DATA SAINS
BIDANG KEAMANAN INFORMASI

Disusun oleh:
Dr. Eng. Rini Wisnu Wardhani, M.T.

Bogor
2026

DATA SAINS
BIDANG KEAMANAN INFORMASI

Dr. Eng. Rini Wisnu Wardhani, M.T.

Data Sains
Bidang Keamanan Informasi

Tim Editor
Yeni Farida, S.Stat., M.Si.
Nur Praptiwi Mita Hapsari, S.I.Pust.

Penerbit:
Politeknik Siber dan Sandi Negara Press
Jl. Raya H. Usa, Ciseeng, Bogor, Jawa Barat 16210
Telp. (0251) 8541752
e-mail: perpustakaan@poltekssn.ac.id

ISBN cetak: 978-623-XX-XXXX-X
ISBN digital: 978-623-XX-XXXX-X
Cetakan I: 2026
viii + 71 halaman; Ukuran:17 x 24 cm

Kata Pengantar

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga Buku *Data Sains Bidang Keamanan Informasi* ini dapat diterbitkan sebagai buku pengantar bagi materi perkuliahan maupun masyarakat umum. Kehadiran buku ini diharapkan dapat memberikan manfaat tidak hanya bagi mahasiswa, tetapi juga sebagai bahan referensi bagi para pengajar dan masyarakat luas yang ingin memahami penerapan data sains dalam konteks keamanan informasi.

Dengan diterbitkannya buku ini, koleksi buku yang dihasilkan oleh para dosen Politeknik Siber dan Sandi Negara semakin bertambah. Hal ini merupakan langkah positif dalam mendukung pengembangan atmosfer akademik dan peningkatan mutu pendidikan di lingkungan Politeknik Siber dan Sandi Negara, sehingga patut disambut dengan penuh apresiasi.

Kami mengucapkan selamat dan penghargaan setinggi-tingginya kepada tim penyusun atas kerja keras dan dedikasinya dalam menyelesaikan buku ini. Semoga karya ini dapat memberikan kontribusi nyata dalam mendukung proses pembelajaran, serta menjadi inspirasi bagi para penulis untuk terus berkarya dan menghasilkan buku-buku lain yang relevan dengan perkembangan zaman, guna melengkapi khazanah kepustakaan yang ada.

Bogor, 5 April 2026

Direktur Politeknik Siber dan Sandi Negara

Ir. Arnoldus Triono, M.M, M.Tr.Opsla, CIQnR

Prakata Penulis

Assalamualaikum warahmatullahi wabarakatuh

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan buku **Data Sains Bidang Keamanan Informasi** ini. Buku ini disusun sebagai bahan pendukung proses belajar mengajar di lingkungan perguruan tinggi, khususnya bagi mahasiswa dan dosen, serta bagi masyarakat umum, profesional, dan praktisi yang ingin memahami penerapan data sains di bidang keamanan siber dan persandian. Dengan demikian, buku ini dirancang untuk berbagai level pembaca, mulai dari pemula yang ingin mengenal konsep dasar, hingga pembaca yang ingin mengembangkan keterampilan analisis data untuk aplikasi nyata di bidang keamanan informasi.

Tujuan penyusunan buku ini adalah menyediakan referensi pembelajaran yang tidak hanya memberikan pemahaman dasar mengenai data sains, tetapi juga menyajikan *insight* praktis terkait pengolahan dan analisis data dalam konteks keamanan siber dan persandian. Keunikan buku ini terletak pada pendekatan yang menggabungkan teori data sains dengan studi kasus nyata di bidang *cybersecurity*, sehingga pembaca tidak hanya memahami konsep, tetapi juga mampu menerapkannya secara langsung dalam lingkungan kerja atau penelitian.

Penyusunan buku ini dapat terlaksana dengan baik berkat bantuan, dukungan, dan kontribusi dari berbagai pihak, baik yang terlibat langsung maupun tidak langsung. Oleh karena itu, tim penulis mengucapkan terima kasih yang sebesar-besarnya kepada seluruh pihak yang telah membantu dalam proses penyusunan buku ini.

Penulis menyadari bahwa buku ini masih memiliki keterbatasan dan belum sempurna. Oleh sebab itu, saran dan masukan dari para pembaca sangat diharapkan demi penyempurnaan di masa mendatang. Semoga buku ini dapat memberikan manfaat praktis dan kontribusi positif bagi pengembangan ilmu pengetahuan, peningkatan kualitas pembelajaran, dan penguasaan kompetensi analisis data yang relevan dengan kebutuhan bidang keamanan informasi.

Wassalamualaikum warahmatullahi wabarakatuh

Bogor, 5 April 2026

Penulis

Daftar Isi

Daftar Gambar

GAMBARAN UMUM

A. Lingkup dan Profesi dalam Data Sains

Di era digital saat ini, data menjadi inti dari hampir seluruh aspek kehidupan modern. Data dapat dipahami sebagai informasi mentah yang dikumpulkan, disimpan, dan diproses untuk menghasilkan pengetahuan yang bermakna. Dalam dunia teknologi, data berperan sebagai fondasi utama sistem informasi—mulai dari aplikasi sederhana hingga infrastruktur kritis. Bahkan, data sering disebut sebagai "bahan bakar" ekonomi digital, karena menjadi dasar pengambilan keputusan di berbagai sektor, baik bisnis, pemerintahan, maupun masyarakat luas.

Perkembangan sains dan teknologi modern sangat ditopang oleh pemanfaatan data secara intensif. Aktivitas bisnis, ekonomi, sosial, politik, budaya, dan bahkan diplomasi internasional kini semakin bergantung pada data sebagai sumber analisis dan dasar perumusan strategi. Data tidak hanya menjadi sekadar fakta mentah, tetapi ketika diolah menjadi informasi terstruktur dan dianalisis dengan tepat, ia dapat menghasilkan pengetahuan strategis yang membantu individu dan organisasi membuat keputusan yang tepat.

Seiring meningkatnya peran data, data sains dan profesi terkait analisis data menjadi semakin vital. Beberapa profesi utama dalam bidang data sains antara lain:

1. *Data Scientist* – bertugas menganalisis data kompleks untuk menghasilkan wawasan dan solusi berbasis data.
2. *Data Analyst* – berfokus pada pengolahan dan analisis data untuk mendukung pengambilan keputusan.
3. *Data Engineer* – bertanggung jawab membangun dan mengelola infrastruktur serta alur data.
4. *Machine Learning Engineer* – mengembangkan dan menerapkan model *machine learning* untuk berbagai aplikasi.
5. *Business Intelligence Analyst* – menyajikan data dan laporan analitis untuk kebutuhan strategis organisasi.
6. *Big Data Engineer* – menangani pengolahan data berskala besar dan berkecepatan tinggi.
7. *Security Data Analyst* – menganalisis data keamanan informasi untuk mendeteksi ancaman dan risiko siber.

Dengan perkembangan ini, mata kuliah Data Sains kini menjadi bagian penting di tingkat perguruan tinggi, tidak hanya untuk pembelajaran dasar tetapi juga untuk penelitian dan aplikasi praktis, terutama dalam bidang *Computer Science, Engineering*, dan Keamanan Informasi.

Buku ini ditujukan bagi pembaca yang telah memiliki pengetahuan dasar dalam matematika, statistika, pemrograman, teknologi informasi, dan basis data, namun dirancang sedemikian rupa agar materi tetap mudah dipahami dan sistematis. Fokus utama buku ini adalah pada penerapan data sains di bidang keamanan informasi, sekaligus memberikan pemahaman umum mengenai konsep, metode, dan profesi data sains. Dengan pendekatan ini, buku ini diharapkan bermanfaat tidak hanya bagi mahasiswa dan profesional, tetapi juga bagi masyarakat luas yang ingin memahami peran data dalam dunia digital dan keamanan siber.

B. Organisasi Buku

Organisasi bab dalam buku ini disusun secara sistematis dan bertahap, sehingga memudahkan pembaca memahami materi dari konsep dasar hingga penerapan lanjutan. Buku ini dimulai dengan pengenalan prinsip-prinsip dasar data dan data sains, dilanjutkan dengan pembahasan metode analisis, pemrograman, dan *machine learning*, hingga implementasi praktis dalam konteks keamanan informasi.

1. **Gambaran Umum**
   Membahas pengertian data sains, ruang lingkup, sejarah perkembangan, serta peran data sains dalam era digital dan Industri 4.0.
2. **Jenis dan Sumber Data**
   Menguraikan konsep data dan informasi, jenis-jenis data, sumber *dataset*, serta karakteristik data yang digunakan dalam data sains.
3. **Dasar-Dasar Statistik untuk Data Sains**
   Menjelaskan konsep statistik dasar, ukuran pemusatan dan penyebaran data, serta visualisasi data untuk mendukung analisis.
4. **Proses dan Tahapan Pengolahan Data**
   Membahas tahapan pengolahan data, termasuk pengumpulan, pembersihan, transformasi, dan integrasi data.
5. **Pemrograman untuk Data Sains**
   Mengenalkan bahasa pemrograman dan alat bantu yang digunakan dalam data sains, seperti Python, R, dan SQL.
6. **Machine Learning dalam Data Sains**
   Menguraikan konsep dasar *machine learning*, jenis-jenis pembelajaran, serta penerapannya dalam berbagai domain.

7. **Latihan dan Materi Pengayaan**

Bagian ini dirancang untuk memberikan gambaran praktis mengenai tahapan dalam data sains, sekaligus melatih pemahaman pembaca melalui studi kasus sederhana di bidang keamanan informasi.

Latihan disusun untuk membantu pembaca menerapkan konsep dan teori yang telah dipelajari, baik untuk pembelajaran formal, pengembangan profesional, maupun aplikasi praktis dalam pekerjaan sehari-hari.

Materi pengayaan berfungsi untuk memperluas wawasan, mendorong kemampuan analisis yang lebih mendalam, dan memberikan *insight* tambahan dalam menyelesaikan permasalahan atau studi kasus lain di bidang data sains. Bagian ini juga membantu pembaca memahami penanganan data berskala besar dan kompleks, sehingga solusi yang dihasilkan menjadi lebih matang dan aplikatif.

Dengan pendekatan ini, bagian Latihan dan Materi Pengayaan diharapkan dapat menghubungkan teori dengan praktik nyata, serta membekali pembaca dari berbagai latar belakang—baik akademik, profesional, maupun masyarakat umum—dengan keterampilan dan pemahaman yang relevan dalam dunia data sains dan keamanan informasi.

BAB 1
PENGERTIAN DATA SAINS

A. Pengertian

Data science, atau data sains, merupakan bidang ilmu yang mempelajari metode dan teknik untuk mengumpulkan, mengolah, menganalisis, serta menginterpretasikan data guna menghasilkan informasi dan pengetahuan yang bernilai. Pada era digital saat ini, data memiliki peran yang sangat penting karena dapat dimanfaatkan oleh organisasi untuk mendukung pengambilan keputusan yang lebih tepat, meningkatkan efisiensi operasional, serta menciptakan nilai tambah yang berkelanjutan.

Tujuan utama data sains adalah mengubah data mentah menjadi informasi yang bernilai dan dapat ditindaklanjuti. Melalui pemahaman data sains, seseorang dapat meningkatkan kemampuan analisis data, mendukung pengambilan keputusan yang lebih tepat, serta menciptakan nilai tambah bagi organisasi. Selain itu, Data sains berperan dalam membantu memahami fenomena yang kompleks, mengidentifikasi peluang baru, dan meningkatkan efisiensi operasional. Oleh karena itu, pemahaman terhadap Data sains menjadi sangat penting, baik bagi individu yang ingin berkarier di bidang ini maupun bagi mereka yang ingin meningkatkan kompetensi analisis data secara profesional.

Data sains merupakan salah satu bidang ilmu yang saat ini berkembang pesat, termasuk dalam penerapannya pada bidang siber dan persandian. Dalam konteks tersebut, data keamanan informasi yang dihasilkan umumnya memiliki volume yang sangat besar serta kecepatan pertumbuhan yang tinggi, sehingga data yang perlu diolah menjadi semakin kompleks dan masif. Oleh karena itu, pemanfaatan data sains menjadi sangat penting untuk mendukung pengolahan, analisis, dan pengambilan keputusan yang efektif di bidang keamanan informasi.

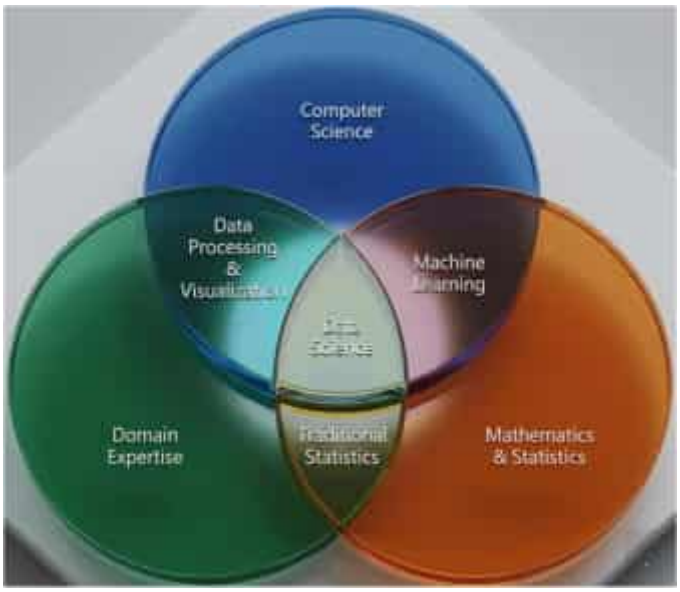
Hingga saat ini, data sains merupakan salah satu bidang ilmu yang berkembang pesat di dunia, termasuk dalam penerapannya pada bidang siber dan persandian. Pemanfaatan data sains memungkinkan pengolahan dan analisis data keamanan informasi secara lebih efektif. Oleh karena itu, penguasaan data sains dapat memberikan nilai tambah yang signifikan bagi individu, serta berpotensi

meningkatkan daya saing dan kinerja perusahaan atau instansi yang menerapkannya.

B. Ruang Lingkup dan Aplikasi Data Sains

Sejak awal perkembangannya hingga saat ini, data sains telah memiliki banyak referensi pembelajaran dari berbagai perspektif keilmuan dan teknik yang digunakan, baik melalui buku cetak maupun media daring. Bidang ini mencakup beragam aplikasi, karena konsep dan metodologinya dapat diterapkan di berbagai disiplin ilmu, mulai dari matematika, statistika, dan pemrograman, hingga teknik optimasi dan analisis lanjutan.

Seperti terlihat pada Gambar 1, keluasan cakupan data sains dapat dilihat dari beragam teknik pemrograman dan aplikasinya di berbagai sektor, mulai dari bisnis, kesehatan, pemerintahan, hingga keamanan informasi. Oleh karena itu, merangkum seluruh aspek data sains secara menyeluruh dalam satu buku atau media pembelajaran merupakan tantangan yang kompleks, mengingat luasnya materi dan variasi pendekatan yang ada. Buku ini hadir untuk memberikan panduan yang ringkas namun sistematis, dengan fokus pada konsep dasar, metode analisis, dan penerapan praktis, khususnya dalam konteks keamanan informasi, sehingga pembaca dapat memperoleh pemahaman yang komprehensif tanpa kehilangan fokus pada aplikasi nyata.



Gambar 1 Cakupan Keilmuan Data Sains [1]

Dalam buku ini, pembahasan difokuskan pada perumusan dan peringkasan materi-materi inti yang perlu dikuasai untuk memahami dan mengaplikasikan data sains. Ruang lingkup materi yang disajikan diharapkan dapat memberikan landasan pengetahuan yang cukup bagi pembaca untuk menguasai konsep dasar data sains beserta penerapannya dalam berbagai bidang.

1. Konsep Data, nilai suatu data, termasuk proses pengambilan dan pengolahannya
   Memahami konsep nilai suatu data serta menjelaskan proses pengambilan, pengolahan, dan pemanfaatan data sebagai dasar dalam menghasilkan informasi yang bernilai.
2. Dasar-dasar statistika untuk visualisasi data serta penerapan perhitungan menggunakan alat (tools) data sains
   Memahami konsep dasar statistika dan menerapkannya dalam visualisasi data serta perhitungan analitis menggunakan berbagai alat (tools) data sains.
3. Tahapan-tahapan dalam proses data sains
   Menjelaskan tahapan-tahapan utama dalam proses data sains, mulai dari pengumpulan data hingga penyajian hasil analisis, serta memahami peran setiap tahapan dalam menghasilkan informasi yang akurat.
4. Bahasa pemrograman yang digunakan dalam data sains
   Mengenal dan memahami peran bahasa pemrograman yang umum digunakan dalam data sains serta mampu menggunakannya secara dasar untuk pengolahan dan analisis data.
5. Hubungan antara data sains dan machine learning
   Memahami keterkaitan antara data sains dan *machine learning* (atau pembelajaran mesin), termasuk peran machine learning dalam proses analisis dan pengambilan keputusan berbasis data.
6. Aplikasi serta perkembangan (advancement) penggunaan machine learning
   Menjelaskan berbagai aplikasi machine learning serta memahami perkembangan dan tren terkini dalam pemanfaatannya di berbagai bidang, termasuk penggunaan data sains untuk analisis big data, penerapan metode MapReduce, serta implementasinya dalam berbagai proyek keamanan informasi.

C. Sejarah dan Peran Data Sains Saat ini

Data sains mulai berkembang sejak dekade 1960-an, berakar dari bidang statistika dan pengolahan data. Salah satu tokoh awal yang berperan penting adalah John W. Tukey, yang pada tahun 1962 memperkenalkan gagasan *data analysis* sebagai

disiplin tersendiri dan menekankan pentingnya analisis eksploratif dalam memahami data. Konsep ini kemudian diperkuat pada akhir 1960-an melalui pengenalan istilah *datalogy* dalam forum internasional, yang memandang data sebagai objek kajian ilmiah. Perkembangan selanjutnya ditandai dengan semakin terintegrasinya metode statistika, komputasi, dan pengolahan data, termasuk munculnya *exploratory data analysis*, *business intelligence*, serta praktik pengumpulan data berskala besar oleh organisasi pada era 1990-an.

Memasuki abad ke-21, data sains berkembang pesat seiring kemajuan *machine learning*, komputasi terdistribusi, dan teknologi *big data* seperti *Hadoop*. Era *big data* menegaskan peran data sebagai aset strategis dengan karakteristik volume, kecepatan, dan keragaman yang tinggi. Hingga saat ini, data sains telah menjadi fondasi utama dalam pengembangan kecerdasan buatan, analitik prediktif, dan pengambilan keputusan berbasis data di berbagai sektor. Fokus pengembangan data sains juga bergeser ke arah otomasi analisis, integrasi AI dalam alur kerja, pemrosesan data *real-time*, serta penerapan solusi data yang siap digunakan pada skala industri.

Data sains merupakan inti dari pertumbuhan bisnis modern saat ini, dengan penerapan yang meluas pada berbagai bidang, mulai dari kesehatan, pemerintahan, hingga periklanan. Pengumpulan dan analisis data melalui pendekatan Data sains memiliki potensi besar untuk meningkatkan kualitas, efektivitas, dan efisiensi hasil kerja, baik untuk keperluan profesional maupun pribadi. Individu yang berkecimpung dalam bidang ini dikenal sebagai *data scientist*.

