Tab 1

NAMA KELOMPOK :

1. Chelsea (5054241010)
2. Wayan Raditya Putra (5054241029)
3. Syauqi Nabil Tasri (5054241040)

Prodi : Rekayasa Kecerdasan Artifisial

- **Attributes and Objects**

Data pada hakikatnya tersusun dari dua elemen utama, yaitu objek dan atribut. Objek adalah entitas yang sedang diamati, bisa berupa manusia, transaksi, dokumen, atau bahkan sebuah peristiwa. Sedangkan atribut adalah sifat atau karakteristik yang melekat pada objek tersebut, misalnya tinggi badan, warna mata, atau status pernikahan. Dengan kata lain, objek adalah "siapa atau apa yang kita amati", sedangkan atribut adalah "bagaimana kita mendeskripsikan objek itu".

Hal penting yang perlu dipahami adalah perbedaan antara atribut dan nilai atribut. Atribut adalah konsep abstrak, sementara nilai atribut adalah wujud konkret dari konsep itu. Contohnya, "tinggi badan" adalah atribut, sementara "170 cm" adalah nilai atribut. Satu atribut bisa diukur dengan berbagai cara (meter, feet, inci), dan beberapa atribut berbeda bisa saja menggunakan bentuk nilai yang sama (misalnya angka pada umur dan ID karyawan), tetapi makna di balik angka itu jelas tidak sama. Inilah alasan kenapa membedakan antara atribut dan nilai atribut menjadi krusial dalam analisis data. Selain itu, data tidak selalu berbentuk tabel sederhana. Data bisa memiliki struktur yang lebih kompleks, misalnya hubungan antar atribut, bagian yang hilang, atau data yang bergantung pada waktu (time series). Dengan memahami objek dan atribut sebagai fondasi, kita dapat membangun analisis data yang lebih kokoh, baik untuk tujuan eksplorasi maupun pemodelan.

- **Kelebihan**

  Pendekatan ini memberikan kerangka dasar yang jelas karena semua data, betapapun kompleks, dapat direduksi menjadi objek dan atribut. Dengan adanya atribut yang terstandarisasi, kita bisa membandingkan berbagai objek secara konsisten. Kerangka ini juga memungkinkan analisis lanjutan, seperti klasifikasi atau clustering, karena setiap objek dipetakan menjadi sekumpulan atribut yang bisa dihitung. Konsep objek dan atribut juga bersifat fleksibel dan umum, sehingga dapat diterapkan pada berbagai jenis data, mulai dari tabel transaksi hingga teks dan gambar, selama dapat direpresentasikan dalam bentuk atribut.

- **Kekurangan**

  Representasi dengan objek dan atribut berpotensi mereduksi kenyataan yang lebih kaya. Tidak semua aspek dari suatu objek dapat dicerminkan hanya dengan atribut.

Contohnya, rasa makanan atau kepuasan pelanggan sulit diukur dengan satu variabel sederhana. Selain itu, hubungan antar atribut bisa hilang jika hanya dipandang sebagai daftar ciri. Misalnya, dalam analisis teks, urutan kata akan hilang bila hanya dihitung frekuensinya. Terdapat juga potensi salah tafsir ketika nilai atribut dianggap memiliki sifat tertentu yang sebenarnya tidak ada, seperti menganggap ID mahasiswa menunjukkan peringkat. Akhirnya, dalam praktik nyata, data sering kali tidak lengkap, mengandung noise, atau ambigu, sehingga membuat representasi objek menjadi tidak sempurna.

- **Types Of Data**

Setelah memahami konsep objek dan atribut, langkah berikutnya adalah mengenali jenis-jenis data. Tidak semua atribut bersifat sama, karena cara kita memperlakukan nilai atribut berbeda tergantung pada sifat dasarnya. Secara umum, data dapat dibedakan menjadi empat tipe utama: nominal, ordinal, interval, dan rasio. Data nominal adalah data kategori yang hanya berfungsi sebagai label atau penanda. Contohnya warna mata, jenis kelamin, atau kode pos. Pada tipe ini, angka atau simbol hanyalah nama, tidak bisa dipakai untuk operasi matematis. Data ordinal lebih kaya, karena selain membedakan kategori, ia juga memberikan urutan. Contoh yang sering digunakan adalah peringkat lomba atau tingkat kepuasan (puas, netral, tidak puas). Namun, jarak antar kategori tidak bisa dipastikan sama. Data interval sudah memungkinkan kita menghitung perbedaan antar nilai. Contohnya adalah suhu dalam Celcius atau tanggal pada kalender. Perbedaan 10 derajat selalu sama di seluruh skala, tetapi titik nol tidak absolut. Sebaliknya, data rasio memiliki semua sifat interval, ditambah dengan makna nol absolut. Contoh yang jelas adalah panjang, umur, berat, atau suhu dalam Kelvin. Pada skala rasio, pernyataan seperti “dua kali lebih besar” menjadi sah secara matematis.

