TUGAS INDIVIDU :

DATA SCIENCE FUNDAMENTAL

Disusun untuk memenuhi tugas Mata Kuliah Pengantar Sains Data

Dosen Pengampu : **Eko Prasetio Widhi, S.Kom., M.Kom.**

Disusun oleh :

Ghulam Mushthofa 442023611060

PROGAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS DARUSSALAM GONTOR

2025

BAB 1:

DEFINISI

1.1. Data Science

Data Science adalah sebuah seni untuk mengubah data mentah menjadi wawasan (*insight*) yang bermanfaat dan dapat ditindaklanjuti. Ini bukan sekadar tentang *coding* atau matematika, melainkan sebuah bidang interdisipliner yang menggabungkan keahlian dari:

- **Ilmu Komputer:** Untuk mengolah data dalam skala besar dan membuat algoritma.
- **Statistika & Matematika:** Untuk merancang model dan memastikan kesimpulan yang diambil valid secara ilmiah.
- **Pengetahuan Domain/Bisnis:** Untuk memahami konteks masalah dan menerjemahkan hasil teknis menjadi solusi di dunia nyata.

Singkatnya, seorang *Data Scientist* itu seperti detektif yang menggunakan data sebagai petunjuk untuk memecahkan masalah bisnis atau menjawab pertanyaan yang kompleks.

Tiga bidang yang dapat memanfaatkan Data Science:

1. **E-commerce & Ritel:** Untuk membuat sistem rekomendasi produk yang dipersonalisasi (seperti di Tokopedia atau Shopee), memprediksi permintaan barang, dan menganalisis sentimen pelanggan dari ulasan produk.
2. **Kesehatan (Healthcare):** Untuk membantu dokter mendiagnosis penyakit lebih awal melalui analisis gambar medis (seperti CT scan), memprediksi wabah penyakit berdasarkan data geografis, dan mengembangkan obat baru.
3. **Keuangan (Finance):** Untuk mendeteksi transaksi penipuan (*fraud detection*) secara *real-time*, melakukan penilaian risiko kredit bagi calon peminjam, dan membuat model prediksi pergerakan harga saham.

1.2. Komponen Utama dalam Data Science Pipeline

Data Science Pipeline adalah alur kerja standar dari awal hingga akhir sebuah proyek data. Ibaratnya seperti resep, ada langkah-langkah yang harus diikuti secara berurutan. Lima komponen utamanya adalah:

1. **Pengumpulan Data (Data Acquisition):** Tahap paling awal di mana kita mengumpulkan data mentah dari berbagai sumber. Sumbernya bisa dari database internal perusahaan, survei, API (misalnya, data dari Twitter), atau bahkan dengan teknik *web scraping*. Kualitas data yang dikumpulkan di sini sangat menentukan hasil akhir.

2. **Pembersihan dan Persiapan Data (Data Cleaning & Preparation):** Data mentah itu sering kali "kotor": ada yang hilang (*missing values*), tidak konsisten, atau formatnya salah. Di tahap ini, kita merapikan data tersebut. Prosesnya meliputi mengisi data yang hilang, mengoreksi kesalahan, dan mengubah data ke dalam format yang siap untuk dianalisis. Tahap ini sering memakan waktu paling banyak.

3. **Analisis Data Eksplorasi (Exploratory Data Analysis - EDA):** Setelah data bersih, kita mulai "berkenalan" dengan data tersebut. Tujuannya adalah untuk menemukan pola, anomali, atau hubungan antar variabel menggunakan statistik deskriptif dan visualisasi data (grafik dan plot). Tahap ini membantu kita merumuskan hipotesis sebelum masuk ke pemodelan.

4. **Pemodelan Data (Data Modeling):** Ini adalah inti dari Data Science, di mana kita menggunakan algoritma *machine learning* untuk membuat model prediktif atau klasifikasi. Misalnya, membuat model untuk memprediksi pelanggan mana yang akan berhenti berlangganan (*churn*) atau mengklasifikasikan email sebagai spam atau bukan.

