalikan data yang kompleks (terutama untuk organisasi yang besar) akan lebih mudah dan terintegrasi jika organisasi memiliki visibilitas terhadap semua data, memasukkannya ke dalam arsitektur (misalnya, arsitektur perusahaan), dan didukung oleh sistem dan infrastruktur yang kuat dan dapat diandalkan.
5. Data *Insight*

Sangat penting untuk menggunakan insight untuk mendukung hasil data yang jelas dan akurat. Tools seperti dashboard dapat membantu organisasi memantau dan memberikan peringatan dini pelanggaran data etika.

6. *Training & Development*
   Dalam hal etika data, orang adalah komponen penting. Organisasi harus memberikan instruksi tentang etika penggunaan dan penyalahgunaan data. Karena data etika terkait, hal ini dapat dilakukan saat organisasi melakukan pelatihan atau sosialisasi tentang Privasi dan Perlindungan Data Pribadi.

Dua faktor utama penting untuk keberhasilan penerapan *Data Ethics*: dukungan dari Pemimpin Tinggi (termasuk penerapan) dan kesadaran dan pengetahuan dari stakeholder (termasuk kami). Data adalah minyak baru. Banyak orang berlomba untuk mendapatkan manfaat terbaik dari data. Namun, prinsip-prinsip pengelolaan dan pemanfaatan data juga harus mempertimbangkan aturan dan prinsip data privasi. Ini akan membantu kita memahami dan menerapkan penggunaan data yang etis.

## Estetika Data

Menurut Kamus Besar Bahasa Indonesia (KBBI), "estetika" adalah subdisiplin ilmu filsafat yang membahas seni dan keindahan, serta bagaimana manusia melihatnya. Ilmu estetika juga adalah ilmu yang membahas keindahan, bagaimana ia terbentuk, dan bagaimana seseorang merasakannya. Filosofi seni dan estetika adalah bidang yang sangat dekat. Fokusnya dalam data estetika adalah memvisualisasikan data.

Visualisasi data adalah proses menampilkan data menggunakan elemen visual seperti diagram, grafik, atau peta.

Estetika data dimaksudkan untuk meningkatkan pemahaman tentang pola, tren, atau hubungan dalam data dengan cara yang lebih mudah dipahami daripada hanya angka atau tabel.

Dengan menggunakan estetika data yang baik, informasi yang kompleks dapat dijelaskan dengan lebih baik kepada audiens yang beragam, termasuk mereka yang mungkin tidak memiliki pengetahuan teknis yang kuat. Memvisualisasikan data berarti mengambil nilai data dan mengubahnya menjadi komponen visual yang membentuk grafik akhir secara logis. Banyak jenis visualisasi data, seperti bar chart, yang biasanya digunakan untuk menunjukkan kenaikan atau penurunan dalam suatu periode waktu tertentu; line chart, yang biasanya digunakan untuk melihat kemajuan suatu data dalam jangka waktu pendek atau panjang; atau pie chart, yang biasanya digunakan untuk menunjukkan komposisi data. Semua visualisasi ini dapat dijelaskan dengan bahasa umum, sehingga orang dapat memahami bagaimana data nilai diubah menjadi grafik yang menarik. Arti estetika terletak pada fakta bahwa setiap visualisasi data mengintegrasikan nilai data ke dalam fitur-fitur yang dapat diukur dari grafik yang dihasilkan.

Untuk memetakan nilai data ke dalam estetika, kita perlu menentukan nilai mana yang sesuai dengan nilai estetika tertentu. Ini dapat dilakukan dengan mengidentifikasi nilai mana yang berada pada posisi tertentu di sepanjang sumbu x dalam grafik kita. Hal yang sama berlaku untuk nilai mana yang diwakili oleh bentuk atau warna tertentu. Skala digunakan untuk mendefinisikan pemetaan unik antara nilai data dan nilai estetika ini. Sangat penting bahwa skalanya satu-ke-satu, sehingga ada satu nilai estetika untuk setiap nilai data dan sebaliknya. Jika tidak, visualisasi data akan menjadi ambigu.

Nilai data dihubungkan dengan estetika melalui skala, di

mana angka dari satu hingga empat dipetakan ke dalam skala posisi, bentuk, dan warna. Untuk setiap skala, setiap angka sesuai dengan posisi, bentuk, atau warna yang berbeda, dan sebaliknya. Kesederhanaan, konsistensi, dan keseimbangan adalah prinsip estetika visual data. Untuk membuat visualisasi data yang efektif, Anda harus menghindari kekacauan visual dan memastikan bahwa komponen desain saling melengkapi. Menyederhanakan informasi yang kompleks tanpa menghilangkan maknanya adalah masalah utama dalam visualisasi data. Desainer perlu menguasai teknik pemilahan data dan selalu berkonsentrasi pada tujuan komunikasi agar mereka dapat mengekstraksi dan menyajikan informasi penting dengan cara yang jelas dan ringkas. Untuk menyampaikan informasi dengan efektif, Anda harus menggunakan diagram, grafik, dan infografis yang tepat. Dalam desain komunikasi visual, mengoptimalkan visualisasi data adalah proses yang melibatkan penggunaan berbagai teknik kreatif dan teknologi modern.

Dengan menggunakan infografis interaktif, psikologi warna, estetika grafis, dan cerita, desainer dapat membuat visualisasi data yang tidak hanya menarik tetapi juga berguna. Untuk memastikan bahwa pesan disampaikan dengan jelas dan efisien, pengguna harus selalu menjadi fokus utama dalam setiap tahap desain. Ada banyak data yang disampaikan dalam visualisasi data. Biasanya ada beberapa poin penting yang harus ditekankan dari berbagai data yang ingin disampaikan. Untuk memastikan bahwa poin-poin penting ini disampaikan dengan benar, poin-poin penting ini harus ditekankan dengan cara yang mudah dipahami oleh pembaca. Sangat disarankan untuk menampilkan data dengan cara yang tidak mengganggu. Menghilangkan berbagai "gangguan" seperti garis, sumbu, dan

label yang tidak penting adalah salah satu cara untuk menyajikan visual yang bebas ditraksi. Selain itu, Anda dapat meneliti penggunaan pola, ukuran, dan warna tertentu untuk menekankan informasi tertentu. Contoh dua alat visualisasi data yang populer adalah:

1. Google Data Studio, alat visualisasi data paling populer yang disediakan oleh perusahaan Google. Ini menawarkan banyak kemampuan, termasuk:
    a. Membuat laporan dan *dashboard*
    b. Menganalisis dan menyajikan hasil data
    c. Membuat keputusan yang lebih baik berdasarkan data. Salah satu keunggulan Google Data Studio adalah mudah digunakan bahkan bagi mereka yang sebelumnya belum pernah mengelola data. Selain itu, Google Data Studio terintegrasi dengan Google Analytics dan bersifat *open-source*.
2. *Tableau* adalah alat visualisasi data lainnya, yang dapat digunakan untuk analisis dan visualisasi data sekaligus. Anda dapat membuat grafik sederhana hingga visualisasi kreatif yang interaktif, semuanya tanpa harus menulis kode atau sintaks. Tableau juga kompatibel dengan banyak sumber data.

# BAB 13 KETERBATASAN DATA SCIENCE

## Pendahuluan

Data science telah merevolusi cara membuat keputusan, dengan kemampuannya untuk mengekstrak wawasan berharga dari data dan menghasilkan prediksi yang kuat (Mr. Ramkumar A, 2023). Namun, penting untuk menyadari bahwa data science bukanlah solusi sempurna dan memiliki beberapa keterbatasan yang perlu dipertimbangkan dengan cermat untuk memastikan penerapannya yang bertanggung jawab.

Salah satu keterbatasan data science adalah ketergantungannya pada kualitas data. Model data science dilatih menggunakan kumpulan data yang besar, dan akurasi hasil sangat ditentukan oleh kualitas data tersebut. Data yang tidak lengkap, bias, atau tidak akurat dapat menghasilkan prediksi yang tidak dapat diandalkan, berpotensi menimbulkan konsekuensi serius dalam situasi di mana keputusan penting dibuat berdasarkan hasil data science.

Kompleksitas algoritma data science juga menjadi tantangan. Algoritma ini bisa sangat rumit dan sulit dipahami, bahkan bagi para ahli sekalipun. Hal ini dapat mempersulit interpretasi hasil dan identifikasi potensi kesalahan atau bias dalam model. Lebih lanjut, data science terbatas pada pola yang ada dalam data. Model data science tidak dapat mengungkapkan informasi baru atau membuat prediksi di luar jangkauan data yang digunakan untuk melatihnya. Artinya, data science tidak dapat digunakan untuk menjawab pertanyaan yang belum pernah diamati sebelumnya atau memprediksi kejadian yang

tidak terduga.

Meskipun memiliki keterbatasan, data science tetap menjadi alat yang berharga dengan potensi luar biasa. Dengan memahami keterbatasannya dan menggunakannya dengan hati-hati, kita dapat memanfaatkan kekuatan data science untuk mendapatkan wawasan yang berharga, meningkatkan pengambilan keputusan, dan mendorong inovasi. Penting untuk selalu menggabungkan hasil data science dengan keahlian dan penilaian manusia untuk memastikan interpretasi dan aplikasi yang tepat.

## Keterbatasan Data Science

Data science memiliki beberapa keterbatasan. Berikut adalah beberapa kendala utama yang perlu diperhatikan:

1. *Data Quality Issues*
   Data science menganut prinsip "masuk sampah, keluar sampah" - kualitas hasil bergantung pada kualitas data yang dimasukkan (Irfan Whyudi, 2019). Dalam aplikasi dunia nyata, data bisa jadi berantakan, tidak lengkap, atau bias, yang mengarah ke model yang menyesatkan atau tidak akurat.
2. *Limited Scope of Inquiry*
   Data science hebat dalam menemukan pola dalam data, tetapi tidak dapat menjawab mengapa pola tersebut ada. Ia tidak dapat menjelaskan faktor eksternal atau perilaku manusia, yang bisa menjadi krusial untuk memahami situasi kompleks.
3. *Domain Knowledge Gap*
   Data scientists perlu memahami konteks spesifik dari masalah yang mereka coba selesaikan. Tanpa pemahaman yang kuat tentang bidangnya, akan sulit untuk

menginterpretasikan hasil secara akurat atau mengajukan pertanyaan yang tepat pada data.

4. *Ethical Concerns*
   Privasi data menjadi perhatian utama. Proyek data science dapat menimbulkan masalah seputar bagaimana data dikumpulkan, disimpan, dan digunakan. Penting untuk memastikan praktik data yang bertanggung jawab selama proses berlangsung.
5. *Misinterpretation of Results*
   Hanya karena model menghasilkan suatu hasil, tidak selalu berarti itu benar. Korelasi tidak sama dengan kausalitas, dan temuan Data science perlu diperiksa dan divalidasi dengan hati-hati oleh pakar.

### Masalah Kualitas Data

Kualitas data merupakan fondasi utama dalam Data science. Data yang berkualitas rendah dapat menghasilkan model yang tidak akurat, menyesatkan, dan bahkan berbahaya. Berikut adalah beberapa aspek penting dari masalah kualitas data:

1. Akurasi
   Tingkat kesesuaian data dengan kenyataan. Contoh: Rekam medis yang salah mencatat dosis obat pasien, data sensor yang terkalibrasi dengan tidak benar, hasil survei yang dipengaruhi oleh bias responden.
2. Kelengkapan
   Tingkat tersedianya semua data yang diperlukan untuk analisis. Contoh: Catatan pelanggan yang tidak memiliki informasi alamat; data pasar saham yang tidak memiliki data harga untuk periode tertentu
3. Konsistensi

Tingkat kesamaan data antar sumber atau waktu. Contoh: Format data yang berbeda antar departemen dalam suatu organisasi; inkonsistensi dalam penamaan file atau folder.

4. Ketepatan waktu
   Tingkat relevansi data dengan waktu saat ini. Contoh: Data harga saham yang tidak diperbarui secara *real-time*, informasi lalu lintas yang tidak mencerminkan kondisi terkini.

Berbagai teknik dapat digunakan untuk mengatasi masalah kualitas data, seperti (Santoso, 2023):

1. Pembersihan data
   Mengidentifikasi dan memperbaiki kesalahan, seperti nilai yang hilang atau tidak konsisten.
2. Integrasi data
   Menggabungkan data dari berbagai sumber dan memastikan konsistensi.
3. Transformasi data
   Mengubah format data agar sesuai dengan kebutuhan analisis.
4. Validasi data
   Memverifikasi keabsahan dan keandalan data.

### Ruang Lingkup Inquiry Terbatas

Meskipun data science menawarkan kemampuan luar biasa untuk mengungkap pola dan wawasan dari data, ia memiliki keterbatasan dalam hal ruang lingkup inquiry. Data science tidak dapat menjawab pertanyaan "mengapa" di balik pola yang ditemukannya, dan tidak dapat menjelaskan faktor eksternal atau perilaku manusia yang kompleks.

1. Ketidakmampuan Menjelaskan Hubungan Kausalitas
   Kemampuan untuk menentukan apakah satu peristiwa

menyebabkan peristiwa lain. Contoh: Model data science dapat menunjukkan korelasi antara penggunaan media sosial dan depresi, tetapi tidak dapat menjelaskan apakah media sosial menyebabkan depresi atau sebaliknya.

2. Keterbatasan dalam Memahami Konteks
   Kemampuan untuk memahami situasi atau peristiwa secara menyeluruh, termasuk faktor-faktor yang tidak dapat diukur secara kuantitatif. Contoh: Algoritma rekomendasi produk mungkin menyarankan item yang sering dibeli bersama dengan produk yang dipilih pengguna, tetapi tidak dapat memahami alasan di balik pembelian tersebut, seperti preferensi pribadi, pengaruh sosial, atau faktor ekonomi.
3. Ketidakmampuan untuk Memperhitungkan Perilaku Manusia
   Memahami dan memprediksi bagaimana manusia akan berperilaku dalam situasi tertentu. Contoh: Model prediksi Churn pelanggan mungkin gagal memperhitungkan faktor-faktor kualitatif seperti kepuasan pelanggan atau loyalitas merek, yang dapat memengaruhi keputusan mereka untuk tetap menggunakan layanan.

Mengatasi keterbatasan ruang lingkup *inquiry*:

1. Menggabungkan data kualitatif data kuantitatif dengan data kualitatif
2. Berkolaborasi dengan pakar di bidang yang relevan untuk mendapatkan wawasan dan interpretasi atas hasil data science.
3. Mempertimbangkan implikasi etis dari penggunaan data science dan potensi bias dalam data dan algoritma.

## Kesenjangan Pengetahuan Domain dalam Keterbatasan Data Science

Data science, meskipun kuat, memiliki tantangan signifikan yang disebut Domain Knowledge Gap. Data scientist membutuhkan pemahaman mendalam tentang bidang atau masalah yang mereka coba selesaikan untuk menginterpretasikan hasil data secara akurat dan mengajukan pertanyaan yang tepat (Science, 2020) (Suharto, 2023).

1. Kesalahpahaman Terminologi dan Konsep
   Kurangnya pemahaman tentang istilah dan konsep yang spesifik untuk bidang tertentu. Contoh: Data scientist yang tidak memiliki latar belakang medis mungkin salah mengartikan hasil analisis data kesehatan.
2. Ketidakmampuan Mengidentifikasi Faktor Penting
   Contoh: Ilmuwan data yang menganalisis data kriminalitas tanpa memahami faktor-faktor sosial dan ekonomi yang mendasarinya mungkin gagal mengidentifikasi penyebab sebenarnya dari tingkat kejahatan yang tinggi.
3. Kesalahan Interpretasi Hasil
   Menarik kesimpulan yang salah dari hasil analisis data karena kurangnya pengetahuan tentang konteks dan implikasi. Contoh: Ilmuwan data yang menafsirkan korelasi antara penggunaan media sosial dan depresi sebagai bukti kausalitas tanpa mempertimbangkan faktor lain mungkin memberikan saran yang keliru untuk mengatasi depresi.

   Mengatasi Kesenjangan Pengetahuan Domain:

1. Bekerja sama dengan Pakar Domain
   Berkolaborasi dengan pakar di bidang yang relevan untuk mendapatkan wawasan dan interpretasi atas hasil data

science.

2. Memahami Konteks Bisnis
   Memahami tujuan dan strategi bisnis untuk memastikan bahwa proyek data science selaras dengan kebutuhan.
3. Belajar Secara Berkelanjutan
   Terus belajar tentang tren dan perkembangan terbaru dalam bidang yang relevan untuk meningkatkan pemahaman domain.

### Masalah Etika dalam Keterbatasan Data Science

Data science memiliki banyak potensi untuk membantu kita menyelesaikan berbagai masalah dan membuat keputusan yang lebih baik. Namun, penting untuk diingat bahwa data science juga memiliki keterbatasan dan dapat menimbulkan masalah etika jika tidak digunakan dengan hati-hati.

Berikut adalah beberapa masalah etika utama dalam keterbatasan data science:

1. Bias
   Data science dapat mencerminkan dan memperkuat bias yang ada dalam data yang digunakan untuk melatih model. Hal ini dapat berakibat pada diskriminasi terhadap kelompok orang tertentu.
2. Privasi
   Data science sering kali melibatkan pengumpulan dan analisis data pribadi. Hal ini menimbulkan kekhawatiran tentang privasi dan bagaimana data tersebut digunakan (Dr. Taufik Hanafi, 2021). Contoh: Perusahaan mungkin menggunakan data science untuk melacak aktivitas online pelanggan mereka tanpa persetujuan mereka.
3. Keamanan
   Model data science dapat diretas atau disalahgunakan

untuk tujuan jahat. Hal ini dapat berakibat pada pencurian identitas, penipuan, atau bahkan kekerasan. Contoh: Peretas dapat menggunakan model data science untuk menargetkan individu dengan serangan phishing atau penipuan.

4. Akuntabilitas
   Model data science bisa sangat kompleks dan sulit dipahami. Hal ini dapat membuat sulit untuk mengetahui siapa yang bertanggung jawab atas hasil model tersebut.

Penting untuk menyadari masalah etika yang terkait dengan data science dan mengambil langkah-langkah untuk memitigasinya. Berikut adalah beberapa cara untuk melakukannya:

1. Gunakan data yang berkualitas tinggi dan beragam
   Data yang digunakan untuk melatih model data science harus berkualitas tinggi dan beragam. Hal ini akan membantu memastikan bahwa model tersebut akurat dan tidak bias.
2. Lindungi privasi
   Penting untuk melindungi privasi individu saat mengumpulkan dan menganalisis data. Ini termasuk mendapatkan persetujuan dari individu sebelum mengumpulkan data mereka dan mengambil langkah-langkah untuk mengamankan data.
3. Buat model yang dapat dijelaskan
   Model data science harus dapat dijelaskan, sehingga orang dapat memahami bagaimana mereka bekerja dan mengapa mereka menghasilkan hasil tertentu.
4. Gunakan data science secara bertanggung jawab
   Penting untuk menggunakan data science secara bertanggung jawab dan etis. Ini berarti mempertimbang-

kan potensi dampak dari model data science dan mengambil langkah-langkah untuk memitigasi risiko.

### Kesalahpahaman Hasil

Meskipun data science menawarkan kemampuan untuk menghasilkan wawasan dan prediksi yang berharga, terdapat risiko misinterpretasi hasil analisis yang dapat berakibat fatal. Kesalahpahaman hasil dapat terjadi karena berbagai faktor, seperti:

1. Kurangnya Pengetahuan Kontekstual
   Ketidakmampuan untuk memahami konteks di balik data dan implikasi dari hasil analisis. Contoh: Model data science yang menunjukkan korelasi antara tingkat pendidikan dan pendapatan dapat disalahartikan sebagai bukti bahwa pendidikan secara langsung menyebabkan peningkatan pendapatan, padahal faktor lain seperti latar belakang keluarga dan akses ke peluang kerja juga dapat berperan.
2. Ketidakmampuan Membedakan Korelasi dan Kausalitas
   Menyamakan korelasi antara dua variabel dengan hubungan sebab-akibat. Contoh: Analisis data yang menunjukkan hubungan antara penggunaan media sosial dan depresi tidak dapat secara langsung menyimpulkan bahwa media sosial menyebabkan depresi. Faktor lain seperti masalah kesehatan mental yang sudah ada, cyberbullying, atau isolasi sosial juga dapat berkontribusi pada depresi.
3. Ketidakpastian Model
   Ketidakmampuan untuk memahami tingkat kepercayaan dan batas-batas model data science. Contoh: Model prediksi harga saham yang memiliki tingkat akurasi tinggi

dalam data pelatihan mungkin tidak akurat dalam data baru karena fluktuasi pasar yang tidak terduga atau faktor lain yang tidak terukur.

4. Bias Algoritmik
   Keberpihakan yang tidak disengaja dalam algoritma data science yang dapat memengaruhi hasil analisis. Contoh: Algoritma perekrutan yang dilatih pada data historis yang bias terhadap kelompok demografis tertentu mungkin secara tidak proporsional menyaring kandidat dari kelompok tersebut.

Mencegah Kesalahpahaman Hasil:

1. Memahami Konteks
   Selalu mempertimbangkan konteks di balik data dan implikasi dari hasil analisis.
2. Mencari Hubungan Kausalitas
   Berhati-hatilah dalam menafsirkan korelasi sebagai hubungan sebab-akibat. Gunakan metode analisis yang tepat dan pertimbangkan faktor-faktor lain yang mungkin berkontribusi.
3. Mengevaluasi Model
   Pahami tingkat kepercayaan dan batas-batas model data science. Lakukan validasi dan pengujian untuk memastikan model akurat dan generalisabel.
4. Mempertimbangkan Bias
   Periksa potensi bias dalam data dan algoritma. Gunakan teknik untuk mengurangi bias dan meningkatkan keadilan algoritmik.
5. Berkolaborasi dengan Pakar
   Bekerja sama dengan pakar domain dan ahli statistik untuk menginterpretasikan hasil data science secara akurat dan bertanggung jawab.

# BAB 14 MODEL REGRESI

## Pendahuluan

Bertahun-tahun yang lalu, Francois Dalton adalah pionir analisis regresi, yang memanfaatkannya sebagai alat untuk memahami transmisi sifat dari satu generasi ke generasi berikutnya. Di bidang statistik, analisis regresi bertujuan untuk menguji korelasi antara dua variabel. Representasi matematis dari hubungan ini disebut sebagai persamaan regresi.

Analisis regresi, sebuah metode statistik, memainkan peran penting dalam penelitian dengan membantu dalam prediksi berbagai fenomena. Metode ini mengkaji pola logis yang ada antara dua variabel atau lebih, dengan satu variabel berperan sebagai variabel terikat dan variabel lainnya sebagai variabel bebas. Nilai variabel terikat dipengaruhi oleh variabel lain sehingga dapat diprediksi atau dijelaskan, sedangkan nilai variabel bebas tidak dipengaruhi oleh variabel lain dan digunakan untuk memprediksi atau menjelaskan variabel lain. Dengan menganalisis data masa lalu dan masa kini, analisis regresi memungkinkan kita menentukan bagaimana variabel terikat dapat diramalkan berdasarkan masing-masing variabel bebas, sehingga meminimalkan kesalahan prediksi. Hasil yang diperoleh dari analisis regresi memberikan wawasan apakah perubahan variabel dependen dapat dicapai dengan memanipulasi variabel independen. Lebih lanjut, analisis ini membantu menetapkan persamaan atau garis matematis yang mewakili hubungan fungsional antar variabel.

## Pengertian Regresi

Tindakan regresi melibatkan penggunaan data historis dan data terkini untuk membuat estimasi yang tepat mengenai kemungkinan kejadian di masa depan, semua dengan tujuan mengurangi ketidakakuratan. Selain itu, regresi juga dapat dipandang sebagai upaya mengantisipasi perubahan. Meskipun peramalan tidak memberikan hasil yang pasti mengenai peristiwa-peristiwa yang akan terjadi, peramalan berusaha untuk menghasilkan perkiraan tentang apa yang berpotensi terjadi di masa depan. Oleh karena itu, regresi berupaya untuk memahami potensi keadaan di masa depan untuk memfasilitasi pengambilan keputusan yang optimal.

Analisis regresi berfungsi sebagai alat yang berharga dalam penelitian, khususnya dalam memprediksi variabel terikat (Y) berdasarkan variabel bebas (X) yang diketahui. Saat mengkaji hubungan antara beberapa variabel, ada dua aspek utama yang perlu dipertimbangkan: bentuk hubungan dan keeratan hubungan. “Analisis regresi memungkinkan kita menentukan bentuk hubungan, sedangkan analisis korelasi membantu menentukan derajat kedekatan. Pendekatan analitis ini terbukti sangat berguna ketika mengeksplorasi fenomena kompleks, dimana modelnya tidak sepenuhnya dipahami atau ketika menyelidiki bagaimana variasi variabel independen berdampak pada variabel dependen. Dalam hal terdapat beberapa variabel bebas (X1, X2, …, Xn) dan satu variabel terikat (Y), hubungan fungsionalnya dapat dinyatakan sebagai berikut: Y = f(X1, X2, …, Xn, e ), dimana Y melambangkan variabel terikat, X melambangkan variabel bebas, dan e melambangkan variabel sisa (istilah gangguan). M. Nazir (1983) menguraikan empat kegiatan yang dapat dilakukan dalam analisis regresi:

1. Memanfaatkan data observasi untuk menghitung nilai

parameter.
2. Menilai sejauh mana fluktuasi variabel independen menjelaskan varians yang diamati pada variabel dependen.
3. Menilai pentingnya estimasi parameter.
4. Pastikan parameter yang diestimasi selaras dengan teori dengan memverifikasi tanda dan besarannya.

## Regresi Linear Sederhana

Analisis regresi adalah teknik yang digunakan untuk mengeksplorasi hubungan antara variabel prediktor dan variabel hasil, dengan menjelaskan sifat hubungan keduanya. Secara spesifik, regresi linier berfokus pada pengujian hubungan antara variabel terikat dan bebas yang dapat diwakili oleh garis lurus (Yasril, dkk: 2009).

Anggaplah analisis regresi sebagai instrumen mistik yang membantu memahami hubungan antara dua variabel. Ambil contoh, dampak berat badan terhadap tekanan darah seseorang. Melalui perhitungan matematis yang rumit, kita dapat memperkirakan tekanan darah seseorang dengan mempertimbangkan berat badannya. Untuk membuat prediksi yang tepat, peneliti sering kali menggunakan metode kuadrat terkecil untuk membuat garis yang paling sesuai dengan data.

$$Y = a + bx$$

Analisis persamaan tersebut di atas merupakan model deterministik yang hanya berlaku pada fenomena alam. Ilustrasi utama dari hal ini adalah hukum gravitasi Isaac Newton, yang beroperasi sebagai model deterministik. Dalam skenario ideal, laju jatuhnya suatu benda (variabel terikat) merupakan fungsi matematis sempurna dari variabel bebas seperti disparitas berat

dan gaya gravitasi. Demikian pula korelasi antara suhu Fahrenheit dan Celsius dapat direpresentasikan dengan persamaan Y = 32 + 9/5X. Jika suhu dalam Celsius (X) diketahui, suhu dalam Fahrenheit (Y) dapat dihitung atau diprediksi dengan sempurna tanpa kesalahan apa pun.