Saat ini, analisis *big data* menjadi salah satu tantangan utama yang mendorong pesatnya perkembangan ilmu data sains serta munculnya berbagai profesi yang berfokus pada pengelolaan dan analisis data. Data sains merupakan disiplin ilmu yang mengintegrasikan matematika, statistika, dan ilmu komputer dengan tujuan melakukan analisis data dari suatu himpunan data, baik dalam skala kecil (sampel) maupun besar (populasi). Proses ini dilakukan dengan menerapkan algoritma tertentu untuk menggali data *(data mining)*, menemukan pola, serta melakukan prediksi *(prediction)* secara akurat. Hasil analisis tersebut dapat dimanfaatkan untuk mendukung pengambilan keputusan dan membangun sistem cerdas *(artificial intelligence)* yang mampu belajar secara mandiri melalui metode *machine learning* [2].

Bidang keilmuan *big data* muncul sebagai akibat dari permasalahan utama dalam pengelolaan data, yang dikenal dengan karakteristik 5V, yaitu *velocity* (kecepatan), *volume* (jumlah), *variety* (keragaman), *value* (nilai), dan *veracity* (keakuratan). Banyaknya faktor yang memengaruhi analisis data, serta beragamnya kebutuhan pengguna dalam mengubah data menjadi informasi, mendorong penggunaan

berbagai metode dan alat bantu. Beberapa di antaranya adalah metode *machine learning*, *artificial intelligence*, serta alat analisis *big data* seperti *MapReduce*.

Gambar 2 Data Sains dan Big Data dalam Industry 4.0 (sumber: diolah kembali)

Seperti terlihat pada Gambar 2, Revolusi Industri 4.0 ditandai dengan integrasi teknologi digital, sistem siber-fisik (*Cyber-physical system)*, *Internet of Things* (IoT), komputasi awan, dan kecerdasan buatan dalam berbagai proses industri. Dalam konteks ini, data sains dan *big data* memegang peran sentral sebagai fondasi utama pengambilan keputusan berbasis data. Berbagai aktivitas industri menghasilkan data dalam jumlah besar, beragam, dan berkecepatan tinggi, yang memerlukan pendekatan analitis canggih untuk mengolahnya secara efektif.

Data sains berperan dalam mengolah dan menganalisis *big data* untuk mengekstraksi pola, wawasan, serta prediksi yang bernilai. Melalui penerapan metode statistika, *machine learning*, dan kecerdasan buatan, data sains memungkinkan optimalisasi proses produksi, peningkatan kualitas produk, pemeliharaan prediktif, serta efisiensi rantai pasok. Sementara itu, *teknologi big data* menyediakan infrastruktur dan kerangka kerja untuk menyimpan, memproses, dan mengelola data berskala besar secara terdistribusi.

Dengan memanfaatkan data sains dan *big data*, industri dapat meningkatkan daya saing, merespons perubahan pasar secara lebih cepat, serta menciptakan inovasi berbasis data. Oleh karena itu, penguasaan data sains dan pemahaman terhadap *big data* menjadi kompetensi kunci bagi sumber daya manusia dalam menghadapi tantangan dan peluang Industri 4.0.

D. Latihan Pemahaman

1. Jelaskan pengertian data sains dan sebutkan disiplin ilmu utama yang dapat mendukung bidang ini!
2. Uraikan perbedaan antara *data analysis, data mining*, dan *machine learning* dalam konteks data sains!
3. Sebutkan dan jelaskan minimal tiga aplikasi data sains dalam kehidupan sehari-hari!
4. Jelaskan peran *data scientist* serta keterampilan utama yang harus dimiliki dalam profesi tersebut!
5. Bagaimana perkembangan data sains memengaruhi proses pengambilan keputusan di era digital saat ini?

BAB 2
JENIS DAN SUMBER DATA

A. Data dan Informasi

Data merupakan sekumpulan fakta mentah (*raw facts*) yang diperoleh dari hasil pengamatan, pengukuran, atau pencatatan. Data dapat berupa angka, kata, simbol, suara, gambar, maupun bentuk lainnya yang masih belum memiliki makna apabila berdiri sendiri. Data tersebut dapat dianalisis dan diolah lebih lanjut untuk memperoleh pemahaman, wawasan, atau kesimpulan tertentu. Contoh data antara lain nilai ujian, usia, jenis kelamin, warna, dan jumlah mahasiswa.

Sementara itu, informasi merupakan hasil pengolahan dan analisis data yang telah diberi konteks sehingga memiliki makna dan dapat digunakan untuk mendukung pengambilan keputusan. Dengan demikian, data menjadi informasi apabila telah diinterpretasikan dan memberikan nilai tambah bagi penggunanya.

Data secara umum merujuk pada kumpulan fakta yang menjadi dasar pembentukan informasi dan pengetahuan. Data dapat diperoleh melalui proses pengamatan, pengukuran, atau pencatatan. Dalam kondisi mentah, data belum memiliki makna yang utuh. Data baru menjadi bernilai ketika diolah, dianalisis, dan diberi konteks sehingga dapat menghasilkan informasi dan pengetahuan yang bermanfaat bagi pengambilan keputusan.

Penguasaan dan kepemilikan data merupakan indikator kemampuan, sumber kekuatan, serta aset strategis yang sangat penting bagi individu maupun organisasi. Data tersebar di seluruh aspek kehidupan manusia, sehingga peradaban manusia dapat dipandang sebagai akumulasi fakta-fakta yang terus berkembang. Fakta-fakta tersebut akan memberikan manfaat optimal apabila ditransformasikan menjadi informasi dan pengetahuan yang mendukung perumusan strategi dan keputusan yang tepat.

B. Jenis Data

Terdapat berbagai cara dan referensi dalam mengelompokkan data. Salah satu pengelompokan data yang paling umum digunakan dalam data sains dan statistika adalah berdasarkan sifatnya, yaitu data kualitatif dan data kuantitatif. Namun terdapat beberapa pengelompokan lain data.

1. Data Kualitatif dan Data Kuantitatif

Data kualitatif adalah data yang tidak berbentuk angka dan digunakan untuk menggambarkan kualitas, karakteristik, atau kategori tertentu. Data ini tidak dapat diolah menggunakan operasi matematika secara langsung. Contoh data kualitatif antara lain:

a. Jenis kelamin: laki-laki, perempuan
b. Warna rambut: hitam, cokelat, pirang
c. Status pernikahan: belum menikah, menikah
d. Merek telepon seluler: Samsung, iPhone, Oppo

Data kuantitatif adalah data yang berbentuk angka dan dapat diukur atau dihitung secara matematis. Data ini dapat diolah menggunakan berbagai operasi matematika, seperti penjumlahan, pengurangan, rata-rata, dan perhitungan statistik lainnya. Contoh data kuantitatif antara lain:

a. Usia: 16 tahun
b. Tinggi badan: 165 cm
c. Jumlah mahasiswa: 32 orang
d. Nilai ujian: 90

2. Data Primer dan Data Sekunder

Data primer adalah data yang diperoleh secara langsung dari sumber utama oleh peneliti atau pengumpul data. Data ini biasanya dikumpulkan melalui observasi, wawancara, survei, atau eksperimen. Contoh data primer antara lain:

a. Hasil kuesioner yang diisi oleh responden
b. Data hasil pengamatan langsung terhadap sistem
c. Data eksperimen laboratorium

Data sekunder adalah data yang diperoleh dari sumber tidak langsung atau pihak lain yang telah mengumpulkan dan mengolah data tersebut sebelumnya. Contoh data sekunder antara lain:

a. Data statistik dari instansi pemerintah

b. Laporan penelitian sebelumnya
c. *Dataset* yang diunduh dari repositori publik

3. Data Nominal dan Ordinal

Data nominal adalah data yang digunakan untuk mengelompokkan objek ke dalam kategori tertentu tanpa adanya urutan atau tingkatan. Contoh data nominal antara lain:

a. Jenis kelamin
b. Agama
c. Merek perangkat lunak
d. Alamat IP (kategori jaringan)

Data ordinal adalah data yang memiliki kategori sekaligus urutan atau tingkatan, namun jarak antar kategori tidak dapat diukur secara pasti. Contoh data ordinal antara lain:

a. Tingkat kepuasan: sangat puas, puas, cukup, tidak puas
b. Tingkat risiko keamanan: rendah, sedang, tinggi
c. Peringkat kelas atau peringkat kinerja

4. Data Diskrit dan kontinu

Data diskrit adalah data kuantitatif yang nilainya berupa bilangan bulat dan diperoleh melalui proses perhitungan *(counting)*. Data ini tidak dapat dinyatakan dalam nilai pecahan. Contoh data diskrit antara lain:

a. Jumlah mahasiswa dalam satu kelas
b. Jumlah kendaraan yang terparkir
c. Jumlah paket data yang diterima server

Data kontinu adalah data kuantitatif yang nilainya diperoleh melalui proses pengukuran *(measurement)* dan dapat dinyatakan dalam bentuk pecahan atau desimal. Contoh data kontinu antara lain:

1. Berat badan: 60,5 kg
2. Tinggi badan: 165,7 cm
3. Waktu akses sistem: 2,35 detik

5. Dataset

Dataset merupakan kumpulan data sekunder, yaitu data yang diperoleh dari sumber tidak langsung atau dari pihak lain yang telah melakukan proses pengumpulan dan pengolahan data sebelumnya. Dalam konteks data sains, penggunaan *dataset*

memiliki peran yang sangat penting, baik untuk tujuan pembelajaran maupun penelitian.

Dataset banyak digunakan sebagai media pembelajaran untuk memahami dan menguji berbagai metode dasar dalam data sains, seperti analisis statistik, *machine learning*, dan visualisasi data. Selain itu, *dataset* juga dimanfaatkan dalam riset berskala nasional maupun internasional untuk menyediakan objek penelitian yang sama, sehingga memungkinkan para peneliti membandingkan, mengevaluasi, dan mengembangkan metode atau algoritma baru secara objektif dan terstandar.

Selain diperoleh dari buku referensi dan media digital, saat ini tersedia berbagai sumber *dataset* terbuka (*open datasets*) yang dapat diakses secara daring. Sumber-sumber *dataset* tersebut sangat membantu analis data dalam memperoleh data yang siap digunakan untuk analisis, eksperimen, maupun pengembangan model.

Beberapa sumber pembelajaran online atau *dataset* diantaranya :

a. **Kaggle Learn**
   URL: https://www.kaggle.com/learn

b. **DataCamp**
   URL: https://www.datacamp.com/

c. **Towards Data Science (Medium)**
   URL: https://towardsdatascience.com/

d. **Real Python**
   URL: https://realpython.com/

e. **Analytics Vidhya**
   URL: https://www.analyticsvidhya.com/

f. **Mode SQL Tutorial**
   URL: https://mode.com/sql-tutorial/

g. **Seeing Theory (Probability & Statistics)**
   URL: https://seeing-theory.brown.edu/

h. **Data Science Central**
   URL: https://www.datasciencecentral.com/

i. **Kaggle Datasets**
   URL: https://www.kaggle.com/datasets

j. **UCI Machine Learning Repository**
   URL: https://archive.ics.uci.edu/ml/

k. **Google Dataset Search**
URL: https://datasetsearch.research.google.com/

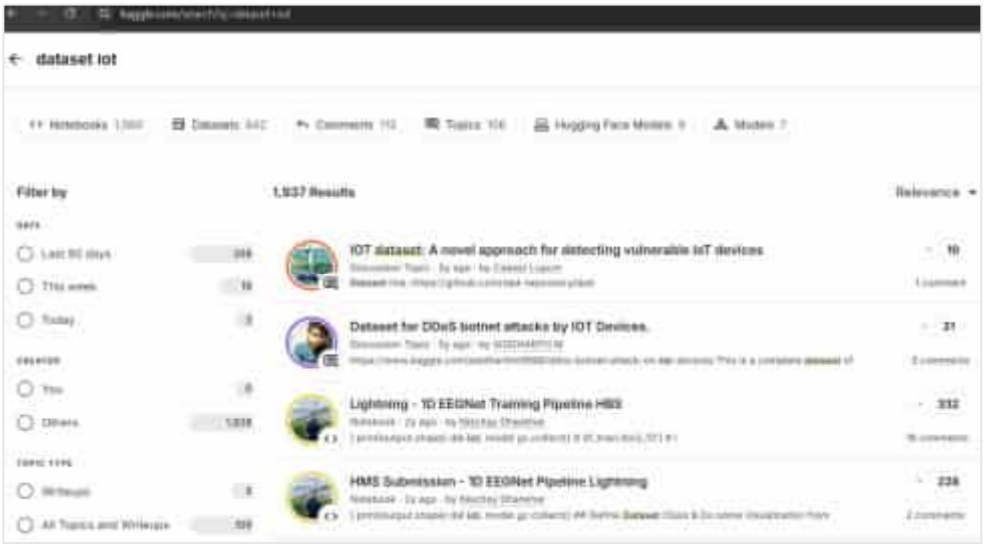
l. **Data.gov (U.S. Government Open Data)**
URL: https://www.data.gov/

m. **FiveThirtyEight Data**
URL: https://data.fivethirtyeight.com/

n. **Awesome Public Datasets (GitHub)**
URL: https://github.com/awesomedata/awesome-public-datasets

o. **OpenML**
URL: https://www.openml.org/

Gambar 2 menampilkan laman Kaggle [3] sebagai salah satu sumber dataset untuk penelitian dengan data sekunder. Contoh sumber lain di antaranya adalah DataCamp [4], Towards Data Science, Real Python, Data Science Central, dan lain sebagainya.



Gambar 3 Laman Kaggle sebagai sumber Data Sekunder [3]

C. Soal Latihan Mandiri

1. Jelaskan perbedaan antara data dan informasi, serta berikan masing-masing satu contoh!
2. Sebutkan dan jelaskan jenis-jenis data berdasarkan: Sifatnya Struktur data dan Sumber data!
3. Apa yang dimaksud dengan *dataset* dalam data sains? Sebutkan komponen utama yang terdapat di dalamnya.!
4. Berikan contoh penggunaan data primer dan data sekunder dalam suatu penelitian!
5. Mengapa pemilihan jenis data yang tepat sangat penting dalam proses analisis data sains?

BAB 3
STATISTIKA UNTUK DATA SAINS

A. Statistika untuk Data Sains

Statistik merupakan salah satu fondasi utama dalam data sains yang berperan penting dalam proses pengumpulan, pengolahan, analisis, dan interpretasi data [5]. Melalui statistik, data yang bersifat mentah dapat diubah menjadi informasi yang bermakna dan dapat digunakan untuk mendukung pengambilan keputusan. Dalam data sains, statistik digunakan untuk memahami karakteristik data, mengidentifikasi pola, mengukur hubungan antar variabel, serta mengevaluasi hasil analisis dan model yang dibangun.

Statistik dalam data sains tidak hanya berfokus pada perhitungan numerik, tetapi juga pada pemahaman konteks data, validitas hasil analisis, serta penyajian informasi dalam bentuk yang mudah dipahami melalui visualisasi data [5].

B. Dasar Statistik dalam Data Sains

1. Mean

Mean merupakan nilai rata-rata dari sekumpulan data yang diperoleh dengan menjumlahkan seluruh nilai data kemudian dibagi dengan jumlah data. Mean digunakan untuk menggambarkan nilai pusat data, namun sensitif terhadap nilai ekstrem (outlier).

2. Median

Median adalah nilai tengah dari sekumpulan data yang telah diurutkan. Median lebih stabil dibandingkan mean ketika data memiliki nilai ekstrem, sehingga sering digunakan untuk menggambarkan kecenderungan pusat data yang tidak berdistribusi normal.

3. Modus

Modus adalah nilai yang paling sering muncul dalam suatu kumpulan data. Modus umum digunakan pada data kategorik maupun data diskrit untuk mengetahui nilai yang paling dominan.

4. Varians

Varians merupakan ukuran penyebaran data yang menunjukkan seberapa jauh nilai data menyebar dari nilai rata-rata. Nilai varians yang besar menunjukkan data lebih bervariasi.

5. Standar Deviasi

Standar deviasi adalah akar kuadrat dari varians dan digunakan untuk mengukur tingkat penyimpangan data terhadap nilai rata-rata dalam satuan yang sama dengan data aslinya.

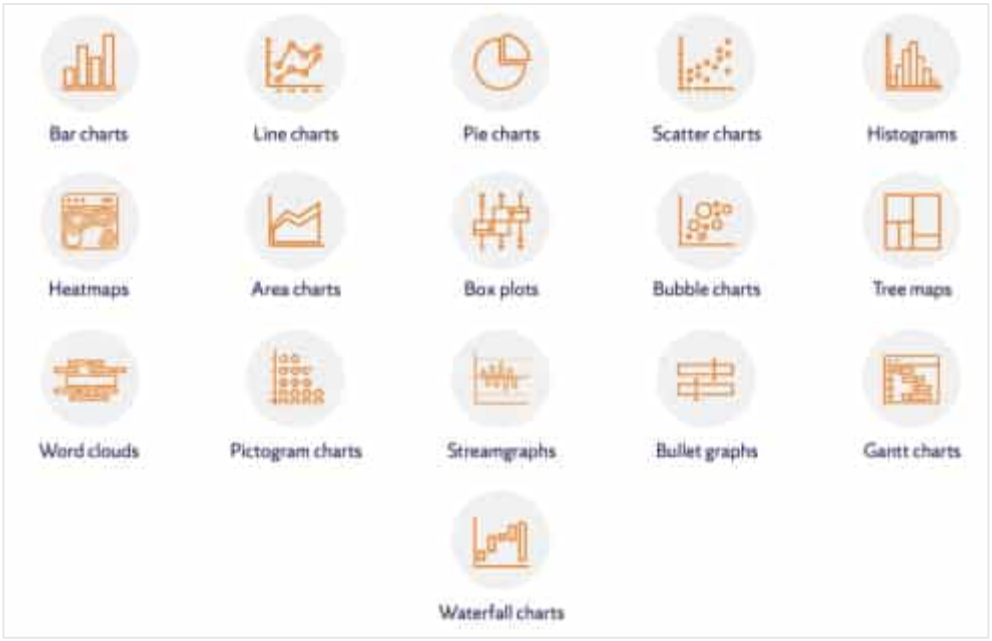
6. *Range* (Rentang)

Range adalah selisih antara nilai maksimum dan nilai minimum dalam suatu kumpulan data. Ukuran ini memberikan gambaran sederhana tentang luas penyebaran data.

7. Statistik untuk Visualisasi Data

Visualisasi data digunakan untuk menyajikan data dan hasil analisis secara grafis agar lebih mudah dipahami. Beberapa bentuk visualisasi yang umum digunakan dalam statistik dan data sains antara lain:

a. Histogram
b. Boxplot
c. Scatterplot



Gambar 4 Contoh Jenis Chart-Statistik dalam Data Sains

Gambar 4 menampilkan ikon berbagai jenis *chart* yang umum digunakan dalam visualisasi data pada data sains. Selain *chart* tersebut, terdapat banyak bentuk visualisasi serta *tools* lain yang juga dapat dimanfaatkan.

C. Penggunaan Statistik untuk Machine Learning pada Data Sains

Statistik memiliki peran yang sangat penting dalam pengembangan dan penerapan *Machine learning* pada data sains. statistik menjadi dasar dalam memahami karakteristik data, mengidentifikasi pola, serta mengevaluasi performa model yang dibangun. Sebelum model *machine learning* diterapkan, analisis statistik digunakan untuk mengeksplorasi data, memahami distribusi variabel, mendeteksi *outlier*, serta menilai hubungan antar fitur melalui ukuran korelasi dan kovariansi.