5. **Komunikasi dan Visualisasi Hasil (Communication & Visualization):** Model yang canggih tidak akan berguna jika hasilnya tidak bisa dipahami oleh pengambil keputusan (misalnya, manajer atau direktur). Di tahap akhir ini, kita menyajikan

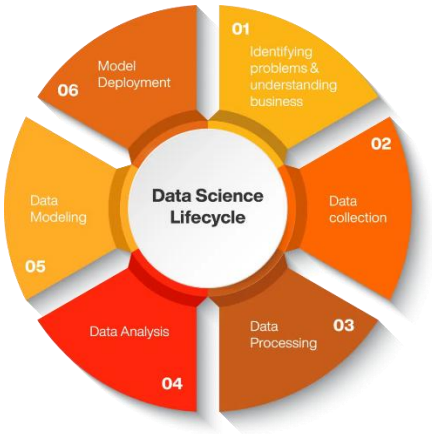
temuan dalam bentuk yang mudah dicerna, seperti dasbor interaktif, laporan, atau presentasi. Kemampuan *storytelling with data* sangat penting di sini.

1.3. Jenis Data

Di dunia nyata, data ada dalam berbagai bentuk. Secara umum, kita bisa membaginya menjadi tiga jenis:

- **Structured Data (Data Terstruktur):** Ini adalah jenis data yang paling rapi dan terorganisir, biasanya dalam format tabel dengan baris dan kolom yang jelas.
    - **Contoh:** Data transaksi penjualan di database sebuah toko. Setiap baris adalah satu transaksi, dan kolomnya berisi informasi seperti ID Transaksi, Tanggal, Nama Pelanggan, Nama Barang, Jumlah, dan Harga. Data di Microsoft Excel atau SQL Database adalah contoh klasiknya.
- **Semi-structured Data (Data Semi-terstruktur):** Data ini tidak sekaku data terstruktur, tetapi masih memiliki struktur atau penanda semantik untuk memisahkan elemen-elemennya.
    - **Contoh:** Data dalam format **JSON** (JavaScript Object Notation) yang sering digunakan oleh API. Ada "kunci" (*key*) dan "nilai" (*value*) yang membuatnya terorganisir, tapi tidak dalam format tabel yang ketat. Contoh lain adalah file XML atau sebuah email yang memiliki struktur (Pengirim, Penerima, Subjek) tetapi isi pesannya tidak terstruktur.
- **Unstructured Data (Data Tidak Terstruktur):** Jenis data ini tidak memiliki model atau struktur yang jelas dan merupakan mayoritas data di dunia.
    - **Contoh:** Isi sebuah postingan di media sosial (teks bebas), sebuah file gambar (kumpulan piksel), rekaman suara, atau video. Menganalisis data jenis ini memerlukan teknik yang lebih canggih seperti *Natural Language Processing* (NLP) untuk teks atau *Computer Vision* untuk gambar.

1.4. Tahapan Proses



Proses Data Science adalah serangkaian langkah sistematis untuk mengubah masalah bisnis menjadi solusi berbasis data. Urutannya adalah sebagai berikut:

1. **Pemahaman Masalah (Business Understanding):** Tahap awal yang krusial. Kita perlu mendefinisikan masalah yang ingin dipecahkan dan tujuan yang ingin dicapai bersama *stakeholder* (pemangku kepentingan). Pertanyaan yang harus dijawab: "Apa masalahnya?" dan "Bagaimana kesuksesan diukur?".

2. **Pengumpulan Data (Data Collection):** Setelah masalah jelas, kita mencari dan mengumpulkan data yang relevan dari berbagai sumber yang tersedia.

3. **Pembersihan Data (Data Cleaning):** Data mentah yang terkumpul dibersihkan dari error, data duplikat, dan nilai yang hilang (*missing values*) agar siap diolah.

4. **Eksplorasi Data (Data Exploration/EDA):** Di sini, kita melakukan analisis awal untuk memahami karakteristik data. Dengan visualisasi dan statistik, kita mencari pola, tren, dan hubungan tersembunyi di dalam data.