Dalam ranah ilmu-ilmu sosial sering dijumpai kesalahan atau penyimpangan dalam hubungan antar variabel. Hal ini menyiratkan bahwa nilai Y yang berbeda dapat diperoleh untuk beberapa nilai X yang sama. Misalnya, ketika memeriksa hubungan antara berat badan dan tekanan darah, menjadi jelas bahwa individu dengan berat badan yang sama belum tentu memiliki tekanan darah yang sama. Karena sifat hubungan antara X dan Y dalam ilmu sosial atau kesehatan masyarakat yang tidak tepat, persamaan garis yang dihasilkan mengambil bentuk yang berbeda:

$$Y = a + bx + e$$

Mari kita uraikan persamaan Model Regresi Linier Sederhana, yang membantu kita menguji hubungan linier antara dua variabel. Dalam model ini, kita memiliki variabel terikat (ý), yang mewakili nilai prediksi, dan variabel bebas (x). Intersep (a) adalah selisih ukuran rata-rata variabel terikat pada saat variabel bebas berada pada angka nol. Koefisien regresi (b) memperkirakan perbedaan antara nilai variabel terikat individu yang diamati dan nilai sebenarnya pada titik tertentu pada variabel bebas. Baik a maupun ß merupakan parameter yang tidak diketahui yang kami perkirakan menggunakan statistik sampel.

$$a = \frac{\sum Y - b\sum X}{N} = \overline{Y} - b\overline{X}$$

$$b = \frac{N \cdot (\sum XY) - \sum X \sum Y}{N \cdot \sum X^2 - (\sum \cdot X)^2}$$

Contoh -1:

Tujuan penelitian ini adalah untuk mengetahui hubungan antara pendapatan keluarga (X) dan pengeluaran konsumsi (Y). Untuk mencapai hal ini, sampel 10 keluarga dipilih secara acak untuk wawancara, dan data yang dikumpulkan dari penelitian menghasilkan informasi berikut.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Konsumsi (y) | 5 | 6 | 8 | 9 | 10 | 12 | 12 | 14 | 15 | 20 |
| Pendapatan (x) | 6 | 8 | 10 | 12 | 13 | 17 | 20 | 22 | 24 | 28 |

Dengan mempertimbangkan data yang diberikan:

1. Hitung persamaan yang memperkirakan regresi populasi.
2. Berikan penjelasan tentang nilai b yang diperoleh.

Kira-kira perkiraan pengeluaran rata-rata sebuah keluarga dengan pendapatan 18

Penyelesaian:

Persamaan regresi populasi akan diduga dengan persamaan regresi sampelnya:

| $X_i$ | $Y_i$ | $X_i^2$ | $Y_i$ | $x_i y_i$ |
|---|---|---|---|---|
| 6 | 5 | 36 | 25 | 30 |
| 8 | 6 | 64 | 36 | 48 |
| 10 | 8 | 100 | 64 | 80 |
| 12 | 9 | 144 | 81 | 108 |
| 13 | 10 | 169 | 100 | 130 |
| 17 | 12 | 289 | 144 | 204 |
| 20 | 12 | 400 | 144 | 240 |
| 22 | 14 | 484 | 196 | 308 |
| 24 | 15 | 576 | 225 | 360 |
| 28 | 20 | 784 | 400 | 560 |
| Σ = 160 | Σ = 111 | Σ = 3046 | Σ = 1415 | Σ = 2068 |

n = 10

$$\overline{X} = \frac{\sum X_i}{n} = \frac{160}{10} = 16$$

$$\overline{Y} = \frac{\sum Y_i}{n} = \frac{111}{10} = 11,1$$

$$b = \frac{N.(\Sigma xy) - \Sigma x \Sigma y}{.N.\Sigma x^2 - (\Sigma x)^2 -} = \frac{10.(2068) - (160)(111)}{.10(3046) - (160)^2} = 0,60$$

$$a = \overline{Y} - b\overline{X} = 11,1 - 0,60(16) = 1,50$$

Estimator regresi populasi yang diwakili oleh persamaan regresi sampel adalah ŷ = a + bx = 1,5 + 0,6x. Dalam persamaan ini, nilai b yaitu 0,6 menunjukkan bahwa untuk setiap kenaikan satu satuan pendapatan maka akan terjadi kenaikan rata-rata pengeluaran konsumsi sebesar 0,6 satuan.

Jika X sama dengan 18, perhitungan Y adalah sebagai berikut: ŷ sama dengan 1,5 ditambah 0,6 dikali X, yang berarti 1,5 ditambah 0,6 dikalikan 18, sehingga menghasilkan 12,3. Oleh karena itu, jika pendapatan keluarga adalah 18, maka perkiraan pengeluaran konsumsi rata-rata adalah 12,3.

## Regresi Linear Berganda

Analisis statistik yang disebut Analisis Regresi Linier Berganda digunakan untuk menguji hubungan antara beberapa variabel bebas atau variabel penduga dan satu variabel terikat. Tujuannya adalah untuk menilai bagaimana dua atau lebih variabel penjelas berdampak pada satu variabel hasil. Model ini beroperasi dengan asumsi adanya hubungan linier antara variabel terikat dan masing-masing prediktor, yang biasanya direpresentasikan dalam bentuk rumus. Dalam kasus khusus yang disebutkan, rumus yang dihasilkan adalah sebagai berikut:

$$Y = a + \beta_1 X1 + B2X2 + B3X3 + B4X4 + B5X5 + \varepsilon_1$$

Di mana:

| | | |
|---|---|---|
| Y | = | Kinerja keuangan /ROA sebagai variabel dependen |
| A | = | Konstanta |
| β1 | = | Koefisien regresi variabel independen |
| X1 | = | Struktur permodalan/CAR sebagai variabel independen |
| X2 | = | Kualitas aset produktif/NPL sebagai variabel |
| X3 | = | Independen Rentabilitas/ROE sebagai variabel independen |
| X4 | = | Efisiensi biaya/OCOR sebagai variabel independen |
| X5 | = | Likuiditas/LDR sebagai variabel independen |

Analisis regresi berganda merupakan lanjutan dari analisis regresi sederhana yang digunakan untuk meramalkan nilai variabel terikat (Y) dalam situasi dimana terdapat dua atau lebih variabel bebas. Tujuan dari analisis regresi berganda ada dua: untuk mengukur dampak dua atau lebih variabel independen terhadap satu variabel dependen, dan untuk memastikan apakah terdapat hubungan fungsional atau sebab akibat di antara beberapa variabel independen (X1, X2, X3, dll.) dan satu variabel terikat (Y). Rumusan persamaan regresi berganda adalah sebagai berikut.

Dua variabel bebas : $\hat{y} = a + b1x1 + b2x2$
Tiga variabel bebas : $\hat{y} = a + b1x1 + b2x2 + b3x3$
Empat variabel bebas : $\hat{y} = a + b1x1 + b2x3 + b3x3 + b4x4$
n variabel bebas : $\hat{y} = a + b1x1 + b2x2 + \ldots\ldots + bnxn$

Untuk menentukan nilai kedua variabel independen dalam persamaan regresi berganda dapat digunakan pendekatan berikut.

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$a = \frac{\sum Y}{n} - b_1\left(\frac{\sum X_1}{n}\right) - b_2\left(\frac{\sum X_2}{n}\right)$$

Persamaan model regresi linier berganda mendefinisikan hubungan antara variabel terikat/respon (Y) dan dua atau lebih variabel/prediktor bebas (X1, X2,...Xn). Tujuan utama analisis regresi linier berganda adalah untuk meramalkan nilai variabel terikat/respons (Y) dengan mengetahui nilai variabel/prediktor bebas (X1, X2,...Xn) yang diketahui. Selain itu analisis bertujuan untuk mengetahui arah hubungan antara variabel terikat dengan variabel bebas. Secara matematis, persamaan regresi linier berganda dinyatakan sebagai:

$$Y = a + b1X1 + b2X2 + \ldots + bnXn$$

Yang mana:

Y = variable tak bebas (nilai yang akan diprediksi)
a = konstanta
b1, b2,.., bn = koefisien regresi
X1, X2,..., Xn = variable bebas

Jika terdapat dua variabel bebas, X1 dan X2, maka persamaan regresinya berbentuk sebagai berikut.

$$Y = a + b1X1 + b2X2$$

Apabila nilai koefisien regresi b1 dan b2 adalah sebagai berikut: jika bernilai 0, maka tidak terdapat pengaruh X1 dan X2 terhadap Y; jika bernilai negatif maka terdapat korelasi terbalik antara variabel bebas X1 dan X2 dengan variabel terikat Y; jika bernilai positif, maka terdapat hubungan satu arah antara

variabel bebas X1 dan X2 dengan variabel terikat Y. Konstanta a dan koefisien regresi b1 dan b2 dapat dihitung dengan menggunakan rumus yang diberikan.

$$a = \frac{(\sum Y) - (b_1 \times \sum x_1) - (b_2 \times \sum x_2)}{n}$$

$$b_1 = \frac{[(\sum x_2{}^2 \times \sum x_1 y) - (\sum x_2 y \times \sum x_1 x_2)]}{[(\sum x_1{}^2 \times \sum x_2{}^2) - (\sum x_1 \times x_2)^2]}$$

$$b_2 = \frac{[(\sum x_1{}^2 \times \sum x_2 y) - (\sum x_1 y \times \sum x_1 x_2)]}{[(\sum x_1{}^2 \times \sum x_2{}^2) - (\sum x_1 \times x_2)^2]}$$

Yang mana:

$$\sum x_1^2 = \sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n}$$

$$\sum x_2^2 = \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n}$$

$$\sum y^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$\sum x_1 y = \sum X_1 Y - \frac{\sum X_1 \sum Y}{n}$$

$$\sum x_2 y = \sum X_2 Y - \frac{\sum X_2 \sum Y}{n}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n}$$

Dengan menyusun suatu persamaan, seseorang dapat menentukan nilai a, b1, dan b2 menggunakan metode matriks atau disebut juga metode kuadrat terkecil sebagai pendekatan alternatif:

$$
\begin{aligned}
an \quad &+b_1\sum X_1 \quad &+b_2\sum X_2 \quad &= \sum Y \\
a\sum X_1 + b_1\sum X_1^{\,2} \quad &+b_2\sum X_1X_2 \quad &= \sum X_1Y \\
a\sum X_2 + b_1\sum X_2X_1 \quad &+b_2\sum X_2^{\,2} \quad &= \sum X_2Y
\end{aligned}
$$

Matriks dengan 3 persamaan 3 variabel:

$$
\begin{aligned}
m_{11}a + m_{12}b_1 + m_{13}b_2 &= h_1 \\
m_{21}a + m_{22}b_1 + m_{23}b_2 &= h_2 \\
m_{31}a + m_{32}b_1 + m_{33}b_2 &= h_3
\end{aligned}
$$

$$
\begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}
\begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} =
\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}
$$

$$
a = \frac{\det M_1}{\det M} \qquad b_1 = \frac{\det M_2}{\det M} \qquad b_2 = \frac{\det M_3}{\det M}
$$

$$
M_1 = \begin{bmatrix} h_1 & m_{12} & m_{13} \\ h_2 & m_{22} & m_{23} \\ h_3 & m_{32} & m_{33} \end{bmatrix}
$$

$$
M_2 = \begin{bmatrix} m_{11} & h_1 & m_{13} \\ m_{21} & h_2 & m_{23} \\ m_{31} & h_3 & m_{33} \end{bmatrix}
$$

Contoh:

$$
\left.\begin{aligned} 2a + b_1 + 4b_2 &= 16 \\ 3a + 2b_1 + b_2 &= 10 \\ a + 3b_1 + 3b_2 &= 16 \end{aligned}\right\} \rightarrow
\begin{bmatrix} 2 & 1 & 4 \\ 3 & 2 & 1 \\ 1 & 3 & 3 \end{bmatrix}
\begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} =
\begin{bmatrix} 16 \\ 10 \\ 16 \end{bmatrix} \Rightarrow M =
\begin{bmatrix} 2 & 1 & 4 \\ 3 & 2 & 1 \\ 1 & 3 & 3 \end{bmatrix}
$$

$$
M_1 = \begin{bmatrix} 16 & 1 & 4 \\ 10 & 2 & 1 \\ 16 & 3 & 3 \end{bmatrix} \quad
M_2 = \begin{bmatrix} 2 & 16 & 4 \\ 3 & 10 & 1 \\ 1 & 16 & 3 \end{bmatrix} \quad
M_3 = \begin{bmatrix} 2 & 1 & 16 \\ 3 & 2 & 10 \\ 1 & 3 & 16 \end{bmatrix}
$$

Nilai a, b1 dan b2 diperoleh dari determinan, yaitu:

$$a = \frac{\det M_1}{\det M} = \frac{26}{26} = 1; \quad b_1 = \frac{\det M_2}{\det M} = \frac{52}{26} = 2; \quad b_2 = \frac{\det M_3}{\det M} = \frac{78}{26} = 3$$

Koefisien determinasi (r2) adalah alat yang berguna untuk menilai dampak variabel independen X1 dan X2 terhadap variabel dependen Y. Koefisien ini mengukur persentase pengaruh variabel-variabel tersebut, dan dapat dihitung menggunakan rumus berikut

$$r\ 2 = ((b1 \sum x1\ y) + (b2 \sum x2y))/(\sum y\ 2)$$

Jika koefisien determinasi r2 sama dengan 0, berarti fluktuasi variabel independen X1 dan X2 tidak mampu menjelaskan varians yang diamati pada variabel dependen Y dalam model persamaan regresi.

Sebaliknya, jika r2 sama dengan 1, hal ini menunjukkan bahwa variasi X1 dan X2 dapat dengan sempurna menjelaskan variabel dependen Y dalam model persamaan regresi.

## Koefisien Korelasi Ganda (r)

1. Untuk mengetahui tingkat korelasi antara variabel X1, X2, ..., Xn dan variabel Y secara bersamaan digunakan koefisien korelasi berganda. Rumus untuk menghitung koefisien korelasi berganda memungkinkan penentuan nilai spesifiknya.
2. Nilai R yang berkisar antara -1 hingga +1 merupakan ukuran kekuatan suatu hubungan. Jika nilainya mendekati +1 atau -1, maka hubungannya semakin kuat. Sebaliknya, ketika nilainya mendekati 0 maka hubungannya menjadi semakin lemah.

## Korelasi Parsial

Korelasi parsial mengacu pada suatu bentuk korelasi yang

menjelaskan hubungan antara satu variabel dengan variabel lain dengan tetap menjaga variabel lainnya tetap konstan. Jenis korelasi ini dapat diklasifikasikan menjadi tiga kategori berbeda:

1. Korelasi antara X1 dan X2 dengan tetap mempertahankan Y konstan (r12.Y).

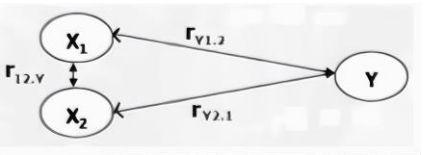
$$r_{12.Y} = \frac{r_{12} - (r_{Y1} r_{Y2})}{\sqrt{(1 - r_{Y1}^2)(1 - r_{Y2}^2)}}$$

2. Korelasi antara Y dengan X1 yang mana X2 dianggap konstan (rY1.2).

$$r_{Y1.2} = \frac{r_{Y1} - (r_{Y2} r_{12})}{\sqrt{(1 - r_{Y2}^2)(1 - r_{12}^2)}}$$

3. Korelasi antara Y dengan X2 yang mana X1 dianggap konstan (rY2.1).

$$r_{Y2.1} = \frac{r_{Y2} - (r_{Y1} r_{12})}{\sqrt{(1 - r_{Y1}^2)(1 - r_{12}^2)}}$$




Yang mana:

$$r_{Y1} = \frac{n \times \sum X_1 Y - (\sum Y \times \sum X_1)}{\sqrt{[(n \times \sum Y^2) - (\sum Y^2)] \times [(n \times \sum {X_1}^2) - (\sum X_1)^2]}}$$

$$r_{Y2} = \frac{n \times \sum X_2 Y - (\sum Y \times \sum X_2)}{\sqrt{[(n \times \sum Y^2) - (\sum Y^2)] \times [(n \times \sum {X_2}^2) - (\sum X_2)^2]}}$$

$$r_{12} = \frac{n \times \sum X_1 X_2 - (\sum X_1 \times \sum X_2)}{\sqrt{[(n \times \sum {X_1}^2) - (\sum {X_1}^2)] \times [(n \times \sum {X_2}^2) - (\sum X_2)^2]}}$$

### Kesalahan Baku Estimasi (*Standart Error Estimate*)

Kecukupan persamaan regresi dalam mengestimasi atau memprediksi variabel respon Y dapat ditentukan dengan mengevaluasi standar kesalahan estimasi. Kesalahan standar yang besar menunjukkan bahwa persamaan regresi yang dihasilkan tidak sesuai untuk tujuan estimasi. Hal ini disebabkan adanya disparitas yang signifikan antara nilai estimasi dan nilai aktual dari variabel respon Y. Ekspresi matematis untuk kesalahan standar estimasi tetap tidak berubah:

$$S_e(S_{yx}) = \sqrt{\frac{\sum Y^2 - (a\sum Y) - (b_1 \sum X_1 Y) - (b_2 \sum X_2 Y)}{N-3}}$$

# BAB 15 MODEL KLASIFIKASI

## Pendahuluan

Model klasifikasi adalah teknik yang sangat penting dalam data science, digunakan untuk memprediksi kategori atau kelas dari data baru berdasarkan data historis. Klasifikasi merupakan bentuk supervised learning dimana model dilatih menggunakan dataset yang berisi input dan label output. Proses ini memungkinkan model untuk belajar dari data historis dan membuat prediksi yang akurat pada data baru. Klasifikasi telah diterapkan dalam berbagai domain, mulai dari kesehatan, keuangan, hingga teknologi informasi, untuk membantu pengambilan keputusan yang berbasis data.

Penggunaan model klasifikasi semakin populer seiring dengan meningkatnya volume data yang dihasilkan setiap hari. Teknologi dan algoritma yang mendukung klasifikasi telah berkembang pesat, memungkinkan para ilmuwan data untuk menangani dataset yang besar dan komplek. Algoritma seperti *Logistic Regression, Decision Tree, Random Forest, Support Vector Machine* (SVM), dan *Neural Networks* menawarkan berbagai pendekatan untuk memecahkan masalah klasifikasi. Masing-masing algoritma memiliki kekuatan dan kelemahan tersendiri, yang membuat pemilihan model menjadi keputusan penting dalam setiap proyek data science.

Selain itu, evaluasi model klasifikasi merupakan langkah krusial untuk memastikan bahwa model yang dibangun memiliki performa yang baik dan dapat diandalkan. Berbagai metrik evaluasi seperti *accuracy, precision, recall, F1 score*, dan ROC-

AUC digunakan untuk mengukur kinerja model. Tantangan seperti *overfitting*, data yang tidak seimbang, dan pemilihan fitur juga harus diatasi untuk mengoptimalkan kinerja model klasifikasi. Dengan pemahaman yang mendalam tentang konsep, teknik, dan tantangan yang terkait, para ilmuwan data dapat membangun model klasifikasi yang efektif dan efisien untuk berbagai aplikasi praktis.

## Konsep Dasar Klasifikasi

Klasifikasi adalah teknik dalam *Machine Learning* yang bertujuan untuk memprediksi kategori atau kelas dari data baru berdasarkan data historis. Ini adalah bentuk supervised learning, di mana model dilatih menggunakan dataset yang berisi input (fitur) dan output (label). Proses klasifikasi melibatkan dua tahap utama: pelatihan (*training*) dan pengujian (*testing*). Pada tahap pelatihan, model belajar dari data yang sudah diberi label untuk mengidentifikasi pola dan hubungan antara fitur dan label. Setelah pelatihan, model diuji dengan data baru yang belum pernah dilihat sebelumnya untuk mengevaluasi seberapa baik model dapat menjeneralisasi pengetahuan yang dipelajari (Bishop, 2006).

Dalam evaluasi model klasifikasi, berbagai metrik digunakan untuk mengukur kinerjanya, seperti akurasi, precision, *recall*, dan F1 score. Akurasi mengukur persentase prediksi yang benar, precision mengukur proporsi prediksi positif yang benar, dan *recall* mengukur proporsi aktual positif yang berhasil diidentifikasi oleh model. *F1 score* adalah harmonic mean dari *precision* dan *recall*, memberikan keseimbangan antara keduanya. Tantangan utama dalam klasifikasi adalah menghindari overfitting, di mana model terlalu disesuaikan dengan data pelatihan dan berkinerja buruk pada

data baru, serta underfitting, di mana model terlalu sederhana untuk menangkap pola yang ada dalam data. Teknik seperti cross-validation dan regularisasi sering digunakan untuk mengatasi masalah ini dan meningkatkan kemampuan generalisasi model.

## Algoritma Klasifikasi Umum

Berbagai algoritma dapat digunakan untuk klasifikasi, tergantung pada jenis data dan masalah yang dihadapi. Beberapa algoritma klasifikasi yang umum meliputi:

### Logistic Regression

Regresi logistik adalah metode statistik yang sering digunakan untuk tugas klasifikasi biner, yaitu untuk memprediksi probabilitas bahwa suatu input masuk ke salah satu dari dua kategori yang mungkin. Metode ini berbeda dari regresi linier yang memprediksi nilai kontinu. Dalam regresi logistik, probabilitas kejadian diprediksi melalui penggunaan fungsi logistik, atau fungsi sigmoid, yang mengubah angka bernilai real menjadi nilai antara 0 dan 1. Hal ini membuat regresi logistik sangat cocok untuk masalah klasifikasi di mana hasil yang diinginkan adalah kategori biner seperti "ya" atau "tidak", "benar" atau "salah" (Pampel, 2000).

Fungsi logistik yang digunakan dalam regresi logistik adalah kurva berbentuk S yang memetakan nilai-nilai input menjadi probabilitas. Rumus fungsi logistik adalah:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

di mana z adalah kombinasi linier dari variabel-variabel independen yang diperoleh dari model regresi. Kombinasi linier ini biasanya melibatkan penjumlahan dari produk antara

koefisien regresi dengan variabel-variabel independen.

Salah satu kelebihan dari regresi logistik adalah kemampuannya untuk memberikan interpretasi yang jelas tentang pengaruh masing-masing variabel independen terhadap probabilitas hasil. Koefisien yang dihasilkan dari model regresi logistik menggambarkan log-odds perubahan dalam variabel dependen untuk setiap unit perubahan dalam variabel independen. Ini berarti bahwa dengan memeriksa nilai koefisien, kita dapat memahami seberapa besar pengaruh masing-masing faktor terhadap kemungkinan kejadian tertentu (Hosmer, dkk, 2013).

Selain itu, regresi logistik dapat diperluas untuk menangani kasus-kasus di mana variabel dependen memiliki lebih dari dua kategori. Hal ini dilakukan melalui pendekatan seperti regresi logistik multinomial dan regresi logistik ordinal. Regresi logistik multinomial digunakan ketika variabel dependen memiliki beberapa kategori yang tidak memiliki urutan tertentu, sementara regresi logistik ordinal digunakan ketika kategori tersebut memiliki urutan alami.

Regresi logistik banyak digunakan dalam berbagai bidang aplikasi praktis. Dalam bidang kesehatan, regresi logistik dapat digunakan untuk memprediksi kemungkinan seorang pasien menderita penyakit tertentu berdasarkan sejumlah faktor risiko. Dalam bidang keuangan, regresi logistik digunakan untuk mendeteksi aktivitas penipuan dengan menganalisis pola transaksi. Kegunaan lainnya termasuk analisis pemasaran, ilmu sosial, dan berbagai aplikasi lain di mana prediksi kategori biner diperlukan.

### *Decision Tree*

*Decision Tree* adalah salah satu algoritma klasifikasi yang

paling intuitif dan mudah dipahami. Algoritma ini bekerja dengan membagi dataset ke dalam subset yang lebih kecil dan berulang kali berdasarkan fitur yang paling informatif. Proses ini diwakili dalam bentuk struktur pohon, dimana setiap node internal mewakili pengujian terhadap suatu atribut, setiap cabang mewakili hasil pengujian, dan setiap daun mewakili label kelas atau nilai prediksi. *Decision Tree* sangat efektif karena kemampuannya untuk menangani data kategori maupun numerik, serta kemampuannya untuk menangani hubungan non-linear antara fitur.

Salah satu keunggulan utama dari *Decision Tree* adalah interpretabilitasnya. Struktur pohon yang dihasilkan mudah dipahami dan diinterpretasikan bahkan oleh non-ahli. Setiap cabang dalam pohon mewakili keputusan berdasarkan nilai atribut, dan setiap daun mewakili hasil prediksi atau kategori. Ini membuat *Decision Tree* sangat berguna dalam situasi di mana penting untuk dapat menjelaskan alasan di balik prediksi atau keputusan model kepada pemangku kepentingan (Hestie, dkk, 2009).

Namun, *Decision Tree* juga memiliki kelemahan, yaitu kecenderungannya untuk overfitting, terutama ketika pohon menjadi sangat dalam dan rumit. Overfitting terjadi ketika model menangkap terlalu banyak detail dari data pelatihan, termasuk noise, yang dapat mengurangi kinerjanya pada data baru yang tidak terlihat sebelumnya. Untuk mengatasi masalah ini, teknik seperti pemangkasan pohon (pruning), penetapan kedalaman maksimum pohon, atau menggunakan ensemble methods seperti *Random Forest* dapat diterapkan untuk meningkatkan generalisasi model (Kuhn & Johnson, 2013).

*Decision Tree* banyak digunakan di berbagai bidang seperti kedokteran, keuangan, dan pemasaran. Misalnya, dalam

bidang kedokteran, *Decision Tree* dapat digunakan untuk mengembangkan model prediktif yang membantu dokter dalam mendiagnosis penyakit berdasarkan gejala pasien. Dalam keuangan, mereka dapat membantu dalam mendeteksi penipuan dengan menganalisis pola transaksi. Kemampuan *Decision Tree* untuk menangani data dengan fitur numerik dan kategorikal serta menghasilkan model yang mudah diinterpretasikan menjadikannya alat yang sangat berharga dalam *Machine Learning*.

### *Random Forest*

*Random Forest* adalah metode ensemble yang terdiri dari banyak pohon keputusan yang dibangun selama pelatihan dan menghasilkan prediksi dengan menggabungkan output dari masing-masing pohon. Algoritma ini meningkatkan akurasi dan stabilitas prediksi dibandingkan dengan model pohon keputusan tunggal dengan mengurangi overfitting. *Random Forest* bekerja dengan memilih subset acak dari fitur dan data pada setiap langkah pembuatan pohon, yang menciptakan berbagai model pohon keputusan yang berbeda (Breiman, 2001).

Salah satu keunggulan utama dari *Random Forest* adalah kemampuannya untuk mengurangi overfitting yang sering terjadi pada pohon keputusan tunggal. Dengan menggabungkan prediksi dari banyak pohon yang tidak berkorelasi, *Random Forest* mampu menghasilkan model yang lebih kuat dan generalisasi yang lebih baik. Setiap pohon dalam hutan dibangun dari sampel acak dari data pelatihan, dan pada setiap split, hanya sejumlah fitur acak yang dipertimbangkan untuk pemisahan terbaik. Hal ini menghasilkan model yang lebih bervariasi dan robust terhadap data baru.