Dalam tahap *preprocessing* data, konsep statistik digunakan untuk menangani nilai yang hilang (*missing values*), melakukan normalisasi atau standarisasi data, serta melakukan transformasi agar data sesuai dengan asumsi algoritma *machine learning*. Ukuran statistik seperti mean, median, standar deviasi, dan varians berperan penting dalam proses ini untuk memastikan data berada pada skala yang seragam dan tidak menimbulkan bias pada proses pembelajaran model.

Statistik juga berperan dalam pemilihan fitur *(feature selection)*, di mana teknik statistik seperti uji korelasi, uji hipotesis, dan analisis varians (ANOVA) digunakan untuk menentukan fitur-fitur yang paling berpengaruh terhadap variabel target. Dengan memilih fitur yang relevan, model *machine learning* dapat bekerja lebih efisien, mengurangi kompleksitas, dan meningkatkan akurasi prediksi.

Pada tahap pembangunan dan evaluasi model, statistik digunakan untuk mengukur performa model melalui berbagai metrik evaluasi seperti *accuracy, precision, recall*, F1-score, serta analisis kesalahan (*error analysis*). Teknik validasi seperti *cross-validation* juga merupakan pendekatan statistik yang penting untuk menilai kemampuan generalisasi model terhadap data baru dan mencegah terjadinya *overfitting*.

Secara keseluruhan, statistik berfungsi sebagai fondasi teoritis dan praktis dalam *machine learning*. Tanpa pemahaman statistik yang baik, proses pengembangan model *machine learning* berisiko menghasilkan kesimpulan yang keliru. Oleh karena itu, integrasi statistik dan *machine learning* dalam data sains memungkinkan pengambilan keputusan yang lebih akurat, dapat dipercaya, dan berbasis data [5]. Beberapa contoh penerapan statistik dalam data sains meliputi:

1. Regresi linear untuk memodelkan hubungan antar variabel
2. Evaluasi performa model menggunakan metrik statistik seperti *Mean Squared Error* (MSE) dan *R-squared* (R^2)
3. Normalisasi dan standarisasi data
4. Analisis distribusi dan uji statistik untuk memahami karakteristik data

Penggunaan metode statistik dalam *machine learning* dapat bervariasi, tergantung pada jenis algoritma dan tujuan analisis yang digunakan. Setiap algoritma *machine learning* memiliki asumsi dan kebutuhan statistik yang berbeda, sehingga pemahaman terhadap konsep statistik menjadi sangat penting dalam proses analisis data.

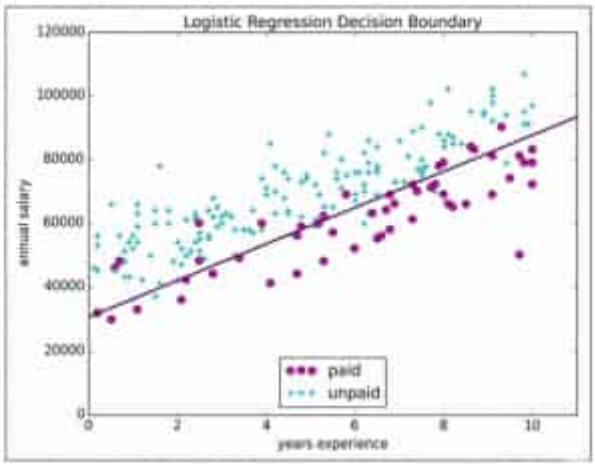
Pertama, regresi linear digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan variabel dependen. Metode ini membantu memahami pola hubungan antar variabel sekaligus digunakan sebagai dasar dalam membangun model prediktif.

Kedua, evaluasi performa model dilakukan menggunakan metrik statistik seperti Mean Squared Error (MSE) dan R-squared (R^2). MSE digunakan untuk mengukur rata-rata kesalahan prediksi model, sedangkan R^2 menunjukkan seberapa besar variasi data yang dapat dijelaskan oleh model. Metrik ini penting untuk menilai kualitas dan keandalan model prediktif.

Ketiga, normalisasi dan standarisasi data merupakan teknik statistik yang digunakan untuk menyamakan skala antar fitur. Proses ini sangat penting pada algoritma tertentu, seperti K-Nearest Neighbor dan Support Vector Machine, agar perbedaan skala data tidak menyebabkan bias dalam proses pembelajaran model.

Keempat, analisis distribusi data dan uji statistik digunakan untuk memahami karakteristik data, seperti bentuk distribusi, tingkat penyebaran, serta keberadaan *outlier*. Uji statistik juga membantu dalam pengambilan keputusan berbasis data, misalnya untuk menguji hipotesis atau menentukan apakah suatu perbedaan antar kelompok data bersifat signifikan.

Dengan demikian, ilmu statistik tidak hanya berfungsi sebagai alat pendukung, tetapi menjadi fondasi utama dalam penerapan *machine learning* pada data sains, mulai dari tahap eksplorasi data hingga evaluasi dan interpretasi hasil model.



Gambar 5 Contoh Penggunaan Statistik dalam Machine Learning-Data Sains

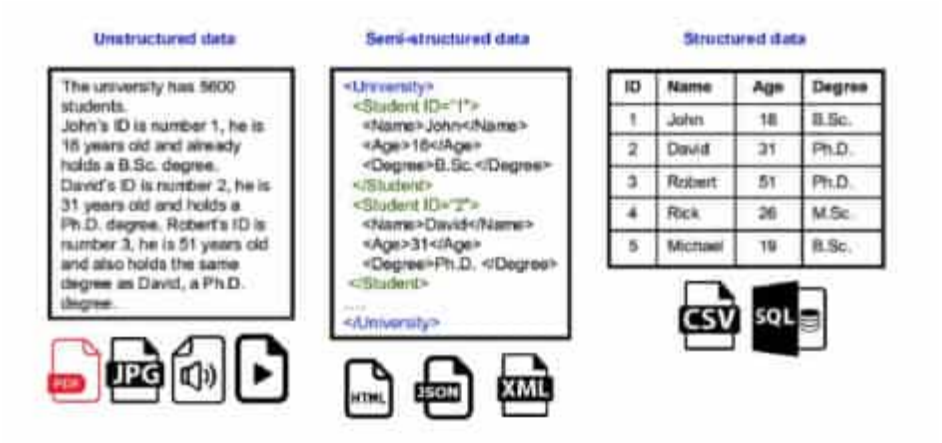
D. Soal Latihan Mandiri

1. Jelaskan peran statistik dalam data sains dan mengapa statistik menjadi dasar penting dalam analisis data!
2. Sebutkan dan jelaskan ukuran pemusatan data (mean, median, modus) serta kegunaannya.!
3. Apa yang dimaksud dengan distribusi data dan mengapa penting dalam analisis Data Sains?
4. Berikan contoh bagaimana konsep statistik digunakan dalam *machine learning*!.

BAB 4
TAHAPAN DATA SAINS

A. Tahapan dalam Data Sains

Data sains merupakan bidang ilmu multidisiplin yang memanfaatkan metode ilmiah, proses, algoritma, dan sistem untuk mengekstraksi pengetahuan serta wawasan dari data terstruktur maupun tidak terstruktur. Data terstruktur adalah data yang tersusun secara sistematis dalam format tertentu, seperti tabel pada basis data relasional, sehingga mudah disimpan, dicari, dan dianalisis (misalnya data nilai, data transaksi, atau data sensus). Sementara itu, data tidak terstruktur adalah data yang tidak memiliki format baku, seperti teks bebas, dokumen, gambar, audio, video, unggahan media sosial, dan data log, yang memerlukan teknik khusus untuk dapat dianalisis. Perbandingan antara data tidak terstruktur dan data terstruktur serta contoh *file extension* dapat dilihat pada Gambar 6.



Gambar 6 Contoh data terstruktur untuk pengolahan data sains

Proses pengolahan data dalam Data sains terdiri atas beberapa tahapan utama, yaitu pengumpulan data, pembersihan data, transformasi data, analisis data, dan visualisasi data. Pada tahap pengumpulan data, data diperoleh dari berbagai sumber, seperti basis data, berkas digital, sensor, maupun sumber daring. Selanjutnya, dilakukan pembersihan data untuk menghilangkan kesalahan, duplikasi, serta ketidakakuratan yang dapat memengaruhi hasil analisis. Data yang

telah diproses kemudian dianalisis untuk menghasilkan informasi yang bermakna, dan hasil analisis tersebut disajikan dalam bentuk visualisasi agar lebih mudah dipahami dan diinterpretasikan.

Dalam melakukan analisis data, berbagai bahasa pemrograman dan alat bantu digunakan untuk mendukung proses pengolahan dan interpretasi data. Bahasa pemrograman yang umum digunakan antara lain R [6] dan Python [1] [7], yang didukung oleh beragam pustaka untuk analisis statistik, penerapan metode machine learning, serta pengolahan data. Selain itu, berbagai teknik dan alat visualisasi, seperti pembuatan grafik dan kurva, digunakan untuk menyajikan hasil analisis secara informatif. Visualisasi data ini bertujuan untuk membantu perancang sistem maupun pengambil keputusan dalam memahami informasi yang dihasilkan dari data.

Pemrosesan data terdiri atas beberapa tahapan yang saling berkaitan. Tahap awal adalah pengumpulan dan penyimpanan data, yaitu proses menghimpun data mentah dari berbagai sumber seperti basis data, perangkat IoT, sensor, web, dan media sosial, kemudian menyimpannya secara efisien untuk keperluan pengolahan lebih lanjut. Selanjutnya, dilakukan pembersihan dan prapemrosesan data untuk menangani data yang hilang, tidak konsisten, atau mengandung gangguan (*noise*), sehingga kualitas dan keandalannya dapat terjaga. Tahap analisis dan interpretasi data bertujuan untuk menerapkan teknik statistik dan analitik guna mengidentifikasi pola, tren, serta hubungan antar data. Agar hasil analisis lebih mudah dipahami, data divisualisasikan dalam bentuk grafik, diagram, dasbor, dan alat visual lainnya.

Tahapan lanjutan meliputi penerapan *machine learning* dan pemodelan prediktif, yaitu pembangunan algoritma yang mampu belajar dari data untuk menghasilkan prediksi atau mengotomatisasi pengambilan keputusan. Selain itu, analitik big data digunakan untuk mengelola dan menganalisis kumpulan data berukuran sangat besar yang melampaui kemampuan alat pemrosesan konvensional. Seluruh proses tersebut dapat diperkuat melalui integrasi kecerdasan buatan, yang memungkinkan sistem meniru kecerdasan manusia dan terus meningkatkan kinerjanya melalui pembelajaran berbasis data. Sepanjang seluruh tahapan pemrosesan, aspek etika dan privasi data harus diperhatikan untuk menjamin keadilan, keamanan, dan penggunaan data secara bertanggung jawab.

B. Pengumpulan dan Penyimpanan Data *(Data Collection)*

Tahap ini merupakan proses awal dalam pemrosesan data, yaitu mengumpulkan data mentah dari berbagai sumber dan menyimpannya agar dapat diakses serta diolah lebih lanjut. Sumber data dapat berasal dari basis data perusahaan, sensor,

perangkat Internet of Things (IoT), aplikasi web, maupun media sosial. Penyimpanan data harus dirancang secara efisien agar mampu menangani volume data yang besar dan menjamin keamanan data.

Contoh:
Data suhu dan kelembapan yang dikumpulkan secara *real-time* dari sensor cuaca disimpan dalam basis data *cloud* untuk dianalisis lebih lanjut.

Pada riset yang membutuhkan data primer, tahap ini sering disebut data generation atau *data acquisition*. Tahap pengumpulan data ini melibatkan proses mendapatkan data mentah secara langsung dari berbagai sumber untuk keperluan analisis.

Tujuan pengumpulan data adalah mendapatkan data yang relevan, lengkap, dan berkualitas tinggi, sehingga hasil analisis menjadi akurat dan dapat dipercaya. Sumber data dapat beragam, termasuk survei, sensor, perangkat IoT, media sosial, *database*, hingga catatan aktivitas di web (*web logs*). Proses ini merupakan langkah awal yang sangat penting dalam data sains, karena kualitas dan relevansi data yang dikumpulkan akan menentukan keberhasilan analisis dan pengembangan model prediktif selanjutnya.

Setelah proses pengumpulan, data dapat disimpan untuk digunakan di masa mendatang. Secara umum, tujuan penyimpanan data adalah menyimpan data yang telah dikumpulkan secara aman dalam *database* atau sistem penyimpanan berbasis *cloud*. Metode penyimpanan bisa dilakukan melalui penyimpanan lokal (*local storage*), media penyimpanan portabel, *cloud storage*, maupun *data warehouse*.

Dalam beberapa kasus, penyimpanan juga dapat dilakukan melalui integrasi dengan database yang menggunakan bahasa pemrograman atau *framework* lain, misalnya SQL, Hadoop, AWS, atau Google Cloud.

Selain untuk menyimpan data, tujuan utama *data storage* adalah untuk memastikan skalabilitas, aksesibilitas, dan keamanan data agar dapat digunakan secara efektif di masa mendatang.

C. Pembersihan dan Pemrosesan Data *(Data Cleaning and Preprocessing)*

Data Cleaning adalah proses memeriksa atau menghapus kesalahan, duplikasi dan tidakkonsistenan untuk meningkatkan kualitas data, dengan tehnik penanganan bagian data yang hilang, normalisasi, integrasi data dan transformasi. Tahap ini

bertujuan menyiapkan data yang bersih, terstruktur dan konsisten untuk analisis selanjutnya.

Pada tahap ini, seperti terlihat pada Gambar 7, data mentah diperbaiki agar layak digunakan. Selain penanganan data yang hilang, proses ini melakukan pengurangan *noise* seperti menghapus duplikasi data dan mengoreksi data yang tidak konsisten agar data sesuai dengan kebutuhan analisis.

Contoh:
Pada *dataset* penjualan, nilai transaksi yang kosong diisi dengan nilai rata-rata, dan menghapus duplikasi data akibat kesalahan pencatatan.

Gambar 7 Contoh Data Cleaning dalam Data Sains

D. Data Transformation

Pada tahap *preprocessing*, data mentah diperbaiki agar layak digunakan. Selain penanganan data yang hilang, penghapusan data duplikat, koreksi data yang tidak konsisten, proses ini juga melakukan pengurangan *noise*, agar transformasi data sesuai dengan kebutuhan analisis. Selanjutnya, tahap analisis dan interpretasi data bertujuan untuk mengekstraksi informasi bermakna dari data melalui teknik statistik, analitik, atau eksplorasi data. Hasil analisis digunakan untuk memahami pola, tren, dan hubungan antar variabel yang terdapat dalam data sehingga dalam beberapa referensi disebutkan sebagai tahap *data transformation* dan analisis.

Data eksplorasi dan analisis adalah tahap di mana data diperiksa secara mendalam untuk menemukan pola, korelasi, dan insight yang terkandung di dalamnya. Pada tahap ini, digunakan berbagai *tools* dan metode prediktif seperti Python [1] [7], R [6], Excel, Tableau, Power BI, Matplotlib, Seaborn, dan lain-lain, untuk mendukung pemahaman data.

Contoh :
Analisis data penjualan dilakukan untuk mengetahui produk yang paling laku pada periode tertentu dan hubungan antara harga dengan jumlah penjualan.

Tujuan dari proses ini adalah:

1. Menyajikan informasi yang akurat sebagai hasil produk data sains.
2. Memahami tren dan hubungan dalam data, sehingga dapat membantu pengambilan keputusan berbasis data.

Proses ini sering pula disebut dengan Exploratory Data Analysis (EDA), dan menjadi langkah penting sebelum membangun model prediktif atau sistem kecerdasan buatan.

Analisis data deskriptif dan ekploratif adalah tahapan dalam data sains yang bertujuan untuk memahami karakteristik dasar data serta menemukan pola awal, hubungan, dan anomali yang terdapat di dalamnya sebelum dilakukan analisis lanjutan atau pemodelan.

Analisis data deskriptif berfokus pada proses merangkum dan menggambarkan kondisi data secara kuantitatif. Analisis ini menjawab pertanyaan apa yang terjadi dengan menggunakan ukuran statistik seperti rata-rata, median, modus, minimum, maksimum, varians, dan standar deviasi, serta visualisasi sederhana seperti tabel dan grafik.

Sementara itu, analisis data eksploratif (Exploratory Data Analysis/EDA) bertujuan untuk mengeksplorasi data secara lebih mendalam guna mengidentifikasi pola, tren, korelasi, dan *outlier*. EDA sering menggunakan visualisasi data dan teknik statistik untuk membantu *data scientist* memahami struktur data, menguji asumsi awal, serta menentukan pendekatan analisis atau model yang paling tepat.

Istilah lain yang sering digunakan dalam *data transformation* antara lain *data integration, data mapping, data extraction, data merging*, dan *data loading*. Secara teknis, istilah-istilah tersebut merujuk pada proses penggabungan dan pengolahan data dari berbagai sumber dengan tujuan untuk menghasilkan informasi yang lebih lengkap dan terpadu.

Proses ini bertujuan untuk menghindari duplikasi data, memudahkan analisis, serta mendukung pengambilan keputusan yang lebih baik. Dengan melakukan transformasi dan integrasi data secara tepat, data yang awalnya terpisah dan tidak seragam dapat disatukan menjadi satu struktur yang konsisten dan siap digunakan untuk analisis lanjutan maupun sistem pendukung keputusan.

E. Visualisasi Data

Selanjutnya adalah tahapan presentasi atau visualisasi data yang merupakan salah satu kemampuan fundamental seorang *data scientist*. Visualisasi data digunakan untuk menyajikan hasil analisis dalam bentuk visual agar lebih mudah dipahami oleh pengguna data. Teknik visualisasi diantaranya meliputi grafik batang, grafik garis, diagram lingkaran, peta panas *(heatmap)*, dan dasbor interaktif.

> **Contoh:**
> Hasil analisis penjualan ditampilkan dalam bentuk grafik bulanan untuk menunjukkan tren peningkatan atau penurunan penjualan.

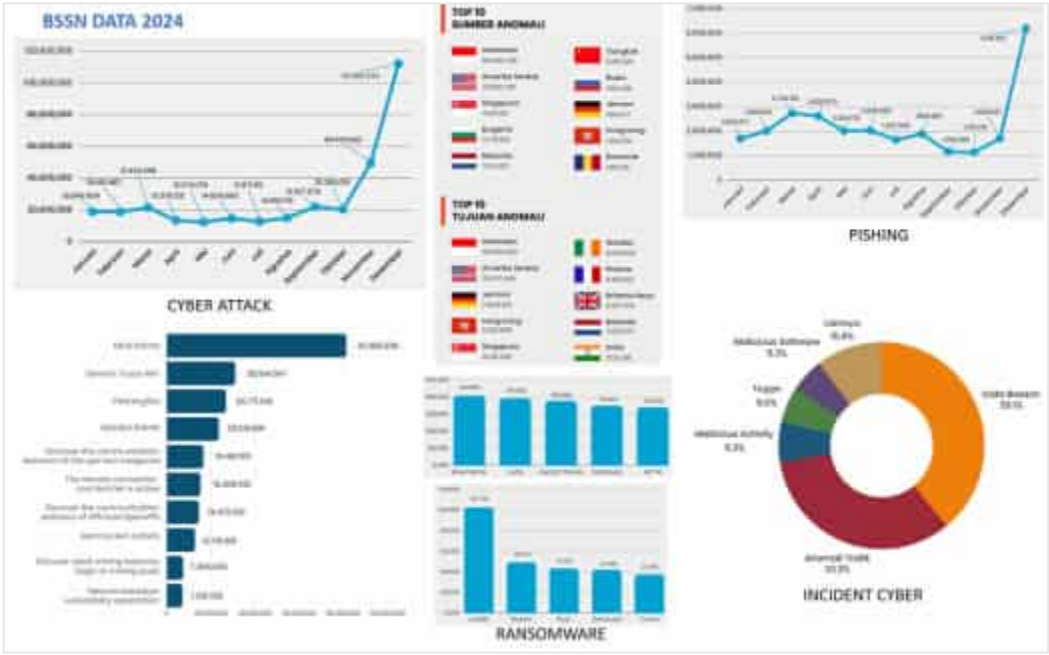
Visualisasi data memiliki dua tujuan utama yaitu eksplorasi data dan komunikasi data ke pengguna data sains. Eksplorasi data *(data exploration)* digunakan untuk memahami pola, tren, dan hubungan dalam data sebelum melakukan analisis lebih lanjut. Komunikasi data *(data communication)* digunakan untuk menyajikan hasil analisis secara jelas dan informatif kepada pengguna atau pemangku kepentingan.

