Anggota kelompok:

1. Ashbar Selle 221011081
2. M. Fauzan Iskandar 221011063
3. Muhammad Fadel Hasyim 221011042

2 Algoritma Pra pemprosesan data

1. Min Max
   Merupakan algoritma yang digunakan untuk mengubah skala sekelompok data, biasanya berada diantara 0 dan 1.

   Rumusnya: $$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   Adapun tujuan mengubah skala suata dataset agar semua fitur memiliki kontribusi yang seimbang dan ramah diolah oleh algoritma yang sensitive terhadap skala data seperti KNN, dan SVM.

   Contoh:
   Ukuran tinggi badan siswa menjadi dataset penerapan ini, dengan data:

Nama	Tinggi badan
Fauzan	150
Ashbar	160
Fadel	170

   Hasilnya:

Nama	Data hasil
Fauzan	0.0
Ashbar	0.5
Fadel	1.0

2. One Hot Encoding
   Merupakan algoritma yang digunakan untuk mengubah data kategorikal menjadi numerik biner, artinya hanya bernilai 0 atau 1. Menjadi suatu cara yang berguna jika ingin menggunakan algoritma machine learning dengan data numerical. Algoritma ini bekerja menggunakan table, jika terdapat suatu kategori unik maka akan dibuatkan sebuah kolom berisi nilai 0 atau 1.

   Contoh:
   Data kategori warna yaitu merah, hijau, dan biru.

Warna	Warna_Merah	Warna_Hijau	Warna_Biru
Merah	1	0	0
Hijau	0	1	0
Biru	0	0	1

Algoritma Pra-Pemrosesan Data

"Standarisasi (Z-Score Normalization)"

- **Pengertian**

Standarisasi atau Z-Score Normalization adalah teknik pra-pemrosesan data yang digunakan untuk menstandarisasi fitur numerik dalam dataset sehingga memiliki mean (rata-rata) = 0 dan standar deviasi = 1 . Teknik ini sangat berguna ketika fitur-fitur dalam dataset memiliki skala yang berbeda-beda, karena standarisasi memungkinkan model pembelajaran mesin untuk memproses data tanpa bias akibat perbedaan skala.

- **Cara Kerja Algoritma Z-Score Normalization**

Proses standarisasi menggunakan rumus matematika berikut:

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

Dimana:

- z: Nilai terstandarisasi (hasil transformasi).
- x: Nilai asli dari fitur.
- μ: Mean (rata-rata) dari fitur tersebut.
- σ: Standar deviasi dari fitur tersebut.

Langkah-langkahnya adalah sebagai berikut:

1. Hitung Mean (μ) :

- Hitung rata-rata dari setiap fitur dalam dataset.
- Rumus mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Hitung Standar Deviasi (σ) :
    - Hitung standar deviasi dari setiap fitur.
    - Rumus standar deviasi:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

3. Transformasi Data :
    - Untuk setiap nilai *x* dalam fitur, hitung nilai terstandarisasi *z* menggunakan rumus di atas.

4. Hasil Akhir :
    - Dataset yang telah distandarisasi akan memiliki distribusi dengan mean = 0 dan standar deviasi = 1.

- **Contoh Penerapan Lengkap**

Kasus : Analisis Performa Siswa Berdasarkan Nilai Ujian

Sebuah sekolah ingin menganalisis performa siswa berdasarkan nilai ujian dua mata pelajaran: Matematika dan Bahasa Inggris. Namun, nilai Matematika memiliki rentang 0–100, sementara nilai Bahasa Inggris memiliki rentang 0–50. Untuk memastikan kedua fitur memiliki pengaruh yang sama dalam analisis, dataset perlu distandarisasi menggunakan Z-Score Normalization.

Dataset Awal :

Nama Siswa	Matematika (X1)	Bahasa Inggris (X2)
Ali	80	40
Budi	60	30
Cici	90	45
Dedi	70	35

- **Langkah 1: Hitung Mean dan Standar Deviasi**

1. Mean (μ) :
    - Matematika ($X1$):

$$\mu X1 = \frac{80 + 60 + 90 + 70}{4} = 75$$

    - Bahasa Inggris ($X2$):

$$\mu X2 = \frac{40 + 30 + 45 + 35}{4} = 37{,}5$$

2. Standar Deviasi (σ) :
    - Matematika ($X1$):

$\sigma X1 = \sqrt{\frac{(80-75)2+(60-75)2+(90-75)2+(70-75)^2}{4}} = \sqrt{\frac{25+225+225+25}{4}} = \sqrt{125}$

≈ 11.18

- Bahasa Inggris ($X2$):

$$\sigma X2 = \sqrt{\frac{(40-37.5)2+(30-37.5)2+(45-37.5)2+(35-37.5)2}{4}}$$

$$= \sqrt{\frac{6.25+56.25+56.25+6.25}{4}} = \sqrt{31.25} \approx 5.59$$

- **Langkah 2: Transformasi Data**

Gunakan rumus $z = \frac{x-\mu}{\sigma}$ untuk setiap nilai:

- Untuk Ali :
  - Matematika:

$$zX1 = \frac{80-75}{11.18} \approx 0.45$$

  - Bahasa Inggris:

$$zX2 = \frac{40-37.5}{5.59} \approx 0.45$$

- Untuk Budi :
  - Matematika:

$$zX1 = \frac{60-75}{11.18} \approx -1.34$$

  - Bahasa Inggris:

$$zX2 = \frac{30-37.5}{5.59} \approx -1.34$$

- Untuk Cici :
    - Matematika:

$$zX1 = \frac{90 - 75}{11.18} \approx 1.34$$

    - Bahasa Inggris:

$$zX2 = \frac{45 - 37.5}{5.59} \approx 1.34$$

- Untuk Dedi :
    - Matematika:

$$zX1 = \frac{70 - 75}{11.18} \approx -0.45$$

    - Bahasa Inggris:

$$zX2 = \frac{35 - 37.5}{5.59} \approx -0.45$$

Dataset Setelah Standarisasi :

Nama Siswa	**Matematika (Z1)**	**Bahasa Inggris (Z2)**
Ali	0.45	0.45
Budi	-1.34	-1.34
Cici	1.34	1.34

Nama Siswa	Matematika (Z1)	Bahasa Inggris (Z2)
Dedi	-0.45	-0.45

Interpretasi Hasil

Setelah standarisasi:

1. Dataset memiliki mean = 0 dan standar deviasi = 1 untuk kedua fitur.
2. Fitur Matematika dan Bahasa Inggris sekarang memiliki skala yang sama, sehingga dapat digunakan secara adil dalam analisis atau model pembelajaran mesin.