Selain itu, *Random Forest* juga memiliki kemampuan

untuk menangani data dengan banyak fitur dan data yang hilang dengan baik. Algoritma ini memberikan ukuran pentingnya fitur, yang memungkinkan kita untuk mengidentifikasi dan memahami kontribusi masing-masing fitur dalam model (Hastie, dkk, 2009).

Namun, salah satu kekurangan dari *Random Forest* adalah kompleksitas komputasinya yang lebih tinggi dibandingkan dengan model tunggal seperti *Decision Tree*. Meskipun paralelisasi dapat mengurangi waktu pelatihan, penggabungan banyak pohon keputusan memerlukan sumber daya komputasi yang lebih besar. Selain itu, interpretasi dari model *Random Forest* bisa menjadi lebih sulit dibandingkan dengan model yang lebih sederhana, meskipun fitur penting dapat diidentifikasi.

### *Support Vector Machine* (SVM)

*Support Vector Machine* (SVM) adalah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, tetapi lebih dikenal karena kinerjanya yang sangat baik dalam tugas klasifikasi biner. SVM bekerja dengan mencari hyperplane optimal yang memisahkan data dari dua kelas dengan margin terbesar. *Hyperplane* ini adalah garis atau permukaan dalam ruang multidimensi yang memisahkan data berdasarkan kelasnya. Dengan demikian, SVM bertujuan untuk menemukan batas keputusan yang memaksimalkan margin antara dua kelas yang berbeda, menjadikannya sangat efektif untuk masalah klasifikasi yang kompleks (Cortes & Vapnik, (1995).

Salah satu konsep utama dalam SVM adalah penggunaan "kernel trick," yang memungkinkan algoritma untuk bekerja dalam ruang fitur yang lebih tinggi tanpa harus secara eksplisit

menghitung koordinat dalam ruang tersebut. Ini sangat berguna ketika data tidak dapat dipisahkan secara linier dalam ruang aslinya. Fungsi kernel seperti Gaussian, polynomial, dan sigmoid dapat digunakan untuk memetakan data ke ruang dimensi yang lebih tinggi di mana hyperplane linier dapat digunakan untuk pemisahan (Hastie, dkk, 2009).

Keunggulan lain dari SVM adalah kemampuannya untuk menangani data dengan dimensi yang sangat tinggi, yang berarti bahwa ia dapat bekerja dengan baik bahkan ketika jumlah fitur lebih besar daripada jumlah sampel. Selain itu, SVM relatif tahan terhadap overfitting, terutama ketika menggunakan regularisasi yang tepat. Parameter regularisasi C dalam SVM memungkinkan pengontrolan trade-off antara memaksimalkan margin dan meminimalkan kesalahan klasifikasi. Dengan demikian, SVM sering digunakan dalam aplikasi seperti pengenalan wajah, teks, dan bioinformatika, di mana akurasi dan generalisasi model sangat penting.

Namun, SVM juga memiliki beberapa kelemahan. Salah satu kelemahan utama adalah kompleksitas komputasi dan memori yang tinggi, terutama untuk dataset yang sangat besar. Latihan model SVM bisa memakan waktu lama dan membutuhkan sumber daya komputasi yang signifikan. Selain itu, pemilihan fungsi kernel yang tepat dan pengaturan hyperparameter yang optimal bisa menjadi tantangan dan membutuhkan pemahaman yang mendalam serta eksperimen yang ekstensif (Bishop, 2006).

### *Neural Networks*

*Neural Networks* adalah algoritma *Machine Learning* yang terinspirasi oleh cara kerja otak manusia. Model ini terdiri dari lapisan-lapisan neuron yang saling terhubung, di mana

setiap neuron dalam satu lapisan terhubung ke neuron-neuron di lapisan berikutnya. Lapisan-lapisan ini biasanya terdiri dari lapisan input, satu atau lebih lapisan tersembunyi, dan lapisan output. Setiap neuron menerima satu atau lebih input, mengaplikasikan bobot tertentu pada input tersebut, dan menggunakan fungsi aktivasi untuk menentukan outputnya. *Neural Networks* sangat kuat dan mampu menangani data yang sangat kompleks dan non-linear (Haykin, 2009).

Salah satu kekuatan utama *Neural Networks* adalah kemampuannya untuk belajar representasi fitur yang sangat komplek. Dengan arsitektur yang cukup dalam (dikenal sebagai *Deep Neural Networks* atau DNN), model ini dapat mengidentifikasi pola dan struktur dalam data yang sulit atau tidak mungkin ditemukan oleh algoritma lain. Misalnya, dalam pengenalan gambar, lapisan pertama mungkin belajar mendeteksi tepi, sementara lapisan yang lebih dalam belajar mendeteksi bentuk yang lebih kompleks seperti wajah atau objek tertentu (Goodfellow, dkk, 2016).

Proses pelatihan *Neural Networks* melibatkan optimisasi bobot-bobot neuron menggunakan algoritma backpropagation, dimana kesalahan prediksi (*loss*) dihitung dan disebarkan kembali melalui jaringan untuk memperbarui bobot. Optimisasi ini biasanya dilakukan dengan menggunakan teknik gradient descent. Dengan iterasi yang cukup banyak dan dataset yang memadai, *Neural Networks* dapat mencapai performa yang sangat tinggi dalam berbagai tugas klasifikasi dan regresi.

Meskipun *Neural Networks* sangat kuat, ada beberapa tantangan yang perlu diperhatikan. Salah satunya adalah kebutuhan akan data yang besar untuk melatih model dengan baik. *Neural Networks* cenderung memerlukan lebih banyak data dibandingkan dengan algoritma lain untuk mencegah

overfitting dan memastikan generalisasi yang baik. Selain itu, pelatihan model yang sangat dalam bisa memerlukan sumber daya komputasi yang signifikan, seperti GPU, untuk mempercepat proses pelatihan.

## Evaluasi Model Klasifikasi

Evaluasi model klasifikasi adalah proses untuk mengukur seberapa baik model yang telah dilatih mampu membuat prediksi yang benar pada data yang tidak terlihat sebelumnya. Beberapa metrik evaluasi umum yang digunakan untuk model klasifikasi meliputi akurasi, *precision, recall, F1-score*, dan *area under the ROC curve* (AUC-ROC). Penjelasannya sebagai berikut.

1. Akurasi
   Akurasi adalah proporsi prediksi yang benar (baik positif maupun negatif) dari total prediksi yang dibuat oleh model. Metrik ini memberikan gambaran umum tentang kinerja model, tetapi dapat menyesatkan jika data tidak seimbang (misalnya, jika satu kelas jauh lebih sering muncul daripada kelas lainnya).
2. *Precision* dan *Recall*
   *Precision* adalah proporsi prediksi positif yang benar dari semua prediksi positif yang dibuat oleh model. Precision tinggi menunjukkan bahwa ketika model memprediksi kelas positif, ia cenderung benar.
   *Recall* (atau *Sensitivity*) adalah proporsi prediksi positif yang benar dari semua kasus positif yang sebenarnya. *Recall* tinggi menunjukkan bahwa model mampu menangkap sebagian besar dari kasus positif yang ada.
   *Precision* dan *recall* sering digunakan bersama karena *trade-off* antara keduanya; meningkatkan precision

biasanya mengurangi *recall*, dan sebaliknya.

3. F1-Score
   F1-score adalah rata-rata harmonis dari precision dan *recall*. Metrik ini berguna ketika Anda membutuhkan keseimbangan antara precision dan *recall*, terutama ketika distribusi kelas tidak seimbang. F1-score memberikan gambaran yang lebih baik tentang kinerja model dibandingkan akurasi dalam situasi ini. Rumus F1-score adalah:

$$F1 = 2\,x\,\frac{precision\,x\,recall}{precision + recall}$$

4. ROC Curve dan AUC-ROC
   ROC (*Receiver Operating Characteristic*) curve adalah plot yang menunjukkan *trade-off* antara *true positive rate* (TPR) dan false positive rate (FPR) pada berbagai threshold klasifikasi. AUC-ROC (*Area Under the ROC Curve*) mengukur area di bawah kurva ROC dan memberikan gambaran tentang kemampuan model untuk membedakan antara kelas-kelas. AUC-ROC berkisar dari 0.5 (tidak ada kemampuan klasifikasi) hingga 1 (klasifikasi sempurna).

## Tantangan dalam Klasifikasi

Berikut ini beberapa tantangan utama dalam klasifikasi, yaitu:

1. Ketidakseimbangan Kelas
   Ketidakseimbangan kelas terjadi ketika satu kelas jauh lebih dominan dibandingkan kelas lainnya dalam dataset. Ini dapat menyebabkan model menjadi bias terhadap kelas yang lebih sering muncul, mengabaikan kelas yang

lebih jarang. Ketidakseimbangan ini bisa membuat metrik seperti akurasi menjadi menyesatkan. Untuk mengatasi masalah ini, teknik seperti *oversampling* kelas minoritas, *undersampling* kelas mayoritas, atau menggunakan metrik evaluasi yang lebih informatif seperti F1-score dan AUC-ROC sering digunakan.

2. *Overfitting* dan *Underfitting*
   *Overfitting* terjadi ketika model terlalu kompleks dan mulai menangkap noise dalam data pelatihan sebagai informasi yang berguna, yang mengakibatkan kinerja yang buruk pada data baru. Sebaliknya, underfitting terjadi ketika model terlalu sederhana untuk menangkap pola yang ada dalam data. Regularisasi, pemangkasan pohon dalam *Decision Tree*s, dan menggunakan teknik validasi silang dapat membantu mengatasi masalah ini.
3. Dimensi Tinggi
   Ketika dataset memiliki banyak fitur (dimensi tinggi), ini bisa menyebabkan masalah yang dikenal sebagai "kutukan dimensi" (*curse of dimensionality*). Dalam ruang dimensi tinggi, jarak antar titik data menjadi lebih seragam, sehingga membuat klasifikasi menjadi lebih sulit. Teknik seperti seleksi fitur, ekstraksi fitur (misalnya, PCA), dan penggunaan algoritma yang secara eksplisit dirancang untuk menangani data dimensi tinggi bisa membantu.
4. Variabilitas dan Kualitas Data
   Kualitas data sangat penting untuk kinerja model klasifikasi. Data yang mengandung banyak noise, missing values, atau data yang tidak representatif dapat mengurangi kinerja model secara signifikan. Data preprocessing yang baik, seperti pembersihan data,

imputasi nilai yang hilang, dan deteksi serta penanganan outliers, sangat penting untuk memastikan bahwa data yang digunakan untuk melatih model adalah data yang berkualitas.

5. Interpretabilitas Model
   Beberapa algoritma klasifikasi, seperti *Neural Networks* dan ensemble methods (misalnya, *Random Forest*), dapat menghasilkan model yang sangat akurat tetapi sulit untuk diinterpretasikan. Dalam banyak aplikasi, terutama dalam bidang yang sangat regulatif seperti kesehatan dan keuangan, penting untuk dapat menjelaskan bagaimana model membuat keputusan. Model yang lebih sederhana seperti *Decision Tree*s atau logistic regression biasanya lebih mudah diinterpretasikan, tetapi mungkin kurang akurat dibandingkan model yang lebih kompleks.

# BAB 16 PREDIKSI DALAM DATA SCIENCE

## Pendahuluan

Ilmu Data (Data Science) tidak bisa terlepas kaitannya dengan Data Besar (*Big Data*) dan Kecerdasaan Buatan (*Artificial Intelligence*/AI). *Big Data* adalah jumlah data yang sangat besar yang tidak bisa diolah dengan metode konvensional. Data besar ini memiliki beberapa karakteristik umum seperti volumenya yang besar, kecepatan (*velocity*) yang bisa diakses secara *real-time*, keakuratan data (*veracity*) dan keberagaman tipe data (*variety*).

Data besar ini perlu diolah dan dianalisa sehingga menghasilkan informasi-informasi berharga yang bisa memberikan rekomendasi bagi institusi. Analisa *Big Data* membutuhkan beberapa metode sehingga data bisa terekstrak dengan baik dan menghasilkan data yang valid. Metode ini bisa dipelajari dalam ilmu data (Data Science). Ilmu data adalah ilmu yang mempelajari metode-metode dalam menganalisa data besar sehingga bisa menghasilkan output yang bernilai.

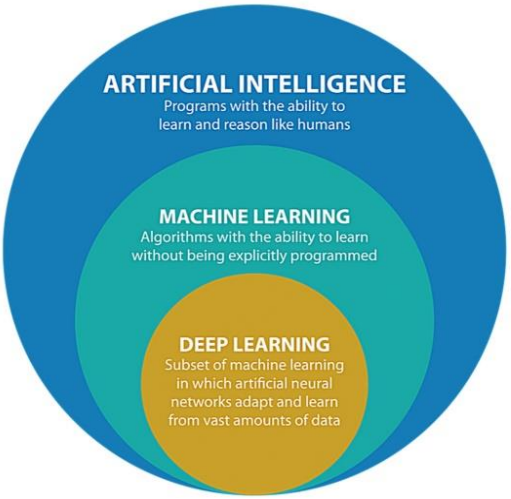
Kemudian, AI merupakan tools atau teknologi yang digunakan untuk memproses analisa data besar tersebut menggunakan metode yang dipelajari dalam ilmu data sehingga menghasilkan output atau prediksi. Itulah keterlibatan antara *Big Data*, Data Science dan AI.

## Pengantar Prediksi

Prediksi adalah proses analitis yang menggunakan data

historis untuk meramalkan kejadian atau nilai di masa depan. Dalam data science, prediksi sangat penting karena memungkinkan pengambilan keputusan yang lebih baik berdasarkan informasi yang diperoleh dari data yang tersedia. Tujuan utama dari model prediktif yaitu memprediksi hasil dimasa depan yang bisa menjadi acuan dalam menentukan kebijakan.

Seperti yang sudah dijelaskan dipendahuluan bahwa AI adalah teknologi yang digunakan tidak hanya untuk memproses dan menganalisa data besar, tapi juga untuk menghasilkan prediksi. Dalam melaksanakan tugasnya, AI terdiri dari Pembelajaran Mesin (*Machine Learning*) yang digunakan untuk memprediksi data yang bersifat struktur seperti data-data yang biasanya disajikan berbentuk tabel dan Pembelajaran Mendalam (*Deep Learning*) yang digunakan untuk memprediksi data yang bersifat tidak terstruktur seperti suara, gambar, dan teks. *Deep Learning* adalah cabang dari *Machine Learning* yang berfokus pada algoritma yang terinspirasi oleh struktur dan fungsi otak yang disebut jaringan saraf tiruan (*Artificial Neural Networks*). *Deep Learning* adalah salah satu teknik paling canggih dalam AI, memungkinkan komputer untuk belajar dari data dan membuat keputusan atau prediksi tanpa diprogram secara eksplisit untuk setiap tugas. Jadi bisa disimpulkan bahwa *Machine Learning* dan *Deep Learning* merupakan bagian dari AI yang melaksanakan tugasnya masing-masing berdasarkan tipe datanya (*variety*) Gambar 16.1.




*Gambar 16.1 Bagian-Bagian AI*

## *Machine Learning*

*Machine Learning* adalah komputer yang diprogram untuk belajar dan membuat prediksi ataupun keputusan berdasarkan yang diinput. Konsep kecerdasan buatan ini didasari pada kerja *Machine Learning* yang belajar dari input data yang kemudian di proses untuk menghasilkan output dimana semakin banyak data yang dipelajari semakin cerdas mesin memprosesnya sehingga menghasilkan prediksi yang semakin akurat, ini lah dasar konsep kerja sehingga disebut dengan kecerdasan buatan.

Dalam pembelajarannya, *Machine Learning* melibatkan fungsi model pembelajaran terawasi (*supervised learning*) yang digunakan untuk prediksi nilai target (Pradnyana,2020) dan pembelajaran tak terawasi (*unsupervised learning*) yang digunakan untuk menemukan kelompok yang bermakna dari data. Supervised Learning adalah teknik pembelajaran mesin dimana model dilatih menggunakan data yang diberi label,

dimana setiap masukan memiliki keluaran yang sesuai. Tujuan model adalah mempelajari hubungan antara masukan dan keluaran untuk membuat prediksi pada data baru yang tidak terlihat sebelumnya. Sedangkan, Unsupervised Learning adalah teknik pembelajaran mesin dimana model dilatih menggunakan data yang tidak diberi label. Tujuan dari unsupervised learning adalah untuk menemukan pola atau struktur yang tersembunyi dalam data. Tidak ada output yang diinginkan dalam data pelatihan, dan model harus belajar sendiri untuk menemukan kelompok yang bermakna dalam data.

### *Supervised Learning*

*Supervised Learning* dibagi menjadi dua kategori utama yaitu Klasifikasi (*Classification*) dan Regresi (*Regression*).

1. *Classification*

   *Classification* merupakan tugas *supervised learning* yang digunakan ketika target yang diprediksi adalah variabel kategorik. Model dilatih untuk mengklasifikasikan input ke dalam salah satu dari beberapa kelas yang telah ditentukan. Contoh aplikasi klasifikasi meliputi:

   a. Deteksi Penipuan: Mengklasifikasikan transaksi sebagai penipuan atau tidak.
   b. Diagnosis Medis: Mengklasifikasikan kondisi medis pasien berdasarkan data kesehatan.
   c. Klasifikasi Email: Mengklasifikasikan email sebagai spam atau tidak spam.

   Algoritma yang digunakan dalam model prediksi klasifikasi antara lain:

   a. Regresi Logistik

      Regresi logistik digunakan untuk prediksi klasifikasi biner, di mana target memiliki dua

kategori. Model ini memperkirakan probabilitas suatu kejadian berdasarkan fitur yang ada. Misalnya, dalam diagnosis medis, regresi logistik dapat digunakan untuk memprediksi apakah seorang pasien memiliki penyakit tertentu atau tidak berdasarkan data kesehatan.

b. *Decision Trees*
   *Decision Trees* adalah model berbasis pohon yang membuat keputusan berdasarkan fitur data. Setiap node dalam pohon mewakili fitur, dan cabang mewakili keputusan yang mengarah ke nilai target. Decision Trees mudah diinterpretasikan dan digunakan dalam berbagai aplikasi prediksi.
c. SVM (*Support Vector Machines*)
   SVM adalah algoritma yang menemukan hyperplane terbaik untuk memisahkan data dalam klasifikasi. Hyperplane ini memaksimalkan margin antara dua kelas data, sehingga menghasilkan prediksi yang lebih akurat.
d. *K-Nearest Neighbors* (KNN)
   KNN adalah algoritma yang memprediksi nilai atau kategori target berdasarkan kemiripan dengan data terdekat. Algoritma ini menghitung jarak antara data baru dan data pelatihan, kemudian mengambil keputusan berdasarkan mayoritas tetangga terdekat.
e. *Random Forest*
   *Random Forest* adalah kombinasi dari banyak pohon keputusan yang digunakan untuk meningkatkan akurasi prediksi. Setiap pohon dalam hutan membuat prediksi, dan hasil akhir

ditentukan berdasarkan mayoritas prediksi pohon.

2. *Regression*

   *Regression* adalah tugas supervised learning dimana model dilatih untuk memprediksi nilai numerik berdasarkan input yang diberikan. Contoh aplikasi regresi meliputi:

   a. Prediksi Harga Rumah
      Memprediksi harga rumah berdasarkan fitur seperti luas tanah, jumlah kamar, dan lokasi.
   b. Prediksi Suhu
      Memprediksi suhu masa depan berdasarkan data historis suhu.

   Algoritma yang digunakan dalam model prediksi regresi antara lain:

   a. Regresi Linear
      Regresi linear adalah model statistik yang berusaha untuk menemukan hubungan linear antara variabel independen (fitur) dan variabel dependen (target). Model ini menggunakan persamaan linear regresi.
   b. Regresi Polinomial
      Regresi polinomial memperluas regresi linear dengan menambahkan derajat polinomial pada fitur-fitur untuk menangkap hubungan non-linear antara fitur dan target.
   c. *Decision Trees for Regression*
      Decision trees adalah model yang menggunakan struktur pohon untuk membuat keputusan berdasarkan nilai fitur. Setiap node dalam pohon mewakili fitur, setiap cabang mewakili aturan keputusan, dan setiap daun mewakili hasil

prediksi.

d. *Random Forest*
   *Random forest* adalah model ensemble yang terdiri dari banyak pohon keputusan. Model ini menggabungkan prediksi dari beberapa pohon untuk meningkatkan akurasi dan mengurangi overfitting dengan membuat banyak pohon keputusan dari sampel acak data dan menggabungkan prediksi dari semua pohon untuk memberikan hasil akhir.
e. *Support Vector Machines for Regression* (SVR)
   *Support Vector Regression* (SVR) adalah versi dari *Support Vector Machines* (SVM) yang digunakan untuk tugas regresi. SVR menggunakan hyperplane untuk memprediksi nilai kontinu dalam ruang fitur. Model ini menggunakan kernel untuk mengubah data ke dalam ruang fitur yang memaksimalkan margin antara data dan hyperlane.

### *Unsupervised Learning*

*Unsupervised Learning* adalah metode pembelajaran mesin mesin di mana model dilatih menggunakan data yang tidak diberi label. Model mencoba menemukan pola atau struktur yang tersembunyi dalam data. *Unsupervised learning* dibagi menjadi dua kategori utama: *Clustering* dan *dimensionality reduction*.

1. *Clustering*
   *Clustering* adalah tugas unsupervised learning di mana data di kelompokkan ke dalam beberapa kluster berdasarkan kemiripan antar data. Model mencoba

menemukan kelompok atau segmen dalam data.

a. Contoh Kasus:
    1) Segmentasi Pelanggan
       Mengelompokkan pelanggan ke dalam segmen yang berbeda berdasarkan perilaku belanja.
    2) Analisis Genom
       Mengelompokkan gen berdasarkan ekspresi genetik mereka.
b. Algoritma yang digunakan:
    1) K-Means
       Mengelompokkan data ke dalam k klaster berdasarkan jarak terdekat.
    2) Hierarchical *Clustering*
       Mengelompokkan data dalam hierarki klaster.
    3) DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)
       Mengelompokkan data berdasarkan kepadatan.

2. *Dimensionality Reduction*
   *Dimensionality reduction* adalah tugas *unsupervised learning* di mana data dengan dimensi tinggi direduksi menjadi data dengan dimensi lebih rendah sambil mempertahankan sebanyak mungkin informasi penting. Tujuannya adalah untuk mengurangi kompleksitas data dan meningkatkan efisiensi pemrosesan.
    a. Contoh Kasus:
        1) Visualisasi Data
           Mengurangi dimensi data yang tinggi untuk memudahkan visualisasi (misalnya, dari

100 fitur menjadi 2 atau 3 fitur).

2) Pra-pemrosesan Data
Mengurangi jumlah fitur sebelum menerapkan algoritma pembelajaran mesin untuk mengurangi overfitting dan meningkatkan kinerja.

b. Algoritma yang Digunakan:

1) *Principal Component Analysis* (PCA)
Mengurangi dimensi data dengan mempertahankan variansi sebanyak mungkin.

2) *t-Distributed Stochastic Neighbor Embedding* (t-SNE)
Mengurangi dimensi data untuk visualisasi yang lebih baik dalam ruang dua atau tiga dimensi.

3) *Linear Discriminant Analysis* (LDA)
Mengurangi dimensi data dengan memaksimalkan separasi antar kelas.

## Tahapan Penerapan Model Prediksi

Model Prediksi melibatkan dua komponen utama: variabel independen (fitur) dan variabel dependen (target). Variabel independen adalah faktor atau fitur yang digunakan untuk memprediksi nilai target. Misalnya, dalam memprediksi harga rumah, fitur bisa berupa luas rumah, jumlah kamar, dan lokasi, sementara target adalah harga rumah tersebut.

Pada dasarnya, proses prediksi dapat dibagi menjadi beberapa tahap:

1. Pengumpulan Data
Langkah pertama dalam pengembangan model prediksi

adalah mengumpulkan data yang relevan dan cukup untuk analisis. Data ini bisa berasal dari berbagai sumber seperti *database* perusahaan, sensor, atau data publik.

2. Praproses Data
   Setelah data dikumpulkan, langkah selanjutnya adalah praproses data. Proses ini mencakup pembersihan data, penanganan missing values, dan transformasi data. Misalnya, data yang hilang bisa diisi dengan rata-rata atau median, dan data yang tidak relevan bisa dihapus.
3. Pembagian Data
   Untuk mengevaluasi kinerja model, data biasanya dibagi menjadi set pelatihan dan set pengujian. Set pelatihan digunakan untuk melatih model, sementara set pengujian digunakan untuk mengevaluasi kinerja model.
4. Pemilihan Model
   Pemilihan model adalah langkah penting dalam pengembangan prediksi. Model yang dipilih harus sesuai dengan karakteristik data dan tujuan analisis. Misalnya, untuk prediksi kontinu, regresi linear bisa menjadi pilihan yang baik, sementara untuk klasifikasi, decision trees atau SVM bisa lebih cocok.
5. Pelatihan Model
   Pelatihan model melibatkan penggunaan data pelatihan untuk mengajarkan model tentang pola dalam data. Proses ini mencakup penyesuaian parameter model untuk meminimalkan kesalahan prediksi.
6. Evaluasi Model
   Evaluasi model dilakukan menggunakan data pengujian dan metrik evaluasi untuk mengukur kinerja model. Proses ini memastikan bahwa model dapat membuat prediksi yang akurat pada data yang tidak terlihat

sebelumnya. Tahap ini adalah langkah penting untuk memastikan model yang digunakan cukup akurat dan dapat diandalkan. Beberapa metrik evaluasi yang sering digunakan untuk tugas regresi sebagai berikut.

a. *Mean Absolute Error* (MAE)
   MAE adalah rata-rata selisih absolut antara nilai prediksi dan nilai sebenarnya. Metrik ini memberikan gambaran tentang seberapa besar kesalahan prediksi rata-rata.
b. *Mean Squared Error* (MSE)
   MSE adalah rata-rata kuadrat selisih antara nilai prediksi dan nilai sebenarnya. MSE memberikan penalti lebih besar pada kesalahan yang lebih besar, sehingga lebih sensitif terhadap outliers.
c. *Root Mean Squared Error* (RMSE)
   RMSE adalah akar kuadrat dari MSE, memberikan penalti lebih besar pada kesalahan yang lebih besar, dan sering digunakan untuk mengevaluasi model regresi.
d. *R-Squared*
   *R-squared* mengukur proporsi variansi dalam target yang dapat dijelaskan oleh fitur. Nilai R-squared berkisar antara 0 dan 1, dengan nilai yang lebih tinggi menunjukkan model yang lebih baik.

Sedangkan metrik evaluasi yang sering digunakan untuk tugas klasifikasi sebagai berikut.

a. *Accuracy*
   *Accuracy* adalah persentase prediksi benar dari total prediksi dan sering digunakan dalam evaluasi model klasifikasi.
b. *Confusion Matrix*

Sebuah tabel yang sering digunakan untuk mengukur kinerja model klasifikasi di *Machine Learning*. Tabel ini menggambarkan detail tentang data yang diklasifikan dengan benar maupun salah. Confusion matrix adalah salah satu tools analitik prediktif yang menampilkan dan membandingkan nilai actual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan untuk menghasilkan metrik evaluasi seperti Accuracy (akurasi), Precision, *Recall* dan F1-Score (https://ilmudatapy.com/apa-itu-confusion-matrix/).

