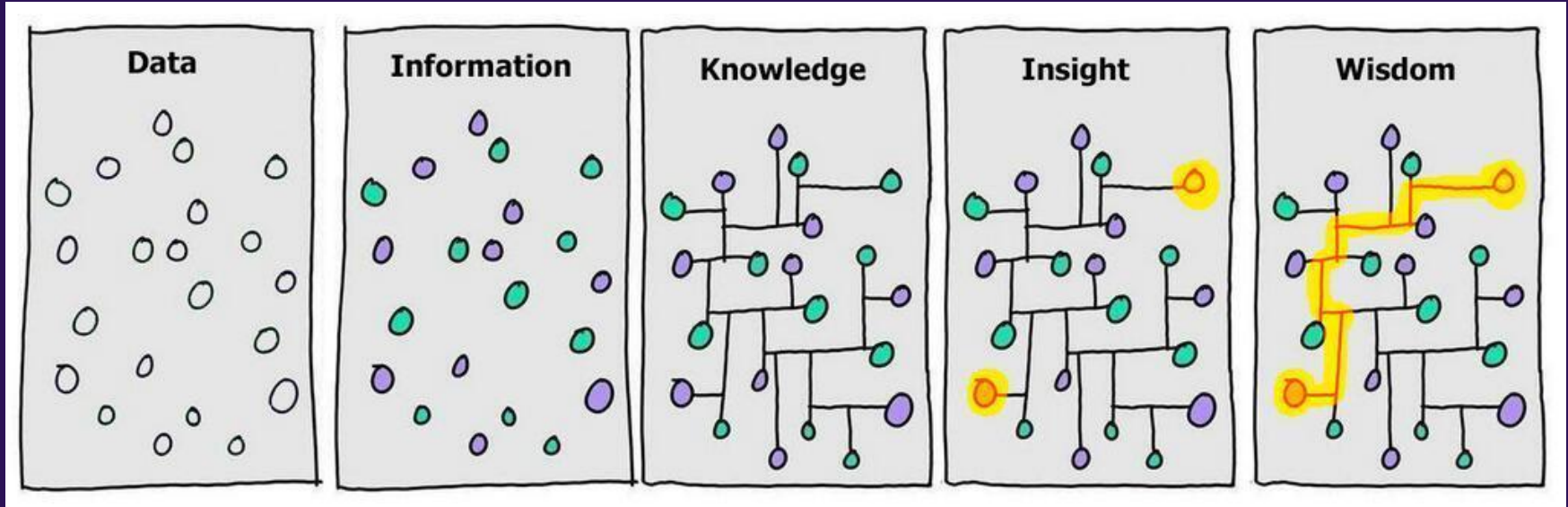


Pengantar *Data Mining* #2: Data & Weka

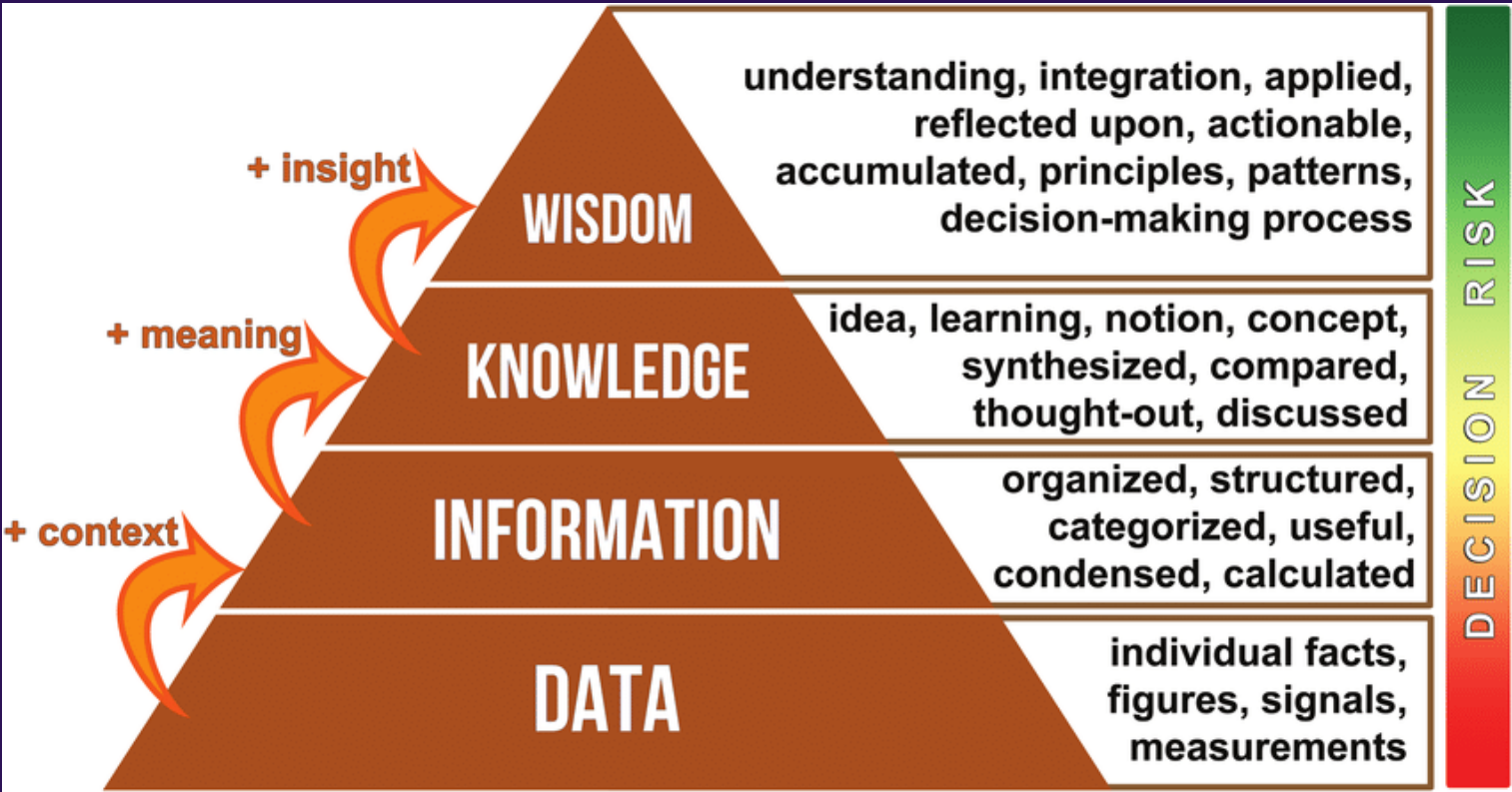
Isnan Mulia, S.Komp, M.Kom

Dari Data Menjadi Pengetahuan



Sumber: <https://www.theifactory.com/wp-content/uploads/2019/01/Data-Wisdom.jpg>

Dari Data Menjadi Pengetahuan



Sumber: <https://www.researchgate.net/publication/332400827/figure/fig6/AS:747208399912965@1555159773957/The-data-information-knowledge-wisdom-DIKW-hierarchy-as-a-pyramid-to-manage-knowledge.ppm>

Objek & Atribut Data

- Kumpulan data tersusun dari objek data, yang mewakili entitas tertentu
- Objek data dideskripsikan oleh atribut data
- Atribut data: *field* data yang mewakili karakteristik/fitur dari objek data

Contoh:

- *Database RS*
 - Entitas: pasien → atribut: nama, tanggal lahir, riwayat berobat
- *Database Samsat*
 - Entitas: kendaraan → atribut: merk, tipe, jumlah roda, kode mesin

