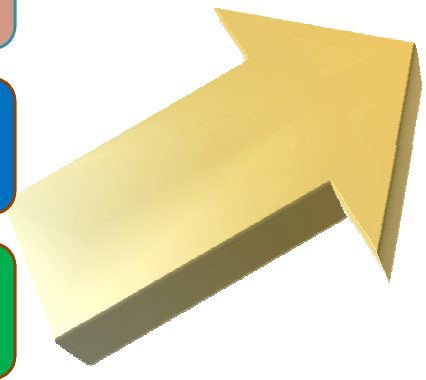
Data Mining

Data Preprocessing

Tutik Khotimah, S.Kom, M.Kom

Tujuan Instruksional

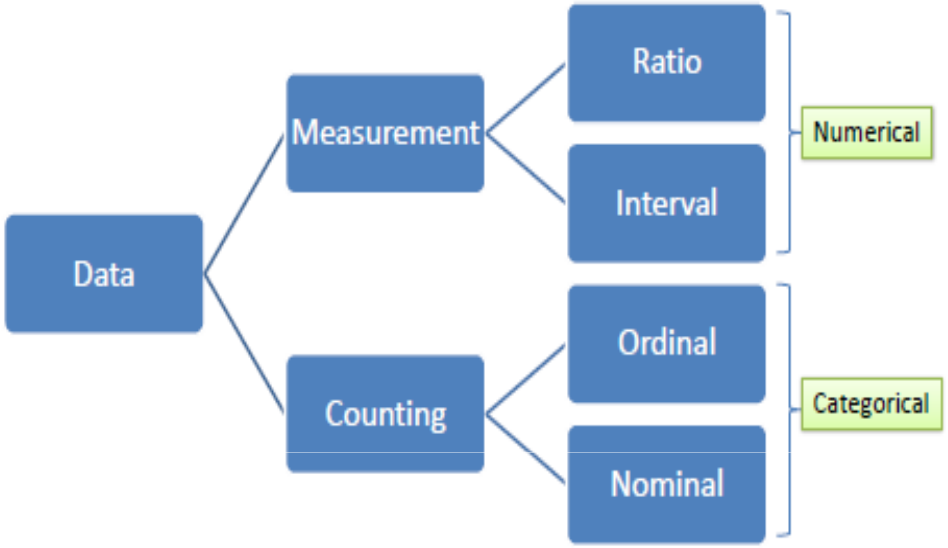
- Menjelaskan macam-macam variabel
- Memberi ilustrasi pentingnya data preprocessing
- Menjelaskan tentang data cleaning
- Menjelaskan tentang data transformation



Menjelaskan data preprocessing

Data dan Variabel

- Data dapat berupa hasil pengukuran (numerik) atau penghitungan (kategori)
- Variabel berfungsi sebagai placeholder untuk data