## Studi Kasus Prediksi

Untuk memperjelas konsep prediksi, bab ini akan menyajikan beberapa studi kasus, misalnya:

1. Prediksi Harga Rumah
   Menggunakan regresi linear untuk memprediksi harga rumah berdasarkan fitur seperti luas tanah, jumlah kamar, dan lokasi.
2. Klasifikasi Penyakit
   Menggunakan regresi logistik atau SVM untuk mengklasifikasikan apakah seorang pasien menderita penyakit tertentu berdasarkan data medis.

## Tantangan dalam Prediksi

Meskipun prediksi dapat sangat berguna, ada beberapa tantangan yang perlu diperhatikan, seperti:

1. Overfitting dan Underfitting
   Ketidakseimbangan antara model yang terlalu kompleks

atau terlalu sederhana.

2. Data Bias
   Ketidakcocokan data pelatihan dengan data nyata.
3. Ketergantungan Temporal
   Prediksi yang memerlukan data historis yang relevan dan terbaru.

Prediksi adalah komponen kunci dalam data science yang memungkinkan kita membuat keputusan berdasarkan data historis. Dengan memahami teknik, proses, dan evaluasi yang tepat, kita dapat mengembangkan model prediksi yang andal dan berguna dalam berbagai aplikasi nyata.

# BAB 17 *NATURAL LANGUAGE PROCESSING*

## Pendahuluan

Natural Language atau Bahasa Alami adalah bahasa yang digunakan oleh manusia untuk berkomunikasi sehari-hari, baik dalam bentuk lisan maupun tulisan. Bahasa alami bersifat kompleks, penuh dengan aturan dan struktur yang berkembang secara alami melalui interaksi manusia dalam berbagai konteks sosial, budaya, dan historis. *Natural Language Processing* (NLP) merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*, AI) yang berfokus pada interaksi antara komputer dan bahasa manusia (Jurafsky & Martin, 2021).

NLP bertujuan untuk mengembangkan teknologi yang memungkinkan mesin untuk memahami, memproses, dan menghasilkan bahasa alami dengan cara yang lebih efektif. Teknologi ini memainkan peran penting dalam berbagai aplikasi, mulai dari asisten virtual seperti Siri dan Alexa, hingga penerjemah otomatis dan analisis sentimen di media sosial (Goldberg, 2017). Perkembangan NLP dipengaruhi oleh kombinasi antara linguistik komputasional, pembelajaran mesin (*Machine Learning*), dan kecerdasan buatan yang lebih luas (Jurafsky & Martin, 2021).

## Sejarah dan Perkembangan NLP

Sejarah NLP dimulai pada pertengahan abad ke-20, seiring dengan perkembangan komputer dan kecerdasan buatan. Pada masa awal, pendekatan NLP didasarkan pada aturan dan

logika yang ditulis secara eksplisit oleh manusia. Pendekatan ini, yang dikenal sebagai *Rule-Based* NLP, mengandalkan serangkaian aturan linguistik yang rumit untuk menganalisis dan memproses bahasa (Manning & Schütze, 1999). Namun, pendekatan ini memiliki keterbatasan dalam menangani ambiguitas dan variasi bahasa.

Pada akhir 1980-an dan awal 1990-an, pendekatan Statistical NLP mulai muncul, yang menggunakan teknik statistik dan probabilitas untuk mengatasi beberapa tantangan dalam analisis bahasa (Manning & Schütze, 1999). Pendekatan ini kemudian digantikan oleh metode *Machine Learning* pada dekade 2000-an, yang memungkinkan mesin untuk belajar dari data dalam jumlah besar tanpa perlu aturan eksplisit. Saat ini, metode *Deep Learning* dan model bahasa besar seperti BERT dan GPT telah merevolusi NLP dengan menghasilkan hasil yang jauh lebih akurat dan kontekstual (Devlin et al., 2019; Vaswani et al., 2017).

## Peran NLP dalam Kehidupan Sehari-hari

NLP telah menjadi bagian integral dari kehidupan sehari-hari dalam berbagai aplikasi, baik disadari maupun tidak. Chatbot dan asisten virtual kini digunakan secara luas untuk membantu dalam berbagai tugas, seperti menjawab pertanyaan, mengatur jadwal, dan menyediakan rekomendasi (Jurafsky & Martin, 2021). NLP juga digunakan dalam aplikasi penerjemahan otomatis, yang memungkinkan terjemahan bahasa secara real-time antara berbagai bahasa. Di bidang bisnis, NLP membantu perusahaan menganalisis umpan balik pelanggan dan menentukan sentimen dari komentar yang diungkapkan di media sosial atau survei (Goldberg, 2017).

NLP juga memiliki dampak besar dalam dunia kesehatan, misalnya dalam analisis catatan medis elektronik (*Electronic Health Records*, EHR) untuk membantu dokter dan profesional kesehatan dalam memahami tren dan pola yang mungkin tidak terlihat secara langsung (Hirschberg & Manning, 2015).

## Komponen *Natural Language*

*Natural Language* atau Bahasa Alami memiliki struktur yang sangat kompleks. Untuk memahami dan memproses bahasa alami, baik oleh manusia maupun mesin, terdapat beberapa komponen linguistik penting yang perlu diperhatikan. Menurut Jurafsky & Martin (2021), komponen ini membantu dalam memahami bahasa alami secara komprehensif.

### Fonologi

Fonologi adalah studi tentang bunyi dalam bahasa. Dalam konteks *Natural Language Processing* (NLP), fonologi berfokus pada cara komputer mengenali dan memproses suara bahasa. Ini penting untuk aplikasi seperti pengenalan suara (*speech recognition*) dan *text-to-speech* (TTS) (Jurafsky & Martin, 2021). Fonologi mencakup dua aspek utama:

1. Fonem
   Unit terkecil dari suara dalam bahasa yang dapat membedakan makna. Misalnya, dalam bahasa Indonesia, perbedaan fonem antara "p" dan "b" dapat membedakan kata "palu" dan "balu" (Manning & Schütze, 1999).
2. Prosodi
   Studi tentang pola intonasi, tekanan, dan durasi dalam ucapan. Dalam NLP, prosodi sangat penting dalam memahami konteks emosional dan makna dari ucapan

(Hirschberg & Manning, 2015).

## Morfologi

Morfologi adalah studi tentang struktur kata dan bagaimana kata terbentuk dari unit terkecil yang disebut morfem. Morfologi berkaitan dengan cara kata-kata dalam bahasa alami dibentuk dan diubah (Manning & Schütze, 1999).

Morfem: Bagian terkecil dari sebuah kata yang memiliki makna. Ada dua jenis morfem, yaitu morfem bebas (yang dapat berdiri sendiri sebagai kata) dan morfem terikat (yang memerlukan imbuhan). Sebagai contoh, dalam kata "makanan", "makan" adalah morfem bebas dan "-an" adalah morfem terikat (Jurafsky & Martin, 2021).

Dalam NLP, *stemming* dan *lemmatization* adalah teknik penting yang digunakan untuk memproses morfologi (Goldberg, 2017). *Stemming* memotong kata menjadi bentuk dasarnya, sementara lemmatization mempertahankan bentuk dasar kata dengan mempertimbangkan aturan tata bahasa.

## Sintaksis

Sintaksis adalah studi tentang aturan dan struktur kalimat dalam bahasa. Ini mengatur bagaimana kata-kata digabungkan untuk membentuk kalimat yang bermakna (Jurafsky & Martin, 2021). Dalam NLP, sintaksis berperan dalam parsing, yaitu proses untuk mengidentifikasi struktur gramatikal kalimat.

Parsing: Teknik parsing digunakan untuk menguraikan struktur gramatikal kalimat, membantu komputer memahami hubungan antara kata-kata (Manning & Schütze, 1999).

*Part-of-Speech Tagging* (POS Tagging): Metode ini mengidentifikasi kategori gramatikal (seperti kata benda atau

kata kerja) dari setiap kata dalam sebuah kalimat, yang penting untuk pemrosesan bahasa alami (Jurafsky & Martin, 2021).

Sebagai contoh, kalimat "Kucing itu makan ikan" memiliki struktur sintaksis yang terdiri dari subjek (kucing), predikat (makan), dan objek (ikan).

### Semantik

Semantik berkaitan dengan makna kata, frasa, dan kalimat dalam konteks tertentu. Ini adalah aspek kunci dalam memahami bahasa alami karena menentukan apa yang sebenarnya dimaksud oleh penutur atau penulis (Hirschberg & Manning, 2015). Dalam NLP, semantik mencakup tugas seperti:

1. *Named Entity Recognition* (NER)
   Teknik ini digunakan untuk mengidentifikasi entitas penting seperti nama orang, tempat, dan tanggal dalam teks (Jurafsky & Martin, 2021).
2. *Word Sense Disambiguation* (WSD)
   Proses ini membantu menentukan makna kata yang benar dalam konteks tertentu, terutama ketika kata tersebut memiliki lebih dari satu makna (Goldberg, 2017).

Sebagai contoh, kata "bisa" dalam kalimat "Dia bisa menyelesaikan masalah" berarti "mampu", sementara dalam kalimat "Bisa ular itu berbahaya" berarti "racun".

### Pragmatik

Pragmatik mempelajari bagaimana konteks mempengaruhi makna bahasa. Tidak semua makna dalam bahasa dapat ditentukan hanya melalui kata-kata dan struktur kalimat. Pragmatik sering kali dipengaruhi oleh konteks sosial, budaya, dan situasional (Jurafsky & Martin, 2021). Dalam NLP, pragmatik berperan dalam:

1. Resolusi Coreference
   Teknik ini mengidentifikasi referensi yang sama dalam teks. Sebagai contoh, dalam kalimat "Ani pergi ke sekolah. Dia membawa buku," kata "Dia" merujuk pada Ani (Goldberg, 2017).
2. Natural Language Understanding (NLU)
   Memahami maksud penutur atau penulis dalam konteks tertentu, termasuk elemen-elemen implisit yang tidak langsung disampaikan (Jurafsky & Martin, 2021).

Sebagai contoh:

A: "Apakah Anda punya waktu?"

B: "Saya sedang terburu-buru."

Di sini, secara pragmatis, kita dapat memahami bahwa B sebenarnya tidak punya waktu meskipun jawabannya tidak eksplisit.

## Teknik dalam Natural Language Processing (NLP)

*Natural Language Processing* (NLP) adalah bidang interdisipliner yang memadukan linguistik, ilmu komputer, dan kecerdasan buatan untuk memungkinkan komputer memahami, menafsirkan, dan memproses bahasa manusia. Dalam pengembangan aplikasi NLP, berbagai teknik telah digunakan untuk menangani kerumitan bahasa alami. Teknik-teknik ini mencakup metode statistik, pembelajaran mesin (*Machine Learning*), serta model berbasis aturan. Berikut adalah beberapa teknik utama dalam NLP:

1. Tokenisasi
   Tokenisasi adalah teknik dasar dalam NLP yang memecah teks menjadi unit-unit yang lebih kecil, yang disebut token. Token bisa berupa kata, frasa, atau bahkan karakter tergantung pada pendekatan yang digunakan

(Jurafsky & Martin, 2021). Tokenisasi penting karena banyak model NLP bekerja pada level kata atau token untuk analisis lebih lanjut.

Sebagai contoh, kalimat "Kucing makan ikan" dapat dipecah menjadi token: ["Kucing", "makan", "ikan"].

Dalam aplikasi NLP, tokenisasi sangat penting untuk tahap-tahap berikutnya seperti analisis sintaksis dan semantik. Teknik ini biasanya digunakan sebagai langkah awal dalam berbagai tugas NLP, seperti machine translation, text summarization, dan information retrieval (Goldberg, 2017).

2. *Stemming* dan *Lemmatization*

   Stemming dan lemmatization adalah teknik untuk mereduksi kata-kata ke bentuk dasarnya. Stemming melibatkan pemotongan akhiran dari kata untuk menemukan bentuk dasarnya, sedangkan lemmatization mempertahankan bentuk dasar kata dengan memperhatikan aturan tata bahasa (Manning & Schütze, 1999).

   Sebagai contoh, dalam stemming, kata "berlari" akan dipotong menjadi "lari". Sementara dalam lemmatization, algoritme akan mempertimbangkan konteks untuk memastikan bahwa "berlari" benar-benar bentuk kata dari "lari."

   Dalam NLP, kedua teknik ini membantu dalam mengurangi variasi kata sehingga mesin dapat lebih mudah memahami hubungan antara berbagai bentuk kata yang berbeda (Goldberg, 2017).

3. *Part-of-Speech Tagging* (POS Tagging)

   *Part-of-Speech* (POS) Tagging adalah teknik untuk menetapkan kategori gramatikal pada setiap kata dalam

teks, seperti kata benda, kata kerja, atau kata sifat. POS tagging membantu mesin memahami peran kata-kata dalam kalimat dan menentukan makna yang lebih mendalam dari teks (Jurafsky & Martin, 2021).

Sebagai contoh, dalam kalimat "Kucing makan ikan," kata "kucing" diberi tag sebagai kata benda (noun), "makan" sebagai kata kerja (*verb*), dan "ikan" sebagai kata benda (*noun*). POS tagging merupakan langkah penting dalam syntactic parsing dan named entity recognition.

Teknik ini umumnya diterapkan menggunakan algoritme berbasis pembelajaran mesin, seperti *Hidden Markov Models* (HMM) atau Conditional Random Fields (CRF), yang mengidentifikasi pola gramatikal dalam teks (Manning & Schütze, 1999).

4. *Named Entity Recognition* (NER)

   Named Entity Recognition (NER) adalah teknik untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks, seperti nama orang, organisasi, lokasi, tanggal, dan lain-lain. Teknik ini memungkinkan mesin untuk mengenali elemen-elemen penting dalam teks yang memiliki relevansi tinggi terhadap konteks (Hirschberg & Manning, 2015).

   Sebagai contoh, dalam kalimat "Barack Obama lahir di Hawaii," NER akan mengenali "Barack Obama" sebagai entitas "orang" dan "Hawaii" sebagai entitas "lokasi."

   NER sering digunakan dalam aplikasi seperti ekstraksi informasi, analisis teks, dan pemrosesan data besar (*Big Data*) untuk membantu mengidentifikasi informasi yang relevan secara otomatis (Jurafsky & Martin, 2021).

5. *Syntactic Parsing*

   *Syntactic Parsing* adalah teknik untuk menganalisis

struktur gramatikal dari sebuah kalimat dengan cara membangun pohon parsing yang menunjukkan hubungan antar kata dalam kalimat tersebut (Manning & Schütze, 1999). Parsing sintaksis membantu dalam memahami hubungan antara subjek, predikat, dan objek dalam kalimat.

Sebagai contoh, dalam kalimat "Kucing itu mengejar tikus," parsing sintaksis akan mengidentifikasi "kucing" sebagai subjek, "mengejar" sebagai predikat, dan "tikus" sebagai objek.

Parsing sangat penting dalam aplikasi seperti machine translation, di mana struktur kalimat harus dipahami secara mendalam sebelum diterjemahkan ke dalam bahasa lain (Jurafsky & Martin, 2021).

6. *Word Embeddings*

   *Word Embeddings* adalah teknik untuk merepresentasikan kata-kata sebagai vektor numerik di dalam ruang dimensi yang lebih rendah, yang memungkinkan mesin untuk mengenali hubungan semantik antar kata (Goldberg, 2017). Model populer seperti Word2Vec dan GloVe menggunakan teknik ini untuk menangkap arti dan konteks kata dalam teks.

   Word embeddings memungkinkan komputer untuk memahami bahwa kata-kata yang memiliki makna serupa, seperti "raja" dan "ratu", atau "mobil" dan "kendaraan", berada dalam vektor yang berdekatan dalam ruang dimensi. Word embeddings telah meningkatkan performa dalam banyak tugas NLP, termasuk text classification dan machine translation (Devlin et al., 2019).

7. *Transformers* dan *Attention Mechanism*

Model transformer adalah salah satu perkembangan terbaru dalam NLP, menggunakan arsitektur yang lebih efisien dalam memproses teks dibandingkan dengan metode sebelumnya, seperti *Recurrent Neural Networks* (RNNs) (Vaswani et al., 2017). Komponen kunci dari transformer adalah attention mechanism, yang memungkinkan model untuk memperhatikan bagian teks yang relevan saat membuat prediksi.

Sebagai contoh, model seperti BERT (*Bidirectional Encoder Representations from Transformers*) dan GPT (*Generative Pre-trained Transformer*) telah merevolusi cara kita menangani tugas-tugas NLP, seperti *text summarization, question answering*, dan *machine translation* (Devlin et al., 2019). Model ini mampu menangkap konteks dua arah (*bidirectional*) dalam teks, yang berarti mereka dapat memahami kata dalam konteks sebelum dan sesudahnya.

Teknik-teknik dalam NLP berkembang pesat, mulai dari teknik dasar seperti tokenisasi hingga model transformer canggih. Dengan teknik-teknik ini, NLP dapat memberikan solusi yang lebih akurat dan efisien untuk berbagai aplikasi seperti machine translation, sentiment analysis, dan chatbot. Setiap teknik memiliki peran penting dalam memproses dan memahami bahasa alami, sehingga memungkinkan mesin untuk berinteraksi dengan manusia dalam bahasa sehari-hari (Jurafsky & Martin, 2021; Vaswani et al., 2017).

# BAB 18 *DEEP LEARNING*

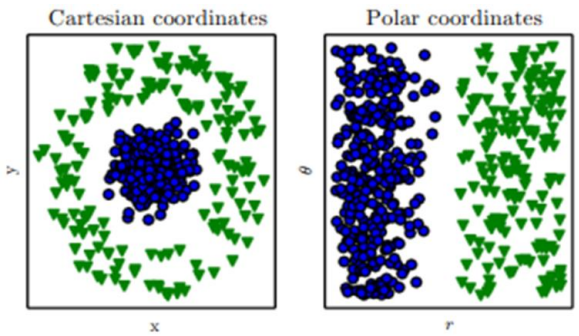
## Pendahuluan

Para penemu telah lama mengimpikan untuk menciptakan mesin yang mampu berpikir, dan hasrat ini telah ada sejak era Yunani kuno. Ada tokoh mitos seperti Pygmalion, Daedalus, dan Hephaestus yang dianggap sebagai penemu legendaris, sementara Galatea, Talos, dan Pandora dianggap sebagai kehidupan buatan. Ketika komputer yang dapat diprogram pertama kali ditemukan, muncul pertanyaan apakah mesin seperti itu dapat menjadi cerdas. Sekarang, kecerdasan buatan (AI) merupakan bidang yang berkembang cepat dengan berbagai aplikasi praktis dan topik penelitian yang dinamis. AI digunakan dalam memahami ucapan atau gambar, mengotomatiskan pekerjaan rutin, membuat diagnosis medis, dan mendukung penelitian ilmiah dasar. Pada masa awal AI, bidang ini berhasil menyelesaikan masalah yang sulit secara intelektual bagi manusia tetapi mudah bagi komputer, seperti masalah yang dapat dijelaskan dengan aturan matematika formal. Tantangan sebenarnya bagi AI adalah memecahkan masalah yang mudah bagi manusia tetapi sulit dijelaskan secara formal, seperti pengenalan ucapan dan wajah dalam gambar.

## Konsep *Deep Learning*

Goodfellow, dkk membahas tentang solusi untuk masalah-masalah yang lebih intuitif dalam kecerdasan buatan. Solusi yang diajukan adalah mengizinkan komputer belajar dari

pengalaman dan memahami dunia melalui hirarki konsep. Setiap konsep didefinisikan dalam relasi dengan konsep yang lebih sederhana. Dengan pendekatan ini, komputer dapat membangun pengetahuan dari pengalaman serta menghindari kebutuhan akan pengetahuan yang formal yang harus ditentukan oleh operator manusia. *Deep Learning* AI, seiring dengan grafik hirarki konsep yang kompleks, memungkinkan komputer memahami konsep yang rumit melalui konsep yang lebih sederhana. Meskipun banyak pencapaian AI awal terjadi pada lingkungan formal dan steril, seperti sistem permainan catur Deep Blue yang mengalahkan Garry Kasparov, tantangan sebenarnya tidak terletak pada kompleksitas permainan catur itu sendiri, tetapi pada kemampuan komputer dalam menggambarkan dan memahami konsep secara abstrak dan formal. Beberapa proyek kecerdasan buatan telah mencoba mengkodekan pengetahuan tentang dunia dalam bahasa formal. Namun, mereka menghadapi kesulitan dalam menyusun aturan formal yang cukup kompleks untuk menggambarkan dunia secara akurat. Dalam hal ini, pembelajaran mesin menjadi penting untuk memperoleh pengetahuan dari data mentah dan membuat keputusan subjektif. Untuk contoh visual yang sederhana, lihat gambar 18.1. Fitur-fitur yang tepat dapat digunakan dalam tugas kecerdasan buatan, seperti perkiraan ukuran saluran vokal untuk mengidentifikasi pembicara. Namun, sulit untuk menentukan fitur yang tepat dalam tugas seperti mendeteksi mobil dalam foto.

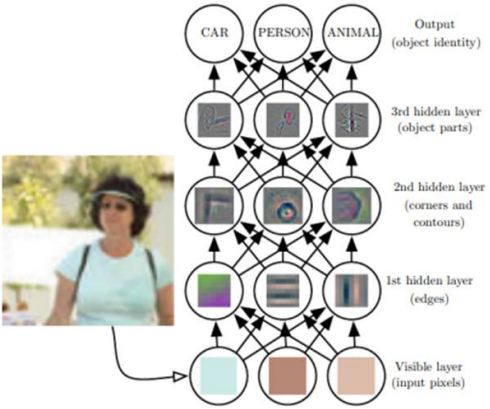



*Gambar 18.1 Contoh Visual Sederhana yang Merepresentasikan Kategori yang berbeda*

Gambar di atas merupakan contoh representasi yang berbeda, misalkan kita ingin memisahkan dua kategori data dengan menarik garis di antara keduanya dalam diagram pencar. Pada plot di sebelah kiri, kita merepresentasikan beberapa data menggunakan koordinat Kartesius, dan tugas tersebut tidak mungkin dilakukan. Di dalam plot di sebelah kanan, kami merepresentasikan data dengan koordinat kutub dan tugasnya menjadi mudah diselesaikan dengan garis vertikal. Gambar di atas dibuat oleh Goodfellow atas kerja sama dengan David Warde-Farley. Pendekatan pembelajaran representasi menggunakan algoritma seperti autoencoder dapat menghasilkan representasi yang lebih baik secara otomatis dengan sedikit campur tangan manusia. Pembuatan fitur manual membutuhkan waktu dan upaya manusia yang banyak. Kesulitas dalam aplikasi AI adalah faktor variasi dalam data yang sulit diuraikan. Mendapatkan representasi yang tepat menjadi masalah terpisah yang sulit diatasi.

*Deep Learning* adalah metode pembelajaran representasi yang menggunakan representasi yang lebih sederhana untuk membangun konsep yang kompleks. Contoh umum dari metode

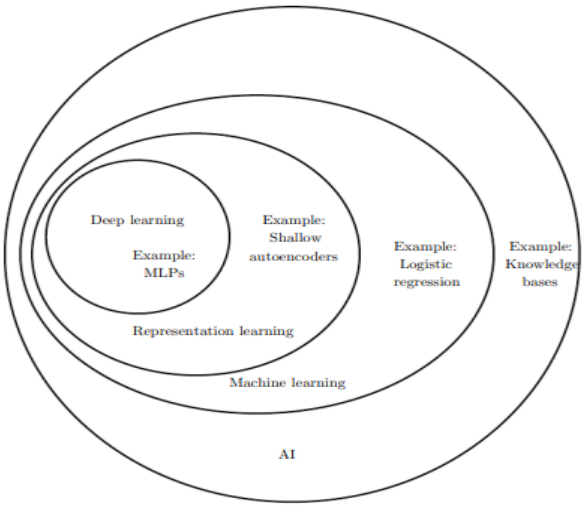
ini adalah jaringan feedforward deep atau multilayer perceptron (MLP), di mana setiap lapisan representasi dianggap sebagai keadaan memori komputer setelah mengeksekusi serangkaian instruksi secara paralel. Dengan menggunakan metode *Deep Learning*, komputer dapat mempelajari program komputer multi-langkah. Kedalaman yang lebih besar pada jaringan juga memungkinkan eksekusi lebih banyak instruksi berurutan. Pembelajaran mendalam memberikan perspektif baru tentang bagaimana komputer dapat mempelajari representasi yang tepat untuk data, dan bagaimana hal ini dapat digunakan untuk pemecahan masalah yang lebih kompleks.




*Gambar 18.2 Contoh Model Deep Learning*

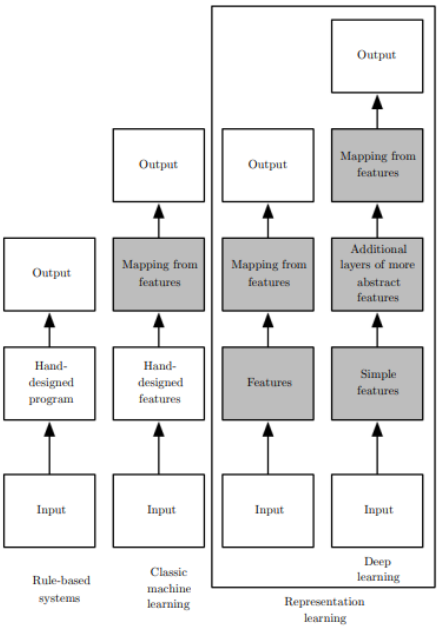
*Deep Learning* merupakan metode pembelajaran mesin yang mengatasi kesulitan dalam memahami data input sensorik mentah. Metode ini menguraikan pemetaan kompleks menjadi serangkaian pemetaan sederhana yang diwakili oleh berbagai lapisan model. Input dimasukkan pada lapisan yang terlihat, lalu lapisan tersembunyi mengekstrak fitur yang semakin abstrak dari gambar. Nilai pada lapisan tersembunyi ini tidak diberikan dalam data. Sebaliknya, model harus menentukan konsep yang

berguna untuk menjelaskan hubungan dalam data yang diamati. Setiap unit tersembunyi mewakili jenis fitur yang dihasilkan. Lapisan pertama dapat mengidentifikasi tepi dengan membandingkan kecerahan piksel tetangga. Lapisan tersembunyi kedua dapat mencari sudut dan kontur yang diperluas. Lapisan tersembunyi ketiga dapat mendeteksi seluruh bagian dari objek tertentu. Dengan memisahkan pemetaan kompleks menjadi serangkaian pemetaan sederhana, *Deep Learning* memungkinkan komputer untuk memahami dan mengenali objek dalam data sensorik mentah.



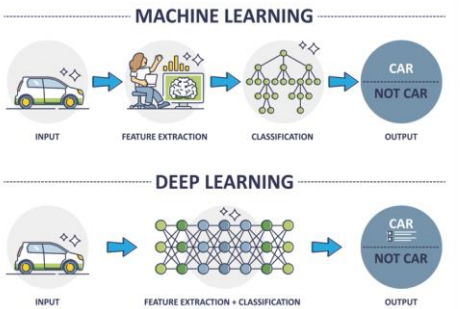

*Gambar 18.3 Jenis-Jenis Representasi Learning*

Gambar 18.3 Diagram Venn yang menunjukkan bagaimana *Deep Learning* adalah sejenis pembelajaran representasi, yang pada gilirannya merupakan jenis pembelajaran mesin, yang digunakan untuk banyak tetapi tidak semua pendekatan untuk AI. Setiap bagian dari diagram Venn menyertakan contoh teknologi AI.