Selain empat tipe ini, data juga dapat dibagi menjadi diskrit dan kontinu. Data diskrit adalah data yang hanya bisa mengambil jumlah nilai terbatas, seperti jumlah anak atau kode pos. Sedangkan data kontinu bisa mengambil nilai riil tanpa batas yang praktis, seperti tinggi badan atau berat badan. Ada pula data asimetris, misalnya dalam analisis keranjang belanja, di mana keberadaan suatu item lebih penting daripada ketiadaannya.

- **Kelebihan dan Kekurangan per Tipe Data**

1. Nominal

Data nominal adalah data kategori yang hanya berfungsi sebagai label atau penanda identitas. Contoh: warna mata (biru, cokelat), jenis kelamin (pria, wanita), kode pos.

- ★ Kelebihan: sederhana, mudah dipahami, cocok untuk klasifikasi dasar. Tidak memerlukan skala atau ukuran.

- ★ Kekurangan: tidak ada urutan, jarak, atau perbandingan yang bermakna. Tidak bisa digunakan untuk operasi matematis seperti rata-rata.

2. Ordinal

Data ordinal tidak hanya membedakan kategori, tapi juga memiliki urutan. Contoh: peringkat lomba (juara 1, 2, 3), tingkat kepuasan (sangat puas, puas, netral).

- ★ Kelebihan: bisa menunjukkan urutan atau ranking, sehingga lebih informatif daripada nominal.

- ★ Kekurangan: jarak antar kategori tidak seragam atau tidak bisa diukur. Peringkat 1 dan 2 mungkin sangat dekat, sedangkan peringkat 2 dan 3 bisa jauh, tapi data ordinal tidak menangkap hal ini.

3. Interval

Data interval memungkinkan kita menghitung perbedaan antar nilai. Contoh: suhu dalam Celsius/Fahrenheit, tanggal kalender.

- ★ Kelebihan: perbedaan antara nilai bermakna, bisa digunakan untuk operasi seperti rata-rata atau standar deviasi.

- ★ Kekurangan: tidak memiliki nol absolut. Nol pada skala interval tidak berarti “tidak ada”. Karena itu, pernyataan rasio seperti “dua kali lebih panas” tidak valid.

4. Rasio

Data rasio memiliki semua sifat interval, ditambah dengan nol absolut. Contoh: panjang, berat, umur, suhu dalam Kelvin.

- ★ Kelebihan: paling kaya informasinya. Bisa digunakan untuk semua operasi matematis termasuk rasio. Nol absolut membuat pernyataan “dua kali lipat” bermakna.

- ★ Kekurangan: lebih jarang tersedia untuk fenomena sosial atau psikologis, karena sulit menemukan atribut dengan nol yang benar-benar absolut.

5. Diskrit

Data diskrit adalah data yang hanya bisa mengambil jumlah nilai terbatas atau dapat dihitung. Contoh: jumlah anak, jumlah transaksi, kode pos.

- ★ Kelebihan: cocok untuk data yang berbentuk hitungan. Mudah disimpan dan diproses karena biasanya berupa bilangan bulat.

- ★ Kekurangan: tidak bisa menggambarkan variasi yang halus atau kontinu.

6. Kontinu

Data kontinu dapat mengambil nilai riil dalam suatu rentang. Contoh: tinggi badan, berat badan, suhu.

- Kelebihan: bisa menggambarkan variasi dengan detail, cocok untuk model matematis yang membutuhkan presisi.

- Kekurangan: secara praktis tidak bisa diukur dengan presisi sempurna karena keterbatasan alat ukur, sehingga selalu ada pembulatan atau error.