Macam-macam Variabel

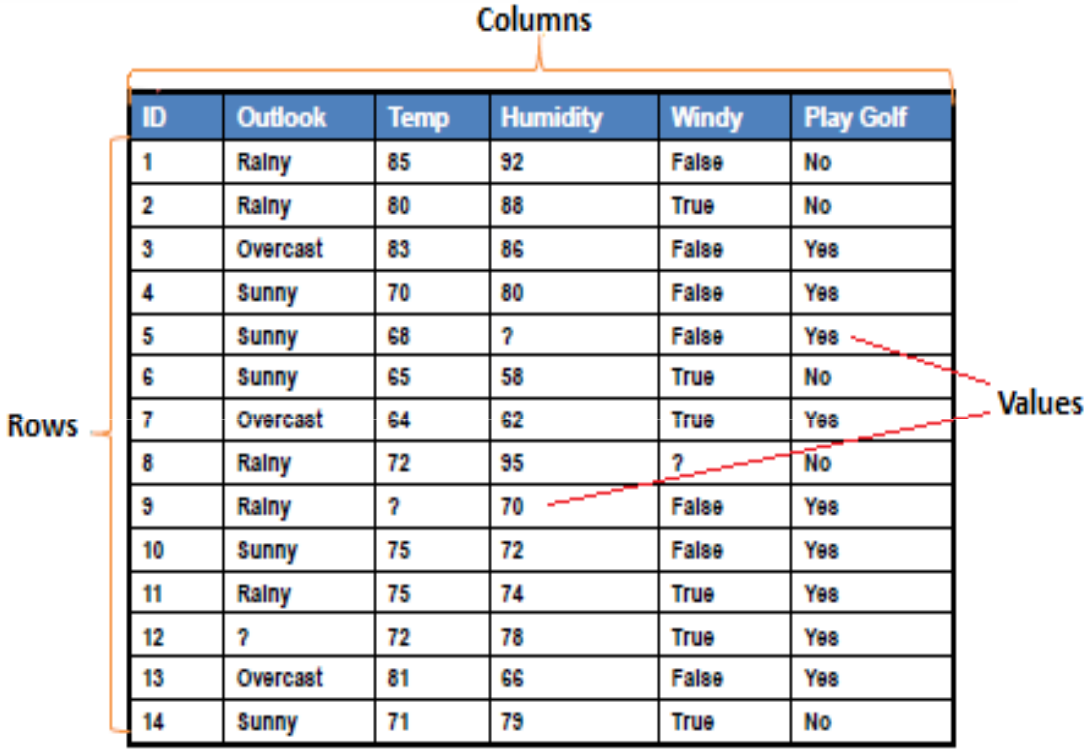
Numerik

- Interval adalah variabel yang nilainya bisa diurutkan dan diukur dengan tetap dan nilai yang sama. Nilai nol tidak didefinisikan secara mutlak. Contohnya temperatur yang diukur dalam derajat Fahrenheit
- Rasio adalah variabel yang mempunyai nilai nol yang mutlak. Semua operasi matematika dapat dilakukan pada variabel ini. Contoh jarak yang diukur dalam centimeter

Kategori

- Nominal adalah variabel yang nilainya berupa simbol tetapi tidak dapat diurutkan atau diukur jaraknya. Contoh: jenis kelamin: pria dan wanita
- Ordinar adalah variabel yang nilainya berupa simbol tetapi dapat diurutkan atau diukur jaraknya. Contoh: jarak (dekat, sedang, jauh)

Data Set



ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	85	92	False	No
2	Rainy	80	88	True	No
3	Overcast	83	86	False	Yes
4	Sunny	70	80	False	Yes
5	Sunny	68	?	False	Yes
6	Sunny	65	58	True	No
7	Overcast	64	62	True	Yes
8	Rainy	72	95	?	No
9	Rainy	?	70	False	Yes
10	Sunny	75	72	False	Yes
11	Rainy	75	74	True	Yes
12	?	72	78	True	Yes
13	Overcast	81	66	False	Yes
14	Sunny	71	79	True	No