Visualisasi data membantu menyampaikan informasi kompleks dalam bentuk visual yang lebih mudah dipahami. Visualisasi yang efektif memudahkan pengguna dalam memahami *insight* dari data dan mendukung pengambilan keputusan berbasis data. Beberapa teknik visualisasi yang umum digunakan antara lain:

1. Grafik Batang *(Bar Chart)*
   Digunakan untuk menampilkan perbandingan antar kategori, baik secara horizontal maupun vertikal.
2. Grafik Garis *(Line Chart)*
   Menggambarkan perubahan nilai dari waktu ke waktu, sehingga baik digunakan untuk analisis tren.
3. Diagram Lingkaran *(Pie Chart)*
   Digunakan untuk menunjukkan proporsi atau persentase suatu kategori terhadap keseluruhan data.
4. Peta Panas *(HeatMap)*
   Digunakan untuk mengilustrasikan intensitas nilai atau korelasi antar variabel melalui gradasi warna.
5. Dasbor Interaktif *(Dashboard)*
   Menggabungkan beberapa jenis visualisasi dalam satu tampilan dengan kemampuan interaksi, sehingga memungkinkan analisis data yang lebih mendalam dan dinamis.
6. Visualisasi Pemodelan Data Prediktif
   Visualisasi ini menyajikan hasil pemodelan prediktif, seperti prediksi tren atau klasifikasi, yang umumnya dihasilkan menggunakan teknik *machine learning*

Visualisasi jenis ini membantu pengguna dalam memahami performa model dan hasil prediksi secara intuitif.

Pada Gambar 8 berikut, dapat kita lihat contoh visualisasi data dari Lanskap Keamanan Siber 2024 yang diterbitkan oleh Badan Siber dan Sandi Negara. Visualisasi ini bertujuan untuk menyajikan informasi kompleks mengenai ancaman dan tren keamanan siber secara lebih mudah dipahami, sehingga pembaca dapat dengan cepat melihat pola, risiko, dan prioritas mitigasi yang diperlukan. Teknik visualisasi seperti grafik, diagram, atau peta panas membantu menyederhanakan data sehingga lebih informatif dan komunikatif.



Gambar 8 Contoh Jenis Visualisasi Data

Pada tahap pengolahan data dalam data sains, data terlebih dahulu divisualisasikan secara umum, kemudian melakukan proses/tahapan berikutnya, antara lain *data evaluation, data validation, data interpretation*, dan pengambilan keputusan *(decision making)*. Tahapan-tahapan ini umumnya diterapkan ketika *data scientist* menggunakan *tools* atau metode lanjutan, seperti *machine learning*.

Seiring dengan perkembangan data sains, bidang ini semakin erat kaitannya dengan ilmu komputer dan pemrograman, khususnya dalam penerapan berbagai teknik yang digunakan dalam *machine learning*. Oleh karena itu, pemahaman yang mendalam mengenai evaluasi data, validasi data, interpretasi hasil, serta proses

pengambilan keputusan menjadi sangat penting dalam pengembangan sistem analitik dan prediktif. Pembahasan lebih lanjut mengenai *data evaluation, data validation, data interpretation*, serta tahapan pengambilan keputusan akan dijelaskan secara khusus pada bagian pembahasan *machine learning*.

F. Soal Latihan Mandiri

1. Jelaskan tahapan utama dalam data sains secara berurutan!
2. Mengapa tahap *data collection* sangat menentukan keberhasilan analisis data sains?
3. Sebutkan dan jelaskan proses yang dilakukan pada tahap *data cleaning* dan *preprocessing*!
4. Apa yang dimaksud dengan *data transformation* dan berikan contoh penerapannya!
5. Jelaskan tujuan visualisasi data dan sebutkan minimal tiga jenis visualisasi yang umum digunakan dalam data sains!

BAB 5
BAHASA PEMROGRAMAN DATA SAINS

A. Bahasa Pemrograman dalam Data Sains

Bahasa pemrograman adalah sekumpulan instruksi yang digunakan untuk memberi tahu komputer bagaimana melakukan tugas tertentu. Bahasa ini memungkinkan manusia menulis kode yang dapat dipahami dan dijalankan oleh mesin. Dalam data sains, bahasa pemrograman digunakan untuk membersihkan data, menganalisis informasi, membuat model statistik, serta menyajikan visualisasi data. Contoh bahasa pemrograman populer di data sains adalah Python [1] [7] [8], R [6], dan MATLAB, masing-masing memiliki keunggulan tersendiri dalam kemudahan penggunaan, analisis statistik, dan komputasi numerik.

Dalam dunia data sains, beberapa bahasa pemrograman menjadi pilihan utama karena kemampuannya dalam analisis dan visualisasi data. Python sangat populer karena sintaksnya sederhana dan mudah dibaca, membuatnya ramah bagi pemula [1] [7] [8]. Selain itu, ekosistemnya sangat kaya dengan pustaka seperti NumPy, pandas, scikit-learn, matplotlib, dan TensorFlow, serta mudah diintegrasikan dengan database, API, dan lingkungan pemrograman lain. Python juga serbaguna, cocok untuk pembersihan data, pemodelan statistik, hingga pengembangan aplikasi kecerdasan buatan.

Sementara itu, R dirancang khusus untuk komputasi statistik dan visualisasi data [6]. Dengan paket seperti ggplot2 dan shiny, R memungkinkan pembuatan grafik dan dasbor interaktif yang canggih, sehingga banyak digunakan di lingkungan akademik untuk penelitian, eksplorasi data, dan pemodelan statistic [6].

Di sisi lain, MATLAB unggul dalam komputasi numerik, terutama untuk operasi matriks, simulasi, dan pengembangan algoritma. MATLAB juga menyediakan *toolbox* siap pakai untuk pemrosesan sinyal, *machine learning*, dan analisis citra, sehingga banyak dipakai di industri teknik dan penelitian ilmiah.

Secara umum, pemilihan bahasa pemrograman dalam data sains disesuaikan dengan tujuan analisis, karakteristik data, dan kebutuhan pengguna, sehingga hasil

pengolahan data dapat diperoleh secara efektif dan optimal. Secara umum, Gambar 9 menunjukkan pengelompokan kemampuan yang diperlukan dalam keilmuan data sains dan data analis.

Gambar 9 Kemampuan (Skills) dalam Keilmuan Data Sains

Untuk memudahkan lingkungan pemrogramman, dalam data sains sering menggunakan aplikasi Jupyter Notebook dari JupyterLab. Jupyter Notebook adalah lingkungan pengembangan interaktif berbasis web yang digunakan untuk bekerja dengan *notebook*, kode, dan data. Antarmuka yang fleksibel memungkinkan pengguna mengonfigurasi dan mengatur alur kerja sesuai kebutuhan, khususnya dalam bidang data sains, komputasi ilmiah, jurnalisme komputasional, dan *machine learning*. Dengan desain yang modular, JupyterLab mendukung penggunaan berbagai ekstensi yang dapat memperluas dan memperkaya fungsionalitas, sehingga menjadi alat yang sangat efektif untuk analisis dan eksplorasi data secara interaktif.

B. Penggunaan Database dalam Data Sains

Database adalah sistem penyimpanan data yang terstruktur sehingga data dapat diakses, dikelola, dan diperbarui secara efisien. Database memungkinkan penyimpanan informasi dalam jumlah besar, umumnya dalam bentuk tabel yang terdiri dari baris dan kolom, sehingga data mudah diorganisasi dan diproses.

Dalam data sains, *database* berperan penting sebagai tempat penyimpanan data mentah, data hasil pembersihan, maupun hasil analisis. Data yang tersimpan dalam *database* dapat diambil kembali untuk keperluan penelitian, pemodelan, analisis statistik, dan visualisasi data [9].

Berdasarkan strukturnya, *database* dalam data sains dapat dibedakan menjadi dua jenis utama. *Database* relasional, seperti MySQL dan PostgreSQL, menggunakan struktur tabel dengan relasi antar data. Sementara itu, *database* non-relasional (NoSQL), seperti MongoDB dan Cassandra. Jenis *database* ini lebih fleksibel dalam menangani data tidak terstruktur dan berskala besar. Masing-masing jenis database memiliki keunggulan tersendiri sesuai dengan kebutuhan pengelolaan data.

Penggunaan basis data dalam data sains umumnya melibatkan bahasa pemrograman SQL (Structured Query Language), yang digunakan untuk melakukan pengambilan data *(query)*, penyaringan, penggabungan, serta manipulasi data. Penguasaan SQL menjadi keterampilan penting bagi *data scientist* karena memungkinkan pengelolaan data secara efektif sebelum data digunakan pada tahap analisis dan pemodelan lanjutan [9].

Penggunaan *database* yang lebih lanjut *(advanced)* dalam data sains umumnya berkaitan dengan pemanfaatan *cloud database*, terutama untuk kebutuhan analisis *big data*, analisis data yang berasal dari berbagai lokasi, atau penggunaan alat analisis data yang tersebar *(distributed analytics tools)*.

Dalam kondisi tersebut, diperlukan sistem *database* dan jaringan *(network)* yang andal agar proses penyimpanan, pengolahan, dan analisis data dapat berjalan secara efisien, cepat, dan optimal. *Cloud database* memungkinkan data diakses secara *real-time* oleh berbagai sistem dan pengguna, mendukung skalabilitas yang tinggi, serta memudahkan kolaborasi dalam lingkungan data sains modern.

Dengan dukungan infrastruktur jaringan yang baik, cloud database menjadi fondasi penting dalam membangun arsitektur data sains terdistribusi, sehingga analisis data dapat dilakukan secara konsisten dan terintegrasi meskipun menggunakan sumber data dan alat analisis yang berbeda-beda.

C. Soal Latihan Mandiri

1. Mengapa tahap *data collection* sangat menentukan keberhasilan analisis data sains?
2. Jelaskan pengertian bahasa pemrograman dan perannya dalam data sains!
3. Mengapa Python menjadi bahasa pemrograman yang paling populer dalam data sains?
4. Jelaskan fungsi *database* dalam data sains dan bagaimana hubungannya dengan bahasa pemrograman!
5. Apa perbedaan antara *database* relasional dan non-relasional, serta contoh penggunaannya dalam analisis data?

BAB 6
MACHINE LEARNING

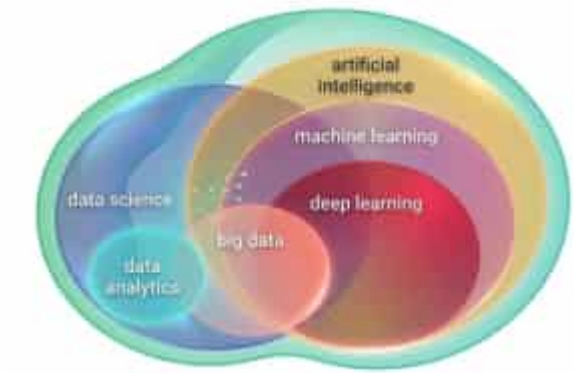
A. Machine Learning dalam Data Sains

Data sains adalah disiplin ilmu yang menggabungkan matematika, statistika, dan ilmu komputer untuk menganalisis data. Tujuan utamanya adalah melakukan analisis data pada berbagai skala, mulai dari sampel kecil hingga populasi besar. Dalam praktiknya, data sains menggunakan algoritma khusus untuk:

1. Menemukan atau mengekstrak informasi yang berguna *(data mining)*.
2. Mengidentifikasi pola yang ada dalam data.
3. Membuat prediksi akurat guna mendukung pengambilan keputusan.

Selain itu, data sains juga berperan dalam pengembangan sistem kecerdasan buatan *(Artificial Intelligence/AI)* yang mampu belajar secara otomatis dan meningkat kinerjanya dari waktu ke waktu melalui *machine learning*.

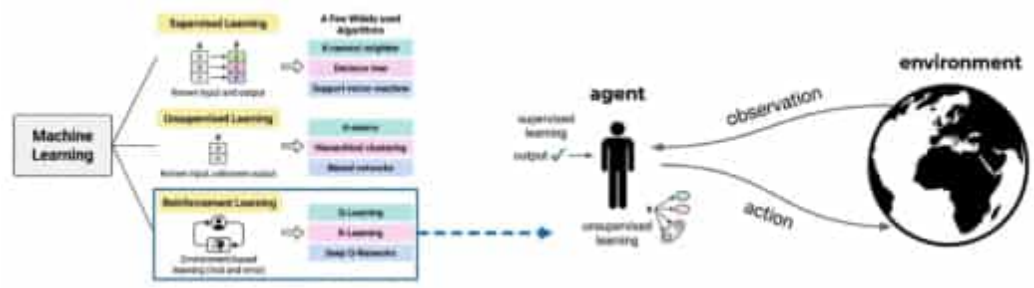
Secara umum, sebagaimana digambarkan pada Gambar 10, hubungan antara data sains dan *machine learning* dapat dijelaskan bahwa data sains menyediakan data, metode, dan analisis statistik, sedangkan *machine learning* memanfaatkan data tersebut untuk membangun model prediktif atau sistem yang belajar sendiri, sehingga hasil analisis data sains dapat digunakan untuk pengambilan keputusan yang lebih cerdas dan otomatis.



Gambar 10 Hubungan Keilmuan Data Sains dengan *Machine Learning*

B. Kategori *Machine Learning*

Machine learning merupakan bidang keilmuan yang luas dan terus berkembang seiring pesatnya kemajuan data Sains. Dalam buku ini, pembahasan difokuskan secara umum dan konseptual pada kategori utama *machine learning*, disertai contoh penerapan sederhana yang relevan dengan domain keamanan informasi dan keamanan siber.



Gambar 11 Kategori Machine Learning

Sesuai ilustrasi pada Gambar 11, *machine learning* dapat dikategorikan sebagai berikut:

1. Supervised Learning

Supervised Learning merupakan metode *machine learning* yang menggunakan data berlabel, di mana setiap data masukan (input) telah memiliki nilai keluaran (output) atau target yang diketahui. Model dilatih untuk mempelajari hubungan antara fitur dan label sehingga mampu melakukan prediksi terhadap data baru yang belum pernah dilihat sebelumnya. Metode ini umumnya digunakan untuk tugas klasifikasi (misalnya deteksi serangan, klasifikasi email spam) dan regresi (misalnya prediksi harga atau jumlah permintaan). Contoh algoritma *supervised learning* antara lain Linear Regression, Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes, Support Vector Machine (SVM), dan Random Forest.

Contoh studi kasus:

Dalam bidang keamanan siber, *supervised learning* digunakan untuk mendeteksi serangan pada sistem Smart Home IoT, di mana data lalu lintas jaringan telah diberi label sebagai aktivitas normal atau jenis serangan tertentu (misalnya DoS, DDoS, atau Brute Force). Algoritma seperti K-Nearest Neighbor (KNN), Naive Bayes, dan Random Forest dilatih menggunakan data berlabel tersebut untuk mengklasifikasikan aktivitas jaringan secara otomatis.

2. Unsupervised Learning

Unsupervised learning merupakan metode *machine learning* yang menggunakan data tanpa label, di mana model tidak mengetahui kategori atau target tertentu. Tujuan utama metode ini adalah untuk menemukan pola tersembunyi, struktur data, atau pengelompokan *(clustering)* berdasarkan kemiripan karakteristik data. *Unsupervised learning* sering digunakan dalam eksplorasi data, segmentasi pelanggan, deteksi anomali, dan reduksi dimensi. Contoh algoritma yang umum digunakan antara lain K-Means Clustering, Hierarchical Clustering, DBSCAN, serta teknik reduksi dimensi seperti Principal Component Analysis (PCA).

Contoh studi kasus:

Dalam konteks *smart city*, *unsupervised learning* dapat digunakan untuk mengelompokkan pola konsumsi energi rumah tangga tanpa mengetahui kategori sebelumnya. Algoritma K-Means atau DBSCAN dapat mengidentifikasi kelompok rumah dengan pola penggunaan energi yang serupa, sehingga pemerintah atau pengelola kota dapat merancang kebijakan efisiensi energi yang lebih tepat sasaran. Selain itu, teknik seperti Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data sensor IoT agar lebih mudah dianalisis dan divisualisasikan.

3. Reinforcement Learning

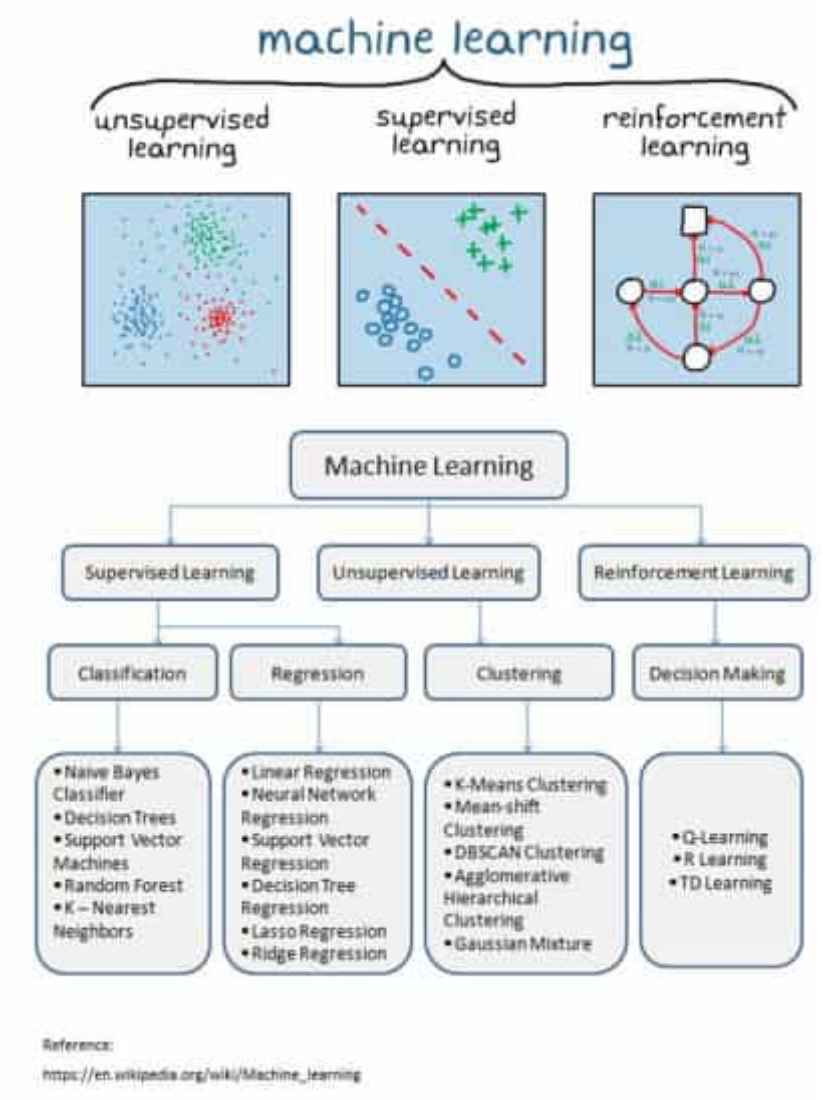
Metode *machine learning* di mana model belajar melalui interaksi dengan lingkungan, dengan cara menerima umpan balik berupa *reward* atau *punishment* untuk menentukan tindakan terbaik secara bertahap. *Reinforcement learning* merupakan metode *machine learning* di mana model (agen) belajar melalui interaksi langsung dengan lingkungan. Dalam metode ini, agen tidak menggunakan data berlabel, tetapi memperoleh umpan balik berupa *reward* atau *punishment* berdasarkan tindakan yang diambil. Tujuan utama *reinforcement learning* adalah mempelajari strategi atau kebijakan *(policy)* terbaik untuk memaksimalkan total *reward* dalam jangka panjang. Metode ini banyak digunakan pada sistem pengambilan keputusan berurutan, seperti robotika, game AI, kendaraan otonom, serta optimasi sistem cerdas.