- **DATA QUALITY**

Kualitas data yang buruk berdampak negatif pada banyak proses pengolahan data. Menurut Thomas C. Redman (DM Review, Agustus 2004), kualitas data yang buruk merupakan "bencana yang terus berkembang" dan biasanya menyebabkan perusahaan kehilangan setidaknya 10% dari pendapatan, bahkan 20% lebih realistis. Contoh di data mining adalah model klasifikasi untuk mendeteksi risiko kredit: jika dibangun menggunakan data yang buruk, beberapa calon peminjam yang layak dapat ditolak, sementara lebih banyak pinjaman diberikan kepada individu yang akhirnya gagal membayar.

Beberapa masalah kualitas data meliputi noise, nilai yang hilang, data duplikat, dan data yang salah. Noise dapat berupa objek yang tidak relevan atau perubahan nilai asli, misalnya suara yang terdistorsi saat berbicara melalui telepon yang buruk atau "salju" pada layar televisi. Nilai yang hilang bisa terjadi karena informasi tidak dikumpulkan (misalnya seseorang menolak memberikan umur atau berat badan) atau atribut tidak berlaku untuk semua kasus (misalnya pendapatan tahunan untuk anak-anak). Penanganannya bisa dengan menghapus objek atau variabel, memperkirakan nilai yang hilang (misalnya dalam seri waktu suhu atau sensus), atau mengabaikan nilai yang hilang saat analisis. Nilai yang hilang dapat dikategorikan menjadi tiga: Missing Completely at Random (MCAR), Missing at Random (MAR), dan Missing Not at Random (MNAR), masing-masing memiliki metode pengisian dan dampak bias yang berbeda.

Data duplikat adalah masalah ketika satu set data memiliki objek yang sama atau hampir sama, terutama saat menggabungkan data dari sumber yang berbeda. Contohnya adalah satu orang dengan beberapa alamat email. Proses untuk menangani masalah ini disebut data cleaning, meskipun terkadang data duplikat tidak selalu harus dihapus tergantung konteks penggunaannya.

Kelebihan Data Quality:

- Meningkatkan akurasi analisis dan pengambilan keputusan.
- Meningkatkan efisiensi operasional perusahaan.
- Meningkatkan kepuasan pelanggan melalui layanan yang lebih baik.
- Memastikan kepatuhan terhadap regulasi dan standar industri.

Kekurangan Data Quality

- Mengarah pada keputusan bisnis yang tidak tepat.
- Berdampak negatif pada banyak proses pengolahan data
- Meningkatkan biaya operasional akibat perbaikan data yang terus-menerus.
- Menyebabkan hilangnya kepercayaan dari pelanggan dan mitra bisnis.
- Meningkatkan risiko hukum dan kepatuhan.

SIMILARITY DATA

Dalam data mining, **similarity** adalah ukuran numerik yang menyatakan seberapa mirip dua objek data, sedangkan **dissimilarity** atau jarak (distance) menyatakan seberapa berbeda keduanya. Semakin besar nilai similarity, semakin mirip dua objek tersebut, biasanya berada dalam rentang [0,1]. Sebaliknya juga, semakin kecil nilai dissimilarity, semakin mirip objek, dengan nilai minimum sering kali 0. Kedua ukuran ini sering disebut secara umum sebagai **proximity**.

Ada berbagai metode untuk mengukur similarity atau dissimilarity tergantung jenis data. Untuk data numerik, ukuran yang paling umum adalah **Euclidean Distance**, yang menghitung jarak lurus antara dua titik:

Euclidean Distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Ukuran ini adalah kasus khusus dari **Minkowski Distance**, yaitu generalisasi jarak dengan parameter rrr:

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r}$$

Jika r = 1, disebut **Manhattan/City block Distance**.

Jika r=2, sama dengan **Euclidean Distance**.

Jika r→∞, disebut **Supremum (L∞ norm)**, yaitu selisih terbesar di antara semua atribut.

Untuk data dengan korelasi antar atribut, digunakan **Mahalanobis Distance**:

$$\textbf{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})$$

Rumus ini mempertimbangkan kovariansi antar variabel sehingga lebih akurat untuk data multivariat.

Setiap ukuran jarak memiliki sifat umum, yaitu positif (≥0), simetris, dan memenuhi aturan segitiga (**triangle inequality**). Jika tiga sifat ini terpenuhi, ukuran tersebut disebut **metric**.

