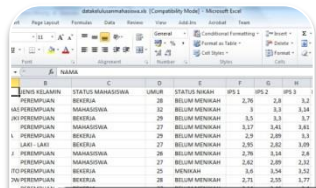


2. Data & Proseses Datamining

Data

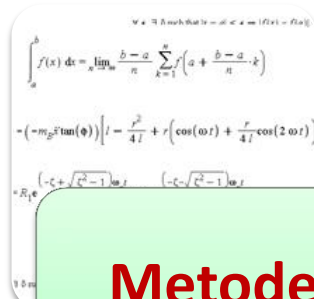
1. Input (Dataset)
2. Pengolahan Data Awal
3. Metode Learning

Tahapan Utama Proses Data Mining

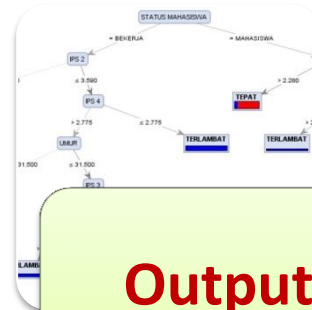


	A	B	C	D	E	F	G	H
1	STATUS MAHASISWA	STATUS MAHASISWA	UMUR	STATUS MAHASISWA	IPS 1	IPS 2	IPS 3	1
2	MAHASISWA	MAHASISWA	20	BEKUM MAHASISWA	2.75	2.8	3.2	1
3	MAHASISWA	MAHASISWA	20	BEKUM MAHASISWA	3	3.3	3.4	1
4	MAHASISWA	MAHASISWA	20	BEKUM MAHASISWA	3.3	3.3	3.7	1
5	MAHASISWA	MAHASISWA	27	BEKUM MAHASISWA	3.17	3.41	3.81	1
6	MAHASISWA	MAHASISWA	29	BEKUM MAHASISWA	3.3	3.09	3.1	1
7	MAHASISWA	MAHASISWA	27	BEKUM MAHASISWA	3.95	3.82	3.89	1
8	MAHASISWA	MAHASISWA	26	BEKUM MAHASISWA	2.76	3.14	2.8	1
9	MAHASISWA	MAHASISWA	27	BEKUM MAHASISWA	2.82	2.89	3.32	1
10	MAHASISWA	MAHASISWA	25	BEKUM MAHASISWA	3.6	3.54	3.52	1
11	MAHASISWA	MAHASISWA	28	BEKUM MAHASISWA	3.75	3.20	3.77	1

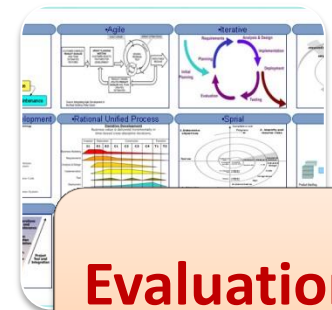
Input
(Data)


$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \sum_{k=1}^n f\left(a + \frac{b-a}{n}k\right)$$
$$r^2 \sin^2(\theta) = r^2 \left(\frac{1 - \cos(2\theta)}{2} \right) = \frac{r^2}{2} (1 - \cos(2\theta))$$
$$= \frac{r^2}{2} \int_0^{2\pi} (1 - \cos(2\theta)) d\theta = \frac{r^2}{2} \left[\theta - \frac{\sin(2\theta)}{2} \right]_0^{2\pi} = \frac{r^2}{2} (2\pi - 0) = \pi r^2$$

Metode
(Algoritma
Data Mining)



Output
(Pola/Model/
Knowledge)



Evaluation
(Akurasi, AUC,
RMSE, etc)

1. Input (Dataset)

- Dataset: (Data
Record/point/vector/pattern/event/case)
Kumpulan obyek data berserta atributnya.
- Atribut: (field/karakteristik, fitur)
Sifat/property/karakteristik obyek data.

Atribut, Class dan Tipe Data

- Atribut(variabel) adalah **faktor atau parameter yang menyebabkan** class/label/target terjadi
- Class adalah atribut yang akan dijadikan **target**, sering juga disebut dengan **label(AtributTarget)**
- Tipe data untuk variabel pada statistik terbagi menjadi empat: nominal, ordinal, interval, ratio
- Tapi secara praktis, tipe data untuk atribut pada data mining hanya menggunakan dua:
 - 1. Nominal** (Kategorikal)
 - 2. Numeric** (Kontinyu)

Tipe Atribut

Kualitatif/Kategoris

Kuantitatif/Numerik

Nominal

[distinctness =, #]

Misl: NIM, KodePos, JenisKelamin

Ordinal

[Order < . <=, >, >=]

Misl:

tk kelulusan: [cumlaude, sangat
memuaskan, memuaskan]

Suhu: [dingin, normal, panas]