Contoh studi kasus:

Dalam bidang transportasi cerdas *(smart transportation)* pada *smart city*, *reinforcement learning* dapat digunakan untuk mengatur sistem lampu lalu lintas secara adaptif. Agen *machine learning* mempelajari pola kepadatan kendaraan dan menyesuaikan durasi lampu hijau atau merah untuk mengurangi

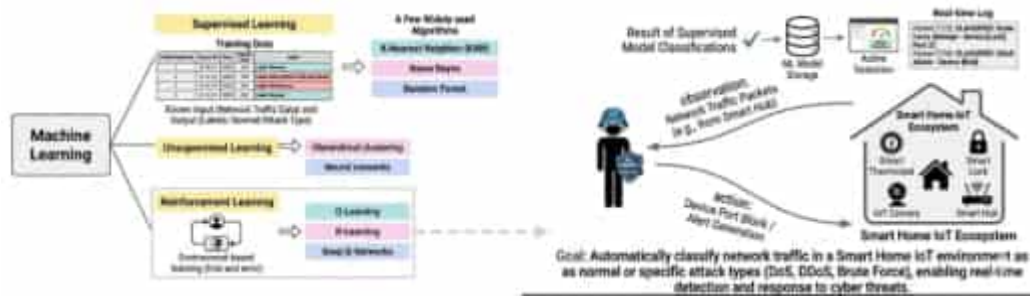
kemacetan. Contoh lainnya adalah penerapan pada robotika, *game AI*, dan kendaraan otonom, di mana sistem harus mengambil keputusan secara dinamis berdasarkan kondisi lingkungan.

Dalam setiap kategori *machine learning*, terdapat algoritma atau metode yang dapat digunakan sebagai *tools* analisis dalam data sains. Visualisasi data dapat dilakukan sesuai dengan kebutuhan analisis dan kedalaman informasi yang diproses. Gambar 12 memberikan skema kategori *machine learning* beserta beberapa algoritma yang termasuk dalam masing-masing kategori tersebut.



Gambar 12 Contoh Metode dan Algoritma dalam setiap Kategori *Machine Learning*

Perbedaan metode, algoritma, dan visualisasi dapat menghasilkan hasil analisis akhir yang berbeda. Oleh karena itu, seperti diilustrasikan pada Gambar 13, seorang *data scientist* harus memiliki kemampuan untuk menentukan metode yang tepat pada setiap tahapan analisis, serta mampu mengeksplorasi berbagai *tools* yang tersedia guna mengoptimalkan hasil analisis dalam mendukung pengambilan keputusan.

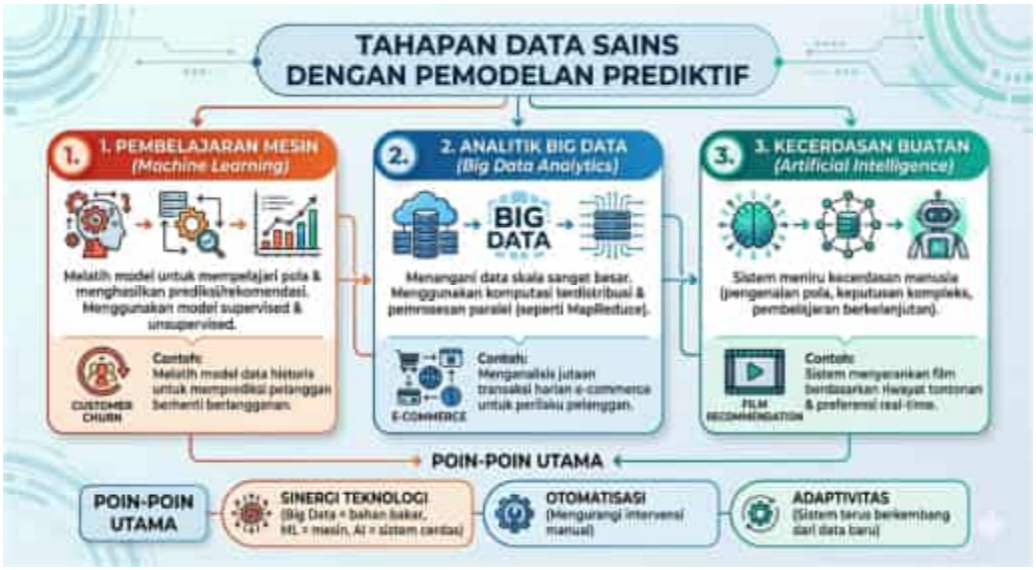


Gambar 13 Ragam Metode dan Algoritma dalam Proses Data Sains

C. Machine Learning dan Pemodelan Prediktif

Tahap *machine learning* dan pemodelan prediktif menggunakan data untuk melatih model yang mampu mempelajari pola dan menghasilkan prediksi atau rekomendasi. Model ini dapat digunakan untuk otomatisasi pengambilan keputusan. Sebagai contoh, model *machine learning* dapat dilatih menggunakan data historis pelanggan untuk memprediksi kemungkinan pelanggan berhenti berlangganan. Pada tingkat yang lebih lanjut, tahap berikutnya adalah analitik *big data*, yang menangani pemrosesan dan analisis data dalam skala sangat besar yang tidak dapat ditangani oleh sistem tradisional. Teknologi khusus, seperti komputasi terdistribusi dan pemrosesan paralel, digunakan untuk mengelola data tersebut. Misalnya, perusahaan *e-commerce* dapat menganalisis jutaan data transaksi harian untuk memahami perilaku pelanggan dan mengoptimalkan strategi pemasaran.

Selanjutnya, integrasi kecerdasan buatan memungkinkan sistem meniru kecerdasan manusia, termasuk pengenalan pola, pengambilan keputusan, dan pembelajaran berkelanjutan dari data baru. Contohnya, sistem rekomendasi film menggunakan kecerdasan buatan untuk menyarankan film berdasarkan riwayat tontonan pengguna. Perbedaan pemodelan prediktif antara *machine learning*, analisis *big data*, dan *artificial intelligence* disajikan pada Gambar 14.



Gambar 14 Data Sains dengan Pemodelan Prediktif

Dalam beberapa kondisi atau analisa data sains, tahap terakhir yang perlu dilakukan adalah tahapan perlindungan etika dan privasi data *(data ethics and privacy)*. Tahap ini menekankan pengelolaan data secara etis dan aman, termasuk perlindungan data pribadi, transparansi dalam penggunaan data, serta kepatuhan terhadap regulasi yang berlaku. Sebagai ilustrasi, data pengguna pada aplikasi kesehatan dienkripsi dan hanya digunakan untuk tujuan yang telah disetujui oleh pengguna.

D. Model Evaluation dan Data Validation

Model evaluation dan *data validation* merupakan tahapan penting dalam *machine learning* yang bertujuan untuk mengevaluasi apakah model yang dibangun telah memprediksi dengan benar dan bekerja secara optimal. Proses ini dilakukan dengan mengukur kinerja model menggunakan data yang tidak digunakan pada saat pelatihan, sehingga hasil evaluasi bersifat objektif dan dapat dipercaya.

1. ***Train-Test Split***
   Membagi *dataset* menjadi data latih *(training data)* dan data uji *(testing data)* untuk mengukur kemampuan generalisasi model.
2. ***Cross-Validation***
   Teknik validasi yang membagi data ke dalam beberapa bagian *(fold)* dan melakukan pelatihan serta pengujian secara berulang untuk mendapatkan hasil evaluasi yang lebih stabil.

3. ***Confusion Matrix***
   Digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi dan nilai sebenarnya.

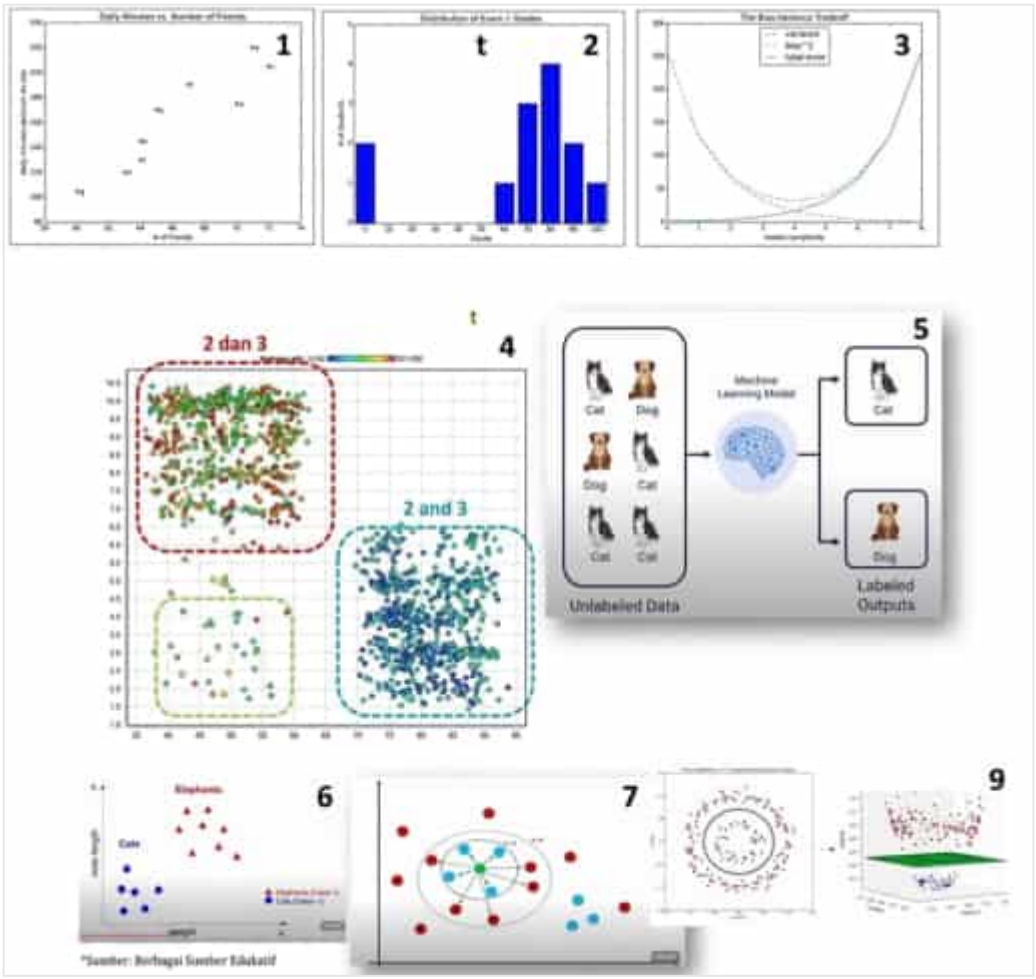
4. ***Evaluation Metrics***
   Metode pengukuran kinerja model, seperti akurasi, presisi, *recall*, *F1-score*, dan ROC-AUC, yang dipilih sesuai dengan jenis permasalahan.

E. Data Interpretation dan Decision Making

Tahap *data interpretation* dan *decision making* dilakukan dengan memanfaatkan hasil analisis *(analytical results)* untuk mengarahkan tindakan atau strategi yang akan diambil. Tujuan utama dari tahap ini adalah mendukung pengambilan keputusan berbasis bukti *(evidence-based decision making)* serta mendorong peningkatan kinerja dan perbaikan proses bisnis.

Secara umum, analisis data pada tahap ini dapat dikelompokkan ke dalam empat kategori utama:

1. ***Descriptive Analysis***
   Analisis yang bertujuan untuk menggambarkan kondisi saat ini dan memberikan ringkasan mengenai apa yang telah terjadi berdasarkan data historis.
2. ***Diagnostic Analysis***
   Analisis yang menggunakan metode statistik untuk menjelaskan penyebab suatu kejadian, sehingga dapat menjawab pertanyaan mengapa hal tersebut terjadi.
3. ***Predictive Analysis***
   Analisis yang digunakan untuk memprediksi tren atau kejadian di masa depan berdasarkan pola yang ditemukan dalam data historis.
4. ***Prescriptive Analysis***
   Analisis yang berfokus pada pemberian rekomendasi tindakan atau solusi, dengan tujuan mencapai hasil yang optimal dalam menyelesaikan permasalahan



Gambar 15 Contoh Hasil Data Visualisasi Prediktif Machine Learning

Gambar 13 menunjukkan contoh hasil visualisasi prediktif dari beberapa algoritma *machine learning*. Selain beragamnya pilihan metode dalam pengolahan data, proses pengambilan hasil analisis dan penarikan kesimpulan merupakan faktor penting karena dapat memengaruhi keputusan. Ketelitian (*attention to detail*) dan sikap kritis (*critical thinking*) seorang *data scientist* pada setiap tahap pengolahan menjadi hal yang sangat penting untuk memastikan bahwa setiap proses dilakukan dengan benar sebelum hasil analisis digunakan sebagai dasar pengambilan keputusan.

BAB 7
LATIHAN-STUDI KASUS DATA SAINS BIDANG KEAMANAN INFORMASI

A. Bidang Penggunaan Data Sains

Data Sains digunakan secara luas di berbagai bidang keilmuan dan sektor industri karena kemampuannya dalam mengolah data menjadi informasi yang bernilai untuk mendukung pengambilan keputusan. Pada Gambar 16, ditampilkan identifikasi beberapa contoh penerapan data sains, antara lain di bidang ilmu bisnis dan pemasaran, kesehatan, keuangan, dan perbankan.

Dalam bidang bisnis dan pemasaran, data sains dimanfaatkan untuk melakukan segmentasi pelanggan, personalisasi layanan, memprediksi tren pasar, membangun sistem rekomendasi, serta menganalisis sentimen pelanggan melalui media sosial. Pada sektor kesehatan, data sains berperan penting dalam prediksi penyakit dan diagnosis dini, analisis citra medis, penemuan obat, serta penelitian genetik guna meningkatkan kualitas layanan dan keselamatan pasien.

Di bidang keuangan dan perbankan, data sains digunakan untuk mendeteksi penipuan, mengelola risiko, melakukan perdagangan algoritmik, optimasi portofolio, penilaian kredit, serta pemodelan profil nasabah. Sementara itu, dalam sektor pemerintahan dan layanan publik, data sains mendukung pengembangan *smart city*, perencanaan infrastruktur, analisis pola kejahatan, prediksi keamanan, serta evaluasi kebijakan berbasis data.

Dalam bidang pendidikan, data sains dimanfaatkan untuk mendukung pembelajaran yang dipersonalisasi, menganalisis kinerja peserta didik, memprediksi capaian akademik, serta mengoptimalkan penggunaan sumber daya pendidikan. Pada sektor manufaktur dan industri 4.0, data sains berperan dalam pemeliharaan prediktif mesin, pengendalian kualitas berbasis data real-time, serta optimasi rantai pasok.

Gambar 16 Bidang Keilmuan Pengguna Data Sains

Pada bidang keamanan siber, data sains digunakan untuk deteksi intrusi dan anomali berbasis model *machine learning*, analisis ancaman secara *real-time*. serta sistem autentikasi berbasis perilaku. Dalam konteks media sosial dan aplikasi internet, data sains mendukung sistem rekomendasi konten, penargetan iklan, serta analisis keterlibatan pengguna. Secara keseluruhan, penerapan data sains pada berbagai bidang tersebut menunjukkan perannya yang strategis dalam mendukung inovasi, efisiensi, dan pengambilan keputusan berbasis data.

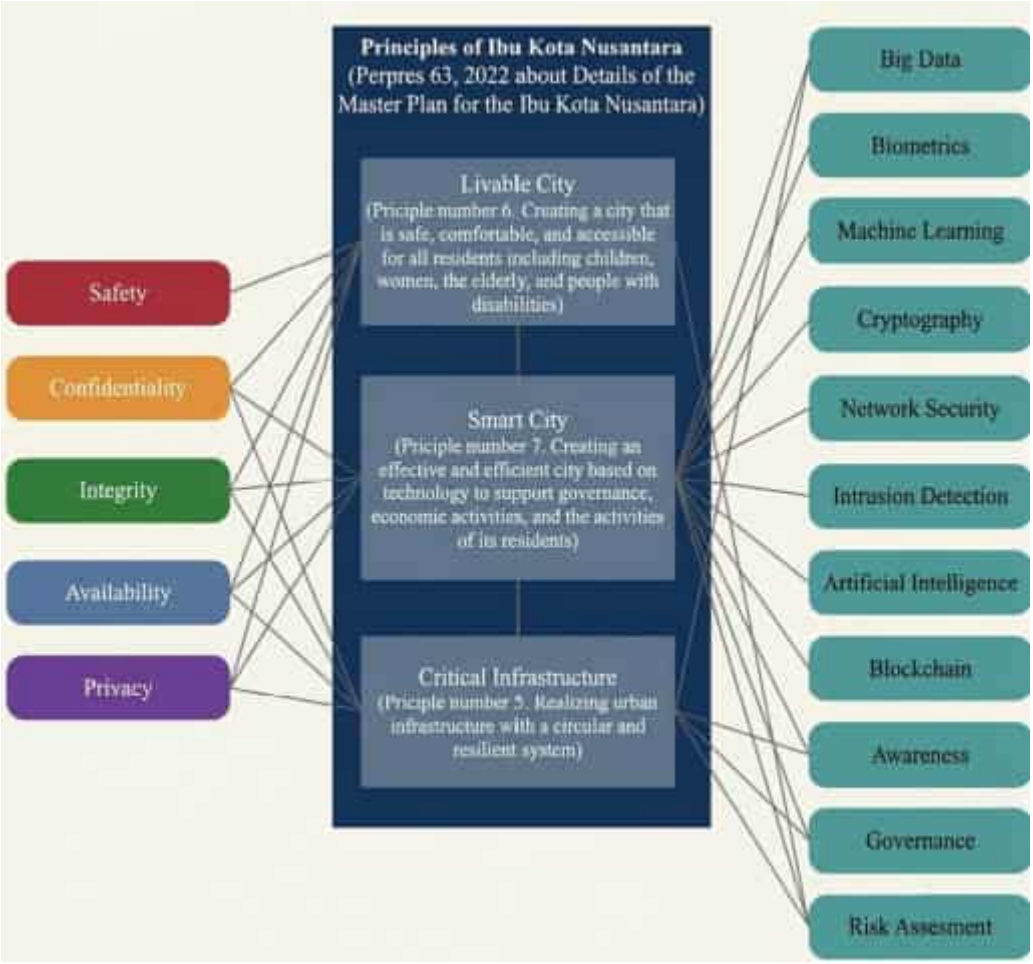
B. Penggunaan Data Sains pada Keamanan Informasi

Data sains kini menjadi alat penting dalam menjaga keamanan informasi di era digital. Dengan kemampuan menganalisis volume data yang besar dan kompleks, manfaat data sains untuk organisasi adalah sebagai berikut:

1. Mendeteksi ancaman dan anomali pada jaringan, sistem, atau pengguna.
2. Menganalisis log dan insiden untuk menemukan pola serangan atau kelemahan keamanan.
3. Memprediksi serangan dan risiko siber sebelum terjadi.
4. Otomatisasi respons berbasis analisis data untuk mitigasi cepat.
5. Melindungi data dan privasi, melalui monitoring, enkripsi, dan kontrol akses.