Untuk data biner (0/1), similarity bisa dihitung dengan dua ukuran populer. Pertama, **Simple Matching Coefficient (SMC)** yang memperhitungkan jumlah kesamaan baik pada 0 maupun 1:

- SMC = number of matches / number of attributes
  $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

Kedua, **Jaccard Coefficient** yang hanya memperhitungkan kesamaan pada nilai 1:

J = number of 11 matches / number of non-zero attributes
$= (f_{11}) / (f_{01} + f_{10} + f_{11})$

Pada data berbentuk vektor atau dokumen, sering digunakan **Cosine Similarity**, yang mengukur sudut kosinus antara dua vektor: $\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|$,

Selain itu juga ada **Extended Jaccard (Tanimoto)** yang merupakan variasi Jaccard untuk data yang kontinu. Similarity juga dapat diukur dengan **Correlation**, yang menyatakan hubungan linear antara variabel dengan nilai antara -1 hingga +1. Namun korelasi memiliki kelemahan, misalnya pada data non-linear bisa saja menunjukkan nilai nol meskipun sebenarnya ada hubungan kuat.

Metode yang lebih canggih adalah **information-based measures**, seperti **Entropy**, **Mutual Information (MI)**, dan **Maximal Information Coefficient (MIC)**. Ukuran ini berasal dari teori informasi, berguna untuk mendeteksi hubungan non-linear yang tidak bisa ditangkap oleh korelasi biasa.

Ketika data memiliki berbagai tipe atribut, similarity dari setiap atribut dapat digabungkan. Prosesnya dilakukan dengan menghitung similarity per atribut, lalu dirata-rata, dengan kemungkinan memberikan bobot yang berbeda sesuai tingkat kepentingan atribut.

Secara umum, pemilihan ukuran similarity tergantung pada **jenis data** (numerik, kategorikal, binary, dokumen, dll.) dan **domain aplikasi**. Tidak ada satu ukuran yang cocok untuk semua kasus; ukuran yang dipilih harus mampu menangkap kemiripan yang bermakna sesuai konteks data yang dianalisis.

5. Data Preprocessing

- Aggregation
  Aggregation dijelaskan sebagai proses menggabungkan dua atau lebih atribut (atau objek) menjadi satu atribut (atau objek). Tujuan utamanya ada tiga: reduksi data (mengurangi jumlah atribut/objek), perubahan skala (misalnya kota → provinsi/negara atau hari → minggu/bulan/tahun), dan mencapai data yang lebih stabil karena variabilitas cenderung berkurang setelah diagregasi. Contoh yang diberikan adalah curah hujan Australia periode 1982–1993: histogram simpangan baku (standard deviation) rata-rata bulanan dibandingkan dengan rata-rata tahunan pada grid lokasi yang sama menunjukkan bahwa variasi tahunan lebih kecil daripada variasi bulanan. Semua pengukuran (termasuk standard deviation) dinyatakan dalam sentimeter.

  Kelebihan:
  - Mengurangi kompleksitas & ukuran data; mempermudah analisis tingkat tinggi.
  - Menurunkan variabilitas (lebih "stabil"), sehingga pola makro lebih jelas (contoh curah hujan: variasi tahunan < bulanan).
  - Mengubah skala (spasial/temporal) sehingga cocok dengan kebutuhan model/analisis.

  Kekurangan:
  - Kehilangan granularitas; pola lokal/outlier bisa tersembunyi.
  - Risiko *ecological fallacy*/Simpson's paradox akibat penyamaan konteks pada level agregat.
  - Pemilihan jendela/level agregasi yang keliru dapat mendistorsi relasi antarfaktor.

- Sampling
  sampling diposisikan sebagai teknik utama untuk **reduksi data**, berguna pada investigasi awal maupun analisis akhir ketika memproses seluruh data terlalu mahal atau memakan waktu. Prinsip kunci yang ditekankan: sampel akan bekerja hampir sama baiknya dengan keseluruhan data **jika sampel itu representatif**, yaitu memiliki properti penting yang kira-kira sama dengan populasi asal. Materi juga menyinggung persoalan desain ukuran sampel, misalnya: "berapa besar sampel yang diperlukan agar **setidaknya satu objek** dari **masing-masing 10 kelompok** berukuran sama terambil?"—sebuah contoh yang menyoroti aspek cakupan antar-kelompok.