Interval

Misl: Tanggal, Suhu

Rasio

Misl: Umur, panjang, tinggi

Jenis Dataset

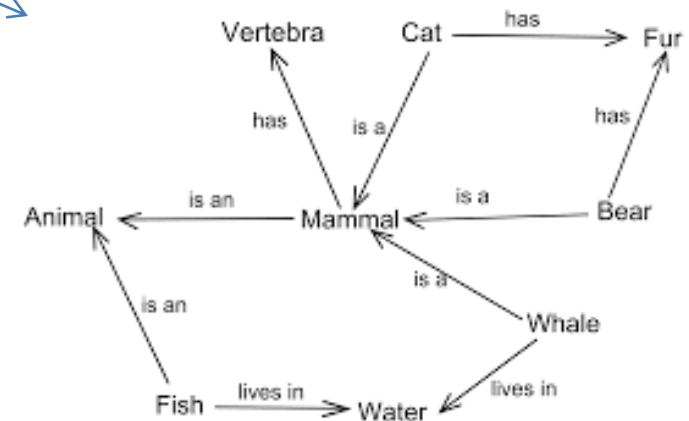
- Jenis dataset ada dua: **Private** dan **Public**
- **Private Dataset**: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
 - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- **Public Dataset**: data set dapat diambil dari repositori publik yang disepakati oleh para peneliti data mining
 - **UCI Repository** (<http://www.ics.uci.edu/~mlearn/MLRepository.html>)
 - **ACM KDD Cup** (<http://www.sigkdd.org/kddcup/>)
- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: **comparable**, **repeatable** dan **verifiable**

Tipe Dataset

- Data Record
 - Data Matrix
 - Data Transaksi
 - Data Graph
- Data Terurut

	1	2	3	4	5	6
1	0	1	1	0	0	1
2	1	0	1	1	0	1
3	1	1	0	1	1	0
4	0	1	1	0	1	0
5	0	0	1	1	0	1
6	1	1	0	0	1	0

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke



Kualitas Data

- Kesalahan Pengukuran: Nilai yg dicatat berbeda dg nilai sebenarnya (noise,bias,precission,acuracy)
- Kesalahan Pengumpulan: spt hilangnya obyek data/nilai dr atribut/lingkup obyek data yg tdk tetap
- Duplicate Data: obyek data ganda

Kesalahan Pengumpulan

- Outliers: obyek data yg memiliki sifat yg berbeda sekali dari kebanyakan obyek data.
- Missing Value: nilai pd suatu atribut yg tdk ditemukan/kosong.
 - Bisa krn responden menolak memberikan informasi
 - Atribut tdk bisa diterapkan ke semua kasus
 - Diatasi dg mengurangi obyek data, memperkirakan missing value, mengganti dg nilai yg memungkinkan

Dataset with Attribute and Class

Attribute

Class/Label

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>
104	6.3	2.9	5.6	1.8	<i>Iris virginica</i>
105	6.5	3.0	5.8	2.2	<i>Iris virginica</i>
...					

Estimasi Waktu Pengiriman Pizza

Customer	Jumlah Pesanan (P)	Jumlah Lampu Merah (L)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