Dalam buku ini, latihan dan studi kasus difokuskan pada keamanan informasi, khususnya yang terkait dengan perangkat IoT dalam sistem *smart city*. Salah satu proyek unggulan di Indonesia adalah pembangunan Ibu Kota Nusantara (IKN), yang

mengadopsi konsep *smart city* untuk meningkatkan kualitas layanan publik, efisiensi operasional, serta keamanan dan kenyamanan masyarakat.



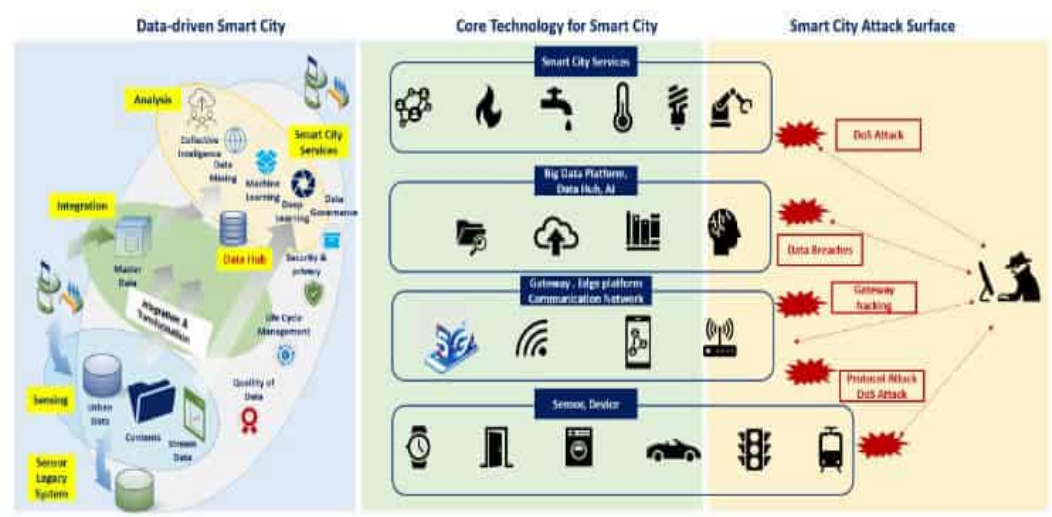
Gambar 17 Diagram alur keterkaitan antara prinsip keamanan informasi (CIA Triad & Privacy) dengan pilar pembangunan Ibu Kota Nusantara (IKN) serta implementasi teknologi pendukung berdasarkan Perpres No. 63 Tahun 2022 [10].

Berdasarkan penelitian terkait *framework cybersecurity* untuk IKN [10] (Gambar 17), pembangunan *smart city* memerlukan perhatian khusus pada berbagai aspek, termasuk diantaranya :

1. Penggunaan teknologi IoT, seperti sensor, kamera, dan aktuator yang terhubung melalui jaringan internet (Gambar 18).
2. Pengelolaan arus data yang terus bergerak di seluruh sistem *smart city* (Gambar 18).

3. Mekanisme pengamanan jaringan dan data, termasuk Intrusion Detection System (IDS), untuk melindungi sistem dari ancaman siber (Gambar 17) [10].

Keamanan yang baik menjadi fondasi penting agar *smart city* dapat beroperasi secara andal, aman, dan berkelanjutan. IoT sendiri adalah sistem perangkat tertanam *(embedded systems)* yang saling terhubung melalui internet, memungkinkan pengumpulan, pertukaran, dan pemrosesan data secara otomatis, misalnya pada *smart home* atau *smart building* dalam konsep *data-driven smart city* (Gambar 18).



Gambar 18 Contoh Penggunaan Data pada *Smart City*

C. Keamanan Informasi pada IoT

Perangkat Internet of Things (IoT) sangat rentan terhadap berbagai serangan siber, seperti Denial of Service (DoS), Distributed Denial of Service (DDoS), brute force, dan spoofing. Keterbatasan sumber daya perangkat IoT membuat pendekatan keamanan tradisional kurang efektif. Oleh karena itu, algoritma Machine Learning (ML) menjadi solusi yang relevan untuk sistem deteksi intrusi (IDS), karena mampu mengklasifikasikan aktivitas menjadi kategori normal atau ancaman, sehingga meningkatkan efektivitas proteksi pada *smart home* maupun *smart building*.

Faktor keamanan memegang peranan penting dalam pemanfaatan IoT, mengingat besarnya volume data yang terhubung ke internet dan tingginya kerentanan terhadap serangan siber. Ancaman jaringan dapat menyebabkan gangguan layanan, akses data tidak sah, hingga pengambilalihan sistem. Dengan penerapan *machine learning*, sistem IDS dapat mendeteksi dan mencegah serangan

secara *real-time*, sehingga meningkatkan keamanan dan keandalan operasi *smart city*.

Sesuai dengan riset terkait *framework cybersecurity* untuk Ibu Kota Nusantara (IKN) [10] pada Gambar 17, kebutuhan akan mekanisme keamanan jaringan yang andal semakin meningkat seiring dengan perkembangan IoT dalam implementasi sistem *smart city*. Salah satu mekanisme penting adalah *intrusion detection*, yang berperan untuk mendeteksi dan mencegah ancaman secara efektif.

Penerapan nyata menunjukkan bagaimana data sains dapat membuat keamanan informasi menjadi lebih proaktif, cerdas, dan berbasis bukti, sekaligus mendukung pengambilan keputusan yang cepat dan tepat. Contohnya termasuk deteksi malware, analisis trafik jaringan, dan monitoring aktivitas pengguna.

Bagian selanjutnya akan membahas contoh dan latihan sederhana penerapan data sains dalam keamanan informasi, dengan fokus pada perangkat pengamanan data IoT berupa IDS, yang banyak digunakan di smart home maupun smart building dalam lingkungan *smart city*.

D. Latihan-Studi Kasus Data Sains Bidang Cybersecurity: Data IoT pada Smart City

Praktik yang disajikan sebagai contoh penelitian Data Sains dalam buku ini bertujuan untuk mengklasifikasikan aktivitas normal dan berbagai jenis serangan pada sistem Smart Home IoT, serta membandingkan performa algoritma K-Nearest Neighbor, Gaussian Naive Bayes, dan Random Forest menggunakan *dataset* ClassWise Balanced CICIoT 2023 [11] [12]. Metode penelitian yang digunakan meliputi tahapan *preprocessing data*, pembagian data menjadi data latih dan data uji, *feature scaling,* pelatihan model klasifikasi, serta optimasi parameter model.

Evaluasi kinerja model dilakukan menggunakan metrik performa seperti *accuracy, precision, recall*, dan *F1-score* untuk menentukan algoritma dengan kemampuan klasifikasi terbaik . Selain itu, dilakukan pula analisis interpretatif terhadap hasil model melalui *feature importance* serta visualisasi struktur data menggunakan Principal Component Analysis (PCA) guna memberikan pemahaman yang lebih mendalam terhadap pola data dan proses pengambilan keputusan model.

E. Latihan: Pengumpulan dan Penyimpanan Data (Data Collection)

Sebagai tahap awal dalam proses data sains, dilakukan pengumpulan dan penyimpanan data mentah dari berbagai sumber agar dapat diakses dan diolah pada tahap selanjutnya. Pada praktik ini, dataset yang digunakan diperoleh dari

Kaggle, berupa rekaman lalu lintas jaringan dari puluhan perangkat IoT yang mengalami berbagai jenis serangan siber. *Dataset* ini dirancang oleh Canadian Institute for Cybersecurity (CIC) dan terdiri dari sekitar 46 juta baris data pada versi lengkapnya. *Dataset* yang digunakan dalam praktik ini merupakan versi yang telah diseimbangkan *(balanced dataset)*, sehingga tidak diperlukan penerapan teknik penyeimbangan data tambahan, *capture dataset* dapat dilihat pada Gambar 19.

Dataset praktik berisi 82.195 baris data dengan 42 fitur lalu lintas jaringan, yang mencakup informasi dasar koneksi, indikator TCP Flag, protokol jaringan, serta statistik pengiriman paket. Data telah diklasifikasikan secara seimbang ke dalam delapan kelas aktivitas jaringan, yaitu RECON, MIRAI, DOS, DDOS, SPOOFING, BRUTEFORCE, WEB_BASED, dan Normal, sehingga sesuai untuk keperluan analisis dan pemodelan deteksi intrusi berbasis data sains.

```
df = pd.read_csv('CleanWise_Labeled_Balanced_Dataset.csv')
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 82195 entries, 0 to 82194
Data columns (total 42 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0.1     82195 non-null  int64
 1   Unnamed: 0       82195 non-null  int64
 2   Header_Length    82195 non-null  float64
 3   Protocol Type    82195 non-null  int64
 4   Time_To_Live     82195 non-null  float64
 5   Rate             82195 non-null  float64
 6   fin_flag_number  82195 non-null  float64
 7   syn_flag_number  82195 non-null  float64
 8   rst_flag_number  82195 non-null  float64
 9   psh_flag_number  82195 non-null  float64
 10  ack_flag_number  82195 non-null  float64
 11  ece_flag_number  82195 non-null  float64
 12  cwr_flag_number  82195 non-null  float64
 13  ack_count        82195 non-null  int64
 14  syn_count        82195 non-null  int64
 15  fin_count        82195 non-null  int64
 16  rst_count        82195 non-null  int64
 17  HTTP             82195 non-null  float64
 18  HTTPS            82195 non-null  float64
 19  DNS              82195 non-null  float64
...
 40  Variance         82193 non-null  float64
 41  Label            82195 non-null  object
dtypes: float64(36), int64(11), object(1)
memory usage: 26.3+ MB
```

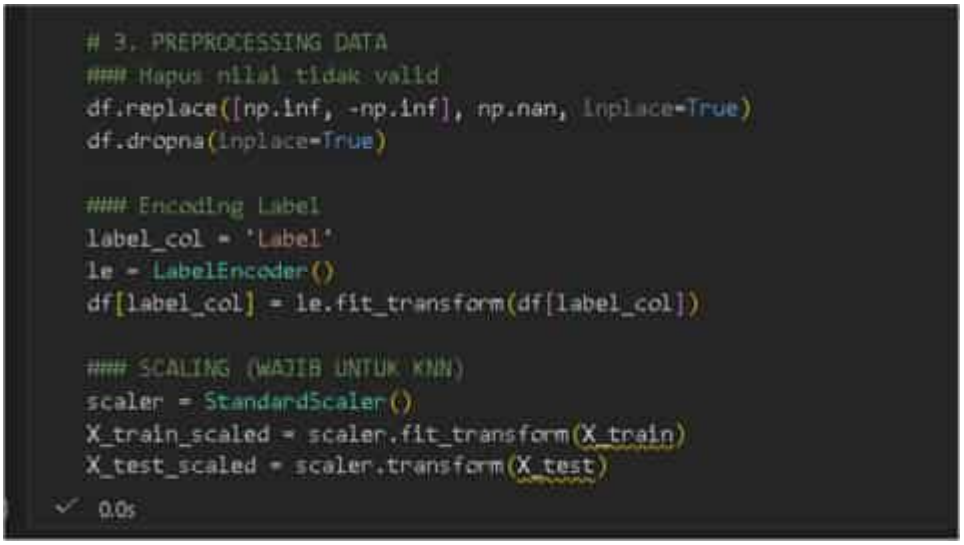
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Gambar 19 Tabel Dataset CIC

F. Latihan: Pembersihan dan Pemrosesan Data (Data Cleaning and Preprocessing)

Data cleaning merupakan proses untuk memeriksa, memperbaiki, atau menghapus kesalahan, duplikasi, serta ketidakkonsistenan data guna meningkatkan kualitas *dataset*. Proses ini mencakup penanganan data hilang, normalisasi, integrasi, dan transformasi data. Tujuan utama tahap ini adalah menghasilkan data yang bersih, terstruktur, dan konsisten sehingga siap digunakan pada tahap analisis dan pemodelan selanjutnya. Pada praktik ini, proses data *cleaning* difokuskan pada penghapusan nilai-nilai yang tidak valid.

Selanjutnya, data *preprocessing* menjadi tahapan penting dalam proses *data mining* untuk meningkatkan kualitas dan keandalan *dataset*. Seperti terlihat pada Gambar 20. tahap ini meliputi tiga proses utama, yaitu penghapusan nilai tidak valid, *label encoding*, dan *scaling data*. *Dataset* diperiksa untuk mendeteksi nilai tak hingga *(infinite)* dan nilai kosong (NaN) yang umumnya muncul akibat kesalahan perhitungan atau proses ekstraksi data yang tidak sempurna. Baris data yang mengandung nilai NaN kemudian dihapus untuk memastikan *dataset* yang digunakan dalam analisis bebas dari anomali dan siap untuk tahap pemodelan *machine learning*.



Gambar 20 Preprocessing Data

Pada proses *encoding* label berfungi untuk mengubah label kelas yang masih berbentuk kategori *string* menjadi bentuk numerik. Proses ini dilakukan karena algoritma *machine learning* seperti K-Nearest Neighbor (KNN), Naive Bayes, dan Random Forest bekerja menggunakan operasi matematis.

G. Latihan: Data Transformation-Data Sampling

Pada contoh praktik ini, algoritma *machine learning* yang digunakan adalah K-Nearest Neighbor (KNN). Setelah tahap preprocessing, dilakukan proses data sampling untuk menentukan arah pembelajaran model serta memastikan evaluasi kinerja dilakukan secara terukur dan tidak bias.

Metode yang digunakan adalah *train-test split*, yaitu membagi dataset menjadi data latih dan data uji untuk keperluan pelatihan dan pengujian model klasifikasi (Gambar 21). Proses sampling ini juga menerapkan parameter stratifikasi, sehingga

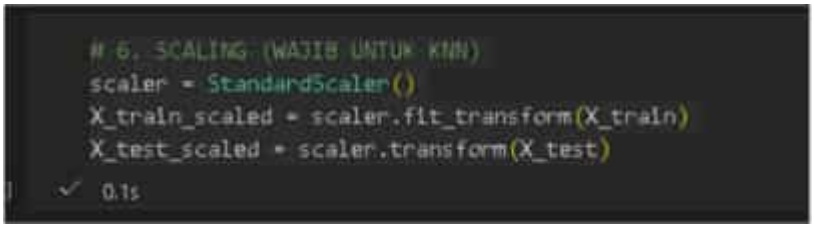
distribusi setiap kelas pada data latih dan data uji tetap proporsional dan merepresentasikan distribusi kelas pada dataset asli. Pendekatan ini penting untuk menjaga keadilan evaluasi model, terutama pada *dataset* dengan banyak kelas.

Gambar 21 Metode Train Test Split

H. Latihan: Data Transformation-Data Scaling

Tahap selanjutnya dalam pengolahan data adalah *data scaling*, diberikan pada Gambar 22, yaitu proses transformasi fitur untuk menyamakan skala antarvariabel. Pada penelitian ini, proses *data scaling* dilakukan menggunakan metode StandardScaler, yang berfungsi menormalkan distribusi setiap fitur sehingga memiliki nilai rata-rata nol dan simpangan baku satu.

Penerapan scaling menjadi langkah yang penting, khususnya pada algoritma K-Nearest Neighbor (KNN), karena algoritma ini sangat bergantung pada perhitungan jarak antar data. Tanpa scaling, fitur dengan rentang nilai yang lebih besar dapat mendominasi perhitungan jarak dan menyebabkan hasil klasifikasi menjadi bias. Dengan melakukan *data scaling*, seluruh fitur berada pada skala yang setara sehingga model dapat menghasilkan prediksi yang lebih akurat dan objektif.



Gambar 22 Tahap Scaling

I. Latihan: Data Klasifikasi dengan Machine Learning

Setelah data latih melalui tahapan *preprocessing*, *sampling*, dan *scaling*, tahap training model dilakukan untuk membangun model klasifikasi yang mampu mengenali pola serta memprediksi kelas data secara akurat, proses diperlihatjan pada Gambar 23. Proses pelatihan ini menggunakan pendekatan *supervised*

learning, di mana algoritma mempelajari hubungan antara fitur masukan (*X_train_scaled*) dan label kelas (*y_train*). Selama proses *training*, model menyesuaikan parameter internalnya agar dapat memetakan pola fitur terhadap kategori target secara optimal.

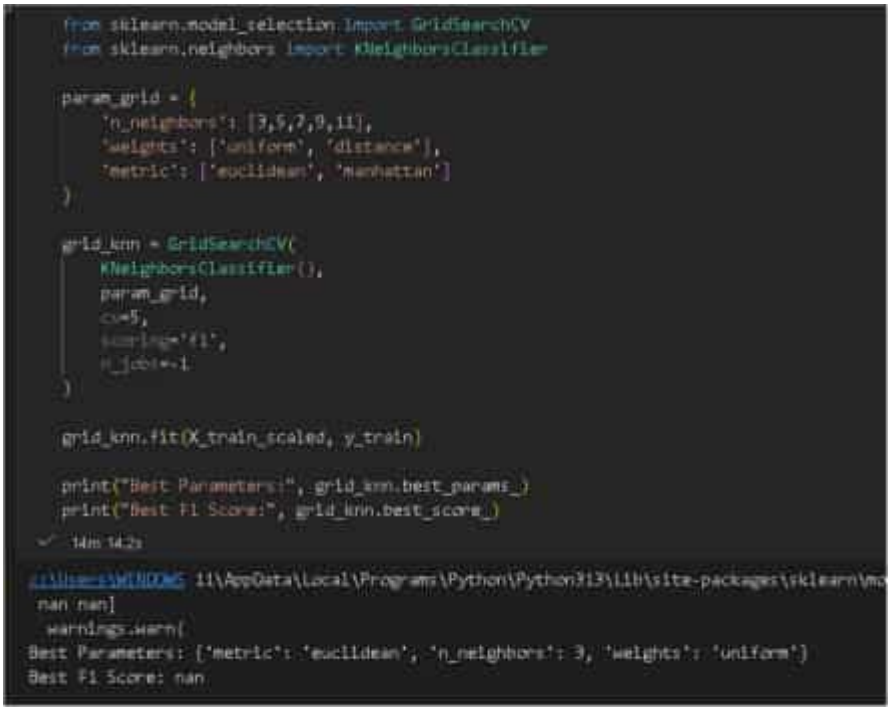
Pada penelitian ini, digunakan tiga algoritma klasifikasi untuk memprediksi data uji, yaitu K-Nearest Neighbor (KNN), Naive Bayes, dan Random Forest, guna membandingkan kinerja masing-masing model dalam mendeteksi dan mengklasifikasikan aktivitas jaringan pada perangkat IoT.

```
# TRAINING MODEL
knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train_scaled, y_train)
y_pred_knn = knn.predict(X_test_scaled)

nb = GaussianNB()
nb.fit(X_train_scaled, y_train)
y_pred_nb = nb.predict(X_test_scaled)

rf = RandomForestClassifier(
    n_estimators=300,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)
rf.fit(X_train_scaled, y_train)
y_pred_rf = rf.predict(X_test_scaled)
✓ 23.4s
```

Gambar 23 Training Model

J. Latihan: Model Evaluation (Optimasi Model Machine Learning)

Pada tahap ini dilakukan proses optimasi dan evaluasi model *machine learning* untuk meningkatkan performa klasifikasi. Optimasi model pada contoh praktik ini difokuskan pada algoritma K-Nearest Neighbor (KNN) dengan menggunakan metode *grid search* yang dikombinasikan dengan *cross-validation*, diperlihatkan pada Gambar 24. Pendekatan ini bekerja dengan menguji berbagai kombinasi nilai *hyperparameter* secara sistematis, kemudian memilih konfigurasi yang menghasilkan performa terbaik berdasarkan metrik evaluasi yang telah ditentukan.