  Jenis-jenis sampling yang dicakup meliputi **simple random sampling**—dengan peluang sama bagi setiap item—yang bisa dilakukan **tanpa pengembalian** (item yang terambil dikeluarkan dari populasi) atau **dengan pengembalian** (item tetap di populasi sehingga bisa terambil berulang), serta **stratified sampling** yang membagi data ke dalam beberapa

partisi (strata) lalu mengambil sampel acak dari masing-masing partisi untuk menjaga keterwakilan.

Kelebihan:

- Reduksi biaya komputasi; memungkinkan eksplorasi/iterasi cepat.
- Jika representatif, hasilnya mendekati memakai seluruh data.
- Relevan saat seluruh data mahal/sulit diakses.

Kekurangan:

- Risiko *bias* (tidak representatif) → kesimpulan menyesatkan.
- Kelas langka/pola minor bisa terlewat.
- Hasil dapat ber-variasi tinggi jika ukuran sampel kecil (isu cakupan antarkelompok).

- Dimensionality Reduction

  Saat jumlah dimensi (atribut) meningkat, data menjadi semakin **jarang (sparse)** dalam ruang yang ditempatinya. Konsep-konsep seperti **kepadatan** dan **jarak antar titik**, yang krusial untuk klastering dan deteksi outlier, menjadi kurang bermakna. Bahkan, selisih antara jarak maksimum dan minimum antar-pasangan titik cenderung menyempit relatif, sehingga kemampuan pembedaan berdasarkan jarak menurun.

  Sebagai penangkalnya, **dimensionality reduction** diajukan dengan beberapa tujuan: menghindari kutukan dimensi, mengurangi waktu dan memori yang dibutuhkan algoritme, memudahkan **visualisasi**, serta berpotensi mengeliminasi fitur yang **irrelevan** atau mengurangi **noise**. Teknik yang disebutkan antara lain **Principal Components Analysis (PCA)**, **Singular Value Decomposition (SVD)**, serta teknik-teknik lain yang bersifat tersupervisi maupun non-linier. Khusus PCA, tujuannya adalah mencari **proyeksi** yang menangkap **variasi terbesar** dalam data, sehingga beberapa komponen teratas dapat merangkum informasi penting data asli.

  Kelebihan:

  - Mengatasi *curse of dimensionality*; mempercepat pelatihan/inferensi.
  - Dapat membuang fitur tidak relevan & mereduksi noise.
  - Memungkinkan visualisasi 2D/3D dan hemat memori/penyimpanan.

  Kekurangan:

  - Interpretabilitas turun (komponen = kombinasi linear fitur).
  - Bisa menghapus informasi yang relevan untuk tugas bila “variasi terbesar ≠ paling prediktif”.

    - Metode linear (seperti PCA) kurang menangkap struktur non-linier; sensitif ke skala/outlier.

- Feature Subset Selection
  memfokuskan pada pemilihan subset fitur alih-alih memproyeksikannya. Dua tipe fitur yang lazim dibuang adalah **redundant** (menggandakan informasi fitur lain—contohnya harga beli dan besaran pajak penjualan) dan **irrelevant** (tidak berguna untuk tugas yang dikerjakan—contohnya student ID ketika memprediksi GPA). Materi juga menyebut bahwa banyak teknik seleksi fitur dikembangkan khususnya untuk skenario **klasifikasi**.

  Kelebihan:
    - Tetap memakai fitur asli → lebih mudah diinterpretasi.
    - Mengurangi overfitting & biaya komputasi.
    - Menghilangkan fitur **redundan**/**irrelevan** (lebih fokus pada sinyal berguna).

  Kekurangan:
    - Ruang pencarian kombinatorial → rawan solusi suboptimal.
    - Interaksi fitur lemah namun komplementer bisa ikut terbuang.
    - Hasil pemilihan bisa tidak stabil antar *split*/resampling.

- Feature Creation
  feature creation membahas cara **menciptakan atribut baru** yang menangkap informasi penting data secara lebih efisien. Terdapat tiga pendekatan umum: **feature extraction** (misalnya mengekstraksi **tepi/edge** dari citra), **feature construction** (contohnya membentuk **densitas = massa/volume**), dan **mapping** data ke **ruang baru** (misalnya melalui **Fourier** atau **wavelet** untuk merepresentasikan komponen frekuensi dan memisahkan sinyal dari noise).

  Kelebihan:
    - Menyuntikkan pengetahuan domain; meningkatkan separabilitas kelas/pola.
    - Dapat memadatkan informasi bermakna (mis. Fourier/wavelet) dan memisahkan sinyal vs noise.
    - Memungkinkan hubungan non-linier terwakili lebih baik.