*Gambar 18.4 Bagian dari sistem AI yang Terintegrasi*

Pada gambar 18.4 menjelaskan bahwa banyak disiplin ilmu perangkat lunak termasuk Computer Vision, signal dan audio processing, NLP, robotika, bioinformatika dan kimia, permainan video, mesin pencari, periklanan dan keuangan online.




*Gambar 18.5 Perbedaan Machine Learning dengan Deep Learning*

Gambar di atas menjelaskan perbedaan antara *Machine Learning* dan *Deep Learning*. *Machine Learning* (ML) dan *Deep Learning* (DL) merupakan dua subbidang kecerdasan buatan (AI) yang berbeda dalam tingkat kompleksitas dan aplikasinya. *Machine Learning* adalah teknik yang memungkinkan komputer belajar dari data dan membuat keputusan berdasarkan pola yang ditemukan, sering menggunakan algoritma seperti regresi linier, klastering k-means, atau support vector machines. Di sisi lain pembelajaran Mendalam (*Deep Learning*) adalah cabang dari *Machine Learning* yang lebih khusus menggunakan jaringan saraf tiruan yang dalam (*deep neural networks*) dengan banyak lapisan (*layers*) untuk menganalisis data. Sementara *Machine Learning* dapat berfungsi dengan baik pada dataset yang lebih kecil dan menggunakan fitur yang direkayasa secara manual, *Deep Learning* unggul dalam menangani data besar dan kompleks seperti gambar, suara, dan teks dengan menemukan fitur secara otomatis melalui proses pelatihan yang mendalam.

## *Convolutional Neural Network* (CNN)

CNN adalah jenis jaringan saraf yang digunakan untuk memproses data yang memiliki struktur topologis. Contohnya adalah data deret waktu dan data gambar. Jaringan ini sukses digunakan dalam berbagai aplikasi praktis. Konvolusi adalah operasi matematika linier khusus yang menjadi dasar dalam CNN. *Convolutional Network* menggunakan konvolusi sebagai pengganti perkalian matriks dalam setidaknya satu lapisan. Pada sesi ini, dijelaskan mengenai konvolusi, motivasi penggunaannya dalam jaringan saraf tiruan, serta operasi pooling yang umum digunakan. Biasanya, operasi konvolusi dalam jaringan konvolusional tidak sesuai dengan definisi konvolusi dalam bidang lain. Beberapa varian fungsi konvolusi

yang umum digunakan juga dijelaskan. Selain itu, penjelasan tentang aplikasi konvolusi pada berbagai jenis data dengan jumlah dimensi yang berbeda serta cara membuat konvolusi yang lebih efisien juga disertakan.

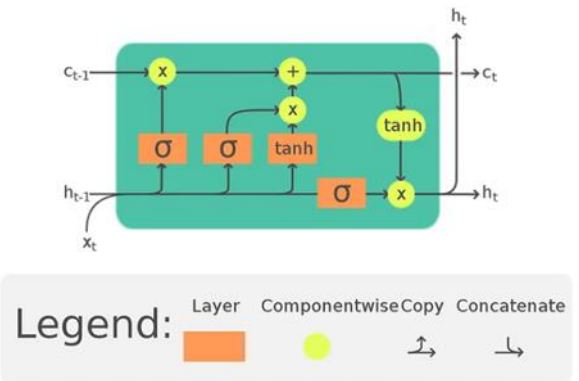
Dalam konteks ini, konvolusi adalah operasi matematika pada dua fungsi yang menghasilkan argumen berupa bilangan real. Sebagai contoh, kita dapat menggunakan sensor laser untuk melacak posisi pesawat luar angkasa. Sensor ini memberikan keluaran tunggal yang merepresentasikan posisi pada waktu tertentu. Namun, sensor laser bisa memberikan pembacaan yang tidak akurat karena kebisingan. Untuk mendapatkan estimasi posisi yang lebih akurat, kita dapat merata-ratakan beberapa pembacaan dengan memberikan bobot yang lebih besar pada pembacaan yang lebih baru. Hal ini dapat dilakukan dengan menggunakan fungsi pembobotan yang bergantung pada usia pembacaan. Dengan menerapkan operasi rata-rata tertimbang ini secara kontinu, kita dapat memperoleh estimasi posisi yang lebih akurat dari sensor laser.

Banyak pustaka *Machine Learning* menggunakan konvolusi sebagai implementasi korelasi silang. Dalam teks ini, kita menggunakan istilah konvolusi dan membahas apakah kita perlu membalik kernel dalam konteks tertentu. Dalam *Machine Learning*, algoritma pembelajaran akan menentukan nilai-nilai kernel yang sesuai dengan posisinya. Oleh karena itu, penggunaan konvolusi dengan membalik kernel akan menghasilkan pembelajaran kernel yang dibalik relatif terhadap kernel aslinya. Biasanya, konvolusi digunakan bersama fungsi lain, dan kombinasi fungsi-fungsi ini tidak memerlukan pembalikan kernel. Konvolusi diskrit dapat dipandang sebagai perkalian dengan matriks yang memiliki entri yang dibatasi agar sama satu sama lain. Misalnya, dalam konvolusi diskrit

univariat, setiap baris dari matriks adalah versi geser dari baris sebelumnya, yang dikenal sebagai matriks Toeplitz. Sedangkan dalam dua dimensi, matriks sirkuler blok ganda sesuai dengan konvolusi. Terlepas dari pembatasan elemen-elemennya, konvolusi umumnya terkait dengan matriks yang sangat jarang.

## *Long Short-Term Memory* (LSTM)

LSTM dibuat untuk mengatasi masalah vanishing gradient pada RNN saat memproses data berurutan yang panjang. Jaringan ini menggunakan mekanisme berulang untuk menyimpan representasi dari peristiwa input terbaru dalam bentuk aktivasi. Hal ini berguna dalam aplikasi seperti pemrosesan ucapan, kontrol non-Markovian, dan komposisi musik. Namun, algoritma yang umum digunakan untuk mempelajari apa yang harus dimasukkan ke dalam memori jangka pendek membutuhkan terlalu banyak waktu atau tidak memberikan hasil yang baik ketika jeda waktu antara input dan sinyal yang sesuai cukup lama. Terdapat beberapa metode yang ada saat ini, namun tidak memberikan keuntungan praktis yang jelas.



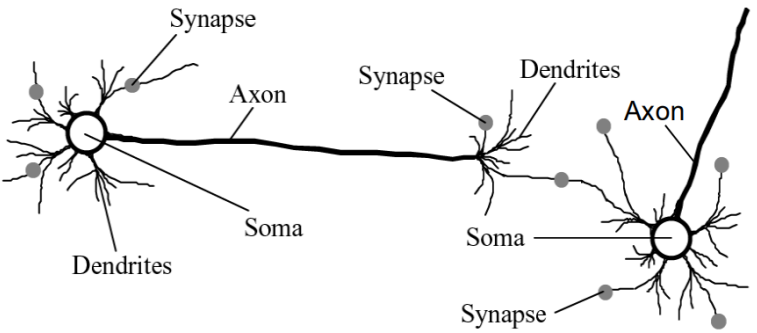

*Gambar 18.6 Arsitektur Long Short-Term Memory*

Konsep utama LSTM melibatkan cell state dan empat gates. Cell state berperan sebagai pembawa informasi penting yang telah diproses melalui semua gates dalam satu cell LSTM ke cell berikutnya. Forget gate menentukan informasi mana yang akan disimpan atau dibuang. Gate ini menerima hidden state dari cell sebelumnya serta informasi baru dari input saat ini, kemudian menggabungkannya dan memprosesnya menggunakan fungsi sigmoid yang menghasilkan nilai antara 0 dan 1. Hasil yang mendekati 0 menunjukkan bahwa informasi akan dibuang, sementara hasil yang mendekati 1 menunjukkan bahwa informasi akan disimpan. Input gate menerima informasi dari hidden state cell sebelumnya dan informasi baru dari input saat ini. Informasi ini digabungkan dan diproses menggunakan fungsi sigmoid dan tanh. Fungsi sigmoid menghasilkan nilai antara 0 dan 1 untuk menentukan informasi yang akan diperbarui; nilai mendekati 0 menunjukkan informasi kurang penting, sedangkan nilai mendekati 1 menunjukkan informasi penting. Fungsi tanh menghasilkan nilai antara -1 dan 1, membantu cell mempelajari informasi dengan lebih efektif. Output gate menentukan hidden state yang akan dikirim ke cell berikutnya. Gate ini menerima hidden state dari cell sebelumnya serta informasi baru dari input saat ini, menggabungkannya, dan memprosesnya menggunakan fungsi sigmoid. Cell state yang baru diproses melalui fungsi tanh, dan hasil dari fungsi tanh dikalikan dengan hasil fungsi sigmoid untuk menentukan informasi yang akan disimpan dalam hidden state yang baru. Hidden state dan cell state yang baru ini kemudian diteruskan ke cell berikutnya.

## *Artificial Neural Network* (ANN)

*Neural network* adalah model komputasi yang

terinspirasi dari otak manusia. Otak manusia terdiri dari jaringan neuron, yaitu unit dasar pemrosesan informasi yang saling terhubung.

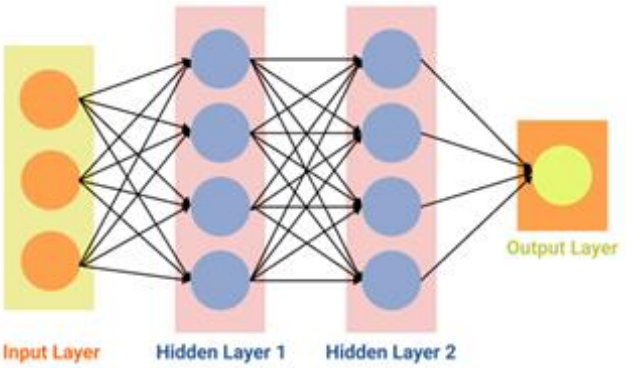



*Gambar 18.10 Biological Neural Network*

*Artificial Neural Network* terdiri dari sejumlah prosesor yang sangat sederhana, disebut juga neuron, yang mirip dengan neuron biologis di otak. Neuron-neuron ini terhubung oleh tautan berbobot yang mengirimkan sinyal dari satu neuron ke neuron lainnya. Sinyal output dikirimkan melalui koneksi keluar dari neuron tersebut. Koneksi keluar ini bercabang menjadi beberapa bagian yang mengirimkan sinyal yang sama. Cabang-cabang ini kemudian berakhir pada koneksi masuk ke neuron lain dalam jaringan.

*Artificial Neural Network* (ANN) memiliki kemampuan luar biasa untuk mengekstraksi informasi dari data yang kompleks atau ambigu, sehingga dapat menangani masalah yang tidak terstruktur dan sulit didefinisikan. Selain itu, ANN dapat melakukan komputasi secara paralel, meningkatkan kecepatan proses, dan menghasilkan representasi informasi secara otomatis selama proses pembelajaran. Namun, ANN memiliki keterbatasan dalam operasi numerik yang membutuhkan presisi tinggi dan sering memerlukan waktu pelatihan yang lama saat

mengolah jumlah data yang besar.




*Gambar 18.11 Arsitektur Artificial Neural Network*
(Sumber: Medim.com)

Neuron dalam tubuh manusia memiliki tiga komponen utama: dendrit, badan sel, dan akson. Dendrit menerima sinyal input dan dipengaruhi oleh bobot, badan sel melakukan komputasi dengan menggabungkan sinyal masuk dan bobot untuk menghasilkan sinyal output, sedangkan akson mengirimkan sinyal output ke neuron lain yang terhubung. Konsep ini memungkinkan *Artificial Neural Network* (ANN) untuk direpresentasikan dalam tiga bagian: lapisan input, lapisan output, dan lapisan tersembunyi yang mengolah input dari lapisan input menjadi format yang dapat diinterpretasikan oleh lapisan output. Hasil dari setiap lapisan tersembunyi ini dikenal sebagai aktivasi atau nilai dari node.

## *Recurrent Neural Network* (RNN)

*Recurrent Neural Network* (RNN) adalah salah satu jenis algoritma *Deep Learning* yang mengadopsi pendekatan berurutan atau sequential. RNN termasuk dalam kelompok *Artificial Neural Network* (ANN) dan sering digunakan dalam

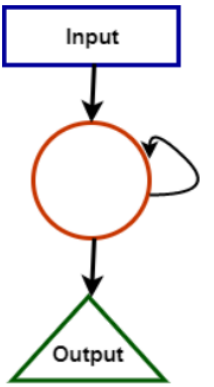
aplikasi seperti pengenalan suara (*speech recognition*) dan Pemrosesan Bahasa Alami (NLP). Dalam konteks *Deep Learning*, RNN digunakan untuk membangun model yang mirip dengan aktivitas neuron dalam otak manusia.

*Recurrent Neural Network* (RNN) pada dasarnya serupa dengan neural network konvensional, namun memiliki kemampuan tambahan berupa state memori pada setiap neuronnya. RNN dapat menyimpan memori atau ingatan (melalui feedback loop) yang membantu dalam mengenali pola data secara efektif dan menggunakan informasi tersebut untuk membuat prediksi yang tepat.

RNN menyimpan informasi dari waktu sebelumnya dengan menggunakan struktur yang mengulang, di mana output dari langkah sebelumnya digunakan kembali sebagai input, memungkinkan informasi dari waktu sebelumnya untuk tetap tersimpan. Ini menjelaskan mengapa algoritma ini disebut *Recurrent Neural Network* atau jaringan saraf berulang, karena melakukan komputasi matematis secara berurutan atau sequential.

### Cara Kerja Algoritma RNN

Algoritma *Recurrent Neural Network* bekerja dengan prinsip perulangan, dimana output dari suatu lapisan disimpan dan diumpankan kembali sebagai input untuk memprediksi output lapisan berikutnya.




*Gambar 18.12 Cara Kerja Algoritma RNN*
(Sumber: javapoint.com)

Pada neural network konvensional, setiap input dan output dianggap mandiri. Namun, dalam konteks tertentu seperti memprediksi kata berikutnya dalam sebuah kalimat, informasi dari kata-kata sebelumnya sangat penting. Untuk mengatasi hal ini, kita perlu menyimpan informasi mengenai kata-kata sebelumnya. RNN mengatasi masalah ini dengan menggunakan lapisan tersembunyi (hidden layer). Salah satu fitur kunci dari RNN adalah keberadaan hidden state, yang menyimpan informasi mengenai urutan data.

# BAB 19 PROYEKSI MASA DEPAN DATA SCIENCE

## Pendahuluan

Di era atau zaman digital dan globalisasi saat ini, data telah menjadi bagian dari komoditas yang cukup bernilai atau berharga. Hampir dan juga kebanyakan seluruh sektor atau lini dibidang kehidupan manusia menghasilkan data dari medsos (media sosial), mulai dari kegiatan seperti misalnya saja adalah belanja online sampai dengan informasi kesehatan. Jika data ini diproses dan dianalisa dengan benar, tentunya bisa memberikan informasi dan wawasan serta pandangan yang bermanfa'at untuk berbagai tujuan atau destinasi. Dengan demikian dapat kita sadari bahwa informasi telah menjadi sumber daya (*resource*) yang sangat berharga dan berperan penting dalam berbagai bidang termasuk diantaranya adalah bisnis, penelitian dan pengambilan keputusan (*decision making*).

Di sinilah peran penting informatika. Ilmu Data (Data science) yang mengkombinasikan banyak disiplin ilmu seperti misalnya ilmu komputer, statistika, matematika, infrastruktur dan tidak kalah pentingnga adalah pengetahuan domain untuk mengekstrak informasi dan makna dari data. Dengan menggunakan teknik dan algoritma yang berbeda, Data science bisa membantu memecahkan berbagai masalah (*problem solving*), memprediksi trend dan membuat keputusan yang lebih tepat dan akurat. Data science juga dapat membantu meningkatkan efisiensi operasional, meningkatkan kualitas produk dan juga layanan serta tentunya dapat meningkatkan kepuasan pelanggan (*customer*).

Kemajuan ilmu pengetahuan dan teknologi telah membuat masyarakat mudah dalam mengakses dan menggali informasi. Masyarakat dapat secara langsung berinteraksi dan juga saling mempengaruhi sehingga bisa melintasi atau melewati batas negara. Dengan perkembangan teknologi telah kehilangan batas, ruang dan waktu. Hasil dari perkembangan teknologi yang dapat ditemui seperti televisi, alat komunikasi (telepon, internet, dan lain-lain). Untuk televisi sendiri merupakan salah satu contoh dari produk kemajuan teknologi yang adopsinya secara luas menjadikannya sebagai media yang cukup menarik bagi periklanan (*advertisement*). Periklanan saat ini menjadi bagian cukup krusial dan penting yang digunakan sebagai strategi pemasaran. Bisa dikatakan jika tanpa iklan dan promosi, sudah pasti merek (*brand*) tidak akan dikenal oleh masyarakat luas. Periklanan punya peran sangat penting dalam memperkenalkan merek dan produknya kepada publik, sasaran serta untuk memperkuat posisi merk.

Selain itu, data science punya penerapan luas diberbagai sektor atau bidang, termasuk diantaranya adalah bisnis, ekonomi, kesehatan, pendidikan, juga sains dan teknologi. Dalam sektor bisnis, data science membantu serta memahami trend dan preferensi pelanggan (*customer*), meningkatkan produk, serta layanan dan bisa juga meningkatkan kepuasan para pelanggan. Di bidang ekonomi, data science sangat membantu untuk dapat memprediksi inflasi, menyiapkan perkiraan atau estimasi ekonomi dan menganalisis faktor yang bisa mempengaruhi perekonomian. Bidang lainnya seperti medis, data science sangat membantu dalam menganalisa data pasien dan bisa menentukan penyakit, mendiaknosa, pengobatan dan pencegahan. Untuk bidang pendidikan, data science dapat membantu memantau atau memonitor dan

meningkatkan prestasi siswa/i, membuat perencanaan dalam pembelajaran yang efektif dan menentukan langkah atau strategi dalam meningkatkan hasil pembelajaran. Tidak kalah penting juga adalah ilmu pengetahuan dan teknologi, data science membantu untuk menelaah juga memahami dan memecahkan masalah yang cukup kompleks di berbagai bidang ilmu pengetahuan dan teknologi.

Data Science dalam era atau zaman saat ini memainkan peran cukup penting dan signifikan dalam menguak tabir wawasan dan makna yang tersembunyi dari data. Oleh sebab itu, kedepannya data science menjadi hal yang krusial dalam mempersiapkan diri dan memanfaatkan potensinya secara maksimal. Data Science akan secara kontinu atau terus menerus meningkatkan perannya dalam memaksimalkan serta mengoptimalkan operasi, memprediksi trend dan dapat membuat keputusan yang tentunya lebih terinformasi serta bisa membantu dalam memecahkan masalah yang cukup kompleks atau rumit dalam berbagai bidang ilmu pengetahuan dan teknologi.

Data science secara umum dapat didefinisikan sebagai suatu studi tentang bagaimana cara untuk mengumpulkan, membersihkan, memproses, menganalisis, dan juga memvisualisasikan data untuk memperoleh informasi dan wawasan yang berguna. Pada Proses ini melibatkan beberapa langkah mulai dari pengumpulan data, pembersihan, pemetaan data, pemodelan hingga evaluasi model. Data science tidak hanya tentang pengumpulan data (*data collection*), akan tetapi melibatkan analisis dan interpretasi data dalam memperoleh atau menggali informasi yang bermakna dan tentunya berguna.

Data Science secara umum dapat diterapkan pada berbagai jenis data, baik itu data terstruktur seperti misalnya

data tabular maupun data tidak terstruktur seperti teks, audio dan gambar. Data terstruktur termasuk didalamnya seperti data statistik, data transaksi juga data kesehatan bisa dianalisis menggunakan suatu teknik statistik dan analisis data yang mengarah ke tradisional. Sementara itu data yang tidak terstruktur seperti teks, gambar dan audio memerlukan teknik analisis yang lebih rumit/kompleks dan juga spesifik, sebagai contoh analisis gambar, analisis teks dan analisis audio.

*Output* dari analisis data science dapat ditampilkan dalam format atau bentuk, seperti report, dashboard dan visualisasi data yang menarik. Laporan dapat berbentuk analisis yang cukup detail dan teknis, sementara itu dashboard dapat berupa visualisasi yang lebih interaktif dan gampang dipahami. Visualisasi data yang menarik bisa membantu dalam memahami pola dan trend yang terjadi dalam data, serta dapat membantu dalam membuat suatu keputusan yang lebih terinformasi. Informasi dan Teknologi telah menjadi bidang penting dan menarik di era digital ini yang penuh atau sarat dengan data. Data science terdiri dari berbagai metode, algoritma, dan teknik untuk memproses, menganalisis, dan mengekstraksi wawasan berharga dari data yang sudah ada. Dapat dilihat lebih dekat data science dan penerapannya diberbagai sektor atau bidang, serta bagaimana data science tersebut dapat membantu memecahkan masalah yang rumit dan cukup kompleks dan dapat meningkatkan efisiensi operasional.

Data science punya peran penting dalam berbagai sektor dan bidang kehidupan, termasuk didalamnya seperti bisnis, kesehatan, manajemen, sains dan tentunya juga penelitian. Dalam bisnis, data science dapat membantu perusahaan untuk lebih memahami pelanggannya, meningkatkan operasinya, mengembangkan produk dan tidak kalah pentingnya adalah

layanan baru dan membuat suatu keputusan yang lebih strategis. Di bidang kesehatan, data science sangat membantu membantu dalam mendiagnosis penyakit, memprediksi risiko kesehatan, mengembangkan obat baru dan meningkatkan kualitas pelayanan kesehatan.

Di pemerintahan, data science dapat juga membantu untuk membuat suatu kebijakan yang lebih efektif, meningkatkan layanan untuk publik dan tidak kalah pentingnya adalah dapat memerangi kejahatan. Dalam sains dan penelitian, data science juga sangat membantu untuk menganalisis data ilmiah yang cukup rumit atau kompleks, mengungkap/menyingkap pola (pattern) dan trend baru, juga membuat suatu penemuan baru. Data science selain itu juga bisa membantu dalam meningkatkan kualitas penelitian, kemudian dapat pula meningkatkan efisiensi operasional serta meningkatkan kepuasan masyarakat.

Hal lainnya bahwa data science berperan penting dalam peningkatan efisiensi dari operasional, meningkatkan kualitas produk dan juga layanan, serta meningkatkan kepuasan pelanggan dan komunitas. Data science sering kali digunakan untuk membantu membuat suatu keputusan yang cukup strategis, meningkatkan daya saing dan meningkatkan keuntungan juga. Oleh sebab itu, data science punya peran penting dalam berbagai sektor atau bidang kehidupan dan punya potensi yang cukup besar dan potensial untuk meningkatkan kualitas hidup masyarakat.

## Tren/Update Terkini Data Science

Bidang atau sektor dari data science terus mengalami perkembangan yang boleh dibilang cukup pesat dengan terbitnya berbagai trend baru yang mentransformasi bagaimana

caranya data dapat dikumpulkan, diolah dan dianalisis. Di bawah ini adalah contoh beberapa trend terkini atau terupdate yang patut diperhatikan.

### *Artificial Intelligence* (AI) dan *Machine Learning* (ML)

AI dan ML adalah 2 (dua) teknologi untuk saat ini yang cukup berkembang pesat dan memberikan efek yang cuku signifikan dan besar pada data science. AI dan ML memungkinkan komputer dapat belajar dari data yang telah diberikan dan bisa mengambil keputusan sendiri (own decision) tanpa harus diprogram lagi secara terpisah. Hal ini akhirnya dapat membuka kemungkinan baru untuk analisis data yang tentunya lebih cerdas dan efisien.

1. Mengenal lebih dekat tentang Pola dan Anomali
   AI dan ML, kebanyakan digunakan dalam hal pengidentifikasian pola dan anomali dalam data yang susah/sulit ditemukan dengan menggunakan metode tradisional. Dengan penggunaan algorithma seperti pengelompokan, regresi, dan jaringan saraf, AI dan ML juga dapat membuka dan mengungkap pola yang sebelumnya tak terlihat dan mengidentifikasi anomali yang bisa membantu dalam mengambil keputusan yang lebih baik.
2. Prediksi dan Personalisasi
   AI dan ML, bisa juga digunakan untuk memprediksi trend dan kejadian atau peristiwa dimasa depan, seperti misalnya saja adalah perkiraan/estimasi penjualan, risiko kredit dan penyakit. Dengan menggunakan algorithma ARIMA, LSTM, data historis dan pohon keputusan, AI dan ML dapat membuat prediksi yang cukup akurat dan dapat membantu dalam mengambil suatu keputusan yang

lebih strategis.

3. Aplikasi AI dan ML dalam Data Science
   AI dan ML memiliki atau punya aplikasi yang cukup luas dalam data science, seperti contoh di bawah ini.
   a. Pengenalan pola dan anomali → AI dan ML bisa digunakan untuk mengidentifikasi pola dan anomali dalam data yang cukup sulit ditemukan dengan metode tradisional.
   b. Prediksi → AI dan ML dapat juga digunakan untuk memprediksi trend dan kejadian dimasa depan, seperti prediksi penjualan, risiko kredit, dan penyebaran penyakit.
   c. Personalisasi → AI dan ML dapat digunakan dalam memberikan suatu rekomendasi dan personalisasi yang cukup akurat kepada user, seperti misalnya rekomendasi produk konten dan layanan.
   d. Optimasi → AI dan ML selain dari aplikasi yang di atas dapat digunakan juga dalam pengoptimalan bisnis proses, seperti misalnya optimasi routing, inventory management dan SCM (*Supply Chain Management*).
   e. Analisis → AI dan ML digunakan juga dalam melakukan analisis yang lebih cerdas dan efisien, seperti misalnya analisis text, analis sentiment dan analisis image.

   Jika dilihat dari beberapa tahun terakhir ini, kecerdasan buatan (AI) dan ML telah menjadi suatu bagian integral dari ilmu informasi yang banyak membantu untuk peningkatan efisiensi operasi, peningkatan kualitas produk dan peningkatan layanan, juga dapat meningkatkan kepuasan pelanggan dan masyarakat.

Dengan demikian, AI dan ML punya potensi yang besar dalam meningkatkan kualitas hidup masyarakat dan dapat membantu dalam memecahkan masalah (*problem solving*) yang cukup kompleks.

## Big Data dan Analisis Big Data

Seperti diketahui bersama bahwa jumlah data yang dihasilkan tiap detik, tiap menit, tiap jam dan seterusnya kian bertambah secara eksponensial dan mengarah pada era *Big Data*, data ini bisa berasal dari bermacam - macam sumber termasuk didalamnya adalah media sosial, market place, sensor, audio, video, stream, log file, IoT dan lain sebagainya, sebagai contoh saja adalah "Google" untuk saat ini dapat memproses 24 petabytes, 1 petabytes itu sama dengan 1000 terabytes data pertahun, selanjutnya lebih dari 800 juta pengguna Youtube mengunggah (*upload*) lebih dari 1 jam video per detiknya, belum lagi pengguna Facebook mengunggah 10 juta foto perjam, maka jangan heran jika laju pertumbuhan informasi yang dihasilkan dan disimpan adalah 4 (empat) x pertumbuhan ekonomi dunia.