Gambar 24 Optimasi Model

Setelah proses optimasi selesai, dilakukan evaluasi model untuk mengukur kinerja seluruh algoritma klasifikasi yang digunakan, yaitu KNN, Gaussian Naive Bayes, dan Random Forest. Evaluasi dilakukan dengan membandingkan hasil prediksi model terhadap label sebenarnya pada data uji. Beberapa metrik evaluasi yang digunakan meliputi accuracy, precision, recall, F1-score, serta *classification report*.

Evaluasi model klasifikasi umumnya didasarkan pada confusion matrix, yang terdiri dari empat komponen utama, yaitu:

1. **True Positive (TP)**: jumlah data positif yang diprediksi benar sebagai positif
2. **True Negative (TN)**: jumlah data negatif yang diprediksi benar sebagai negatif
3. **False Positive (FP)**: jumlah data negatif yang salah diprediksi sebagai positif
4. **False Negative (FN)**: jumlah data positif yang salah diprediksi sebagai negatif

Berdasarkan komponen tersebut, beberapa metrik evaluasi yang digunakan dalam penelitian ini dijelaskan sebagai berikut.

1. **Accuracy**

   Accuracy mengukur tingkat ketepatan model secara keseluruhan dalam melakukan klasifikasi, yaitu perbandingan antara jumlah prediksi yang benar terhadap total seluruh data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision**

   Precision menunjukkan tingkat ketepatan prediksi positif yang dihasilkan oleh model, atau seberapa banyak prediksi positif yang benar dibandingkan seluruh prediksi positif.

$$Precision = \frac{TP}{TP + FP}$$

3. **Recall**

   Recall mengukur kemampuan model dalam menemukan seluruh data yang benar-benar positif, atau seberapa banyak data positif yang berhasil terdeteksi oleh model.

$$Recall = \frac{TP}{TP + FN}$$

4. **F1-Score**

   F1-score merupakan rata-rata harmonik antara *precision* dan *recall*, yang digunakan untuk memberikan keseimbangan antara kedua metrik tersebut, terutama pada *dataset* yang tidak seimbang.

$$F1 - Score = 2x \frac{Precision\ x\ Recall}{Precision + Recall}$$

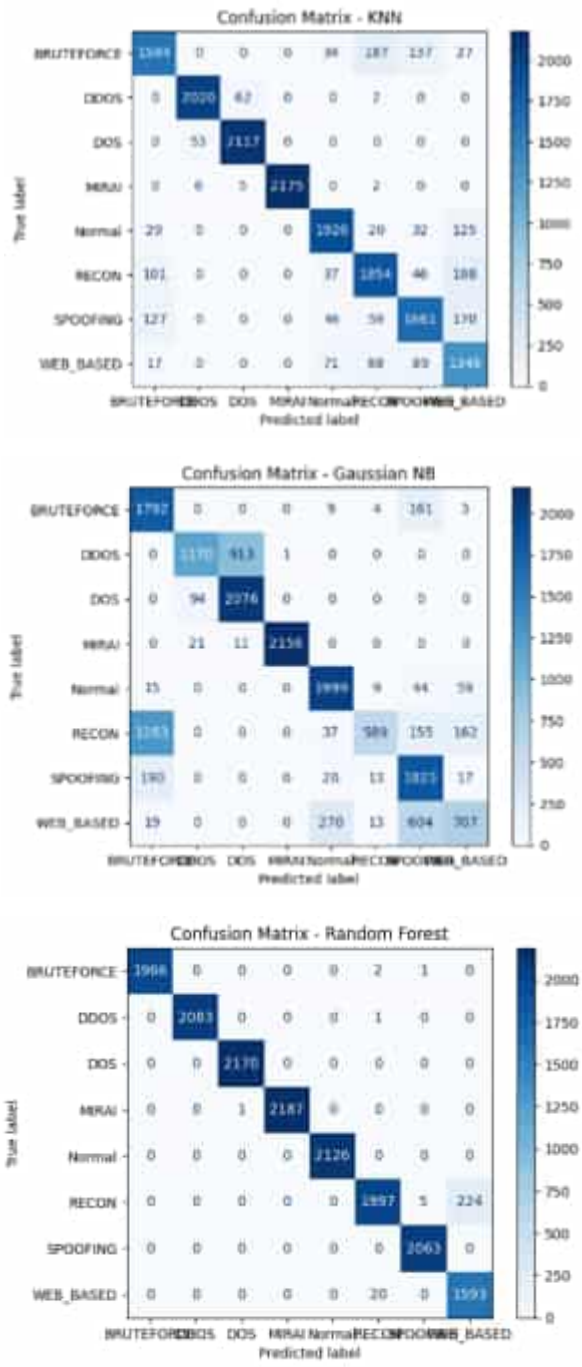
5. **Classification**

   Classification report merupakan ringkasan hasil evaluasi yang menyajikan nilai precision, recall, F1-score, dan support untuk setiap kelas. Laporan ini memberikan gambaran performa model secara lebih rinci pada masing-masing kategori data, sehingga memudahkan analisis kekuatan dan kelemahan model dalam melakukan klasifikasi.

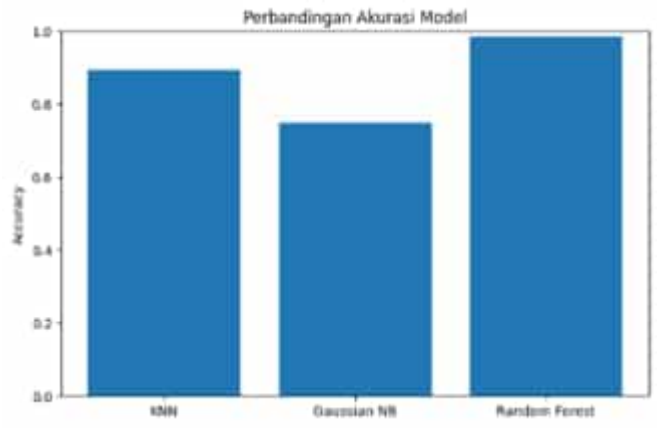
Melalui penggunaan metrik-metrik evaluasi tersebut, diperoleh gambaran kuantitatif yang komprehensif mengenai kualitas, efektivitas, dan kemampuan generalisasi masing-masing model klasifikasi dalam memproses dan memprediksi data. Hasil perhitungan metrik ditampilkan pada Gambar 25, sedangkan hasil *confusion matrix* diperlihatkan pada Gambar 26. Selanjutnya, prediksi data atau metode paling efektif untuk memprediksi data disajikan pada Gambar 27.

```
KNN Accuracy: 0.892937526613541
NB Accuracy: 0.7489506660989111
RF Accuracy: 0.9845489384999088
===== Classification Report KNN =====
              precision    recall  f1-score   support

  BRUTEFORCE       0.85      0.80      0.83      1969
        DDOS       0.97      0.97      0.97      2084
         DOS       0.97      0.98      0.97      2170
       MIRAI       1.00      0.99      1.00      2188
      Normal       0.91      0.90      0.91      2126
       RECON       0.84      0.83      0.84      2226
    SPOOFING       0.85      0.81      0.82      2063
   WEB_BASED       0.73      0.84      0.78      1613

    accuracy                           0.89     16439
   macro avg       0.89      0.89      0.89     16439
weighted avg       0.89      0.89      0.89     16439

===== Classification Report Gaussian NB =====
              precision    recall  f1-score   support

  BRUTEFORCE       0.54      0.91      0.68      1969
        DDOS       0.91      0.56      0.69      2084
         DOS       0.69      0.96      0.80      2170
       MIRAI       1.00      0.99      0.99      2188
      Normal       0.86      0.94      0.90      2126
       RECON       0.94      0.26      0.41      2226
    SPOOFING       0.65      0.88      0.75      2063
   WEB_BASED       0.75      0.44      0.55      1613

    accuracy                           0.75     16439
   macro avg       0.79      0.74      0.72     16439
weighted avg       0.80      0.75      0.73     16439

===== Classification Report Random Forest =====
              precision    recall  f1-score   support

  BRUTEFORCE       1.00      1.00      1.00      1969
        DDOS       1.00      1.00      1.00      2084
         DOS       1.00      1.00      1.00      2170
       MIRAI       1.00      1.00      1.00      2188
      Normal       1.00      1.00      1.00      2126
       RECON       0.99      0.90      0.94      2226
    SPOOFING       1.00      1.00      1.00      2063
   WEB_BASED       0.88      0.99      0.93      1613

    accuracy                           0.98     16439
   macro avg       0.98      0.99      0.98     16439
weighted avg       0.99      0.98      0.98     16439
```

Gambar 25 Evaluasi Model



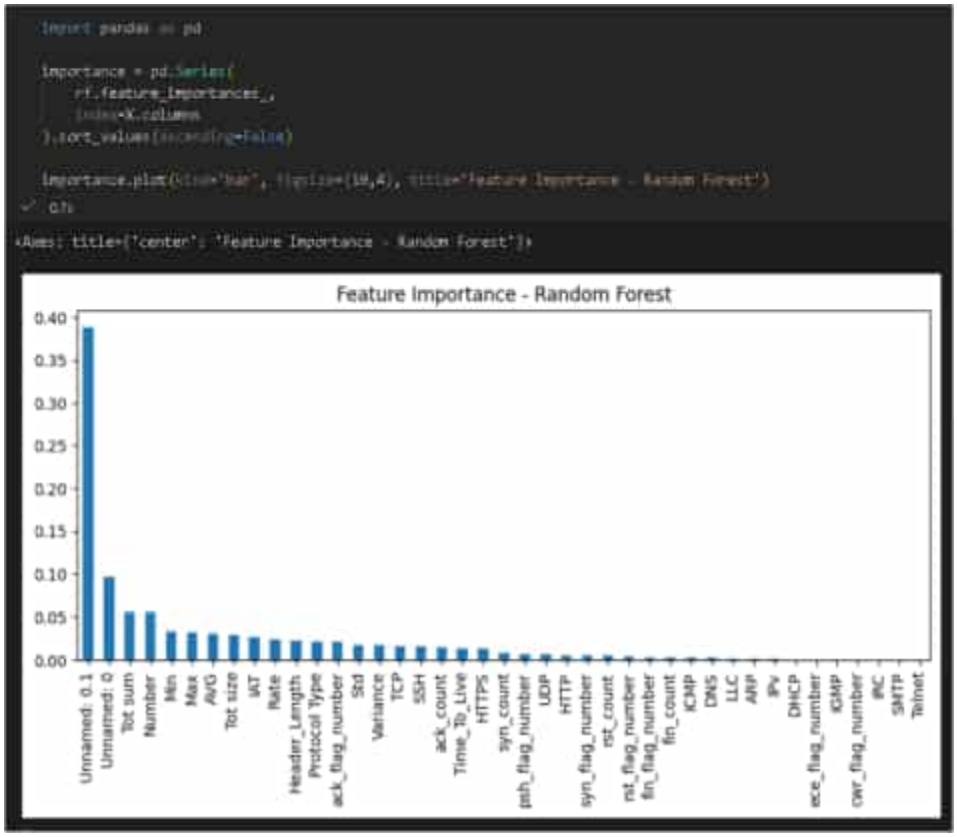
Gambar 26 Confusion Matrix



Gambar 27 Grafik Perbandingan Akurasi 3 Teknik *Machine Learning*

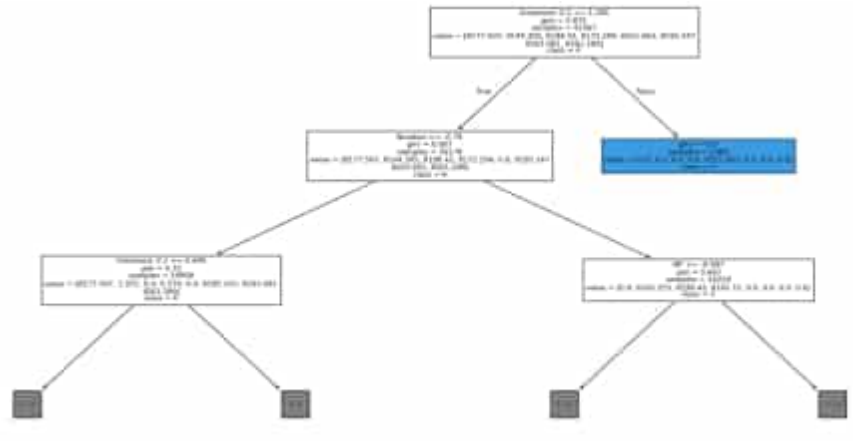
K. Latihan: Analisis Data

Berdasarkan hasil perbandingan akurasi dari tiga teknik *machine learning* yang digunakan, diperoleh bahwa Random Forest menunjukkan nilai akurasi tertinggi dibandingkan dengan algoritma lainnya. Oleh karena itu, analisis lanjutan pada Gambar 28 difokuskan pada model Random Forest, khususnya untuk mengidentifikasi tingkat kepentingan *(feature importance)* dari setiap fitur terhadap hasil prediksi.



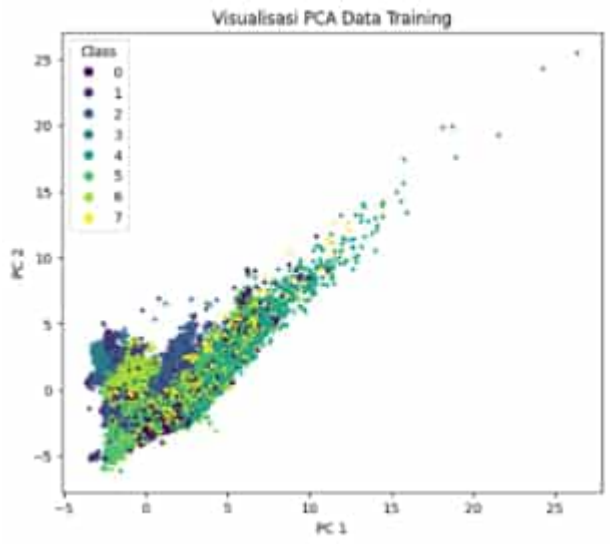
Gambar 28 Grafik Feature Importance-Tehnik Random Forest

Model Random Forest secara alami mampu menghitung kontribusi relatif masing-masing variabel dengan mengukur sejauh mana suatu fitur berperan dalam mengurangi ketidakpastian (impurity) selama proses pembentukan pohon keputusan. Analisis *feature importance* ini memberikan wawasan penting mengenai fitur-fitur yang paling berpengaruh dalam proses klasifikasi, sehingga dapat membantu dalam interpretasi model dan pengambilan keputusan berbasis data.



Gambar 29 Visualisasi Decision Tree

Selanjutnya, pada Gambar 29, ditampilkan visualisasi *decision tree* yang merepresentasikan struktur salah satu pohon keputusan di dalam model Random Forest. Meskipun Random Forest tersusun atas banyak pohon keputusan (*ensemble*), visualisasi satu pohon sudah cukup untuk memberikan gambaran umum mengenai bagaimana proses pengambilan keputusan dilakukan oleh model. Melalui visualisasi ini, dapat dipahami alur pemisahan data berdasarkan fitur-fitur tertentu hingga menghasilkan prediksi kelas, sehingga membantu dalam interpretasi dan pemahaman cara kerja model secara konseptual.



Gambar 30 Principal Component Analysis (PCA)

Visualisasi ini bertujuan untuk membantu memahami pola distribusi data, hubungan antar fitur, serta tingkat keterpisahan antar kelas dalam ruang fitur yang lebih sederhana. Principal Component Analysis (PCA) merupakan metode transformasi linier yang digunakan untuk mereduksi dimensi data dengan tetap mempertahankan sebagian besar variansi informasi yang terkandung dalam *dataset*. Teknik ini banyak dimanfaatkan pada tahap eksplorasi data dan interpretasi struktur data untuk mempermudah analisis dan visualisasi data berdimensi tinggi.

]CA adalah teknik reduksi dimensi yang digunakan untuk mengubah kumpulan data besar menjadi bentuk yang lebih sederhana (2D atau 3D) tanpa menghilangkan informasi pentingnya. Dalam grafik, seperti terlihat pada Gambar 30, ribuan data poin dipetakan ke dalam dua sumbu komponen utama : PC 1 dan PC 2. PC 1 (Sumbu Horizontal): Komponen utama pertama yang menangkap variasi atau perbedaan paling besar antar data. Sedangkan PC 2 (Sumbu Vertikal): Komponen utama kedua yang menangkap sisa variasi yang tidak tertangkap oleh PC 1.

Pada Gambar 30 pewarnaan berdasarkan Kelas (*Class*) dimana terdapat legenda di sisi kiri atas yang menunjukkan angka 0 sampai 7. Ini mewakili label atau kategori data:

1. Setiap warna mewakili kelompok data yang berbeda (misalnya: jenis serangan siber, kategori produk, atau jenis dokumen).
2. Warna gelap (ungu/biru) cenderung berkumpul di area nilai PC 1 dan PC 2 yang rendah (kiri bawah).

3. Warna terang (hijau/kuning) terlihat lebih menyebar ke arah nilai PC 1 dan PC 2 yang lebih tinggi (kanan atas).

   Interpretasi Sebaran Data pada Gambar 30 :

   1. Klaster (Pengelompokan): Jika titik-titik dengan warna yang sama berkumpul erat di satu tempat, artinya algoritma Machine Learning akan lebih mudah mengenali dan membedakan kelas tersebut.
   2. Overlapping (Tumpang Tindih): Di area kiri bawah (sekitar koordinat 0,0), terlihat banyak warna yang tercampur. Ini menunjukkan bahwa antar kelas tersebut memiliki karakteristik yang mirip, sehingga model mungkin akan lebih sulit membedakannya secara akurat pada area tersebut.
   3. Outliers (Pencilan): Titik-titik hijau tua di bagian kanan atas (di atas nilai 20 pada PC 1) adalah data yang memiliki karakteristik sangat berbeda dibanding mayoritas data lainnya.

Kesimpulannya, visualisasi ini sangat berguna bagi *Data Scientist* untuk melihat apakah data mereka dapat dipisahkan dengan jelas sebelum melatih model klasifikasi. Semakin terpisah kelompok warnanya, semakin baik performa model yang akan dihasilkan.