Kelompok Atribut Data

Kuantitatif

Tipe data yang bisa dihitung,
berupa data numerik

- Tipe data interval
- Tipe data rasio

Kualitatif

Tipe data yang tidak bisa dihitung,
berupa data non-numerik

- Tipe data nominal
- Tipe data ordinal

Jenis Atribut Data

- Nominal/kategorik → mewakili kode/kategori tertentu
- Biner → atribut nominal yang hanya memiliki 2 nilai: benar/salah
 - Simetris: jika nilainya seimbang & bobotnya sama
 - Asimetris: jika bobotnya berbeda; nilai 1 menunjukkan kondisi yang jarang terjadi
- Ordinal → memiliki nilai yang bermakna urutan/peringkat
- Numerik → kuantitas yang bisa diukur, menggunakan bilangan
 - Skala interval: diukur dengan skala unit berukuran sama, bisa bernilai negatif, tidak memiliki titik nol sejati, operasi perkalian & pembagian tidak berlaku
 - Skala rasio: memiliki titik nol sejati, sebuah nilai dapat dinyatakan sebagai perkalian dari nilai lain

Contoh:

- Nominal/kategorik: pendidikan terakhir, pekerjaan, jenis mobil
- Biner:
 - Simetris: jenis kelamin
 - Asimetris: hasil test COVID-19, status penyakit TB
- Ordinal → tingkat kepuasan pelanggan
- Numerik :
 - Skala interval: suhu dalam Celsius/Fahrenheit, waktu
 - Skala rasio: tinggi benda, kecepatan, gaji

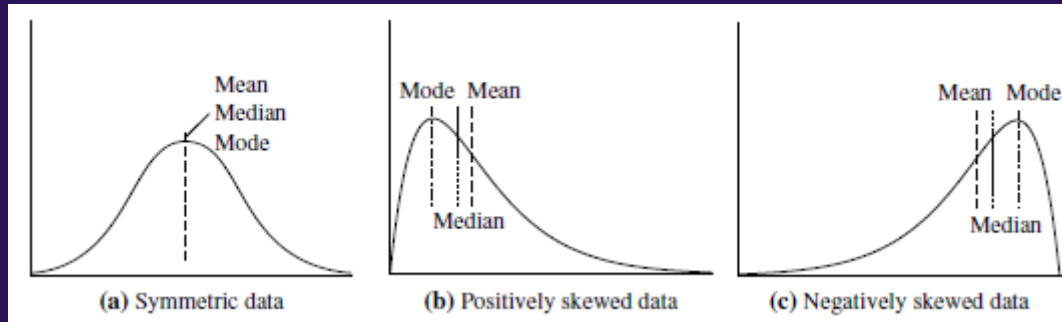
- Atribut Diskret:
 - ✓ memiliki kumpulan nilai yang jumlahnya terbatas atau terhitung-tak-terbatas (*countably infinite*)
 - ✓ bisa dicacah/dihitung (*countable*)
 - ✓ bisa direpresentasikan sebagai bilangan atau non-bilangan
- Contoh: warna rambut, ID user, jenis kelamin, jumlah roda, usia
- Atribut Kontinu:
 - ✓ bisa dinyatakan dalam bilangan real/desimal
 - ✓ bisa diukur (*measurable*)
- Contoh: suhu, gaji, waktu

Statistik Deskriptif Data

- Kecenderungan pusat data
 - Rata-rata, median, modus, *midrange*
- Persebaran data
 - *Quartile*, *interquartile range*, varian, standar deviasi
- Tampilan grafis
 - Histogram, *scatter plot*, *boxplot*

Kecenderungan Pusat Data

- Rata-rata = $\frac{\text{Jumlah nilai data}}{\text{Jumlah data } (n)}$
- Median = nilai data yang terdapat di posisi tengah, setelah nilai data diurutkan
 - Jika n ganjil, median = data ke- $(\lceil \frac{n}{2} \rceil)$
 - Jika n genap, median = di antara data ke- $(\frac{n}{2})$ & data ke- $(\frac{n}{2})$
- Modus = nilai data yang frekuensi kemunculannya paling banyak
- *Midrange* = rata-rata dari nilai tertinggi & terendah