*Big Data* mengarah atau mengacu pada kumpulan data yang besar sekali serta kompleks yang sulit diproses jika menggunakan metode tradisional. *Big Data* Analytics butuh suatu teknik dan alat khusus untuk memproses data dalam jumlah yang sangat besar dan kompleks ini.

*Big Data* punya ciri khas atau karakteristik yang unik seperti volume yang sangat besar, kecepatan yang sangat tinggi dan variasi yang sangat luas. Oleh sebab itu analisis *Big Data* membutuhkan suatu teknik dan alat khusus dalam menangani kumpulan data yang sangat besar dan kompleks tersebut. *Big Data* punya aplikasi yang cukup luas dalam Data Science, termasuk:

1. Analisis Medsos (Media Sosial)
   Digunakan untuk menganalisis opini publik (public opinion), memantau /monitoring trend, memahami sifat dan juga perilaku konsumen. Dengan bantuan analisis *Big Data*, organisasi/perusahaan/lembaga bisa memahami preferensi dari konsumen, kemudian merespons perubahan pasar dengan cepat, dan dapat mengidentifikasi celah atau peluang bisnis baru.
2. Analisis Genomik
   Untuk menganalisis data genomika terkait dengan struktur dan fungsi organisme manusia dalam mendiagnosis penyakit, mengembangkan obat - obatan baru dan memahami perkembangan evolusi manusia. Analisis *Big Data* memungkinkan para peneliti dapat menemukan pola dan trend yang tersembunyi dalam data genom manusia sehingga dapat menjadi acuan dalam pengembangan obat -obatan yang lebih efektif dan tepat.
3. Analisis Lalu Lintas (*Traffic*)
   Digunakan juga dalam menganalisis pola lalu lintas, mengoptimalkan rute-rute transportasi dan mengurangi atau mengurai kemacetan. Dengan bantuan atau support dari analisis *Big Data*, organisasi/perusahaan/lembaga bisa memahami pola lalu lintas, mengoptimalkan rute transportasi dan dapat mengurangi atau mengurai kemacetan, sehingga bisa meningkatkan efisiensi operasional dan kepuasan dari pelanggan.
4. Benefit atau keuntungan Analisis *Big Data*
   a. Dapat memberikan pengetahuan atau wawasan cukup dalam mengenai perilaku pelanggan, trend pasar dan juga model bisnis.
   b. Efisiensi operasional, pengoptimalan rantai

pasokan dapat ditingkatkan.

c. Kepuasan pelanggan dapat ditingkatkan juga dengan memahami preferensi konsumen dan merespon perubahan pasar tersebut dengan cepat.
d. Kemampuan dalam mengambil keputusan yang lebih baik menjadi semakin meningkat dan efektif, dengan menggunakan data yang terkumpul.

Jika melihat kebelakang beberapa tahun ini, analisis *Big Data* untuk saat ini sudah tidak dapat dipungkiri lagi menjadi bagian integral dari data science, yang sangat membantu sekali untuk meningkatkan efisiensi operasi, meningkatkan kualitas dari produk dan juga layanan untuk meningkatkan kepuasan pelanggan dan tentunya masyarakat. Karenanya, analisis *Big Data* punya potensi yang cukup besar dalam meningkatkan kualitas hidup masyarakat dan dapat membantu untuk memecahkan masalah yang cukup kompleks atau rumit sekalipun.

### Komputasi Awan (*Cloud Computing*) dan *Internet of Things* (IoT)

Layanan "*cloud computing*" memungkinkan data dapat disimpan dan diproses pada server dilokasi/tempat jarak jauh, sehingga bisa diakses tanpa batas ruang dan waktu, dimana saja dan kapan saja.

Untuk IoT sendiri mengacu (*revert*) pada jaringan perangkat (*device*) yang saling terkoneksi dan mampu menghasilkan data *real-time* online. Kombinasi atau kolaborasi Cloud Computing dan IoT dapat membuka peluang atau bisnis baru dalam mengumpulkan dan menganalisis data yang lebih efisien.

Aplikasi Komputasi Awan dan IoT dalam Data Science,

punya aplikasi yang cukup luas seperti misalnya di bawah ini.

1. Analisis Data Real-Time-Online
   Digunakan untuk memantau proses industri/manufaktur, melacak (*Track*) pergerakan aset dan dapat mendeteksi masalah secara dini.
2. Analisis Data Terdistribusi (*Distributed Data Analysis*)
   Komputasi Awan diberbagai server, bisa meningkatkan skalabilitas dan kinerja analisis data.
3. Analisis Data Sensor (*Sensor Data Analysis*)
   *Internet of Things* (IoT) dapat menghasilkan data sensor analisis yang bisa digunakan dalam memantau berbagai kondisi atau keadaan, seperti misalnya saja adalah suhu, kelembaban dan kualitas udara.

Keuntungan (*advantage*) kombinasi Komputasi Awan dan IoT punya keuntungan yang signifikan, seperti di bawah ini.

1. Meningkatkan efisiensi operasional dengan mengurangi biaya (*cost reduction*) infrastruktur dan juga dapat meningkatkan skalabilitas.
2. Peningkatan kualitas data dengan mengumpulkan data (*Collecting Data*) yang cukup luas dan sangat cepat.
3. Dapat meningkatkan analisis data dengan penggunaan algorithma yang cukup kompleks atau rumit dan tentunya dapat menjadi lebih cepat.
4. Keamanan data dapat terjaga dan meningkat menjadi lebih aman dengan teknologi menggunakan enkripsi dan juga autentikasi yang lebih baik lagi.

Jika kita tela'ah dalam beberapa tahun terakhir ini, kombinasi atau gabungan antara teknologi cloud dan IoT telah menjadi bagian integral dari teknologi informasi dan komunikasi yang cukup membantu dalam meningkatkan efisiensi operasional, meningkatkan kualitas produk dan tentunya adalah

layanan, serta dapat meningkatkan kepuasan pelanggan dan masyarakat. Oleh karenanya gabungan atau kombinasi teknologi Komputasi Awan dan IoT punya peluang dan potensi yang sangat besar dalam meningkatkan kualitas hidup masyarakat dan dapat membantu dalam memecahkan masalah yang kompleks.

## Teknologi yang Berkembang dalam Data Science

Seiring dengan trend yang sudah disebutkan di atas, beberapa teknologi baru juga terus muncul dan berkembang serta berinovasi dalam bidang Data Science.

### Algorithma dan Data Science Model yang Canggih

Para peneliti dan tentunya praktisi Data Science terus melakukan inovasi dan pengembangan algorithma dan model baru yang lebih canggih dan akurat untuk analisis data. Algorithma dan model ini memungkinkan nantinya untuk menangani data yang cukup kompleks dan heterogen serta menghasilkan prediksi/ramalan dan wawasan yang cukup akurat.

### Pengertian Algorithma dan Model Data Science

Adalah suatu metode atau prosedur yang digunakan untuk memproses data ilmiah, ilmu informasi dan teknologi sendiri merupakan ilmu yang punya tujuan untuk mengumpulkan, menganalisis dan mengekstrak pengetahuan yang berharga dari informasi yang ada. Algorithma dan model data science punya peran yang cukup penting dalam mengolah data, khususnya adalah *Big Data*. Dalam hal pemerosesan data, kita tentu saja menghadapi format data yang berbeda-beda dengan masalah yang berbeda-beda juga. Oleh sebab itu,

algorithma yang digunakan juga harus beradaptasi dengan data.

### Jenis Algoritma Data Science

Algorithma dalam data science dapat terbagi menjadi beberapa jenis, seperti pembelajaran:

1. Terawasi (*Supervised*)
2. Tanpa pengawasan (*Unsupervised*) dan
3. Penguatan

Keterangan seperti pembelajaran yang diawasi adalah algorithma yang biasanya digunakan dalam suatu proses klasifikasi dalam operasi *Machine Learning*. Algorithma ini bisa mengidentifikasi hubungan antara 2 variabel untuk memprediksi hasil yang baru. Contohnya adalah algorithma pembelajaran terawasi adalah regresi linier, hutan acak (Random Forest), peningkatan gradien, mesin vektor dukungan (SVM), regresi logistik, jaringan saraf tiruan (JST) dan K-Nearest Neighbour.

Terawasi (*Unsupervised*) learning adalah algorithma yang kebanyakan digunakan untuk mengidentifikasi suatu pola dalam kumpulan data yang berdasarkan persamaan atau perbedaan, walaupun tak ada pengidentifikasi yang diberikan. Contohnya seperti algorithma unsupervised learning, antara lain adalah K-Means *Clustering*, Hierarchical *Clustering* dan DBSCAN.

Sedangkan untuk reinforcement learning yaitu algorithma yang dipakai untuk mempelajari berdasarkan pengalaman dan mengembangkan suatu strategi dalam mencapai tujuan. Contoh implementasi dari algorithma reinforcement learning yaitu Q-learning, SARSA dan policy gradient.

Kelebihan (*advantage*) Algorithma dan Model Data

Science yang signifikan, jika dilihat di antaranya adalah:

1. Dapat meningkatkan analisis data dengan menggunakan algoritma yang lebih kompleks atau rumit dan lebih cepat.
2. Keamanan data semakin meningkat dengan penggunaan teknologi enkripsi dan juga autentikasi yang lebih baik.
3. Dapat meningkatkan juga, efisiensi operasional dengan mengurangi biaya infrastruktur dan meningkatkan skalabilitas.
4. Selain itu dapat meningkatkan juga, kualitas data dengan mengumpulkan data yang lebih luas dan lebih cepat.

Dalam kurun waktu tahun terakhir, algorithma dan model data science sudah menjadi suatu bagian integral dari sistem dan informatika, yang dapat membantu untuk meningkatkan efisiensi operasi, meningkatkan kualitas produk dan juga layanan, serta meningkatkan kepuasan kepada pelanggan dan juga tentunya bagi masyarakat. Oleh sebab itu, algorithma dan model data science punya potensi besar dalam meningkatkan kualitas hidup masyarakat dan membantu memecahkan masalah yang cukup kompleks dan rumit.

## Platform dan Alat Data Science yang Mutakhir

Bermacam-macam platform dan juga alat Data Science yang baru terus bermunculan, saling menawarkan beragam kemudahan penggunaan, skalabilitas, dan fitur yang canggih. Platform dan alat ini memungkinkan para pelaku seperti profesional Data Science untuk bekerja dengan lebih efisien dan produktif lagi.

### Pengertian Platform dan Alat Data Science

Pengertian Platform dan alat data science, adalah suatu

aplikasi yang digunakan dalam memproses dan untuk menganalisis data. Ilmu informasi sendiri adalah merupakan suatu ilmu yang dibuat dan dibangun mengacu atau berdasarkan pada ilmu matematika, statistika dan tentunya komputer. Pengetahuan yang sekarang populer ini bermanfa'at untuk perusahaan dalam mengambil suatu keputusan dalam meningkatkan strategi bisnisnya. Begitu banyak industri dan perusahaan/organisasi/lembaga yang sudah mulai mempelajari dan menerapkan algorithma data.

### Jenis Platform dan Alat Data Science

Terbagi menjadi beberapa jenis, seperti contoh di bawah ini.

1. BigML
   Merupakan platform data science yang menyediakan lingkungan GUI (Graphical User Interfaces) berbasis cloud untuk menangani algorithma pembelajaran mesin (*Machine Learning*). BigML memungkinkan perusahaan /lembaga/organisasi menggunakan algorithma untuk melakukan analisis risiko, perkiraan penjualan, dan pemodelan prediktif seperti misalnya saja pengelompok-kan, klasifikasi, analisis deret waktu dan lain sebagainya.
2. Tableau
   Alat visualisasi data yang dirancang atau didisain terutama untuk analisis bisnis. Tableau punya misi mempercepat penciptaan visualisasi interaktif pemrosesan data dan punya fitur yang cukup penting kemampuannya untuk berinteraksi dengan DB (*Database*), spreadsheet dan OLAP (*Online Analytical Processing*).
3. Pandas

Pandas merupakan library Python yang dipakai biasanya untuk manipulasi dan pengolahan data. Pandas ini memungkinkan pengguna (*user*) untuk mengumpulkan, menganalisis dan juga menggali wawasan berharga dari data yang besar dan kompleks.

4. Apache Superset
   Suatu platform data science yang menyediakan dan punya kemampuan untuk membuat dashboard yang cukup interaktif dan visualisasi data. Superset memungkinkan pengguna (user) dapat membuat laporan yang lebih detail dan analisis yang mendalam.
5. Dash
   Merupakan suatu library Python yang kebanyakan dipakai dan digunakan untuk membuat aplikasi web yang interaktif untuk analisis data. Dash juga memungkinkan pengguna (user) untuk membuat aplikasi yang sangat cepat dan efisien.

### Kelebihan (Advantage) Platform dan Alat Data Science

Memiliki atau punya kelebihan yang signifikan, seperti di bawah ini.

1. Dapat meningkatkan efisiensi operasional dengan mengurangi biaya infrastruktur dan meningkatkan skalabilitas.
2. Dapat meningkatkan kualitas data dengan cara mengumpulkan data yang lebih luas lagi dan lebih cepat.
3. Dapat meningkatkan analisis data dengan metode algorithma yang cukup kompleks dan lebih cepat.
4. Dapat meningkatkan keamanan data dengan teknologi enkripsi dan autentikasi yang lebih baik.
   Contoh aplikasi platform dan alat data science, punya

aplikasi yang cukup luas dalam berbagai bidang, seperti di bawah ini.

1. Analisis data *real-time* online, untuk membantu dan memantau proses industri/manufaktur, melacak pergerakan aset dan bisa mendeteksi masalah secara dini.
2. Analisis data terdistribusi, kebanyakan digunakan untuk mengoptimalkan rute transportasi dan dapat mengurangi kemacetan.
3. Analisis data sensor, dipakai untuk memonitor berbagai situasi dan kondisi, seperti misalnya suhu, kelembapan dan kualitas dari udara.

Platform dan alat Data science untuk saat ini telah menjadi bagian penting dari Data science dalam beberapa tahun terakhir, dan sangat membantu dalam meningkatkan efisiensi operasi, meningkatkan kualitas produk dan layanan, serta dapat meningkatkan kepuasan pelanggan dan masyarakat. Oleh sebab itu, platform dan alat data science punya potensi besar dalam meningkatkan kualitas hidup masyarakat dan juga membantu memecahkan masalah yang cukup kompleks.

## Visualisasi Data yang Interaktif dan Menarik

Visualisasi data yang efektif memiliki peran penting dalam mengkomunikasikan hasil dari analisis data kepada pemangku kepentingan (*stakeholders*). Visualisasi data yang interaktif dan komprehensif memungkinkan pengguna (*user*) dapat menjelajahi data dengan sangat mudah, memahami pola dan trend, juga mendapatkan pengetahuan dan wawasan yang cukup mendalam.

### Pengertian Visualisasi Data yang Interaktif dan Menarik

Merupakan proses menggambarkan suatu data dalam bentuk atau format visual yang memungkinkan pengguna (*user*) untuk dapat berinteraksi langsung dengan elemen visual tersebut. Visualisasi data yang interaktif dan menarik bisa berbentuk grafik, diagram, peta, atau bahkan visualisasi khusus lainnya yang memungkinkan pengguna (*user*) untuk memfilter, memanipulasi dan membandingkan data untuk mendapatkan wawasan yang lebih mendalam.

### Kelebihan Visualisasi Data yang Interaktif dan Menarik

Visualisasi data yang interaktif dan menarik punya beberapa kelebihan, seperti di bawah ini.

1. Memperkaya Pengalaman Pengguna (*user*)
   Pengguna (*user*) disini dapat berpartisipasi aktif dalam proses pencarian informasi, sehingga dapat menyesuaikan layar pencitraan sesuai dengan preferensi, memilih parameter yang mau dilihat dan mendapatkan informasi yang dibutuhkan. Hal ini memungkinkan pengalaman yang lebih kaya dan lebih personal saat memahami data.
2. Memungkinkan Analisis Mendalam
   Visualisasi data yang interaktif dan juga mendalam memungkinkan pengguna (*user*) dapat melihat data dari berbagai perspektif atau pandangan dan mengeksplornya lebih jauh lagi. Para pengguna (*user*) dapat memilih subkumpulan data tertentu, membandingkan variabelnya, dan menemukan pola atau trend yang mungkin tidak dapat terlihat dalam visualisasi statis.
   Dengan fitur interaktif ini, analisis data dapat lebih mendalam dan bisa memberikan gambaran yang lebih serbaguna.

3. Menyoroti Pola dan Hubungan yang Tersembunyi
   Visualisasi data yang interaktif dan komprehensif dapat juga mengungkap pola dan hubungan tersembunyi dalam data. Dengan menggunakan fitur interaktif seperti misalnya filter, animasi, atau kontrol lainnya, pengguna (user) dapat memfilter data guna memperoleh informasi yang lebih detail lagi dan memahami pola serta hubungan tersembunyi.

## Tantangan dalam Menggunakan Visualisasi Data yang Interaktif dan Menarik

Seperti diketahui meskipun visualisasi data yang interaktif dan menarik ini punya beberapa kelebihan, ternyata ada beberapa tantangan yang perlu diatasi, seperti misalnya:

1. Kompleksitas Pengembangan
   Memerlukan suatu keterampilan teknis yang lebih tinggi daripada visualisasi statis. Pemahaman mengenai bahasa pemerogramman dan visualisator data sangat penting dalam menciptakan visualisasi yang interaktif dan responsif.
2. Keterbatasan Teknologi dan Infrastruktur
   Penggunaan visualisasi data interaktif dan menarik, dibatasi juga oleh ketersediaan teknologi dan infrastruktur yang memadai. Koneksi internet yang lambat atau perangkat keras (*hardware*) yang tidak memadai dapat menjadi penghambat pengalaman pengguna (*user*) saat berinteraksi dengan visualisasi data.
3. Rendahnya Ketersediaan Data yang Berkualitas
   Meskipun kenyataannya data interaktif bisa memberikan wawasan yang lebih baik, akses ke data berkualitas tinggi

sepertinya masih terbatas. Tidak semua sumber data menyediakan versi interaktif yang dapat digunakan untuk visualisasi. Hal ini akhirnya dapat menjadi kendala atau penghalang untuk pengguna (*user*) yang mau mengeksplorasi dan memvisualisasikan data secara interaktif.

Untuk membuat visualisasi data yang interaktif dan menarik, berikut ini adalah cara atau beberapa tahapan yang perlu dilakukan yaitu:

1. Pilih Alat (*device*) atau Platform yang sesuai
   Platform visualisasi data yang sesuai dengan kebutuhan dan kemampuan yang akan dipilih, beberapa conntoh diantaranya adalah Tableau, Power BI, D3.js (untuk pengembangan web), dan lain sebagainya.
2. Persiapkan Data
   Yakinkan dan pastikan bahwa data yang akan divisualisasikan sudah berada dalam format yang sudah tepat dan juga bersih. Apabila data masih berantakan maka lakukanlah pembersihan data seperti menghapus (*delete*) duplikat, mengisi nilai yang hilang, atau merubah format data.
3. Pilih jenis visualisasi
   Tentukanlah jenis visualisasi yang paling cocok dan sesuai dengan data dan tujuan analisis. Contohnya, dapat dipilih diagram batang, grafik garis, peta, atau bahkan visualisasi khusus lainnya.
4. Desain Layout Visualisasi
   Rancangan atau desain layout visualisasi dengan mempertimbangkan kejelasan, mulai dari tata letak elemen, dan juga cara pengguna (*user*) akan berinteraksi dengan data. Tentukan kembali apakah akan ada elemen

interaktif seperti filter, animasi, atau kontrol lainnya.

5. Tambahkan Interaktivitas
   Gunakanlah selalu alat/fitur interaktif yang telah disediakan oleh platform untuk menambahkan elemen interaktif kedalam bentuk visualisasi. Ini mungkin saja termasuk diantaranya adalah filter, tooltip, animasi, atau kontrol lain yang memungkinkan pengguna (*user*) berinteraksi dengan data.
6. Uji dan Evaluasi
   Sebelum diimplementasikan dan dipresentasikan, lakukanlah dulu uji visualisasi data untuk memastikan bahwa interaktivitas sudah berfungsi sebagaimana yang diinginkan. Pastikan juga visualisasi dapat memberikan wawasan yang jelas dan bermanfaat.
7. Sajikan dan Bagikan
   Yang terakhir, sajikanlah visualisasi data secara efektif kepada pemangku kepentingan (stakeholder) atau audiens target. Pastikan juga bawha visualisasi dapat dengan mudah diakses dan dapat diinterpretasikan dengan jelas.

Dengan demikian, visualisasi data yang interaktif dan menarik punya potensi besar dalam meningkatkan efisiensi operasional, dapat meningkatkan kualitas produk dan juga tentnya layanan, serta meningkatkan kepuasan pelanggan dan Masyarakat.

## Potensi Aplikasi Data Science di Berbagai Sektor

Data Science punya potensi aplikasi yang cukup luas diberbagai sektor atau bidang, seperti di bawah ini.

1. Kesehatan dan Kedokteran
   Data Science saat ini dapat digunakan untuk

mendiagnosis penyakit, memprediksi risiko kesehatan, mengembangkan obat-obatan baru dan dapat meningkatkan kualitas layanan kesehatan. Misalnya saja:

a. Analisis genomik → Data Science bisa digunakan untuk menganalisis data genom manusia dalam mendiagnosis penyakit, mengembangkan obat-obatan baru, dan juga memahami evolusi manusia.
b. Analisis gambar medis → Data Science bisa digunakan juga untuk menganalisis gambar medis seperti sinar X dan MRI (Magnetic Resonance Imaging) untuk mendeteksi penyakit dan kelainan pada pasien.
c. Pemantauan kesehatan pasien → Data Science selain itu juga dapat digunakan untuk memantau kesehatan pasien secara *real-time online* dan memberikan peringatan (*alert*) secara dini jika nanti terjadi masalah kesehatan.

2. Keuangan dan Bisnis

   Di bidang keuangan dan bisnis, data science juga memiliki peran penting dalam penerapan yang luas dan potensi besar untuk meningkatkan keuntungan dan efisiensi operasional, seperti contoh di bawah ini.

   a. Analisis Risiko Kredit

      Data science pada analisis risiko kredit biasanya digunakan untuk menganalisis data keuangan pelanggan (*customer*) untuk menilai risiko kredit dan membuat keputusan dalam pemberian pinjaman. Dengan menggunakan algorithma *Machine Learning* dan penggunaan data analysis, data science bisa membantu mengidentifikasi pola risiko kredit dan dapat membuat prediksi yang

tentunya lebih akurat.

b. Deteksi Penipuan
   Data Science ternyata dapat juga digunakan dalam mendeteksi aktivitas penipuan dalam transaksi keuangan. Dengan menggunakan analisis data dan algorithma *Machine Learning*, data Science bisa membantu dalam mengidentifikasi transaksi yang tidak normal dan dapat menghentikan penipuan sebelumnya.
c. Penetapan harga yang optimal
   Tidak hanya seperti sebelumnya di atas, data science ternyata dapat digunakan juga untuk penentuan harga produk dan layanan yang optimal dalam memaksimalkan keuntungan. Dengan menggunakan data analysis dan algoritma *Machine Learning*, data science bisa membantu dalam mengidentifikasi pola harga yang cukup efektif dan dapat membuat prediksi yang lebih akurat.

Benefit atau keuntungan dari data science dalam keuangan dan bisnis, misalnya saja adalah:

1. Meningkatkan efisiensi operasional
   Dapat membantu dalam meningkatkan efisiensi operasional dengan cara mengoptimalkan proses bisnis dan mengurangi biaya.
2. Meningkatkan Keuntungan
   Membantu untuk meningkatkan keuntungan dengan mengembangkan produk dan juga layanan yang efektif dan mengoptimalkan harga.
3. Meningkatkan Kualitas Produk dan Layanan
   Membantu dalam meningkatkan kualitas produk dan juga

layanan dengan cara mengumpulkan dan menganalisis data pelanggan dan meningkatkan pengalaman pelanggan.

4. Meningkatkan Keamanan
   Membantu dalam hal peningkatan keamanan dengan mendeteksi penipuan dan juga menghentikan aktivitas tidak sah.

Dengan demikian, dapat disimpulkan bahwa data science memiliki potensi besar dalam meningkatkan keuangan dan bisnis, serta bisa membantu dalam mengembangkan strategi yang lebih efektif untuk dapat meningkatkan efisiensi operasional.

## Industri dan Manufaktur

Dalam dunia industri dan manufaktur, data science punya peranan juga untuk membantu dalam peningkatan efisiensi produksi, menekan bahkan mengurangi biaya operasional, dan juga bisa meningkatkan kualitas produk. Dalam industri dan manufaktur, data science punya aplikasi yang cukup luas dan potensi yang amat besar didalam peningkatan keefektifan operasional dan juga tentunya bisa meningkatkan kualitas produk, berikut ini adalah salah satu keuntungannya:

1. Prediktif Maintenance.
   Digunakan biasanya untuk memprediksi kegagalan mesin (*machine failure*) dan juga dapat melakukan pemeliharaan preventif untuk menghindari downtime dan kerugian atau kegagalan saat proses produksi. Dengan penggunaan analisis data dan juga algorithma *Machine Learning*, data science bisa membantu untuk mengidentifikasi pola kegagalan mesin dan tentunya dapat juga membuat prediksi yang lebih akurat.

2. Optimasi Proses.
   Pada optimasi proses, data science disini berperan atau digunakan untuk mengoptimalkan proses produksi mulai dari bahan mentah (raw material) sampai dengan barang jadi (finish good) dan meningkatkan efisiensi. Dengan menggunakan analisis data dan algorithma optimasi, data science dapat membantu dalam mengidentifikasi area/wilayah yang dapat ditingkatkan efisiensi dan membuat keputusan yang lebih strategis.
3. Kontrol Kualitas.
   Dalam mengontrol kualitas, data science juga ternyata dapat membantu dalam memonitor atau memantau kualitas produk dan mendeteksi cacat produk (*product defects*) secara dini atau awal. Dengan penggunaan analisis data dan juga algorithma *Machine Learning*, data science dapat banyak membantu dalam menyelesaikan permasalahan dan bisa mengidentifikasi pola cacat produk dan membuat prediksi yang tentunya lebih akurat.

Di bawah ini adalah beberapa keuntungan (*advantage*) dari data science dalam manufaktur dan dunia industri, dimana data science punya beberapa kelebihan dan benefit dalam manufaktur dan juga dunia industri, seperti:

1. Meningkatkan Efisiensi Produksi
   Membantu untuk meningkatkan efisiensi produksi dengan mengoptimalkan proses produksi dan mengurangi biaya operasional (*cost reduction*).
2. Meningkatkan Kualitas Produk
   Membantu untuk peningkatan kualitas produk dengan memonitor atau memantau serta dapat mendeteksi cacat produk secara dini.
3. Meningkatkan Keamanan (*Security*)

Membantu dalam peningkatan keamanan dengan memprediksi atau meramal kegagalan mesin (*Machine Failure*) dan melakukan pemeliharaan preventif (*Preventive Maintenance*).

4. Meningkatkan Keputusan (*Decision*)
   Membantu untuk meningkatkan keputusan (*decision*) dengan melakukan identifikasi area atau wilayah tertentu yang dapat ditingkatkan efisiensinya dan juga membuat suatu keputusan yang cukup strategis.