$$\text{Waktu Tempuh (T)} = 0.48P + 0.23L + 0.5J$$

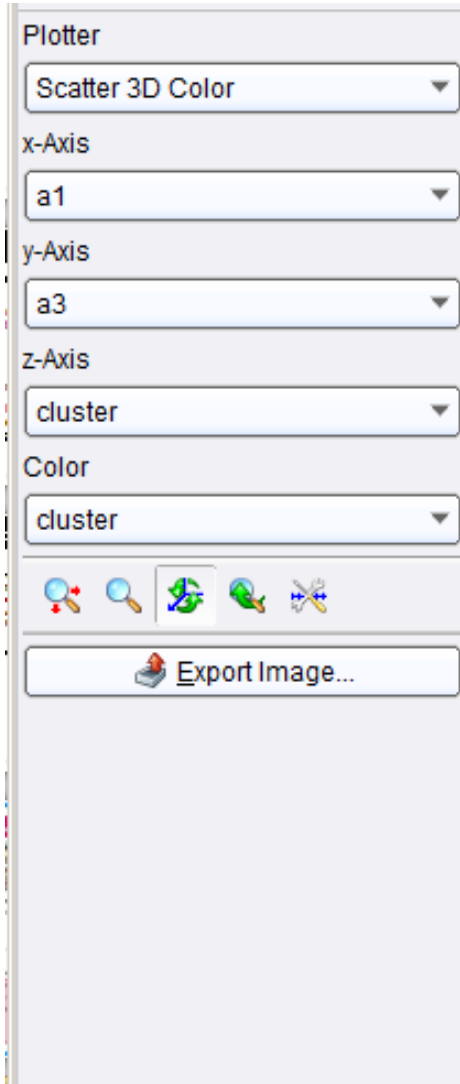
Penentuan Kelulusan Mahasiswa

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

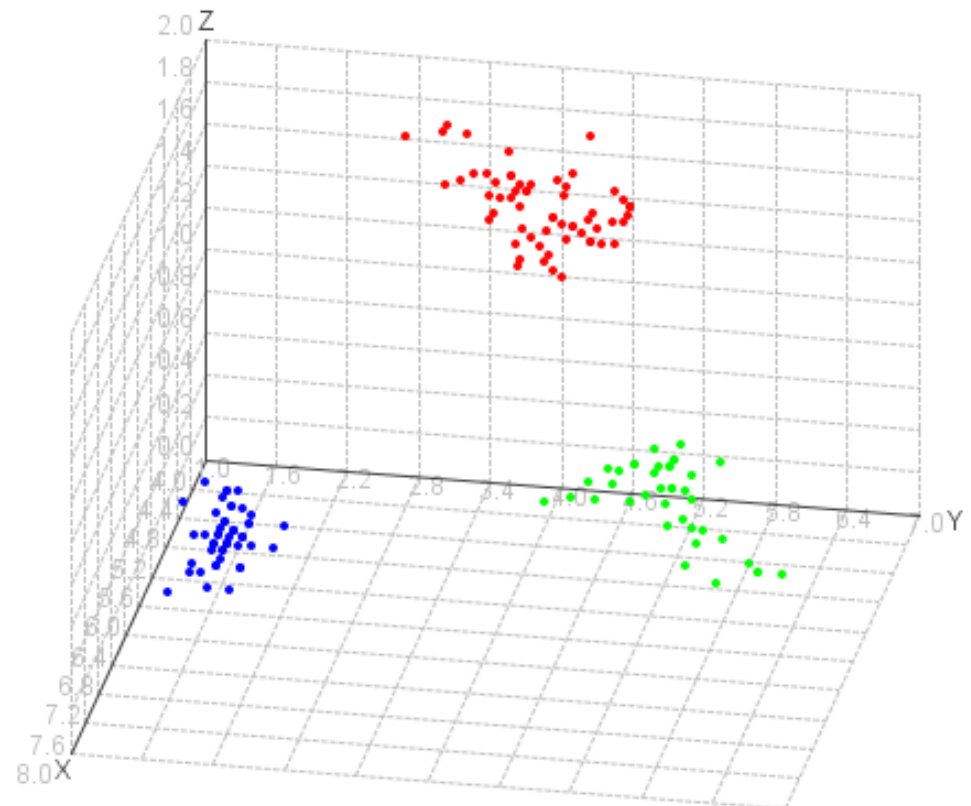
Klastering Bunga Iris

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)						
Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.200
24	id_24	Iris-setosa	5.100	3.300	1.700	0.500

Klastering Bunga Iris



cluster ● cluster_0 ● cluster_1 ● cluster_2



2. Pemrosesan Awal Data

- Agregasi
- Sampling
- Binerisasi dan Diskretisasi
- Pengurangan Dimensi
- Pemilihan Fitur
- Transformasi Variabel

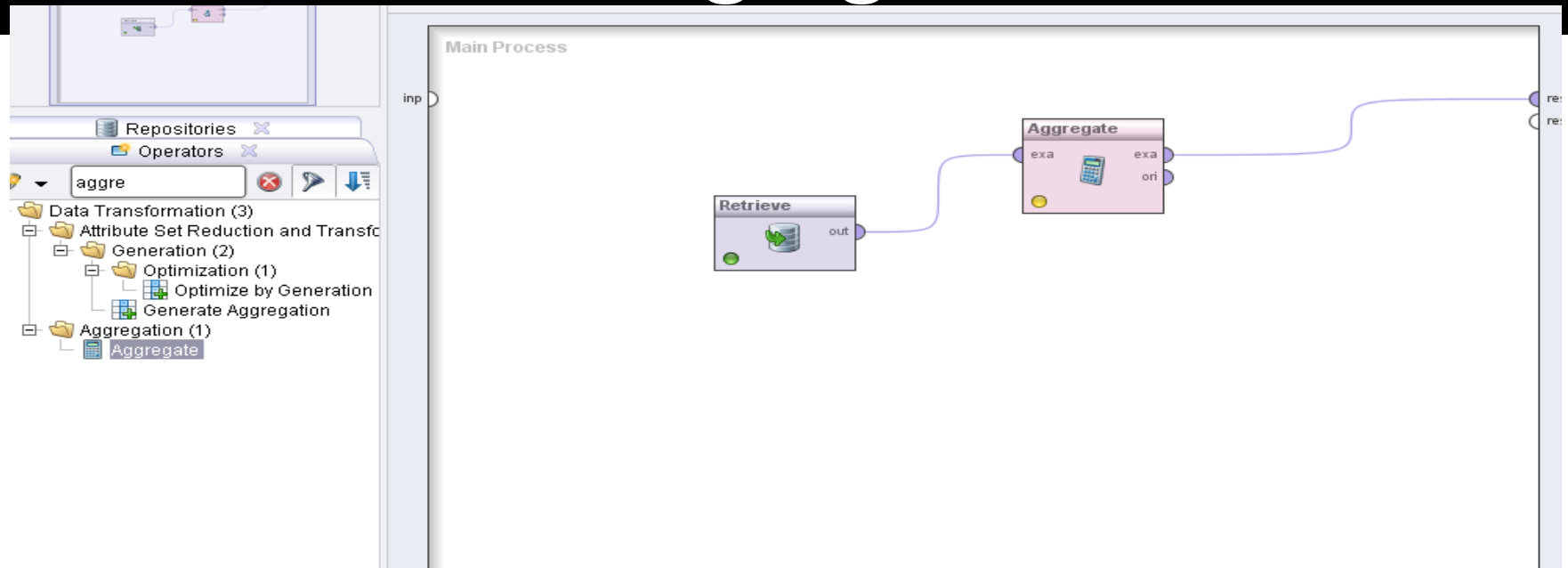
Agregasi

- Penggabungan obyek ke dalam sebuah obyek tunggal
- Sum,average,min,max

Cabang	IDTX	Tanggal	Total
Gresik	2012102	30-01-2013	250,000
Gresik	2012103	30-01-2013	300,000
Surabaya	2012201	30-01-2013	500,000
Surabaya	2012202	30-01-2013	450,000
Surabaya	2012203	31-01-2013	350,000