5. **Pembuatan Model (Modeling):** Berdasarkan wawasan dari EDA, kita memilih dan menerapkan algoritma *machine learning* untuk membangun model. Proses ini melibatkan pelatihan (*training*) dan pengujian (*testing*) model untuk memastikan performanya baik.

6. **Evaluasi Model (Model Evaluation):** Model yang sudah dibuat diukur kinerjanya menggunakan metrik tertentu (misalnya, akurasi, presisi). Jika

performanya belum memenuhi tujuan awal, kita mungkin perlu kembali ke tahap sebelumnya untuk melakukan penyesuaian.

7. **Komunikasi Hasil (Result Communication):** Tahap terakhir adalah mempresentasikan hasil temuan dan model kepada *stakeholder*. Hasil ini harus disajikan dalam bahasa yang mudah dimengerti dan fokus pada bagaimana solusi ini dapat menjawab masalah bisnis yang didefinisikan di awal.

BAB 2:

DATA SCIENTIST VS DATA ANALYST

Meskipun sering dianggap sama, peran *Data Scientist* dan *Data Analyst* memiliki perbedaan yang cukup mendasar:

Aspek	Data Analyst	Data Scientist
Tujuan Pekerjaan	Menjawab pertanyaan bisnis dengan menganalisis data historis. Fokus pada "apa yang terjadi?" dan "mengapa terjadi?".	Menggunakan data untuk membuat prediksi dan membangun sistem cerdas. Fokus pada "apa yang akan terjadi?" dan "bagaimana kita bisa membuatnya terjadi?".
Tugas Utama	Membersihkan data, membuat laporan rutin, membangun dasbor (dashboard) visual, dan melakukan analisis deskriptif.	Melakukan analisis eksplorasi, membangun dan menguji model machine learning, merancang eksperimen (A/B testing), dan membuat produk berbasis data.
Keterampilan	Kuat di SQL, Excel, dan tools visualisasi data seperti Tableau atau Power BI. Memiliki pemahaman statistik yang baik.	Memiliki semua keterampilan Data Analyst, ditambah kemampuan pemrograman (Python/R), pemahaman mendalam tentang algoritma machine learning, statistika tingkat lanjut, dan seringkali software engineering.

Secara sederhana, *Data Analyst* menjelaskan masa lalu, sementara *Data Scientist* mencoba memprediksi dan membentuk masa depan.

BAB 3:

BIG DATA VS SMALL DATA

Perbedaan utama keduanya bukan hanya soal ukuran, tetapi juga kompleksitas.

- **Big Data:**

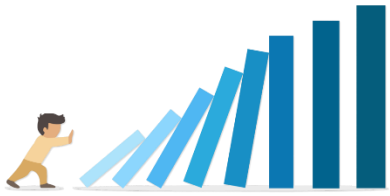
Merujuk pada data yang sangat besar, cepat berubah, dan beragam sehingga tidak dapat diolah dengan *tools* tradisional.

Big Data sering dijelaskan dengan konsep **3V**:

1. **Volume:** Ukurannya sangat masif (terabyte hingga petabyte).
2. **Velocity:** Dihasilkan dan masuk dengan kecepatan sangat tinggi (misalnya, data *streaming*).
3. **Variety:** Jenisnya sangat beragam (terstruktur, semi-terstruktur, dan tidak terstruktur).

    - **Contoh Penggunaan:** Analisis data sensor dari jutaan perangkat *Internet of Things* (IoT) secara *real-time* untuk memprediksi kapan sebuah mesin pabrik akan rusak.

- **Small Data:**



Merujuk pada data yang volumenya cukup kecil dan strukturnya cukup sederhana untuk disimpan dan dianalisis di satu komputer atau server.

  - **Contoh Penggunaan:** Data penjualan bulanan sebuah UMKM yang disimpan dalam file Excel. Data ini digunakan untuk menganalisis produk mana yang paling laku dan membuat laporan keuangan bulanan.