Sehingga dapat dikatakan bahawa, data science disini punya potensi yang sangat besar untuk peningkatan keefektifan operasional dan juga dapat meningkatkan kualitas produk dalam baik itu dalam dunia industri maupun manufaktur.

## Pemerintahan (*Government*) dan Kebijakan Publik (*Public Policy*)

Seperti sudah disinggung sebelumnya di atas, ternyata data science bisa juga membantu pemerintah dalam membuat kebijakan yang lebih efektif dan strategis guna untuk meningkatkan pelayanan publik, dan juga memerangi kejahatan. Didalam pemerintahan dan kebijakan publik, data Science ternyata punya aplikasi yang cukup luas dan memiliki potensi besar dalam peningkatan efektivitas kebijakan dan meningkatkan kualitas pelayanan publik.

1. Analisis Kebijakan Publik (*Public Policy*)
   Pada analisis kebijakan publik, data science bisa digunakan untuk mengevaluasi efektivitas kebijakan publik dan membuat kebijakan cukup efektif. Dengan menggunakan analisis data dan algorithma *Machine Learning*, data Science bisa membantu untuk

mengidentifikasi pola kebijakan yang efektif dan membuat prediksi/ramalan yang lebih akurat.

2. Penegakan Hukum (*Law Inforcement*)
   Tidak sampai disitu saja, ternyata data Science bisa digunakan dalam menganalisis data kriminalitas dan dapat mengidentifikasi pola kejahatan. Dengan menggunakan analisis data dan algorithma *Machine Learning*, data science cukup membantu sekali dalam mengidentifikasi pola kejahatan dan membuat prediksi/peramalan yang tentunya lebih akurat.
3. Perencanaan Kota (*City Planning*)
   Selain itu, data science juga bisa digunakan dalam perencanaan kota yang lebih efisien dan berkelanjutan. Sekali lagi dengan menggunakan analisis data dan algorithma *Machine Learning*, data science bisa membantu untuk mengidentifikasi area/wilayah yang dapat ditingkatkan efisiensinya dan membuat suatu keputusan yang lebih strategis tentunya.
4. Benefit Data Science dalam Pemerintahan dan Kebijakan Publik
   Data Science punya beberapa benefit atau keuntungan dalam pemerintahan dan juga kebijakan publik seperti di bawah ini.
   a. Meningkatkan Efektivitas Kebijakan
      Dapat membantu dalam meningkatkan efektivitas kebijakan dengan mengevaluasi efektivitas kebijakan dan membuat kebijakan yang lebih efektif.
   b. Meningkatkan Pelayanan Publik
      Membantu untuk meningkatkan pelayanan publik dengan mengidentifikasi area/wilayah yang dapat

ditingkatkan efisiensinya dan juga dapat membuat keputusan yang strategis.

c. Meningkatkan Keamanan
   Membantu untuk meningkatkan keamanan (*Security*) dengan menganalisis data kriminalitas dan juga mengidentifikasi pola (pattern) kejahatan.
d. Meningkatkan Keputusan (*Decision*)
   Data Science bisa juga membantu untuk meningkatkan keputusan dengan mengidentifikasi area/wilayah yang dapat ditingkatkan efisiensinya dan bisa membuat suatu keputusan yang cukup strategis.

Dengan demikian kesimpulan yang dapat ditarik, bahwa data science punya peluang atau potensi yang cukup besar didalam peningkatan efektivitas kebijakan dan meningkatkan kualitas pelayanan publik didalam pemerintahan, juga kebijakan publik.

## Sains dan Penelitian

Bicara tentang data science dalam dunia sains dan peneltian, banyak sekali digunakan dalam menganalisis data ilmiah yang cukup kompleks, menemukan pola dan trend baru, serta membuat suatu penemuan baru. Dalam dunia sains dan penelitian, data science juga punya aplikasi yang cukup luas dan potensi yang besar dalam meningkatkan pengetahuan maupun teknologi, diantaranya adalah:

1. Analisis Data Astronomi
   Untuk saat ini, data astronomi banyak digunakan guna mempelajari alam semesta dengan bantuan data science. Dengan menggunakan analisis data dan algorithma

*Machine Learning*, data science sangat membantu untuk mengidentifikasi pola dan juga trend baru dalam data astronomi, serta bisa membuat prediksi/ramalan yang lebih akurat lagi tentang perilaku bintang dan galaksi yang ada di alam semesta ini.

2. Penelitian Iklim
   Data Science selain dari yang telah disebutkan di atas, digunakan juga dalam menganalisis data iklim untuk memahami perubahan iklim yang terjadi dan mengembangkan solusi yang tepat untuk mengatasinya, tentunya dengan menggunakan analisis data dan algorithma *Machine Learning*, data science bisa banyak membantu untuk mengidentifikasi pola dan trend baru dalam data iklim, serta membuat suatu prediksi yang lebih akurat tentang perubahan iklim.
3. Penelitian Genomik
   Pada penelitian genomik, data science biasanya banyak digunakan dalam menganalisis data genom manusia, untuk mempelajari penyakit dan juga pengembangan obat - obatan baru. Sekali lagi dengan cara menggunakan analisis data dan algorithma *Machine Learning*, data Science bisa membantu untuk mengidentifikasi pola dan trend baru dalam data genom, juga dapat membuat prediksi/peramalan yang sangat akurat tentang suatu penyakit dan mengembangkan obat - obatan baru.
4. Keuntungan (Advantage) Data Science dalam Sains dan Penelitian
   Data science punya beberapa benfit atau keuntungan, seperti di bawah ini.
   a. Meningkatkan Efisiensi Operasional
      Membantu untuk meningkatkan efisiensi

operasional dengan cara mengoptimalkan proses analisis data dan mengurangi biaya operasional.

b. Meningkatkan Kualitas Penelitian
   Dapat membantu untuk meningkatkan kualitas penelitian dengan cara mengumpulkan dan menganalisis data yang lebih akurat dan juga lebih cepat.
c. Meningkatkan Keamanan
   Data Science juga dapat membantu untuk meningkatkan keamanan dengan cara mengidentifikasi pola dan trend baru dalam data, serta membuat prediksi yang cepat dan tentunya lebih akurat.
d. Meningkatkan Keputusan (*Decision*)
   Selain dari keuntungan di atas, data science bisa juga membantu untuk meningkatkan keputusan dengan mengidentifikasi pola dan trend baru dalam data, serta membuat prediksi yang cukup akurat.

Sehingga dapat dikatakan bahwa, data Science punya potensi yang cukup besar untuk peningkatan pengetahuan dan teknologi dalam sains dan penelitian.

## Tantangan Masa Depan

Data Science punya potensi dan kontribusi yang cukup besar untuk mengubah berbagai bidang atau sektor kehidupan dan dapat membawa kemajuan yang signifikan bagi manusia. Dalam era dan perkembangan teknologi yang pesat saat ini dan semakin banyaknya data yang tersedia, data science akan terus berkembang dan punya peran yang semakin penting dimasa depan.

Walaupun punya banyak potensi, data science tentunya tidak terlepas dari beberapa tantangan dimasa yang akan datang, diantaranya adalah:

1. Ketersediaan data → Ketersediaan data yang berkualitas dan relevan masih menjadi tantangan bagi banyak organisasi/perusahaan/institusi/lembaga.
2. Keterampilan dan talenta → Kebutuhan akan tenaga ahli atau expert dibidang data science yang terampil dan berbakat terus saja meningkat, kebalikannya adalah ketersediaan talenta/bakat yang mumpuni masih sangat terbatas.
3. Etika dan privasi data → Penggunaan data science harus dilakukan dengan pertimbangan etika dan juga privasi data individu.
4. Ketidakadilan algorithma → Algoritma Data Science bisa menghasilkan bias/semu dan ketidakadilan apabila tidak dirancang atau didisain dan digunakan dengan hati-hati.

Mitigasi terhadap tantangan ini akan menjadi kunci dalam memastikan data science bisa digunakan secara bertanggung jawab dan bermanfaat tentunya bagi seluruh umat manusia.

# BAB 20 PENGGUNAAN DATA SCIENCE DALAM BERBAGAI INDUSTRI DAN SEKTOR

## Pendahuluan

Inovasi yang terus berkembang di berbagai industri dan sektor membutuhkan teknologi modern untuk meningkatkan kemampuan beradaptasi dalam mengatasi berbagai tantangan bisnis di era digital. Penggunaan teknologi data science di era digital dalam berbagai industri dan sektor antara lain; *Artificial Intelligence*, *Machine Learning*, *Internet of Things*, data mining dan *Big Data* sudah sangat banyak diterapkan, dimana pengaruhnya sangat besar dan signifikan terhadap peningkatan daya saing bisnis.

Perbaikan sistem dan teknologi untuk meningkatkan efektifitas dan efisiensi dalam bisnis adalah tujuan yang harus dicapai untuk meningkatkan daya saing bisnis di berbagai industri dan sektor di indonesia. Data science bekerja dengan cara mengumpulkan, memilah, dan menganalisa data yang diperlukan kemudian dikomunikasikan untuk pengambilan suatu keputusan sehingga otomasi membutuhkan data science.

Perkembangan digital yang sangat pesat membuat data menjadi aset yang sangat penting. Data science menjadi salah satu solusi untuk dapat mengolah data secara cepat dan akurat yang bisa diterapkan di berbagai industri dan sektor antara lain perusahaan yang bergerak di bidang perbankan, manufaktur, jasa, lembaga pemerintahan, dan di berbagai industri dan sektor lainnya.

Di dalam dunia industri dan sektor penerapan data science memberikan banyak manfaat bagi perusahaan dalam menghadapi trend dan tantangan bisnis di masa depan, pemanfaatan teknologi di era digital seperti *Artificial Intelligence*, data mining, *Machine Learning*, *Internet of Things* dan *Big Data* akan memberi dampak besar bagi industri dimana tingkat akurasi data yang mampu disediakan dari teknologi ini sangat tinggi sehingga kemungkinan terjadinya resiko kerugian yang akan dihadapi dapat diprediksi lebih dini atau bahkan dapat dihindari oleh organisasi.

## Pentingnya Penggunaan Data Science di Berbagai Industri dan Sektor

Data Science telah menjadi kekuatan pendorong transformasi bisnis di era digital di berbagai industri dan sektor. Data Science sebagai alat pengolahan data canggih yang mengubah bagaimana proses bisnis berjalan secara otomasi untuk mendukung pengambilan keputusan bisnis di suatu organisasi.

Data Science kombinasi dari disiplin ilmu matematika, statistik, dan komputer yang bertujuan untuk membuat proses pengolahan data lebih cepat dan akurat, dimana data science merupakan penerapan dari berbagai aspek yaitu; metode ilmiah, proses bisnis, algoritma, dan sistem teknologi untuk mengekstraksi wawasan dan pengetahuan dari data yang melibatkan pemahaman mendalam terkait analisis data yang canggih yang digunakan organisasi dalam menyelesaikan berbagai masalah bisnis, , dan pembuatan model prediktif untuk mengoptimalkan pengambilan keputusan organisasi.

Dalam dunia bisnis di era digital, data adalah aset

perusahaan menjadi komponen yang sangat penting. Saat ini suatu keputusan bisnis harus didasarkan pada data yang valid, relevan, serta akurat sehingga perusahaan wajib menyediakan sistem pemrosesan dan penyimpanan data yang baik. Dari data yang telah diolah secara komputerisasi perusahaan dapat memperoleh berbagai informasi yang dibutuhkan untuk meningkatkan strategi pemasaran dan pengambilan keputusan diberbagai lini Perusahaan yang mendorong perusahaan menerapkan data science guna meningkatkan performa bisnis dalam menghadapi tantangan dan persaingan bisnis saat ini.

Penggunaan data science mendorong inovasi produk dan layanan. Otomasi memudahkan perusahaan memahami kebutuhan dan keinginan pelanggan. Melalui analisis data, perusahaan dapat mengembangkan produk yang lebih sesuai dengan harapan pelanggan dengan layanan yang lebih inovatif sehingga dapat memenangkan persaingan pasar.

Data science memungkinkan organisasi untuk mengidentifikasi pola dan tren bisnis yang lebih baik dengan akurasi lebih tinggi dibandingkan dengan pendekatan secara konvensional. Analisis data dilakukan secara menyeluruh sehingga perusahaan dapat memahami berbagai permasalahan yang ada dalam Perusahaan baik masalah internal maupun ekternal, perilaku konsumen, perubahan pasar, serta dinamika bisnis dalam beebagai industri dan sektor secara lebih baik.

### Penggunaan Data Science dalam Industri Perbankan

Data Science telah mengubah lanskap perbankan yang dinamis dengan cara yang canggih, menggabungkan kekuatan pada peningkatan kemampuan analisis data yang mendalam, bank dapat meningkatkan keamanan, dapat membangun model prediktif untuk mendeteksi pola fraud detection, lebih awal

mendeteksi transaksi yang mencurigakan dan memodelkan risiko yang kompleks dengan akurat antara lain risiko pada operasional bank, risiko kredit, dan risiko pasar sehingga bank dapat terhindar dari kerugian.

Penerapan teknik *Machine Learning* dalam mengelola dan menganalisis data konsumen memungkinkan bank memberikan rekomendasi personal berdasarkan perilaku dan preferensi nasabah. *Artificial Intelligence* dalam penerapannya mendukung costumer support, chatbot untuk layanan pelanggan selama 24 jam dapat mengoptimalkan penanganan berbagai keluhan nasabah, memberikan solusi secara otomasi yang lebih efisien dan efektif. Penggunaan data science di industri perbankan dapat meningkatkan pengelolaan data menjadi ebih baik sehingga bank dapat bersaing dalam lingkungan bisnis yang semakin kompetitif dan sangat dinamis.

### Penggunaan Data Science dalam Industri Manufaktur

Industri manufaktur merupakan kontributor perekonomian terbesar di Indonesia dengan kontribusi mencapai 20% dari total PDB menurut data Badan Pusat Statistik (BPS) sehingga sangat penting untuk menjadi fokus pemerintah dalam upaya meningkatkan pertumbuhan sektor manufaktur.

Saat ini penggunaan teknologi modern di industri manufaktur di Indonesia masih belum optimal adaptasi dan transformasi bisnis manufaktur dengan menggunakan perangkat digital dukungan penting untuk berinovasi di era industrialisasi digital saat ini. Pemanfaatan teknologi digital seperti *Artificial Intelligence* dan *Big Data* berdampak besar bagi industri manufaktur akan mempercepat keseluruhan proses kerja dan meningkatkan kualitas produk yang dihasilkan.

Secara keseluruhan proses produksi di industri manufaktur bergantung pada mesin, dengan membuat perencanaan dini akan mengoptimasi proses dengan pengambilan keputusan secara real time antara lain dalam memperkirakan kebutuhan produksi dengan lebih akurat, otomasi penjadwalan, dan visual inspection.

Dalam industri manufaktur downtime berpotensi menimbukan kerugian besar pada biaya produksi. Predictive maintenance membantu mengurangi biaya yang ditimbulkan dari perbaikan yang tidak diperlukan, contohnya biaya akibat terjadinya downtime. Secara keseluruhan proses produksi di industri manufaktur bergantung pada mesin, dengan membuat perencanaan dini akan mengoptimasi proses dengan pengambilan keputusan secara real time antara lain dalam memperkirakan kebutuhan produksi dengan lebih akurat, otomasi penjadwalan, dan visual inspection, dimana hampir secara keseluruhan proses produksi pada industri manufaktur mengandalkan mesin sehingga penggunaan teknologi modern bisa menjadi solusi terbaik yang harus terus dikembangkan untuk memajukan industri dibidang manufaktur.

### Penggunaan Data Science dalam *E- Commerce*

Saai ini bisnis *E-commerce* menghadapi persaingan yang sangat ketat, semakin banyak pemain bersaing untuk mendapatkan perhatian dan kepercayaan dari pelanggan. Saai ini konsumensangat dinamis Dimana perubahan perilaku pelanggan terjadi yang meiliki kecenderungan berbelanja menggunakan platform mobile secara online.

Penerapan data science pada *e-commerce* sangat dibutuhkan untuk meningkatkan efisiensi dan efektivitas operasional dalam meramalkan perilaku konsumen, memahami

tren pasar, dan mengidentifikasi peluang penjualan yang memungkinkan perusahaan untuk mengoptimalkan promosi secara lebih lebih personal bagi setiap pelanggan, merencanakan stok, mengatur harga, serta mendeteksi pola dan anomali untuk melindungi pelanggan pada saat melakukan transaksi online. Implementasi rekomendasi produk yang dipersonalisasi, dapat menciptakan pengalaman belanja secara lebih tepat terhadap kebutuhan dan preferensi unik setiap pelanggan, dengan pola permintaan setiap produk dengan menggunakan teknik forecasting secara efektif, *e-commerce* dapat mengurangi risiko kelebihan atau kekurangan persediaan,

Pebisnis *E-commerce* harus dapat beradaptasi dengan cepat untuk memenuhi kebutuhan dan preferensi konsumen Optimasi penjualan dapat menjadi kunci untuk memenangkan persaingan bisnis yang sangat dinamis. Dengan penggunaan data science, *e-commerce* dapat memahami lebih baik perilaku pelanggan, tren pasar, dan pola permintaan pelanggan, sehingga strategi bisnis dapat dinbuat secara berkelanjutan.

### Penggunaan Data Science dalam Layanan Kesehatan

Data science membantu memantau histori kesehatan pasien dengan merekam data secara real time yang memungkinkan tim medis untuk mendeteksi gejala penyakit sejak dini. Dalam suatu penelitian dijelaskan bahwa tubuh manusia menghasilkan data kurang lebih 2 terabyte per hari yang meliputi aktivitas otak, aktivitas sistem tubuh, detak jantung, dan data kesehatan lainnya. Teknologi modern dibutuhkan untuk membantu memantau histori kesehatan pasien dengan merekam data secara real time, berbagai alat kesehatan dengan teknologi yang inovatif memungkinkan dokter dapat memantau kondisi pasien dari jarak jauh dengan bantuan

data science, dokter dapat mengetahui kondisi pasien melalui sebuah perangkat teknologi kesehatan.

Industri kesehatan mengalami saat ini berkembang sangat pesat terutama didorang masa pandemi Covid-19 dari berbagai sisi antara lain perawatan dan pendeteksian pasien dengan mengkorelasikan setiap data yang berkaitan dengan gejala, kebiasaan, dan penyakit yang dimiliki pasien sebelumnya membantu dokter mengidentifikasi penyakit dan upaya penanganan serta pencegahan yang tepat saat menangani pasien, tindakan operasi pada pasien, pengembangan obat-obatan, laboratorium teknik untuk mengidentifikasi penyakit dari gambar seperti X-Ray, MRI, dan CT scan, serta berpengaruh dalam meminimalkan biaya kesehatan dengan *Electronic Health Records* (EHRs) untuk mengidentifikasi pola kesehatan pasien dan mencegah perawatan atau rawat inap yang tidak perlu, sehingga mengurangi biaya serta , efisiensi waktu baik secara manajemen data, dan mengurangi biaya operasional di bidang kesehatan.

# Daftar Pustaka


Afifah, Lutfia.2024. Apa itu Confusion Matrix di *Machine Learning*?. https://ilmudatapy.com/apa-itu-confusion-matrix/

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

Akhoondzadeh M. *Artificial Neural Network* (ANN) [Internet]. Available from: https://www.researchgate.net/publication/336999943

Babones, S. (2021). Methods for Quantitative Macro-Comparative Research. SAGE Publications Ltd.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. Computational Statistics & Data Analysis, 120, 70-83.

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

Bernardita Calzon. (2023, March). Your Modern Business Guide To Data Analysis Methods And Techniques. Https://Www.Datapine.Com/Blog/Data-Analysis-Methods-and-Techniques/#data-Analysis-Methods.

Bill Franks, 2020. "97 Things About Ethics Everyone in Data Science Should Know", O'Reilly Media, Inc.

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

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Brownlee, J. (2017, April 12). Data Preprocessing for *Machine Learning*. *Machine Learning* Mastery. Retrieved from https://machinelearningmastery.com/data-preparation-machine-learning-data-preprocessing/

Carmichael, I and Marron, J.S. 2018. Data Science vs. Statistics: Two Cultures? Japanese Journal of Statistics and Data Science. Doi: 10.1007/s42081-018-0009-3.

Chirag Shah, 2020. “A hands-on Introduction to Data Science”, Cambridge University Press.
Chollet, F. (2018). *Deep Learning* with Python. Manning Publications.
Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. International Statistical Review, 69(1), 21-26.
Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, 13(6), 377-387.
Contents 1. Why do we need *Recurrent Neural Network*?
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
Coursera staff. (2024, April). Data Analytics: Definition, Uses, Examples, and More. Https://Www.Coursera.Org/Articles/Data-Analytics.
Crook, T., & Esser, A. (2021). Advanced Guide to A/B Testing: How to Improve Your CRO and MVT Skills. Packt Publishing.
Davenport, T. H. , R. R. , W. J. , & N. A. (2018). Feature *Artificial Intelligence* for the real world 108 harvard business review. https://doi.org/10.1126/science.aac4520 .
Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186.
Dillman, D. A., Smyth, J. D., & Christian, L. M. (2022). Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. Wiley.

Dong, Y. 2023. Descriptive Statistics and Its Applications. Highlights in Science, Engineering and Technology. 47: 16 – 23.
Dr. Rr. Nur Fauziyah, SKM, MKM, RD, 2018. Analisis Data Menggunakan Uji Korelasi dan Uji Regresi Linier di Bidang Kesehatan Masyarakat dan Klinis. Politeknik Kesehatan Kemenkes Bandung.
Dr. Taufik Hanafi, D. R. (2021). From Data Science To AI Technology Augmanted Human Capability, Knoledge & Application in Indonesia. Jakarta: Perkumpulan Basis Data Indonesia.
Dr. Vladimir, V. F. (2018). Skala Pengukuran. Gastronomía Ecuatoriana y Turismo Local., 1(69), 5–24.
Drs. Ating Somantri & Sambas Alimuhidin, S.Pd. 2006. Aplikasi Statistika Dalam Penerapan. Bandung: Pustaka Setia.
Dyah Nirmala Arum Janie, 2012. Statistik Deskriptif & Regresi Linier Berganda Dengan SPSS. Semarang University Press: Semarang.
F.Rachmawati, Y.A.Sihombing, T.Septiyani, K.M.Putri, C.Widia, Yunike, A.e.Kusumaningrum, 2022. "Digitalisasi dalam Perawatan Kesehatan", Widina.
Fan, W., & Yan, Z. (2022). Factors affecting response rates of the web survey: A systematic review. Computers in Human Behavior, 26(2), 132-139.
Firmansyah, A., & Data, A. J. (n.d.). MODUL 2 " Jenis Data dan Skala Pengukuran". 1, 1–9.
Galton, F. (1889). Natural inheritance. Macmillan.
Gandomi, A., & Haider, M. (2015). Beyond the hype: *Big Data* concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007
Gede Aditra Pradnyana, K. A. (n.d.). Konsep Dasar Data Mining. Pustaka UT.
Géron, A. (2019). Hands-On *Machine Learning* with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly

Media.
Géron, A., 2019. “Hands-On *Machine Learning* with Scikit-Learn, Keras, and TensorFlow”.
Goldberg, Y. (2017). Neural network methods for *Natural Language Processing*. Synthesis Lectures on Human Language Technologies, 10(1), 1-309.
Goodfellow I, Bengio Y, Courville A. *Deep Learning*.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Grus, J. (2019). Data Science from Scratch Second Edition First Principles with Python. http://oreilly.com
Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2021). *Internet of Things* (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645-1660.
Gunal, M. M. (2019). Simulation and the fourth industrial revolution. In Simulation for Industry 4.0 (pp. 1-17). Springer, Cham.
Gupta, P and Tawar, N.V., 2020. The Impact and Importance of Statistics in Data Science. International Journal of Computer Applications. 176(24): 0975 – 8887.
Han, J., Kamber, M., & Pei, J. (2011). Data Preprocessing. In Data Mining: Concepts and Techniques (3rd ed., pp. 83-124). Morgan Kaufmann.
Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now?. *Big Data*, 6(3), 176-190.
Harris, H., Murphy, S. P., & Vaisman, M. (2021). Analyzing Data with Power BI and Power Pivot for Excel. Que Publishing.
Hartatik, B.Kwintiana, T.A.Nengsih, A.Baradja, 2023. “Data Science for Business (Pengantar dan Penerapan Berbagai

Sektor)", Sonpedia Publishing Indonesia.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
Haykin, S. (2009). Neural Networks and Learning Machines. Prentice Hall.
Heri Retnawati. 2017. Pengantar Analisis Regresi dan Korelasi. Bukittinggi (Makalah disajikan pada kegiatan Workshop Teknik Analisis Data Fakultas Ekonomi dan Bisnis IAIN Batusangkar di Rocky Hotel Bukittinggi, 25 Juli 2017.)
Heryana, A. (2020). Data Dan Skala Ukur Kuantitatif. 1–15.
Hirschberg, J., & Manning, C. D. (2015). Advances in *Natural Language Processing*. Science, 349(6245), 261-266.
Hochreiter S, Urgen Schmidhuber J¨. Long Short-Term Memory.
Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. Science, 349(6245), 253–255. https://doi.org/10.1126/science.aac4520
Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression. John Wiley & Sons.
Hunt-Isaak, I., Russell, J., & Hekstra, D. (2024). mpl-interactions: A Python Package for Interactive Matplotlib Figures. Journal of *Open-source* Software, 9(93), 5651. https://doi.org/10.21105/joss.05651
I made yuliara, 2016.Regresi linier berganda. Jurusan Fisika Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Udayana
Imoore. (2020). Intro to Exploratory Data Analysis (EDA) in Python. In kaggle. https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python
Irfan Whyudi, E. T. (2019). Teori dan Panduan Praktis Data Science dan *Big Data*. Bogor: LPPM Universitas Pakuan.