Cabang	Tanggal	Total
Gresik	30-01-2013	550000
Surabaya	30-01-2013	950000
Surabaya	31-01-2013	350000

Agregasi



Row No. ▲	Cabang	IDT	Tgl	D
1	Semarang	t001	17-03-2016	10000000
2	Kendal	t002	17-03-2016	8000000
3	Kendal	t003	17-03-2016	7000000
4	Semarang	t004	17-03-2016	10000000
5	Semarang	t005	18-03-2016	15000000
6	Semarang	t006	18-03-2016	25000000
7	Semarang	t007	18-03-2016	10000000
8	Kendal	t008	18-03-2016	8000000
9	Kendal	t009	19-03-2016	7000000
10	Semarang	t010	19-03-2016	10000000

Example of example, a special example, a regular example

Row No. ▲	Cabang	Tgl	average(D)
1	Semarang	17-03-2016	10000000
2	Semarang	18-03-2016	16666666.667
3	Semarang	19-03-2016	10000000
4	Kendal	17-03-2016	7500000
5	Kendal	18-03-2016	8000000
6	Kendal	19-03-2016	7000000

Sampling

- Pemilihan bagian obyek data yang akan dianalisis.
- Sample harus representatif (mewakili seluruh data)
- Sample disebut resprentatif jika mempunyai sifat yang sama dengan seluruh data biasa diukur dengan rata-rata/mean
- Penggunaan sample yang baik tidak menjamin bahwa hasil pemrosesan datamining pada sample sama bagusnya dengan pemrosesan pada seluruh data asli

Sampling

- Pendekatan sampling
 - Simple random sampling
 - Tanpa pengembalian
 - Dengan pengembalian

Sampling

ProsesData1* - RapidMiner@yogi-PC

File Edit Process Tools View Help

Overview Process XML Parameters

Main Process

Retrieve Sample

Repositories Operators

sample

Data Transformation (5)

Filtering (5)

Sampling (5)

Sample (Stratified)

Meta Data View Data View Plot View Annotations

ExampleSet (100 examples, 2 special attributes, 4 regular attributes)

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_5	Iris-setosa	5	3.600	1.400	0.200
4	id_6	Iris-setosa	5.400	3.900	1.700	0.400
5	id_7	Iris-setosa	4.600	3.400	1.400	0.300
6	id_8	Iris-setosa	5	3.400	1.500	0.200
7	id_10	Iris-setosa	4.900	3.100	1.500	0.100
8	id_13	Iris-setosa	4.800	3	1.400	0.100
9	id_14	Iris-setosa	4.300	3	1.100	0.100
10	id_16	Iris-setosa	5.700	4.400	1.500	0.400
11	id_17	Iris-setosa	5.400	3.900	1.300	0.400
12	id_18	Iris-setosa	5.100	3.500	1.400	0.300
13	id_20	Iris-setosa	5.100	3.800	1.500	0.300
14	id_21	Iris-setosa	5.400	3.400	1.700	0.200
15	id_26	Iris-setosa	5	3	1.600	0.200
16	id_28	Iris-setosa	5.200	3.500	1.500	0.200
17	id_29	Iris-setosa	5.200	3.400	1.400	0.200
18	id_30	Iris-setosa	4.700	3.200	1.600	0.200
19	id_32	Iris-setosa	5.400	3.400	1.500	0.400
20	id_33	Iris-setosa	5.200	4.100	1.500	0.100
21	id_34	Iris-setosa	5.500	4.200	1.400	0.200
22	id_35	Iris-setosa	4.900	3.100	1.500	0.100
23	id_39	Iris-setosa	4.400	3	1.300	0.200
24	id_40	Iris-setosa	5.100	3.400	1.500	0.200

Sample

sample absolute

sample size 100

Meta Data View Data View Plot View Annotations

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.200
24	id_24	Iris-setosa	5.100	3.300	1.700	0.500

Binerisasi

- Transformasi data dari tipe continue, diskret menjadi tipe biner.
- Algoritma asosiasi membutuhkan data dengan atribut bertipe biner
- Jumlah atribut yg dibutuhkan utk binerisasi adalah $N = \log_2(M)$, M = jml kelas kategori
- Contoh: {rusak, jelek, sedang, bagus, sempurna}, $M=5$
- $N = \log_2(5) = 3$, sehingga tdp 3 atribut x_1, x_2, x_3

Class	Nilai integer	x_1	x_2	x_3
Rusak	0	0	0	0
Jelek	1	0	0	1
Sedang	2	0	1	0
Bagus	3	0	1	1
Sempurna	4	1	0	0