  Kekurangan:
    - Risiko *data leakage* bila konstruksi memakai info target secara tidak tepat.
    - Bisa menambah dimensi & kompleksitas → overfitting bila berlebihan.
    - Butuh keahlian & waktu untuk merancang fitur yang efektif.

- Discretization and Binarization
  Discretization adalah proses mengubah atribut **kontinu** menjadi atribut **ordinal**, memetakan jumlah nilai yang berpotensi tak hingga ke **sedikit kategori**. Diskretisasi umum dipakai pada **klasifikasi**, dan banyak algoritme bekerja lebih baik ketika baik variabel independen maupun dependen memiliki hanya beberapa nilai diskrit. Ilustrasi yang diangkat menggunakan **Iris dataset** dari UCI: tiga kelas bunga (Setosa, Versicolour, Virginica) dengan empat atribut non-kelas (sepal width/length, petal width/length). Dengan aturan sederhana, **petal width atau petal length yang rendah → Setosa**, **sedang → Versicolour**, **tinggi → Virginica**.

  Kelebihan:
  - Menyederhanakan fitur kontinu → kategori; membantu banyak algoritme klasifikasi.
  - Lebih tahan terhadap *noise* kecil; memudahkan pembuatan aturan yang interpretable.
  - Dapat selaras dengan kelas (supervised discretization) sehingga batas lebih informatif.

  Kekurangan:
  - Kehilangan informasi karena *binning*; isu ambang/batas.
  - Metode & jumlah *bin* yang keliru → bias/underfit.
  - Unsupervised *binning* bisa tak sejalan dengan label & salah menangani outlier/kelompok.

  Lalu selanjutnya yaitu binarization: dijelaskan sebagai pemetaan atribut **kontinu atau kategorikal** menjadi **satu atau lebih variabel biner**, lazim digunakan pada **association analysis** yang membutuhkan atribut biner **asimetris** (kehadiran bernilai informasi, ketiadaan tidak dianggap setara nilainya). Praktik umum: mengubah atribut kontinu menjadi kategorikal terlebih dahulu (misal **{rendah, sedang, tinggi}**), lalu melakukan **one-hot** menjadi sekumpulan atribut biner.

  Kelebihan:
  - Diperlukan untuk *association analysis* (atribut biner asimetris).
  - Mengubah kategori ke *one-hot* → mudah dipakai banyak model.
  - Representasi sederhana & jelas kehadiran/ketiadaan suatu kategori.

  Kekurangan:
  - *One-hot* dapat meledakkan dimensi → makin jarang (*sparse*).
  - Mengabaikan informasi ordinal/urutan jika dibinerkan mentah.

    - Fokus pada “kehadiran” bisa menyingkirkan informasi “ketiadaan” yang kadang relevan.

- Attribute Transformation
  Attribute Transformation endefinisikan transformasi atribut sebagai fungsi yang memetakan keseluruhan domain nilai suatu atribut ke domain baru secara konsisten, sehingga setiap nilai lama dapat diidentifikasi pasangannya pada domain baru. Dicontohkan fungsi-fungsi sederhana seperti xkx^kxk, log(x)\log(x)log(x), exe^xex, dan $|x||x||x|$. Di sini juga dibahas **normalisasi**—payung berbagai teknik untuk menyesuaikan perbedaan frekuensi kemunculan, mean, varians, atau rentang antar-fitur—serta menghilangkan sinyal umum yang tidak diinginkan seperti **musiman (seasonality)**. Sementara itu, dalam statistik, **standarisasi** (z-score) merujuk pada mengurangkan mean dan membagi dengan simpangan baku.

  Kelebihan:
    - Menyamakan skala → meningkatkan kinerja metode berbasis jarak/optimasi.
    - Dapat menghilangkan sinyal umum tak diinginkan (mis. musiman); contoh NPP: z-score bulanan menurunkan korelasi semu lintas kota.
    - Menstabilkan varians (mis. log untuk sebaran miring).

  Kekurangan:
    - Unit/makna asli berubah → koefisien/fitur kurang intuitif.
    - Risiko *leakage* jika parameter transformasi dihitung di seluruh data (bukan hanya train).
    - Tidak semua transform cocok (mis. log tidak untuk nilai ≤ 0).