L. Latihan: Simpulan atau Pengambilan Keputusan

Berdasarkan praktik dan penelitian yang telah dilakukan, dapat disimpulkan bahwa pendekatan machine learning mampu digunakan secara efektif untuk mengklasifikasikan aktivitas jaringan pada sistem Smart Home IoT. Seluruh algoritma yang diuji, yaitu K-Nearest Neighbor, Gaussian Naive Bayes, dan Random Forest, terbukti dapat mendeteksi berbagai jenis serangan keamanan yang umum terjadi pada lingkungan IoT, seperti Brute Force, DoS, DDoS, Mirai, dan Spoofing, serta membedakannya dari aktivitas normal.

Hasil pengujian menunjukkan bahwa setiap algoritma memiliki karakteristik performa yang berbeda dalam mempelajari pola lalu lintas jaringan. Di antara ketiganya, algoritma Random Forest menghasilkan performa klasifikasi terbaik dengan tingkat akurasi lebih dari 98%. Nilai precision, recall, dan F1-score yang tinggi pada hampir seluruh kelas menunjukkan bahwa model ini memiliki kemampuan diskriminatif yang baik dalam membedakan aktivitas normal dan berbagai jenis serangan. Selain itu, stabilitas performa pada kelas dengan jumlah sampel besar mengindikasikan kemampuan generalisasi model yang baik terhadap data uji.

Secara keseluruhan, penelitian ini menegaskan bahwa Random Forest merupakan algoritma yang andal dan efektif untuk diimplementasikan sebagai sistem deteksi intrusi pada lingkungan Smart Home IoT, serta berpotensi mendukung peningkatan keamanan siber pada sistem berbasis Internet of Things.

Hasil analisis dan klasifikasi yang diperoleh selanjutnya dapat digunakan sebagai dasar pengambilan keputusan oleh para pemangku kepentingan (stakeholder) terkait. Keputusan tersebut dapat mencakup pemilihan teknik pengolahan dan analisis data yang tepat, penentuan strategi penanganan celah keamanan berdasarkan hasil kategorisasi serangan, serta perumusan dan penerapan kebijakan keamanan yang sesuai. Dalam konteks Smart City, khususnya pada lingkungan smart home atau smart building, keputusan ini dapat digunakan untuk meningkatkan mekanisme perlindungan sistem, memperkuat keamanan jaringan, serta meminimalkan risiko serangan siber agar operasional layanan tetap aman dan andal.

BAB 8
MATERI PENGAYAAN

A. Big Data Analysis

Dalam konteks data sains modern, *big data analysis* menjadi pendekatan penting ketika data yang diolah telah melampaui kemampuan sistem basis data konvensional . Seperti diilustrasikan pada Gambar 31, *big data* didefinisikan sebagai kumpulan data yang memiliki ukuran sangat besar, bertumbuh dengan kecepatan tinggi, serta memiliki variasi struktur yang kompleks sehingga tidak dapat diproses secara efektif menggunakan arsitektur *database* tradisional. Oleh karena itu, diperlukan solusi pemrosesan alternatif yang bersifat terdistribusi, salah satunya melalui cloud-based data solution dan framework MapReduce [13].

Gambar 31 Tantangan dalam BigData [14]

Secara umum, karakteristik *big data* dirangkum dalam konsep 5V, yaitu *velocity* (kecepatan), *volume* (jumlah), *variety* (keragaman), *value* (nilai), dan *veracity* (keakuratan). *Volume* mengacu pada jumlah data yang sangat besar dan terus meningkat, yang berasal dari berbagai sumber seperti media sosial, sensor IoT, dan aktivitas web. *Velocity* menggambarkan kecepatan aliran data yang sangat tinggi sehingga menuntut kemampuan pemrosesan secara hampir *real-time*. *Veracity* berkaitan dengan tingkat keakuratan, konsistensi, dan keandalan data, yang sering kali mengandung *noise* atau ketidakpastian sehingga memerlukan teknik analisis yang cermat. Sementara itu, *Value* menunjukkan nilai atau manfaat yang dapat diperoleh dari data apabila diolah dengan tepat, karena tidak semua data memiliki makna yang sama untuk setiap tujuan analisis.

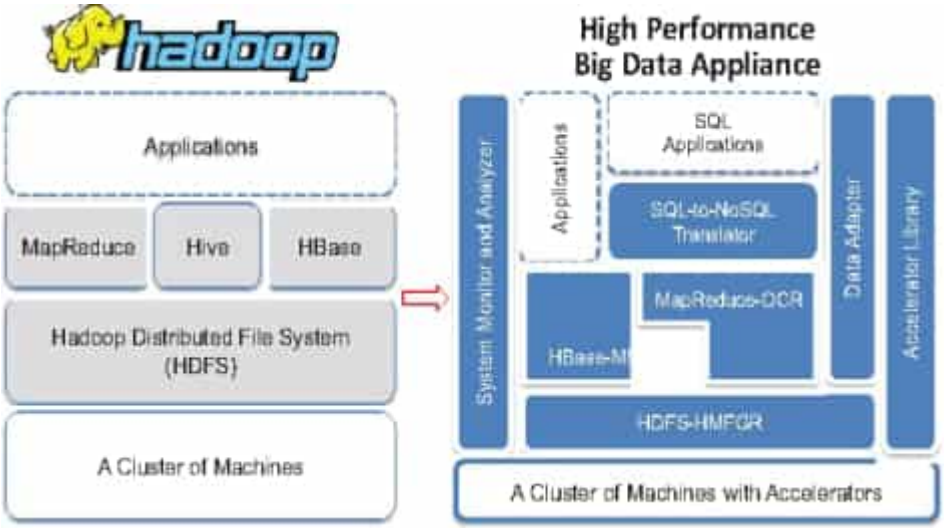
Dalam menghadapi tantangan *big data* tersebut, data sains berperan sebagai pendekatan multidisipliner yang menggabungkan matematika, statistika, dan ilmu komputer untuk melakukan analisis data dalam berbagai skala, mulai dari *dataset* kecil hingga *big data*. Data sains memanfaatkan algoritma dan teknik komputasi untuk mengekstrak informasi yang berguna (*data mining*), mengidentifikasi pola dan hubungan antar data, serta membangun model prediktif guna mendukung pengambilan keputusan berbasis data. Selain itu, data sains juga menjadi fondasi dalam pengembangan sistem cerdas (*Artificial Intelligence*) yang mampu belajar dan meningkatkan performanya secara otomatis melalui machine learning.

Hubungan antara data sains dan *machine learning* bersifat saling melengkapi. Data sains menyediakan proses pengolahan data, eksplorasi, dan analisis statistik, sedangkan *machine learning* memanfaatkan data tersebut untuk membangun model prediktif dan sistem yang mampu belajar secara mandiri. Dengan demikian, hasil analisis *big data* dapat digunakan untuk menghasilkan keputusan yang lebih akurat, adaptif, dan otomatis.

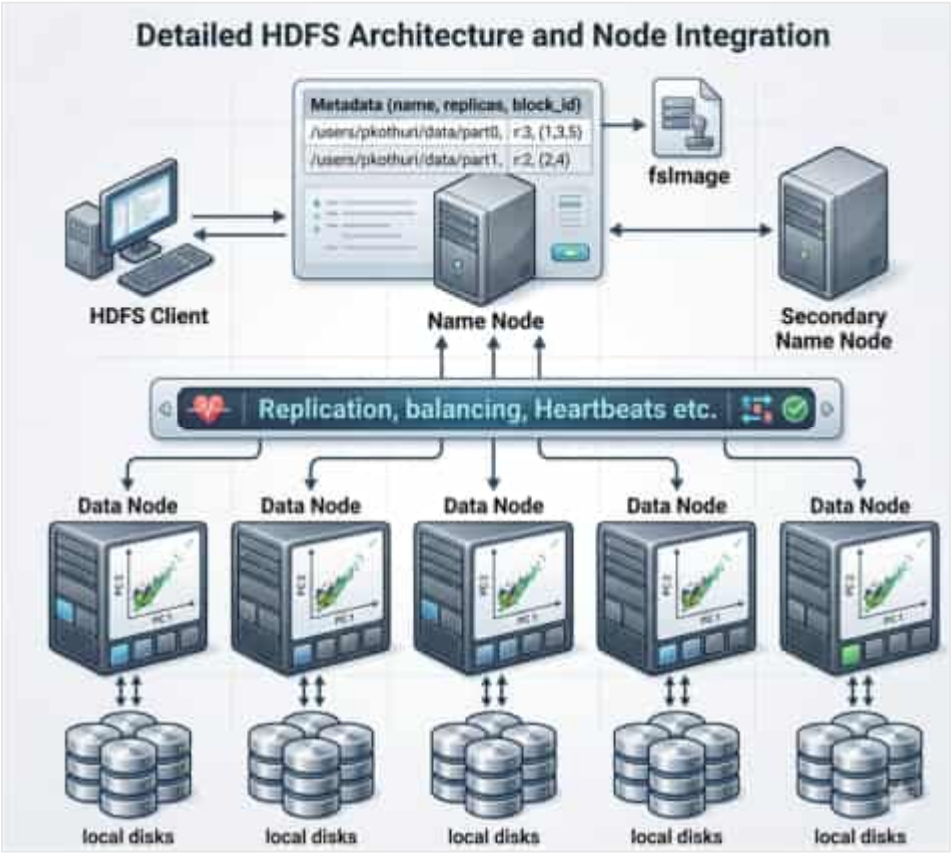
Dalam implementasinya, analisis *big data* banyak memanfaatkan ekosistem Apache Hadoop (Gambar 32), yang menyediakan kerangka kerja pemrosesan data terdistribusi [15]. Hadoop memungkinkan pengolahan data dalam jumlah besar secara paralel melalui model MapReduce, serta menyediakan sistem penyimpanan terdistribusi yang dikenal sebagai Hadoop Distributed File System (HDFS). HDFS dirancang untuk menyimpan data dalam skala besar secara terdistribusi di banyak node, sehingga meningkatkan keandalan, skalabilitas, dan efisiensi akses data dalam proses analisis *big data*.

Dengan dukungan teknologi *big data* dan *cloud-based* data solution, data sains dapat diterapkan secara lebih luas pada berbagai domain, termasuk keamanan siber, *smart city*, Internet of Things, dan analisis sistem skala besar. Oleh karena itu,

pemahaman terhadap *big data* Analysis menjadi materi pengayaan yang penting untuk memperluas wawasan mahasiswa dalam menghadapi tantangan data berskala besar di dunia nyata.



Gambar 32 Tools Hadoop dalam BigData Analysis [15]



Gambar 33 HDFS dalam BigData Analysis

Seperti diperlihatkan pada skema Gambar 33, HDFS adalah sistem file terdistribusi yang dikembangkan sebagai bagian dari ekosistem Apache Hadoop. Sistem ini dirancang untuk menyimpan dan mengelola data dalam jumlah sangat besar (*big data*) secara efisien dengan mendistribusikan data ke banyak node (komputer/servers) dalam sebuah *cluster*.

Beberapa ciri utama HDFS diantaranya adalah sifat terdistribusi, *fault tolerant* dan akses sekuensial. Terdistribusi artinya data dibagi menjadi blok-blok (default 128 MB per blok) dan disimpan di beberapa *node*. Tahan kesalahan *(fault-tolerant)* secara umum menjelaskan setiap blok disalin *(replicated)* ke beberapa *node* agar data tetap aman jika ada *node* yang gagal. HDFS juga dinilai skalabilitas tinggi, artinya bisa menangani petabyte atau lebih data dengan menambah jumlah *node*. Sementara sifat akses yang sekuensial dirancang untuk *throughput* tinggi saat membaca atau menulis file besar, bukan untuk operasi baca-tulis acak kecil.

Dalam data sains, HDFS berperan sebagai pondasi penyimpanan data skala besar sebelum dilakukan analisis. perannya meliputi:

1. Penyimpanan Data Masif
   a. Data sains modern sering menggunakan *dataset* besar (misalnya transaksi e-commerce, log web, sensor IoT).
   b. HDFS memungkinkan penyimpanan data ini secara terdistribusi sehingga tidak terbatas pada kapasitas satu server.
2. Dukungan Analisis Terdistribusi
   a. HDFS bekerja bersama MapReduce, Spark, atau tools analisis lain untuk memproses data langsung di cluster.
   b. Artinya, komputasi dilakukan dekat dengan lokasi data, hal ini akan mengurangi waktu transfer data dan mempercepat analisis.
3. Integrasi dengan Tools Data Sains
   a. HDFS dapat digunakan sebagai sumber data untuk Machine Learning misalnya dengan algoritma Spark MLlib atau TensorFlow dapat membaca data dari HDFS.
   b. Visualisasi dan reporting: data hasil *preprocessing* bisa diekspor dari HDFS untuk *dashboard* atau BI tools.
4. Keamanan dan Keandalan
   Dengan replikasi blok dan mekanisme toleransi kesalahan, HDFS menjaga ketersediaan dan integritas data, yang penting agar hasil analisis data sains valid.

Contoh Penggunaan HDFS di Data Sains misalnya untuk menyimpan *dataset* transaksi bank untuk analisis risiko kredit, menyimpan log sensor IoT untuk prediksi kegagalan mesin. Atau menyimpan data klik dan perilaku pengguna dari situs *web e-commerce* untuk model rekomendasi.

B. Riset Keamanan Informasi-Data Sains

Dalam konteks data sains modern, riset dan praktik di bidang keamanan informasi semakin berkembang dan dapat diterapkan dalam berbagai *field* berikut:

1. Deteksi Serangan Siber Menggunakan Data Sains

Data sains dapat digunakan untuk mendeteksi serangan siber secara proaktif dengan memanfaatkan analisis data besar dan *machine learning*:

a. Analisis log server: mengidentifikasi pola akses abnormal dan potensi intrusi.

b. Deteksi malware menggunakan machine learning: klasifikasi file atau trafik jaringan sebagai benign atau malicious.

c. Identifikasi perilaku abnormal (*anomaly detection*): kendeteksi aktivitas yang menyimpang dari pola normal pengguna atau sistem.

2. Threat Intelligence dan Forensik Digital

Data sains mendukung kegiatan ***threat intelligence*** dan ***digital forensics***, sehingga organisasi dapat memahami, merespons, dan memitigasi ancaman secara lebih efektif:

a. Investigasi insiden keamanan: menggunakan analisis data untuk menemukan penyebab serangan dan jalur infiltrasi.

b. Analisis pola serangan historis: membantu memprediksi jenis serangan yang mungkin terjadi di masa depan.

c. Visualisasi hubungan antar entitas: misalnya menganalisis jaringan botnet, koneksi antar malware, atau hubungan antara pelaku serangan.

3. Visualisasi Serangan Siber

Visualisasi data merupakan elemen penting untuk membaca, memahami, dan menyajikan informasi keamanan:

a. Menampilkan pola serangan dalam bentuk grafik atau *heatmap*.

b. Menunjukkan jalur serangan dan titik rentan sistem secara interaktif.

c. Memudahkan tim keamanan dalam membuat keputusan cepat berdasarkan bukti visual.

4. Topik Riset Lain yang Aplikatif

Beberapa riset lanjutan yang dapat diterapkan di bidang data sains dan keamanan informasi antara lain:

a. Prediksi serangan siber berbasis AI/ML: menggunakan model prediktif untuk memperkirakan potensi serangan sebelum terjadi.

b. Analisis malware polymorphic: menggunakan clustering dan deep learning untuk mendeteksi malware yang terus berubah bentuk.

c. Keamanan IoT dan Smart City: deteksi anomali pada perangkat IoT, sistem smart building, dan sensor kota pintar.

d. Keamanan cloud dan *big data*: analisis akses, audit log, dan deteksi kebocoran data di infrastruktur cloud.

e. *Threat hunting* otomatis: menggabungkan *machine learning*, *network traffic analysis*, dan SIEM (Security Information & Event Management).

Daftar Pustaka

[1] Shaun V. Ault, Soohyun N.L. and Larry Musolino, Principles of Data Science, Texas: Openstax, 2025.

[2] P. Flach, Machine Learning : The Art and Science of Algorithms that Make Sense of Data, Cambridge: Cambridge University Press, 2012.

[3] K. Inc., "Kaggle," [Online]. Available: https://www.kaggle.com/datasets. [Accessed Maret 2026].

[4] D. C. Inc., "Data Lab," Data Camp, 2026. [Online]. Available: https://www.datacamp.com/datalab/datasets. [Accessed Maret 2026].

[5] Dirk P. Kroese, Zdravko I. Botev, Thomas Taimre and Radislav Vaisman, Data Science and Machine Learning : Matematical and Statistical Methods, Brisbane: Chapman and Hall/CRC, 2024.

[6] R. A. Irizarry, Introduction to Data Science, Spanish: CRC Press, 2021.

[7] J. V. Guttag, Introduction to Computation and Programming Using Python, London, England: The MIT Press, 2013.

[8] J. Grus, Data Science from Scratch, Seattle: O'Reilly Media, 2015.

[9] Buck Woody, Danielle Dean, Debraj GuhaThakurta, Gagan Bansal, Matt Conners and Wee-Hyong Tok, Data Science with Microsoft SQL Server 2016, Washington: Microsoft Press, 2016.

[10] Sensuse D., Prasetyo A., Rachmawati R. and Sunindyo W.D., "Initial Cybersecurity Framework in the New Capital City of Indonesia: Factors, Objectives, and Technology," *MDPI,* vol. 13, no. 580, 2022.

[11] Z. Muhammad, J.Nafis, R.Nazilla, R.Nugraha and S.Uyun, "Komputika : Jurnal Sistem Komputer Perbandingan Algoritma Decision Tree dan K-Nearest Neighbour Algorithms for IoT Network Attack Classification," *Komputika,* vol. 13, pp. 245-252, 2024.

[12] Farenza I.R. and Mulia V.Z.D., "Pengklasifikasian Aktivitas Normal dan Serangan Pada sistem Smart Home IoT Menggunakan Algoritma Machine Learning," Internal Report, Indonesia, 2025.

[13] Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Maryland: Morgan & Claypool Synthesis Lectures, 2010.

[14] C. O. Website, "Crunchbase," Crunchbase, 2024. [Online]. Available: https://www.crunchbase.com/. [Accessed January 2024].

[15] T. White, Hadoop : The Definitive Guide, Storage and Analysis at Internet Scale, USA: O'Reilly , 2015.

Biografi Singkat Penulis

Rini Wisnu Wardhani adalah akademisi dan peneliti di bidang keamanan siber dan kriptografi. Ia meraih gelar Doctor of Philosophy (Ph.D.) dalam bidang Computer Science and Engineering dari Pusan National University, Korea Selatan. Latar belakang pendidikannya meliputi Diploma Teknik Kriptografi, Sarjana Teknik Elektro, dan Magister Teknik Elektro yang memperkuat keahliannya dalam teknologi keamanan informasi.

Saat ini, Rini Wisnu Wardhani menjabat sebagai Lektor pada Program Studi Hardware Security di Politeknik Siber dan Sandi Negara (Poltek SSN) serta berkarier sebagai pegawai pemerintah di Badan Siber dan Sandi Negara (BSSN). Ia juga pernah terlibat dalam riset di Blockchain Research Center, Pusan National University, dengan fokus pada keamanan informasi, Artificial Intelligence of Things (AIoT), komputasi kuantum, dan blockchain.

Selain aktif dalam pengajaran dan penelitian, Rini Wisnu Wardhani berkontribusi sebagai penelaah *(reviewer)* pada sejumlah jurnal nasional dan internasional. Minat risetnya mencakup Hardware Security, Information Security, Cryptography, dan Quantum Computing. Beragam kegiatan akademik dan profesional yang dijalaninya menjadi sarana untuk upaya kontribusi nyata dalam memajukan keamanan siber, baik di tingkat nasional maupun global.