Contoh Binerisasi

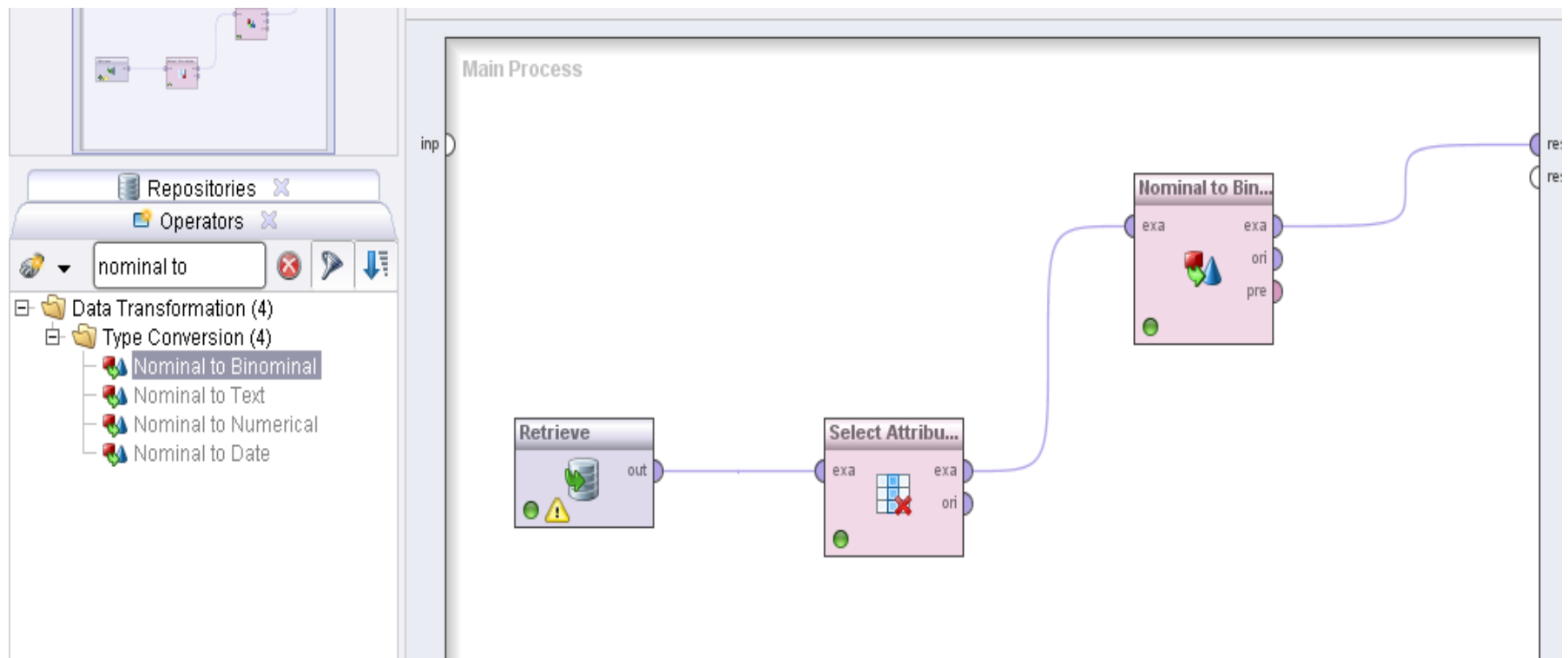
ExampleSet (30 examples, 0 special attributes, 7 regular attributes)

Row No.	IDTX	TGL	IDPASIE	IDOBAT	OBAT	JML	HARGA
1	20150702-1	Jul 2, 2015 7	40402156	011M	CLEANSER-	1	20000
2	20150702-1	Jul 2, 2015 7	32200265	022J	SABUN WAJ	1	40000
3	20150702-1	Jul 2, 2015 7	11800586	011M	CLEANSER-	1	20000
4	20150702-1	Jul 2, 2015 7	30401779	22	SABUN WAJ	1	35000
5	20150702-1	Jul 2, 2015 7	41301650	022J	SABUN WAJ	1	40000
6	20150702-1	Jul 2, 2015 7	30800645	22	SABUN WAJ	1	35000
7	20150702-1	Jul 2, 2015 7	40402156	022J	SABUN WAJ	1	40000
8	20150702-1	Jul 2, 2015 7	31200692	22	SABUN WAJ	1	35000
9	20150702-1	Jul 2, 2015 7	32200316	011M	CLEANSER-	1	20000
10	20150701-1	Jul 1, 2015 7	52501079	22	SABUN WAJ	1	35000
11	20150701-1	Jul 1, 2015 7	30200352	22	SABUN WAJ	1	35000
12	20150701-1	Jul 1, 2015 7	31401342	022J	SABUN WAJ	1	40000
13	20150701-1	Jul 1, 2015 7	11901068	022J	SABUN WAJ	1	40000
14	20150701-1	Jul 1, 2015 7	41801994	022J	SABUN WAJ	1	40000
15	20150701-1	Jul 1, 2015 7	30102734	22	SABUN WAJ	1	35000
16	20150701-1	Jul 1, 2015 7	11200061	022J	SABUN WAJ	1	40000
17	20150701-1	Jul 1, 2015 7	10100392	11	CLEANSER-	1	15000
18	20150701-1	Jul 1, 2015 7	41902888	22	SABUN WAJ	1	35000
19	20150701-1	Jul 1, 2015 7	31401210	022J	SABUN WAJ	1	40000
20	20150701-1	Jul 1, 2015 7	31401244	22	SABUN WAJ	4	35000
21	20150701-1	Jul 1, 2015 7	30102976	022J	SABUN WAJ	1	40000
22	20150701-1	Jul 1, 2015 7	30901589	022J	SABUN WAJ	1	40000
23	20150701-1	Jul 1, 2015 7	30600605	11	CLEANSER-	1	15000
24	20150701-1	Jul 1, 2015 7	10400824	022J	SABUN WAJ	2	40000