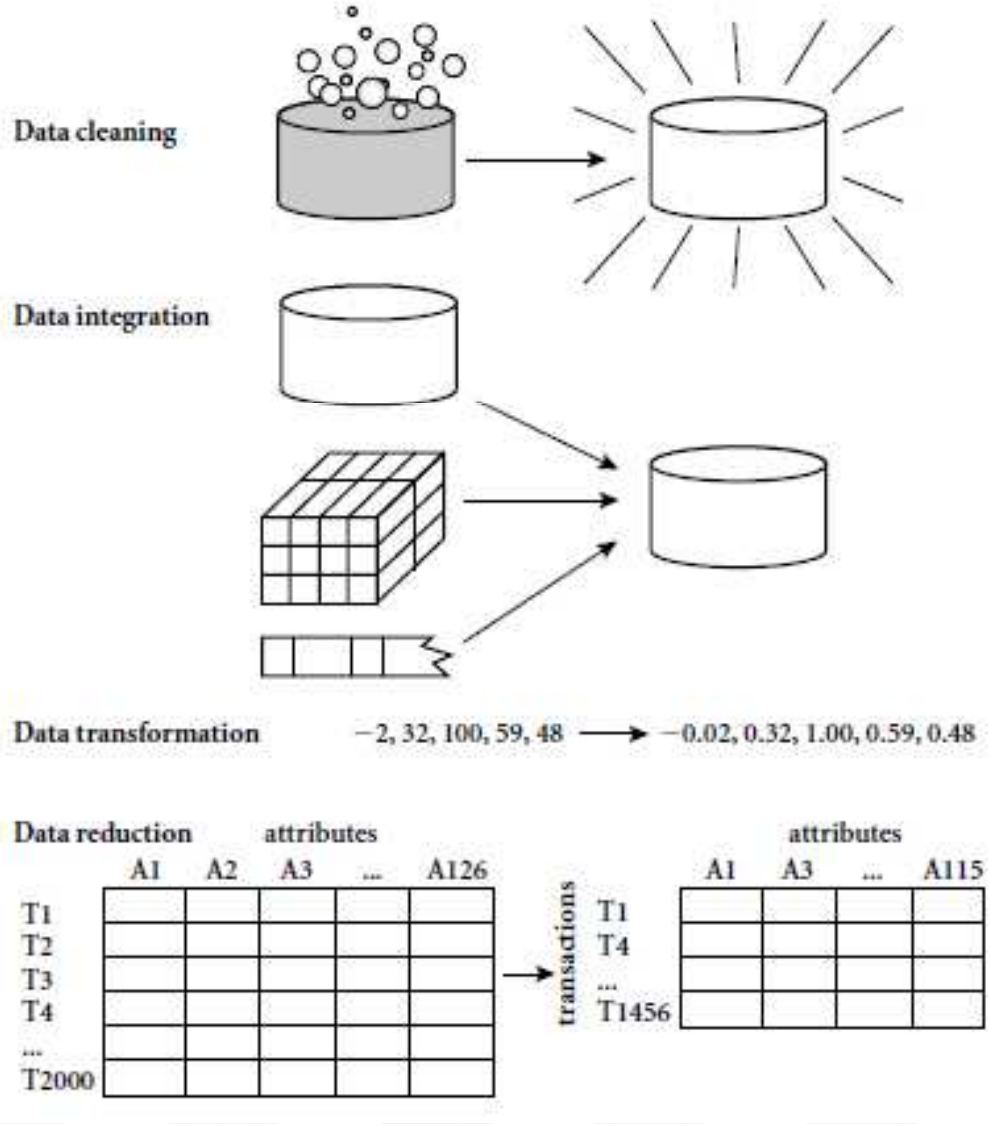
- Dataset adalah kumpulan data yang biasanya disajikan dalam bentuk tabel
- Columns, Fields, Attr ibutes, Variables, Fit urs
- Rows, Records, Obje cts, Cases, Instances, Examples, Vectors
- Values, Data

Perlunya Data Preprocessing

- Raw data banyak yang tidak lengkap, terdapat missing values, redudansi
- Untuk meminimalisasi data sampah
- Untuk mempercepat proses mining data

Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction



Data Cleaning

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.0	8	350.0	165.0
2	31.9	4	89.0	71.0
3	517.0	8	302.0	140.0
4	15.0		400.0	150.0
5	30.5			
6	23.0		350.0	125.0
7	13.0		351.0	158.0
8	14.0	8		215.0
9	25.4	5		77.0
10	37.7	4	89.0	62.0

- Missing Values
  - Menghapus record
  - Mengisi data yang kosong secara manual
  - Menggunakan konstanta global
  - Menggunakan atribut rata-rata

Data Transformation

- Agregation
  - Mengkombinasi 2 atau lebih objek ke dalam objek tunggal
  - Agregasi yang dapat dilakukan: sum (jumlah), average (rata-rata), min (terkecil), max (terbesar)