Kecenderungan Pusat Data

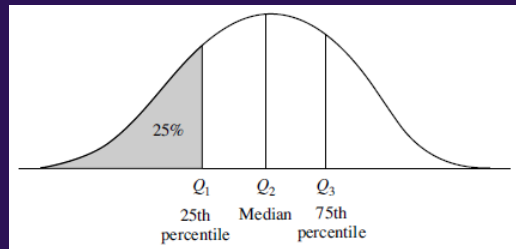
Contoh:

Data = 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 $\rightarrow n = 12$

- Rata-rata = $\frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58$
- Median:
 - $n = 12$, maka median berada di antara data ke-6 & data ke-7
 - Median = $\frac{52 + 56}{2} = 54$
- Modus = 52 & 70
- *Midrange* = $\frac{30+110}{2} = \frac{140}{2} = 70$

Persebaran Data

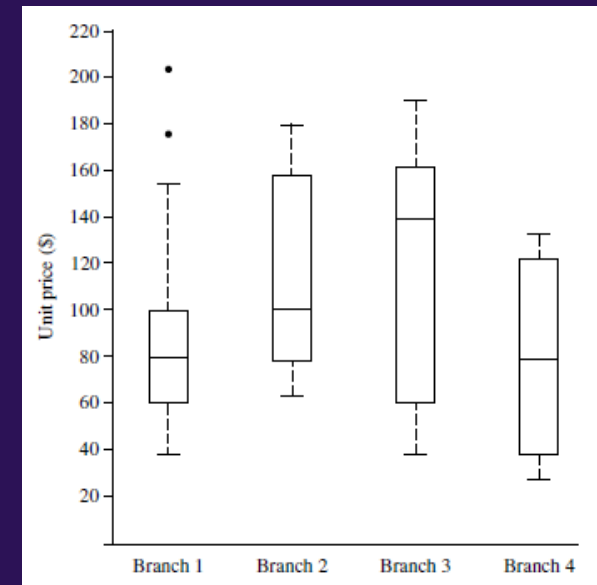
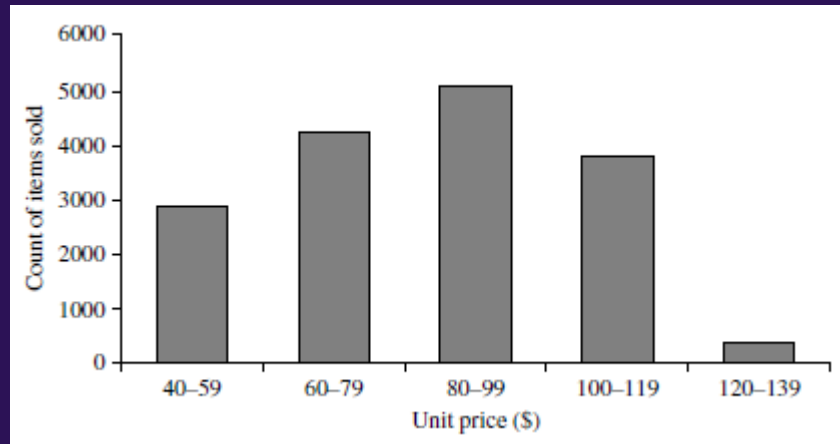
- *Quantile*: titik-titik distribusi data yang diambil pada interval yang teratur
 - Umum digunakan: *quartile* (Q)
 - *Interquartile range* = $Q_3 - Q_1$



- Varian & standar deviasi: ukuran persebaran data
 - Standar deviasi rendah → data cenderung dekat dengan rata-rata
 - Standar deviasi tinggi → data menyebar pada rentang nilai yang luas

Tampilan Grafis

- Histogram → ringkasan distribusi nilai atribut tertentu
- *Boxplot* → merangkum statistik 5 serangkai dari beberapa atribut dalam 1 diagram (minimum, Q_1 , median, Q_3 , maksimum)