Contoh Binerisasi

ExampleSet (30 examples, 0 special attributes, 2 regular attributes)			ExampleSet (30 examples, 0 special attributes, 5 regular attributes)					
Row No.	IDTX	OBAT	Row No.	OBAT = CLEANSER-1M	OBAT = SABUN WAJAH-1M	OBAT = SABUN WAJAH-2	OBAT = CLEANSER-1	IDTX
1	20150702-1	CLEANSER-1M	1	true	false	false	false	20150702-1
2	20150702-1	SABUN WAJAH-2J	2	false	true	false	false	20150702-1
3	20150702-1	CLEANSER-1M	3	true	false	false	false	20150702-1
4	20150702-1	SABUN WAJAH-2	4	false	false	true	false	20150702-1
5	20150702-1	SABUN WAJAH-2J	5	false	true	false	false	20150702-1
6	20150702-1	SABUN WAJAH-2	6	false	false	true	false	20150702-1
7	20150702-1	SABUN WAJAH-2J	7	false	true	false	false	20150702-1
8	20150702-1	SABUN WAJAH-2	8	false	false	true	false	20150702-1
9	20150702-1	CLEANSER-1M	9	true	false	false	false	20150702-1
10	20150701-1	SABUN WAJAH-2	10	false	false	true	false	20150701-1
11	20150701-1	SABUN WAJAH-2	11	false	false	true	false	20150701-1
12	20150701-1	SABUN WAJAH-2J	12	false	true	false	false	20150701-1
13	20150701-1	SABUN WAJAH-2J	13	false	true	false	false	20150701-1
14	20150701-1	SABUN WAJAH-2J	14	false	true	false	false	20150701-1
15	20150701-1	SABUN WAJAH-2	15	false	false	true	false	20150701-1
16	20150701-1	SABUN WAJAH-2J	16	false	true	false	false	20150701-1
17	20150701-1	SABUN WAJAH-2J	17	false	false	false	true	20150701-1
18	20150701-1	SABUN WAJAH-2J	18	false	false	true	false	20150701-1
19	20150701-1	SABUN WAJAH-2	19	false	true	false	false	20150701-1
20	20150701-1	SABUN WAJAH-2	20	false	false	true	false	20150701-1
21	20150701-1	SABUN WAJAH-2J	21	false	true	false	false	20150701-1
22	20150701-1	SABUN WAJAH-2J	22	false	true	false	false	20150701-1
23	20150701-1	CLEANSER-1	23	false	false	false	true	20150701-1
24	20150701-1	SABUN WAJAH-2J	24	false	true	false	false	20150701-1

Binerisasi



Diskretisasi

Transformasi data dari tipe kontinu ke diskrit.

ID	Pajak
1	125
2	100
3	70
4	120
5	95
6	60
7	220
8	85
9	75
10	90

Kategori	range
Rendah	60 – 113
Sedang	114 – 167
Tinggi	168 - 220

ID	Pajak
1	Sedang
2	Rendah
3	Rendah
4	Sedang
5	Rendah
6	Rendah
7	Tinggi
8	Rendah
9	Rendah
10	Rendah

Contoh Diskretisasi

☐ Meta Data View
 ☒ Data View
 ☐ Plot View
 ☐ Annotations

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

☐ Meta Data View
 ☒ Data View
 ☐ Plot View
 ☐ Annotations

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Humidity	Outlook	Temperature	Wind
1	no	range3 [79 - 87.500]	sunny	85	false
2	no	range4 [87.500 - ∞]	sunny	80	true
3	yes	range2 [67.500 - 79]	overcast	83	false
4	yes	range4 [87.500 - ∞]	rain	70	false
5	yes	range3 [79 - 87.500]	rain	68	false
6	no	range2 [67.500 - 79]	rain	65	true
7	yes	range1 [-∞ - 67.500]	overcast	64	true
8	no	range4 [87.500 - ∞]	sunny	72	false
9	yes	range2 [67.500 - 79]	sunny	69	false
10	yes	range3 [79 - 87.500]	rain	75	false
11	yes	range2 [67.500 - 79]	sunny	75	true
12	yes	range4 [87.500 - ∞]	overcast	72	true
13	yes	range2 [67.500 - 79]	overcast	81	false
14	no	range3 [79 - 87.500]	rain	71	true

Diskretisasi

The screenshot displays the Orange3 data mining software interface. On the left, a sidebar lists various operators under the 'Data Transformation' category, with 'Discretize by Size' selected. The main workspace shows a workflow consisting of two connected operators: 'Retrieve' and 'Discretize'. The 'Retrieve' operator has an 'out' port connected to the 'exa' (example) input port of the 'Discretize' operator. The 'Discretize' operator has three output ports labeled 'exa', 'ori', and 'pre'. The 'exa' output port is connected to a 'res' (result) port on the right side of the interface. On the right, a configuration panel for the 'Discretize by Size' operator is visible. It includes settings for 'attribute filter type' (set to 'single'), 'attribute' (set to 'Humidity'), 'invert selection' (unchecked), 'include special attributes' (unchecked), 'size of bins' (set to 4), 'sorting direction' (set to 'decreasing'), and 'range name type' (set to 'long'). At the bottom right, there is a small preview window titled 'Discretize by Size' showing a data visualization.