3.1 Sumber Data

Berikut adalah minimal lima sumber data yang bisa dimanfaatkan dalam Data Science:

1. **Database Internal Perusahaan:**
   - **Contoh:** Database SQL yang berisi data riwayat transaksi pelanggan, informasi produk, dan data karyawan. Ini adalah sumber data utama untuk analisis bisnis internal.
2. **Internet of Things (IoT):**
   - **Contoh:** Data sensor dari *smartwatch* yang merekam detak jantung dan aktivitas fisik pengguna. Data ini bisa digunakan di bidang kesehatan untuk memantau kondisi pasien.
3. **Media Sosial (Social Media APIs):**
   - **Contoh:** Mengambil data cuitan publik dari API Twitter yang mengandung kata kunci tertentu untuk menganalisis sentimen masyarakat terhadap sebuah merek atau kebijakan pemerintah.
4. **Data Terbuka dari Pemerintah (Open Data):**

- **Contoh:** Data kependudukan dari Badan Pusat Statistik (BPS) atau data curah hujan dari BMKG. Data ini dapat digunakan untuk penelitian sosial atau analisis kebijakan publik.

5. **Web Scraping:**
    - **Contoh:** Mengumpulkan data harga dan ulasan produk dari situs *e-commerce* seperti Tokopedia atau Amazon untuk melakukan analisis kompetitor atau tren pasar.

BAB 4:

STUDI KASUS

Bidang: Bisnis (E-commerce)

- **Masalah:** Sebuah platform *e-commerce* mengalami tingkat *customer churn* (pelanggan berhenti berbelanja) yang tinggi. Manajemen ingin memahami mengapa pelanggan pergi dan bagaimana cara mencegahnya.
- **Bagaimana Data Science Membantu:**
    1. **Pengumpulan Data:** Tim *data science* mengumpulkan berbagai data terkait perilaku pelanggan, seperti:
        - **Data demografis:** Usia, lokasi.
        - **Data transaksi:** Frekuensi pembelian, rata-rata nilai belanja, kategori produk yang dibeli.
        - **Data aktivitas:** Seberapa sering *login*, berapa lama waktu di aplikasi, produk apa yang dilihat.
        - **Data interaksi:** Apakah pernah komplain ke *customer service*, ulasan yang diberikan.
    2. **Analisis & Pemodelan:**
        - Pertama, dilakukan **Analisis Data Eksplorasi (EDA)** untuk menemukan pola. Mungkin ditemukan bahwa pelanggan yang tidak pernah membeli dalam 3 bulan terakhir memiliki kemungkinan 80% untuk *churn*.
        - Selanjutnya, dibangun sebuah **model klasifikasi *machine learning*** (misalnya, *Logistic Regression* atau *Random Forest*) untuk memprediksi probabilitas setiap pelanggan akan *churn* dalam waktu dekat. Model ini dilatih menggunakan data historis dari pelanggan yang sudah *churn* dan yang masih aktif.

3. **Solusi dan Tindak Lanjut:**

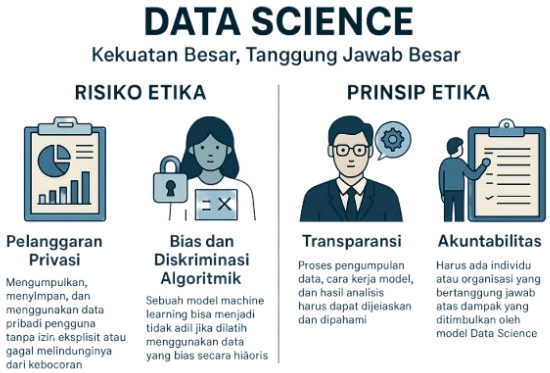
    - Model tersebut menghasilkan "skor risiko *churn*" untuk setiap pelanggan aktif.
    - Berdasarkan skor ini, tim pemasaran dapat meluncurkan kampanye yang ditargetkan. Pelanggan dengan skor risiko tinggi bisa diberikan penawaran khusus seperti *voucher* diskon, notifikasi personal, atau poin loyalitas ekstra untuk mendorong mereka bertransaksi kembali.