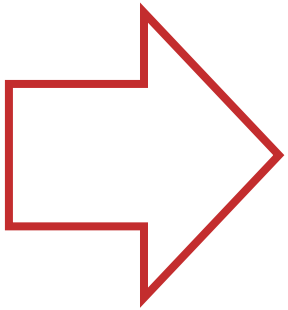
Cabang	IDT	Tanggal	Total
Gresik	102	18-09-2018	250.000
Gresik	103	18-09-2018	300.000
Surabaya	201	18-09-2018	500.000
Surabaya	202	18-09-2018	450.000
Surabaya	203	18-09-2018	350.000

Cabang	Tanggal	Total
Gresik	18-09-2018	550.000
Surabaya	18-09-2018	1.300.000

Data Transformation

- Binarization
  - Mengubah data dari tipe kategori ke atribut biner

Nilai Kategori
Rusak
Jelek
Sedang
Bagus
Sempurna



x1	x2	x3	x4	X5
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

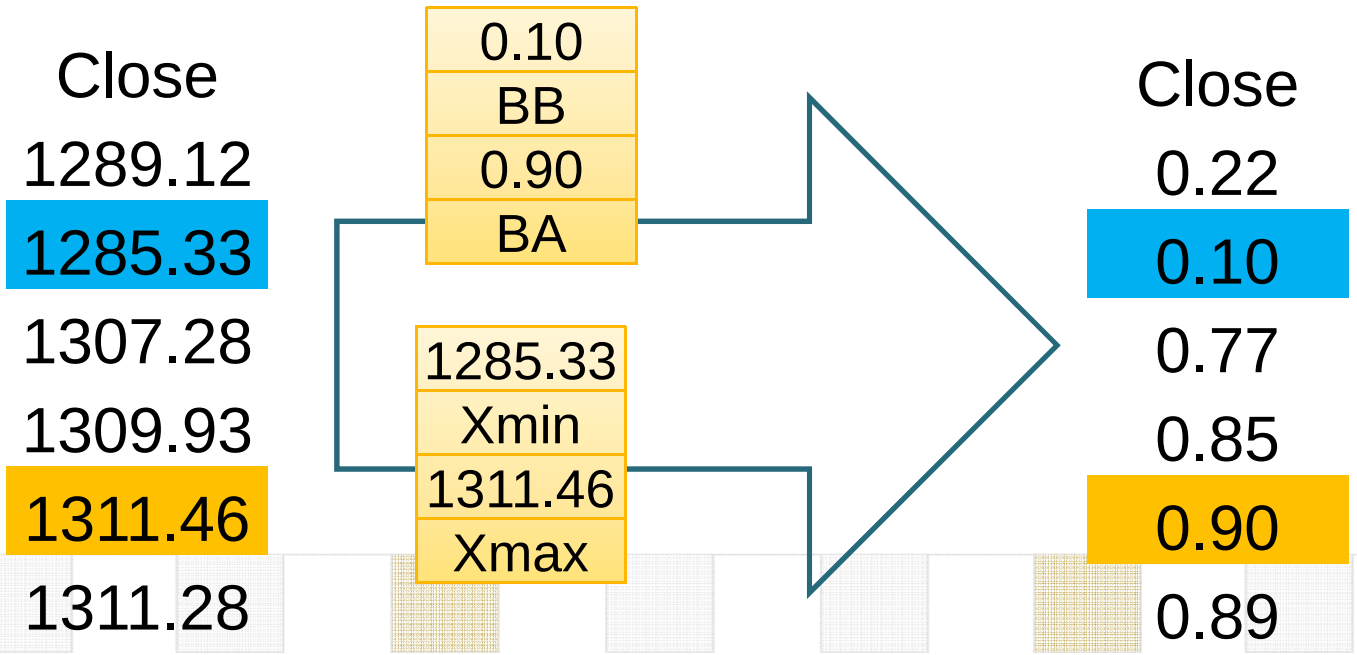
- Discretization
  - Mengubah data dari tipe numerik ke atribut kategori

Data Transformation

- Min-Max Normalization

$$X' = \frac{X - Xmin}{Xmax - Xmin} * (BA - BB) + BB$$

  - BA = Batas Atas
  - BB = Batas Bawah
  - Xmax = nilai maksimum
  - Xmin = nilai minimum



Terima Kasih