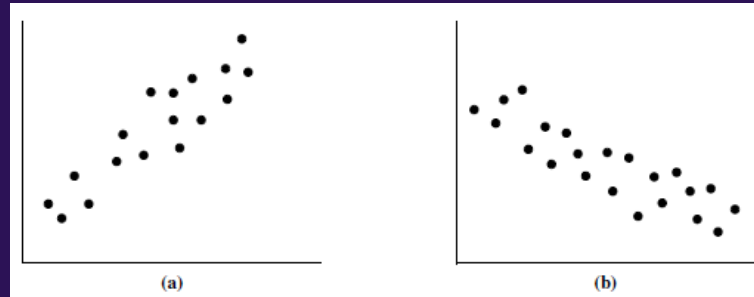


Tampilan Grafis

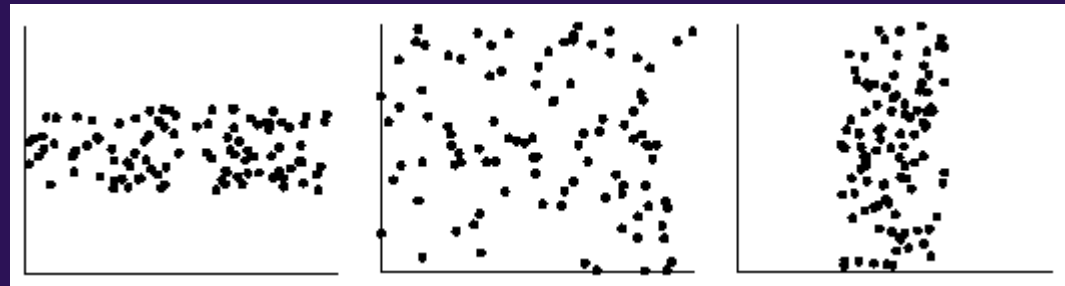
- *Scatter plot* → metode grafik untuk menentukan apakah ada hubungan/korelasi di antara dua atribut numerik
 - Plot setiap titik data pada bidang 2 dimensi

Ada korelasi →

- positif
- negatif



Tidak ada korelasi →



Mengukur Kesamaan Data

- Kesamaan (*similarity*) → bernilai 0 jika dua objek tidak mirip
- Ketidaksamaan (*dissimilarity*) → bernilai 0 jika dua objek mirip

$$sim(i, j) = 1 - d(i, j)$$

$$sim(i, j) + d(i, j) = 1$$

- Metode untuk mengukur kesamaan objek berbeda-beda, untuk setiap jenis atribut data

Matriks Data & Matriks Ketidaksamaan

- Matriks data: susunan dari n objek, masing-masing memiliki p atribut, dalam bentuk matriks berukuran $n \times p$
- Matriks ketidaksamaan: susunan berisi nilai ketidaksamaan dari setiap pasangan objek, dalam bentuk matriks berukuran $n \times n$

$$data = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$dissimilarity = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & d(n,n-1) & 0 \end{bmatrix}$$

$$similarity = \begin{bmatrix} 1 & & & & \\ sim(2,1) & 1 & & & \\ sim(3,1) & sim(3,2) & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ sim(n,1) & sim(n,2) & \cdots & sim(n,n-1) & 1 \end{bmatrix}$$

Atribut Nominal

$$d(i, j) = \frac{p - m}{p}$$

$$\text{sim}(i, j) = 1 - d(i, j) = \frac{m}{p}$$

m = jumlah atribut yang sama pada i & j

p = jumlah atribut yang dimiliki oleh objek

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Matriks ketidaksamaan =
$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Atribut Nominal

Contoh:

Roll No	Marks	Grades
1	96	A
2	87	B
3	83	B
4	96	A

Matriks ketidaksamaan:

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 0,5 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$d(1,1) = \frac{p - m}{p} = \frac{2 - 2}{2} = 0$$

$$d(2,2) = \frac{p - m}{p} = \frac{2 - 2}{2} = 0$$

$$d(3,3) = \frac{p - m}{p} = \frac{2 - 2}{2} = 0$$

$$d(2,1) = \frac{p - m}{p} = \frac{2 - 0}{2} = 1$$

$$d(3,2) = \frac{p - m}{p} = \frac{2 - 1}{2} = 0,5$$

$$d(4,3) = \frac{p - m}{p} = \frac{2 - 0}{2} = 1$$

$$d(3,1) = \frac{p - m}{p} = \frac{2 - 0}{2} = 1$$

$$d(4,2) = \frac{p - m}{p} = \frac{2 - 0}{2} = 1$$

$$d(4,4) = \frac{p - m}{p} = \frac{2 - 2}{2} = 0$$

$$d(4,1) = \frac{p - m}{p} = \frac{2 - 2}{2} = 0$$

Atribut Biner

Tabel kontingensi:

		Objek j		
		1	0	Jumlah
Objek i	1	q	r	$q + r$
	0	s	t	$s + t$
	Jumlah	$q + s$	$r + t$	p

$$p = q + r + s + t$$

Ketidaksamaan biner simetris → semua variabel diperhatikan

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

Ketidaksamaan biner asimetris → variabel t tidak diperhatikan

$$d(i, j) = \frac{r + s}{q + r + s}$$

Atribut BinerContoh (asimetris):

<i>name</i>	<i>gender</i>	<i>fever</i>	<i>cough</i>	<i>test-1</i>	<i>test-2</i>	<i>test-3</i>	<i>test-4</i>
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$d(\text{Jack}, \text{Jim}) = \frac{1 + 1}{1 + 1 + 1} = 0,67$$

$$d(\text{Jim}, \text{Mary}) = \frac{1 + 2}{1 + 1 + 2} = 0,75$$

$$d(\text{Jack}, \text{Mary}) = \frac{0 + 1}{2 + 0 + 1} = 0,33$$

Atribut Numerik

- Jarak Euclidean

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

- Jarak Manhattan (*city block*)

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- Jarak Minkowski → generalisasi dari jarak Euclidean & jarak Manhattan

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

Atribut NumerikContoh:

Bunga	Atribut 1	Atribut 2	Atribut 3	Atribut 4
1	5,0	3,3	1,4	0,2
2	7,0	3,2	4,7	1,4

- Jarak Euclidean

$$d(1,2) = \sqrt{(5,0 - 7,0)^2 + (3,3 - 3,2)^2 + (1,4 - 4,7)^2 + (0,2 - 1,4)^2} = \sqrt{16,34} \approx 4,04$$

- Jarak Manhattan (*city block*)

$$d(1,2) = |5,0 - 7,0| + |3,3 - 3,2| + |1,4 - 4,7| + |0,2 - 1,4| = 6,6$$

Atribut Ordinal

Sebelum diukur ketidaksamaannya, atribut ordinal perlu diubah menjadi bilangan, di mana untuk setiap nilai/*state* diberikan bilangan yang sesuai dengan urutan nilai atribut.

➔ Nilai diskalakan pada rentang yang umum, seperti $[0, 1]$, dengan cara:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

r = *rank*/peringkat dari nilai atribut ordinal

M = jumlah maksimal dari *state* atribut ordinal

➔ Setelah diubah, ketidaksamaan atribut dapat dihitung menggunakan rumus jarak atribut numerik

Atribut Ordinal

Contoh:

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Urutan atribut ordinal:

1 – fair, 2 – good, 3 – excellent → $M_f = 3$

$$z_{fair} = \frac{1 - 1}{3 - 1} = 0$$

$$z_{excellent} = \frac{3 - 1}{3 - 1} = 1$$

$$z_{good} = \frac{2 - 1}{3 - 1} = 0,5$$

$$d(1,1) = \sqrt{(1 - 1)^2} = 0$$

$$d(3,2) = \sqrt{(0 - 0,5)^2} = 0,5$$

$$d(2,1) = \sqrt{(1 - 0)^2} = 1$$

$$d(4,2) = \sqrt{(0 - 1)^2} = 1$$

$$d(3,1) = \sqrt{(1 - 0,5)^2} = 0,5$$

$$d(3,3) = \sqrt{(0,5 - 0,5)^2} = 0$$

$$d(4,1) = \sqrt{(1 - 1)^2} = 0$$

$$d(4,3) = \sqrt{(0,5 - 1)^2} = 0,5$$

$$d(2,2) = \sqrt{(0 - 0)^2} = 0$$

$$d(4,4) = \sqrt{(1 - 1)^2} = 0$$

Matriks ketidaksamaan:

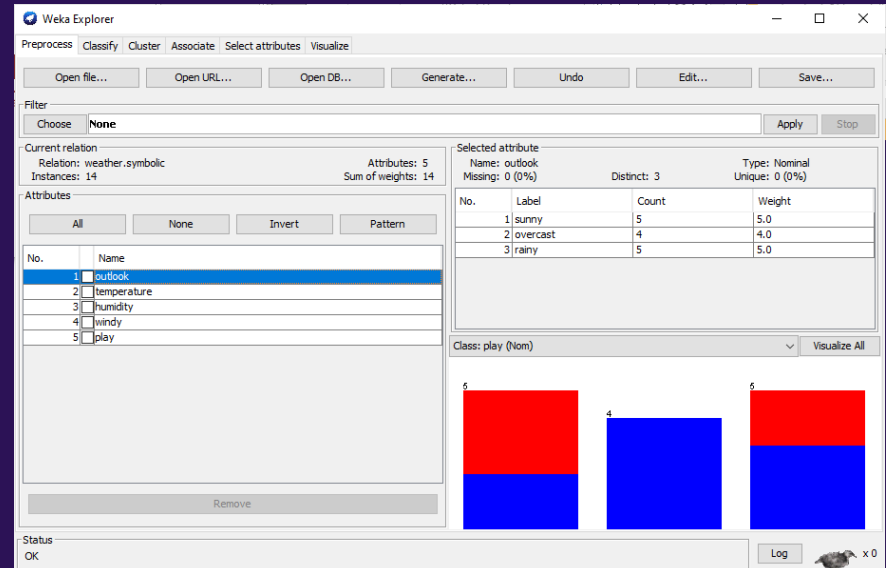
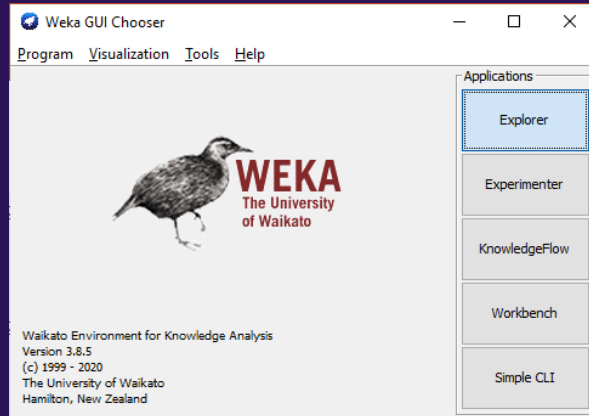
$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0,5 & 0,5 & 0 & \\ 0 & 1 & 0,5 & 0 \end{bmatrix}$$

Sumber Data

- Data hasil pengukuran/pengamatan
- Data hasil survei
- Data transaksi harian
- Dari situs penyedia data
 - UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>)
 - Kaggle Datasets (<https://www.kaggle.com/datasets>)