Dengan pendekatan ini, perusahaan bisa bertindak proaktif untuk mempertahankan pelanggan berharganya, bukan reaktif setelah mereka pergi.

BAB 5:

ETIKA DATA SCIENCE

Dalam praktiknya, Data Science memiliki kekuatan besar, dan dengan kekuatan itu datang tanggung jawab besar.



Berikut adalah dua risiko dan dua prinsip etika yang sangat penting:

Dua Risiko Etika:

1. **Pelanggaran Privasi:** Mengumpulkan, menyimpan, dan menggunakan data pribadi pengguna tanpa izin eksplisit atau gagal melindunginya dari kebocoran. Contohnya adalah skandal Cambridge Analytica yang menggunakan data pribadi pengguna Facebook untuk tujuan politik tanpa persetujuan.
2. **Bias dan Diskriminasi Algoritmik:** Sebuah model *machine learning* bisa menjadi tidak adil jika dilatih menggunakan data yang bias secara historis. Contohnya, sistem rekrutmen otomatis yang ternyata lebih sering menolak kandidat perempuan karena data latihnya didominasi oleh karyawan laki-laki. Algoritma ini hanya mengotomatisasi diskriminasi yang sudah ada.

Dua Prinsip Etika:

1. **Transparansi (Transparency):** Proses pengumpulan data, cara kerja model, dan hasil analisis harus dapat dijelaskan dan dipahami, terutama oleh orang-orang yang terkena dampak keputusan dari model tersebut. Model yang bekerja seperti "kotak hitam" (*black box*) sangat berisiko.

2. **Akuntabilitas (Accountability):** Harus ada individu atau organisasi yang bertanggung jawab atas dampak yang ditimbulkan oleh model Data Science. Jika sebuah model kredit secara keliru menolak pinjaman seseorang, harus ada mekanisme untuk memperbaiki kesalahan tersebut dan pihak yang bisa dimintai pertanggungjawaban.

BAB 6:

KETRAMPILAN DATA SCIENCE

Seorang *Data Scientist* yang andal perlu menguasai berbagai keterampilan. Berikut adalah enam di antaranya:

1. **Pemrograman (Python atau R):**
    - **Mengapa penting?** Ini adalah alat utama untuk membersihkan data, melakukan analisis, dan membangun model *machine learning*. Tanpa kemampuan *coding*, seorang *data scientist* tidak bisa mengimplementasikan idenya.
2. **Statistika dan Probabilitas:**
    - **Mengapa penting?** Statistik adalah fondasi untuk memahami data, merancang eksperimen yang valid (seperti A/B Testing), dan menginterpretasikan hasil model secara benar. Ini membantu membedakan antara sinyal (pola nyata) dan *noise* (kebetulan).
3. **Machine Learning:**
    - **Mengapa penting?** Ini adalah inti dari kemampuan prediktif seorang *data scientist*. Memahami berbagai algoritma, kapan menggunakannya, dan bagaimana mengevaluasi performanya adalah tugas sehari-hari.
4. **Manajemen Database (khususnya SQL):**

- **Mengapa penting?** Sebagian besar data perusahaan disimpan dalam database relasional. SQL (*Structured Query Language*) adalah bahasa standar untuk mengambil, memfilter, dan menggabungkan data dari database tersebut.

5. **Visualisasi dan Komunikasi Data:**
    - **Mengapa penting?** Kemampuan untuk menceritakan sebuah cerita (*storytelling*) menggunakan data sangat krusial. Wawasan yang paling hebat pun tidak akan berguna jika tidak dapat dikomunikasikan secara efektif kepada pengambil keputusan yang mungkin tidak memiliki latar belakang teknis.
6. **Pemahaman Bisnis/Domain (Business Acumen):**
    - **Mengapa penting?** *Data scientist* terbaik adalah mereka yang memahami konteks bisnis dari masalah yang sedang dikerjakan. Ini memungkinkan mereka untuk mengajukan pertanyaan yang tepat, memilih variabel yang relevan, dan menerjemahkan hasil teknis menjadi dampak bisnis yang nyata.