Pengurangan Dimensi

- Mengurangi jumlah waktu dan memory yg dibutuhkan
- Membuat data lebih mudah divisualisasi
- Membantu mengurangi fitur-fitur yang tdk relevan/mengurangi gangguan/derau
- Teknik yang digunakan
 - Principal Component Analysis (PCA)
 - Singular Value Decomposition(SVD)

Pemilihan Fitur (Feature Subset Selection)

- Proses pencarian terhadap semua kemungkinan subset fitur.
 - Menghilangkan fitur yang redundan
misl: harga_jual,pajak,discount
 - Menghilangkan fitur-fitur yang tidak mengandung informasi yang berguna untuk pekerjaan datamining
Misl: tinggi badan mhs pada pekerjaan prediksi kelulusan mhs , tidak relevan

Pemilihan Fitur

- Teknik yang digunakan:

- Brute-force

- Pada proses data mining dilakukan dengan mencoba semua fitur.

- Filtering:

- Memilih fitur sebelum proses datamining dilakukan

- wrapper

- menggunakan algoritma datamining utk memilih sub-set fitur yang paling baik.

Pemilihan Fitur

Proses:

- Melakukan pengukuran untuk evaluasi suatu subset fitur.
- Menggunakan metode pencarian yang mengontrol pemilihan subset-fitur baru
- Menggunakan kriteria untuk melakukan penghentian proses.
- Menggunakan validasi