Jalajakshi, V and Myna, A.N. 2022. Importance of Statistics to Data Science. Global Transitions Proceedings. 3: 326–331.
Jhonson, R.A and Bhattacharyya, G.K. 2010. Statistics: Principles and Methods. United State of America: John Wiley & Sons, Inc.
Jones, R. A., & Bradley, M. P. (2021). Observational Research in Social Science: Theories, Methods, and Ethical Considerations. SAGE Publications.
Joseph M. Tandiallo. (2024). Exploratory Data Analysis (EDA) Using Python. In kaggle. https://medium.com/@teppan_noodle/exploratory-data-analysis-eda-using-python-f85938cb1810
Joseph Santoso. (2023). Ilmu data (Data Science) (Muhammad Sholikan (ed.); Pertama). Yayasan Prima Agus Teknik.
Jurafsky, D., & Martin, J. H. (2021). Speech and language processing (3rd ed.). Prentice Hall.
Kelleher, J. D., & Tierney, B. (2022). Data Science. MIT Press.
Kirk, R.E. 2008. Statistics: An Introduction. United State of America: Thomson Wadsworth.
Koetsier, J. (2020). 6 Future Trends in Data Science. Forbes. Retrieved from https://www.forbes.com/sites/johnkoetsier/2020/02/21/6-future-trends-in-data-science/?sh=5d44e92b5fa4
Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. International Journal of Computer Science, 1(2), 111-117.
Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
Lavrakas, P. J. (Ed.). (2021). Encyclopedia of Survey Research Methods (2nd ed.). SAGE Publications Ltd.
Lee, W. (2019). Data Visualization Using matplotlib. In Python® *Machine Learning* (pp. 67–91). Wiley. https://doi.org/10.1002/9781119557500.ch4
Lewis, R., & Rao, J. M. (2021). On the Use and Misuse of A/B Testing in Data Science. Communications of the ACM, 64(3), 62-71.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical *Natural Language Processing*. MIT Press.
Marr, B. (2018). Data strategy: How to profit from a world of *Big Data*, analytics and the *Internet of Things*. Kogan Page Publishers.
Marr, B. (2018). Data Strategy: How to Profit from a World of *Big Data*, Analytics and the *Internet of Things*. Kogan Page Publishers.
Marr, B. (2018). How Data Science Will Change the Future. Forbes. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/07/how-data-science-will-change-the-future/?sh=7408e0c51c44
Matt Crabtree. (2023, July). What is Data Analysis? An Expert Guide With Examples. Https://Www.Datacamp.Com/Blog/What-Is-Data-Analysis-Expert-Guide.
McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.
McKinney, W. S., 2017. The Art of Data Science.
Meyerson, A. (2022). Web Scraping with Python: Collecting Data from the Modern Web (2nd ed.). O'Reilly Media.
Mirqotussa'adah, Much Aziz Muslim; Budi Prasetiyo; Eva Laily Harum Mawarni; Anisa Juli Herowati; Rukmana, S. H. (2019). Data Mining Algoritma C45. In Eka Listiana; Nova Cahyani (Ed.), Book (Pertama, Vol. 1, Issue 1). http://repositorio.unan.edu.ni/2986/1/5624.pdf%0Ahttp://fiskal.kemenkeu.go.id/ejournal%0Ahttp://dx.doi.org/10.1016/j.cirp.2016.06.001%0Ahttp://dx.doi.org/10.1016/j.powtec.2016.12.055%0Ahttps://doi.org/10.1016/j.ijfatigue.2019.02.006%0Ahttps://doi.org/10.1
Mitchell, R., & Wilson, A. (2021). IoT Applications and Implementations for Industry 4.0. CRC Press.
Montgomery, D.C and Runger, G.C. 2003. Applied Statistics and Probability for Engineers. United State of America: John Wiley & Sons, Inc.
Mr. Ramkumar A, H. B. (2023). Data Science: Foundation &

Fundamentals. Dabra: Xoffencer International Publication.
Nagesh, K., Nageswara, D., & K., S. (2015). Python and MatPlotLib based *Open-source* Software System for Simulating Images with point Light Sources in Attenuating and Scattering Media. International Journal of Computer Applications, 131(8), 15–21. https://doi.org/10.5120/ijca2015907405
Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In Advances in Neural Information Processing Systems (pp. 841-848).
Nilda, janna miftahul. (2020). Variabel dan skala pengukuran statistik. Jurnal Pengukuran Statistik, 1(1), 1–8.
Núria Emilio. (2023). Data Analysis 6 Steps: A Complete Guide Into Data Analysis Methodology. Https://Blog.Bismart.Com/En/Data-Analysis-Steps-Complete-Guide.
Nuryadi, Astuti, T. D., Utami, E. S., & Budiantara, M. (2017). Buku Ajar Dasar-dasar Statistik Penelitian. In Sibuku Media.
Oberoi, A., & Chauhan, R. (2019). Visualizing data using Matplotlib and Seaborn libraries in Python for data science. International Journal of Scientific and Research Publications (IJSRP), 9(3), p8733. https://doi.org/10.29322/IJSRP.9.03.2019.p8733
Pampel, F. C. (2000). Logistic Regression: A Primer. SAGE Publications.
Pang, B., & Ng, V. (2022). Data Science for Social Good: Research and Practice. Cambridge University Press.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(11), 559-572.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: *Machine Learning* in Python. Journal of *Machine Learning* Research, 12, 2825-2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html
Plaue, M. 2023. Data Science: An Introduction to Statistics and

*Machine Learning*. Germany: Springer-Verlag GmbH.
Pradnyana, G. A., Darmawiguna, I. G. M., & Wijaya, I. N. S. W. 2020. Data mining: Menentukan Pengetahuan Dalam Data (Cetakan ke-1). PT RajaGrafindo Persada.
Prasetyo, B., & Si, M. (2018). Pengantar Statistik Sosial. Penerbit Universitas Terbuka. Banten, 1–25.
Press, G. (2013, May 28). A very short history of data science. Forbes. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/
Press, G. (2013, May 28). A Very Short History of Data Science. Forbes. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/
Program Data Science. (2020). Pengantar Data Science dan Aplikasinya bagi Pemula (M. T. A. Vronica S. Moertini (ed.); Edisi Pert). Unpar Press. chrome-extension://mhnlakgilnojmhinhkckjpncpbhabphi/pages/pdf/web/viewer.html?file=https%3A%2F%2Finformatika.unpar.ac.id%2Fwp-content%2Fuploads%2Fsites%2F19%2F2020%2F12%2FPengantarDataScience_dan_Aplikasinya_bagi_Pemula.pdf
Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.
Provost, F., & Fawcett, T. (2021). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (2nd ed.). O'Reilly Media.
Rismayani, W. O. R. A. U. M., Elly Warni, Pasnur, Fanny Ramadhani, A. S., & Abdul Karim, Janner Simarmata, Ilfa

Stephane, Sitti Arni, Muhammad Resha, P. H. P. (2024). Data Science (A. Karim (ed.); Pertama). Yayasan Kita Menulis.
Rugg, G., & Petre, M. (2022). The Unwritten Rules of PhD Research (2nd ed.). Open University Press.
Santoso, D. J. (2023). Ilmu Data (Data Science). Semarang: Yayasan Prima Agus Teknik.
Santoso, J. T., Kom, S., & Kom, M. (2020). Analisis *Big Data* (J. T. Santoso (ed.); Pertama). Yayasan Prima Agus Teknik.
Science, P. D. (2020). Pengentar Data Science dan Aplikasinya Bagi Pemula. Bandung: Unpar Press.
Scikit-learn developers. (2023). scikit-learn: *Machine Learning* in Python. Retrieved from https://scikit-learn.org/stable/index.html
Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.
Shaturaev, J. (2022). Economies and Management as A Result of The Fourth Industrial Revolution: An Education Perspective. Indonesian Journal of Educational Research and Technology, 3(1), 51-58.
Silver, N. (2012). The Signal and the Noise: Why So Many Predictions Fail--but Some Don't. Penguin Books.
Simplelearn. (2024, February 16). What Is Data Analysis: A Comprehensive Guide. Https://Www.Simplilearn.Com/Data-Analysis-Methods-Process-Types-Article.
Smith, J., & Noble, H. (2022). Qualitative Research Methods in Health Science: Interviews and Observations. BMJ Journals.
Soulley, D., Vicente, J., Wolveten,. M & Gilick, M., 2022. “The Limits of Computing for Data Science”.
Suharto, I. A. (2023). Fundamental Data Science dengan Library Pandas Python. Purbalingga: Eurika Media Aksara.
Sunil Ray. (2016). A Comprehensive Guide to Data Exploration. In kaggle. https://medium.com/analytics-vidhya/a-comprehensive-guide-to-data-exploration-d5919167bf6e
Suparyanto. (2020). Skala Pengukuran dan Instrumen

Penelitian. Suparyanto Dan Rosad (2015, 5(3), 248–253.
Suyanto, 2017. “Data Mining Untuk Klasifikasi Dan Klasterisasi Data”, Bandung Informatika.
Tarigan, M., & Frintiana Silaban, D. (2023). Reviu Statistika: Data Dan Skala Pengukuran. JINTAN: Jurnal Ilmu Keperawatan, 3(02), 118–126. https://doi.org/10.51771/jintan.v3i02.658
Vaddhano, N. (2023). Pemasaran Berbasis *Big Data* Dalam Revolusi Industri 4.0: Vol.2, No.2, Januari 2023-910 ULIL ALBAB.
Vallath, K. (2023). API Design Patterns: API Development for the Modern Web. Apress.
van der Aalst, W. M. P. (2016). Data Science in Action. In Process Mining: Data Science in Action (pp. 3-23). Springer.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K. 2012. Probability & Statistics for Engineers & Scientists. United State of America: Prentice Hall.
Williamson, K., Johanson, G., & Raghnaill, M. N. (2021). Research Methods: Information, Systems, and Contexts (3rd ed.). Chandos Publishing.
Wira J, Putra G. Pengenalan Konsep Pembelajaran Mesin dan *Deep Learning* Edisi 1.4 (17 Agustus 2020).
Yovita, 2020. “Kenali Data Science Melalui Penerapannya di Dunia Pemasaran”, DQLab.id.
Yuliansyah, 2016. “Penyelarasan Strategis Organisasi: Teori dan Konsep serta Penerapannya di Industri Perbankan”, Salemba Empat.
Zahriyah, A. (2023). Ekonometrika tekhnik dan Aplikasi dengan SPSS 1 BAB I.

## Tentang Penulis

**Dr. Phie Chyan, ST, M.Cs,** Lahir di Makassar 13 April 1981, Setelah menyelesaikan pendidikan menengah di SMA Katholik Cendrawasih Makassar Tahun 1996. Penulis melanjutkan pendidikan tinggi S1 di Universitas Atma Jaya Makassar program studi Teknik Elektro lalu melanjutkan S2 di Universitas Gadjah Mada Program Studi Ilmu Komputer. Pendidikan terakhir penulis adalah S3 ilmu Elektro setelah lulus studi di Program Doktor di Universitas Hasanuddin pada tahun 2024. Buku ini merupakan salah satu karya dari penulis sesuai bidang minat dan ilmu dari penulis dengan tujuan untuk berbagi ilmu pengetahuan kepada masyarakat.

**Zelvi Gustiana,** buku ini adalah salah satu karya dan inshaa allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku yang ditulis semata-mata untuk berbagi ilmu pengetahuan.

**Sitti Arni.** Bidang penelitian yang diminati adalah pengembangan Sistem Informasi dan Data Sains. Matakuliah yang menjadi perhatian saat ini diantaranya Sistem Basis Data, Analisis Perancangan Sistem Informasi, Teknik Riset Operasi, Data Mining, Pengendalian dan Audit Sistem Informasi serta Manajemen Proyek.

Sejak tahun 2022 aktif menulis buku referensi.

**Amru Yasir, S.Kom., M.Kom,** buku ini adalah salah satu karya dan inshaa allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku yang ditulis semata-mata untuk berbagi ilmu pengetahuan.

**Hartina Husain, S.Si., M.Stat**, merupakan Dosen Program Studi Sains Data Institut Teknologi Bacharuddin Jusuf Habibie. Penulis lahir pada 23 Mei 1996 di Samata provinsi Sulawesi Selatan. Penulis telah meraih gelar sarjana dari Universitas Hasanuddin tahun 2017 pada jurusan Statistika. Kemudian penulis melanjutkan studi magister di Surabaya tepatnya di Institut Teknologi Sepuluh Nopember dengan jurusan yang sama. Penulis juga aktif sebagai tutor online (tuton) di Universitas Terbuka. Adapun topik penulisan artikel penulis terkait pemodelan statistika, regresi non parametrik, dan analisis meta.

**Budi Arif Dermawan, M.Kom.,** adalah dosen Universitas Singaperbangsa Karawang, sebuah perguruan tinggi negeri di Kabupaten Karawang. Penulis mengajar di Program Studi Informatika, Fakultas Ilmu Komputer. Bidang kepakaran penulis meliputi *Machine Learning* dan Computer Vision. Buku ini adalah buku kedua yang ditulis dalam bidang Data Science. Melalui buku ini, penulis berharap

dapat secara konsisten mengembangkan ilmu pengetahuan secara utuh dalam siklus Tridharma, salah satunya melalui buku sebagai media pembelajaran.

**Ade Oktarino, S.Kom., M.S.I** merupakan alumni S1 Teknik Informatika Universitas Surakarta Tahun 2010, Alumni S2 Magister Sistem Informasi Pada Sekolah Tinggi Ilmu Komputer Dinamika Bangsa Jambi Tahun 2013 dan saat ini sedang menempuh Program Doctoral Information Technology di Universitas Putra Indonesia YPTK Padang pada Tahun 2022, sebelumnya pernah sebagai dosen tetap di Politeknik Jambi hingga Tahun 2014 dan sekarang aktif mengajar di Universitas Adiwangsa Jambi pada Program Studi Teknologi Informasi Fakultas Teknik dan Ilmu Komputer. Sebelumnya Dosen Tugas Tambahan sebagai Wakil Rektor II Universitas Adiwangsa periode 2017 – 2023 dan saat ini Dosen Tugas Tambahan sebagai Dekan Fakultas Keguruan Ilmu Pendidikan Universitas Adiwangsa Jambi. Aktif dalam organisasi Relawan TIK Provinsi Jambi segabai salah satu pengurus dan juga aktif sebagai konsultas IT di instansi Polda Jambi, Dinas Lingkungan Hidup Kota Jambi dan Kantor Imigrasi Provinsi Jambi. Kosentrasi penelitian dengan topik artificial intelegency, *Machine Learning*, *Deep Learning* dengan bidang computer vision dan *Natural Language Processing*.

**I Putu Tedy Indrayana, S.Pd., M.Sc**. Penulis lahir di Desa Gunaksa, Kecamatan Dawan, Kabupaten Klungkung Bali pada tahun 1991. Penulis menyelesaikan studi sarjana dari Jurusan Pendidikan Fisika FMIPA Universitas Pendidikan Ganesha Singaraja pada tahun 2013 dan jenjang magister di Program studi S2 Fisika Departemen Fisika FMIPA Universitas Gadjah Mada Yogyakarta pada tahun 2016. Saat ini penulis menjadi Dosen pada Program Studi Fisika, FMIPA Universitas Udayana. Penulis menekuni bidang Fisika Material dan Instrumentasi. Selain meneliti, penulis juga aktif menulis buku bidang fisika, teknologi informasi, filsafat, dan pendidikan fisika. Bagi pembaca yang ingin mengetahui profil penulis lebih detail dapat mengunjungi laman webesite penulis melalui link https://bit.ly/SidewiTedyFisika.

**Dr. Amril Mutoi Siregar, M. Kom**. Lahir di Ujung Padang, Sumatera Utara. Sekolah SD sampai SMP diselesaikan di kota kelahirannya. Kemudian pada tahun 1994 melanjutkan Pendidikan di SMA PGRI 4 Jakarta Jurusan IPA. Tahun 2004 Melanjutkan Pendidikan Program S1 di STMIK MIC Cikarang pada jurusan Teknik Informatika. Tahun 2014 melanjutkan Pendidikan Program S2 di President University Cikarang Jurusan Teknik Informatika. Tahun 2020 melanjutkan Pendidikan Program S3 di IPB University jurusan Ilmu Komputer. Penulis konsentrasi mengajar dan penelitian di bidang: *Machine Learning*, *Deep Learning*, data mining, Data Science dan Algoritma.

**Alfredo Gormantara S.Kom., M.Kom.,** merupakan dosen tetap Universitas Atma Jaya Makassar research interest di bidang Information Systems dan *Machine Learning*. Merupakan alumni sarjana Teknik Informatika Universitas Atma Jaya Makassar dan magister Teknik Informatika Universitas Atma Jaya Yogyakarta. Pokok pembahasan buku ditulis relevan dengan kepakaran dan pengalaman profesional di bidang pengembangan perangkat keras serta diharapkan dapat menjadi sarana berbagi ilmu pengetahuan kepada masyarakat.

**Indah Dwi Mumpuni**, Saat ini penulis aktif mengajar di STMIK PPKIA Pradnya Paramita, Prodi Sistem Informasi. Dengan pendekatan yang praktis dan studi kasus nyata, buku ini dirancang untuk membantu pembaca memahami bagaimana analisis data dapat digunakan untuk mendukung pengambilan keputusan yang lebih baik, mengoptimalkan operasi, dan mengidentifikasi peluang baru. Penulis berharap buku ini dapat menjadi sumber daya yang berarti bagi siapa saja yang tertarik untuk mempelajari dan menganalisis data dalam data science.

**Medy Wisnu Prihatmono, S.Kom., M. Kom.** Lahir di Ujung Pandang, Sulawesi Selatan. Sekolah SD sampai SMA diselesaikan di kota kelahirannya. Melanjutkan Pendidikan Program S1 di STMIK Dipanegara pada jurusan Manajemen Informatika. Tahun 2014 melanjutkan Pendidikan Program S2 di

Universitas AMIKOM Yogyakarta Jurusan Teknik Informatika. Penulis konsentrasi mengajar dan penelitian di bidang: *Machine Learning*, Data Science dan NLP.

**I Putu Gd Sukenada Andisana, S.Kom., M.T.,** saat buku ini diterbitkan, penulis bekerja sebagai dosen di STMIK Bandung Bali dan sebagai tenaga ahli sistem analis di PERUMDA Tirta Mangutama Kabupaten Badung. Penulis menekuni bidang sistem informasi, komputer dan teknologi informasi. Aktif menjadi konsultan sistem informasi beberapa perusahaan swata dan mengembangkan perangkat lunak untuk kebutuhan sistem informasi manajemen pada perusahaan.

**Lenny Maryam AB. Possumah SE. MM,** asal Nambo ini lahir pada 13 Mei 1976. Saat ini menjabat sebagai dosen di Universitas Muhammadiyah Luwuk tepatnya di Fakultas Ekonomi dan Bisnis. Dalam perannya tersebut, beliau memberikan ilmu kepada mahasiswa yang terdaftar di Program Studi Akuntansi, Manajemen, dan Bisnis Digital. Lenny Maryam AB. Possumah memperoleh gelar Sarjana dari Universitas Putra Bangsa Surabaya jurusan Akuntansi pada tahun 1999. Pada tahun 2005 melanjutkan pendidikan dengan menyelesaikan program Sertifikat Mengajar IV di Universitas Terbuka. Terakhir pada tahun 2013, Lenny Maryam AB. Possumah memperoleh gelar Magister dari Sekolah Tinggi Ilmu Ekonomi Malang. Di Universitas Muhammadiyah Luwuk, Fakultas Ekonomi dan Bisnis, penulis telah memberikan ilmunya sejak tahun 2000. Mata kuliah yang dibahas meliputi

mata kuliah pengantar akuntansi, akuntansi manajemen, analisis laporan keuangan, perpajakan, dan kewirausahaan.
Selama ini penulis juga telah menulis beberapa buku, seperti Auditing 2023 terbitan Yayasan Penerbitan Muhammad Zaini, Fiqh Muamalah Kontemporer terbitan CV. Ayrada Mandiri tahun 2023, dan Sistem Pembelajaran Terbuka dan Jarak Jauh yang diterbitkan oleh CV. Pradina Pustaka pada tahun 2024.

**Ibnu Atho'illah, S.T., M.T.** Lahir di Pasuruan, 20 Agustus 1975. Penulis memulai Pendidikan di SD Nahdlatul Ulama Pasuruan Jawa Timur tahun 1982. Alumni SMPN 2 Pasuruan tahun 1991, dan SMAN 1 Pasuruan tahun 1994. Setelah tamat SMA, penulis melanjutkan Pendidikan S1 di Institut Teknologi Nasional (ITN) Malang Jurusan Teknik Kimia lulus tahun 1999. Pada tahun yang sama melanjutkan studi S2 di Universitas Gadjah Mada (UGM) di jurusan yang sama. Setelah lulus pada tahun 2002 sempat bekerja di PT. Kutrindo Indonesia (2002 – 2006). Selepas tahun 2006 penulis banyak berkecimpung di dunia Pendidikan, Pengajaran dan Pelatihan. Pada tahun 2009 – sekarang, penulis merupakan dosen pengajar di STMIK Bandung Bali, juga di tahun 2011 – sekarang sebagai Instruktur IT di LPK Emerald Informatika Bali. Buku ini adalah salah satu karya dan Insya Allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku yang ditulis semata-mata untuk berbagi ilmu pengetahuan.

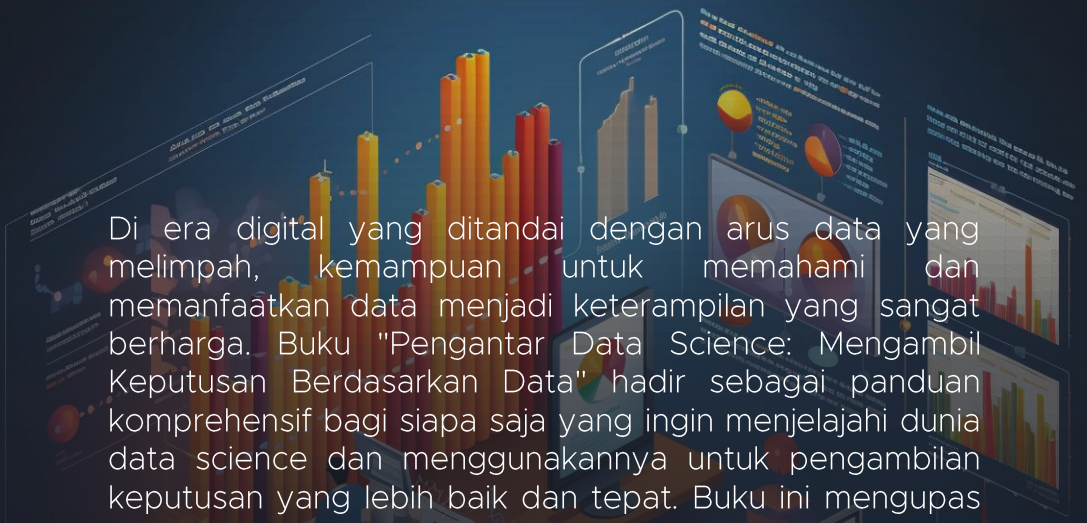
**Siti Aisyah, S.Tr., M.Sc.,** adalah dosen pada Universitas Insan Cita Indonesia (UICI) program studi Sains Data, Trainer untuk Data Analyst dan Data Scientist class, dan konsultan di Kementerian dan Perusahaan. Menyelesaikan Pendidikan sarjana pada Politeknik Negeri Jakarta (PNJ) Jurusan Teknik Informatika, kemudian melanjutkan Pendidikan di University of Stirling, UK untuk gelar Master of Science in *Big Data*. Buku ini adalah salah satu karya dan Inshaa Allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku yang ditulis semata-mata untuk berbagi ilmu pengetahuan.

**Santi Prayudani** lahir di Kotamadya Binjai Sumatera Utara pada tanggal 28 Maret 1986. Penulis menempuh pendidikan S1 di Prodi Ilmu Komputer Universitas Sumatera Utara pada tahun 2004. Kemudian melanjutkan kembali pendidikan S2 di Prodi Teknik Informatika Universitas Sumatera Utara pada tahun 2011. Memulai karir sebagai guru di SDS Al Azhar Medan pada tahun 2010. Kemudian mengajar juga sebagai dosen di AMIK Harapan dan Universitas Pembangunan Panca Budi dari tahun 2011 sampai tahun 2014. Saat ini penulis diberi amanah oleh negara untuk mengabdi sebagai dosen di Politeknik Negeri Medan dari tahun 2015.

**Nuk Ghurroh Setyoningrum,** lahir di Semarang, Jawa Tengah 23 Agustus 1984, Adalah alumni Sarjana Komputer dari Universitas Stikubank Semarang pada tahun 2007 di program studi Sistem Informasi Fakultas Teknologi Informasi dan menyelesaikan program Magister ilmu Komputer di Universitas Gadjah Mada Yogyakarta pada tahun 2010 dengan mengambil konsentrasi Ilmu Komputer Fakultas MIPA. Sekarang Sedang menempuh studi Doktoral Informatika di Universitas AMIKOM Yogyakarta. Penulis mengabdi sebagai Dosen Tetap di Universitas Cipasung Tasikmalaya dan aktif mengajar sebagai Tutor di Universitas Terbuka sejak tahun 2019 sampai sekarang.

**Salman Farizy. S.Kom, M.Kom, MCSE, MVP,** menyelesaikan pendidikan dasar dan menengah di SDN 01 Cipinang Besar, SMPN 25 Cipinang Muara Jakarta Timur dan SMAN 91 Pondok Kelapa Jakarta Timur, sedangkan untuk perguruan tinggi Strata satu (S1) di Universitas Gunadarma dan Pasca Sarjana di Universitas Pamulang, penulis sempat bekerja dibeberapa perusahaan asing (PMA) dan juga lokal seperti Mattel Indonesia (PMA) sebagai MIS Staff AS/400, Frigorex Indonesia (PMA) sebagai IS Dept. Head, PT. Asaba Computer Center (Microsoft Gold Partner) sebagai IS Div. Head, Kaltimex Energy (PMA) sebagai IT Manager dan terakhir sebagai IT Consultant paralel juga dengan menjadi Dosen tetap Universitas Pamulang sampai dengan saat ini. Mudah -mudahan dengan adanya buku ini akan menambah pengetahuan dan juga dapat bermanfaat.

**Vivi Afifah, S. Kom., MMSI** menjadi dosen adalah pilihan karir sejak tahun 2001. Berkiprah sebagai pengajar maupun struktural selama lebih dari 23 tahun di berbagai kampus di wilayah Jawa Barat dan DKI Jakarta. Manajemen Sistem Informasi dan Bisnis Digital merupakan bidang keahlian yang dipilih. Buku ini adalah salah satu karya dan Insyaa Allah secara konsisten akan disusul dengan buku-buku berikutnya. Pokok bahasan buku yang ditulis semata-mata untuk berbagi ilmu pengetahuan.



Di era digital yang ditandai dengan arus data yang melimpah, kemampuan untuk memahami dan memanfaatkan data menjadi keterampilan yang sangat berharga. Buku "Pengantar Data Science: Mengambil Keputusan Berdasarkan Data" hadir sebagai panduan komprehensif bagi siapa saja yang ingin menjelajahi dunia data science dan menggunakannya untuk pengambilan keputusan yang lebih baik dan tepat. Buku ini mengupas berbagai konsep dasar dan teknik penting dalam data science. Pembaca akan diajak untuk memahami proses pengumpulan, pembersihan, dan analisis data, serta cara membangun model pembelajaran mesin yang efektif. Tidak hanya teori, buku ini juga dilengkapi dengan contoh-contoh nyata dan studi kasus yang membantu pembaca melihat aplikasi praktis dari konsep-konsep yang dibahas. "Pengantar Data Science: Mengambil Keputusan Berdasarkan Data" dirancang untuk pembaca dari berbagai latar belakang, baik itu mahasiswa, profesional, maupun siapa saja yang tertarik untuk memahami dan memanfaatkan data dalam pengambilan keputusan. Buku ini memberikan fondasi yang kuat bagi siapa saja yang ingin memulai karier di bidang data science atau meningkatkan keterampilan analitis mereka. Dengan gaya bahasa yang jelas dan mudah dipahami, buku ini tidak hanya memberikan pengetahuan teoretis tetapi juga keterampilan praktis yang dapat langsung diterapkan. Selamat menjelajahi dunia data science dan semoga buku ini menjadi referensi yang berguna dalam perjalanan Anda menuju pengambilan keputusan berbasis data yang lebih baik.

**DITERBITKAN OLEH**
**PT. MIFANDI MANDIRI DIGITAL**

Jln Payanibung Ujung D
Dalu Sepuluh-B, Tanjung Morawa
Kab. Deli Serdang Sumatera Utara

ISBN 978-623-8558-39-1
9 786238 558391