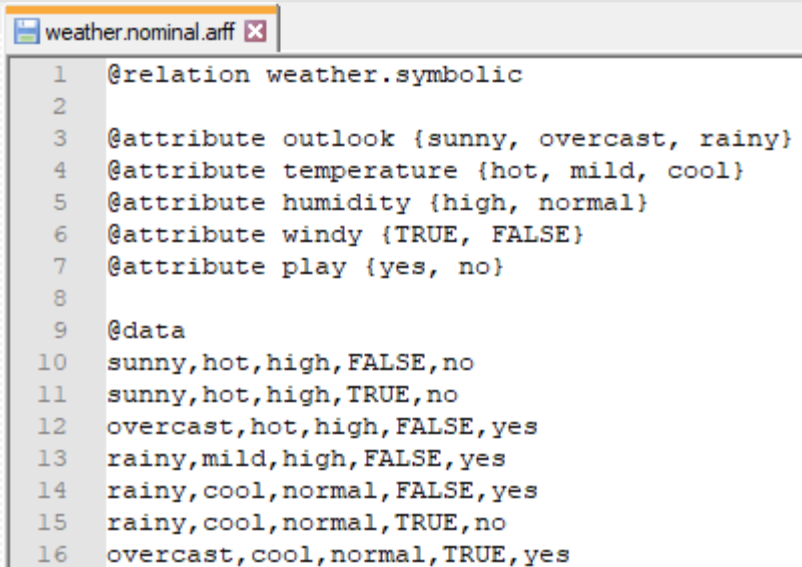
Weka

- “Waikato Environment for Knowledge Analysis”
- Dikembangkan oleh Ian Witten & tim, dari The University of Waikato, Hamilton, NZ
- Basis: bahasa pemrograman Java
- Umum digunakan untuk analisis data



Format Data Weka

- Menggunakan data ARFF (Attribute-Relation File Format) sebagai input
==> Mirip seperti file CSV, dengan tambahan informasi mengenai relasi & jenis atribut data
- @relation → nama relasi pada file
- @attribute → detail atribut data
- @data → bagian data



```
1 @relation weather.symbolic
2
3 @attribute outlook {sunny, overcast, rainy}
4 @attribute temperature {hot, mild, cool}
5 @attribute humidity {high, normal}
6 @attribute windy {TRUE, FALSE}
7 @attribute play {yes, no}
8
9 @data
10 sunny,hot,high,FALSE,no
11 sunny,hot,high,TRUE,no
12 overcast,hot,high,FALSE,yes
13 rainy,mild,high,FALSE,yes
14 rainy,cool,normal,FALSE,yes
15 rainy,cool,normal,TRUE,no
16 overcast,cool,normal,TRUE,yes
```

Membuat *File* ARFF

- Simpan data, yang mungkin semula menggunakan *extension* XLS/XLSX, sebagai *file* CSV
- Buka *file* CSV di *code editor* (Notepad++, VS Code, dll)
- Simpan *file* sebagai file ARFF
- Tambahkan keterangan untuk bagian @relation, @attribute, & @data pada bagian atas *file*
- Simpan *file*
- *File* ARFF sudah siap untuk dimuat di Weka

Soal Latihan

Diketahui data gaji karyawan dari sebuah perusahaan sebagai berikut.
Tentukan jenis dari setiap atribut yang ada pada data tersebut.

No	Nama	JK	Departemen	Grade	Gaji per Hari	Hari Masuk	Gaji Diterima
1	Antoni	L	IT	Staf	200.000	22	4.400.000
2	Yudha	L	Maintenance	Supervisor	420.000	21	8.820.000
3	Bagas	L	Finance	Supervisor	450.000	18	8.100.000
4	Indah	P	Finance	Staf	250.000	18	4.500.000
5	Annisa	P	Marketing	Manager	600.000	20	12.000.000

Recap

- Objek & Atribut Data
- Jenis Atribut Data
- Statistik Deskriptif Data
- Mengukur Kesamaan Data
- Sumber Data
- Weka

Next: mengapa kita perlu melakukan praproses data?

つづく