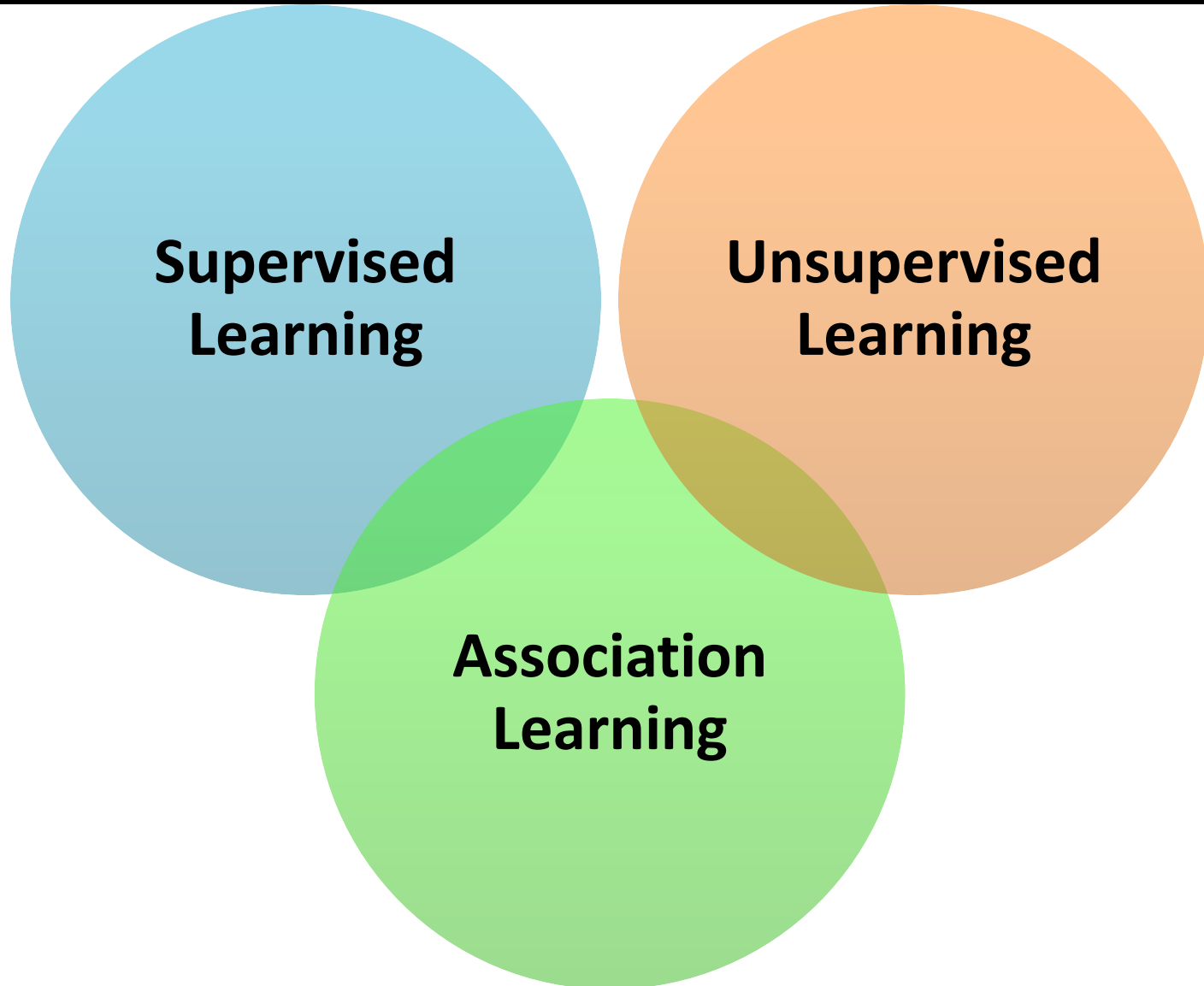
Pembuatan Fitur

- Proses membuat fitur baru yang dapat menangkap informasi penting dalam sebuah himpunan fitur yang lebih efisien daripada fitur-fitur yang ada.
- Metode Pembuatan Fitur:
 - Ekstraksi Fitur
 - Pemetaan menggunakan transformasi fourier/wavelet
 - Konstruksi fitur dengan menggabungkan fitur-fitur yang ada.

Transformasi Fitur

- Merupakan proses yang memetakan keseluruhan himpunan nilai dari fitur-fitur yang diberikan ke suatu subset nilai pengganti sedemikian sehingga nilai yang lama dapat dikenali dengan satu dari nilai-nilai yang baru tersebut.
- Metode dalam transformasi fitur:
 - Standarisasi (median , standar deviasi).
 - Normalization, dimana data sebuah atribut diskalakan ke dalam rentang (kecil) yang ditentukan. Metode: Min-max Normalization, z-score Normalization, Normalization by Decimal Scaling).

3. Metode Learning Pada Algoritma DM



Metode Learning Pada Algoritma DM

1. **Supervised** Learning (Pembelajaran dengan Guru):
 - Sebagian besar algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
 - Variabel yang menjadi target/label/class ditentukan
 - Algoritma melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variable prediktor

Dataset with Attribute and Class

Attribute

Class/Label

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>
104	6.3	2.9	5.6	1.8	<i>Iris virginica</i>
105	6.5	3.0	5.8	2.2	<i>Iris virginica</i>
...					


Metode Learning Pada Algoritma DM

2. **Unsupervised** Learning (Pembelajaran tanpa Guru):

- Algoritma data mining mencari pola dari **semua variable (atribut)**
- Variable (atribut) yang menjadi **target/label/class** tidak **ditentukan (tidak ada)**
- Algoritma **clustering** adalah algoritma unsupervised learning

Dataset with Attribute (No Class)

Attribute



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2
...				

Metode Learning Pada Algoritma DM

3. **Association** Learning (Pembelajaran untuk Asosiasi Atribut)

- Proses learning pada algoritma asosiasi (*association rule*) bertujuan untuk mencari atribut yang muncul bersamaan dalam satu transaksi
- Algoritma asosiasi biasanya untuk analisa transaksi belanja, dengan konsep utama adalah mencari “produk/item mana yang dibeli bersamaan”
- Pada pusat perbelanjaan banyak produk yang dijual, sehingga pencarian seluruh asosiasi produk memakan cost tinggi, karena sifatnya yang kombinatorial
- Algoritma *association rule* seperti a priori algorithm, dapat memecahkan masalah ini dengan efisien

Dataset Transaction

ExampleSet (3 examples, 0 special attributes, 6 regular attributes)

Row No.	CAR = true	APPARTEMENT = true	VILLA = true	POOR = true	AVERAGE = true	RICH = true
1	false	true	false	true	false	false
2	true	true	false	false	true	false
3	true	false	true	false	false	true

Association Rules

Association Rules

Association Rules

```
[VILLA = true] --> [CAR = true] (confidence: 1.000)
[RICH = true] --> [CAR = true] (confidence: 1.000)
[AVERAGE = true] --> [CAR = true] (confidence: 1.000)
[POOR = true] --> [APPARTEMENT = true] (confidence: 1.000)
[AVERAGE = true] --> [APPARTEMENT = true] (confidence: 1.000)
[VILLA = true] --> [RICH = true] (confidence: 1.000)
[RICH = true] --> [VILLA = true] (confidence: 1.000)
[CAR = true, APPARTEMENT = true] --> [AVERAGE = true] (confidence: 1.000)
[AVERAGE = true] --> [CAR = true, APPARTEMENT = true] (confidence: 1.000)
[CAR = true, AVERAGE = true] --> [APPARTEMENT = true] (confidence: 1.000)
[APPARTEMENT = true, AVERAGE = true] --> [CAR = true] (confidence: 1.000)
[VILLA = true] --> [CAR = true, RICH = true] (confidence: 1.000)
[CAR = true, VILLA = true] --> [RICH = true] (confidence: 1.000)
[RICH = true] --> [CAR = true, VILLA = true] (confidence: 1.000)
[CAR = true, RICH = true] --> [VILLA = true] (confidence: 1.000)
[VILLA = true, RICH = true] --> [CAR = true] (confidence: 1.000)
```

