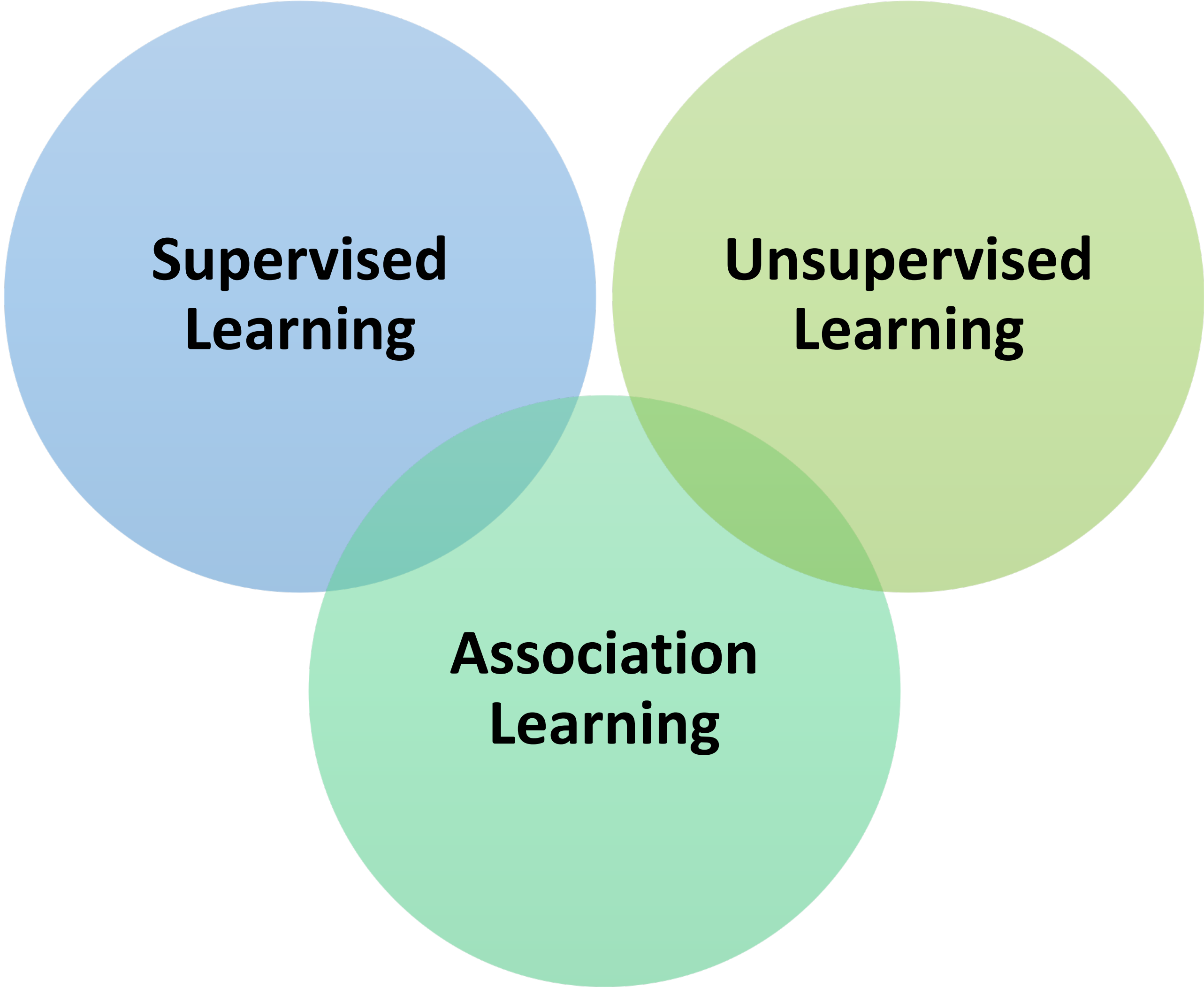
METODE LEARNING DATA MINING



SUPERVISED LEARNING

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

Supervised Learning : Algoritma Estimasi, Prediksi, Klasifikasi

UNSUPERVISED LEARNING

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

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
...				

Dataset with Attribute (No Class)

ASSOCATION LEARNING

Association Learning (Pembelajaran untuk Asosiasi Atribut)

- Proses learning pada algoritma asosiasi (*association rule*) agak berbeda karena tujuannya adalah 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 apriori algorithm, dapat memecahkan masalah ini dengan efisien

ASSOCATION LEARNING (2)

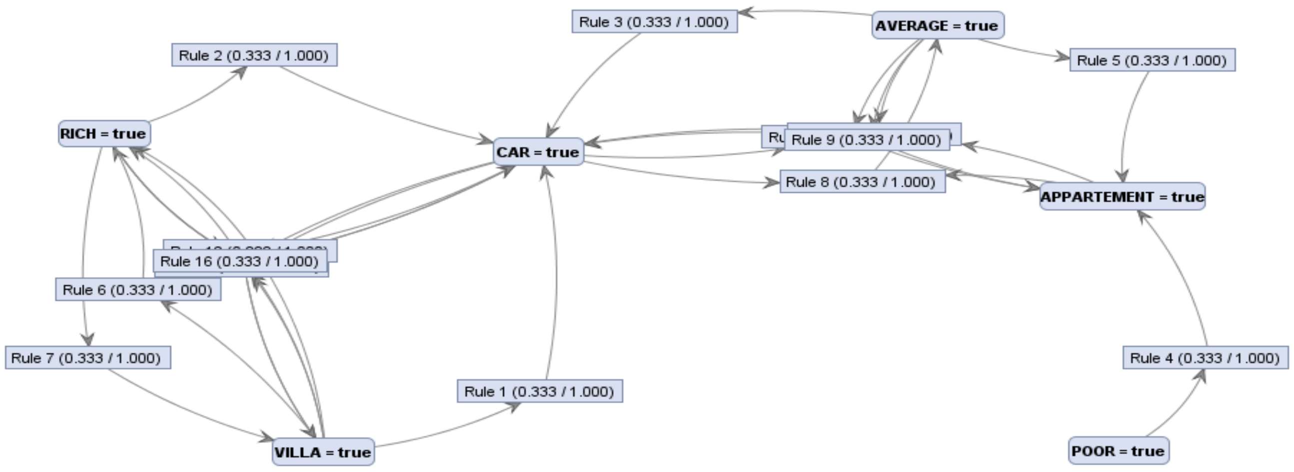
Dataset Transaction

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

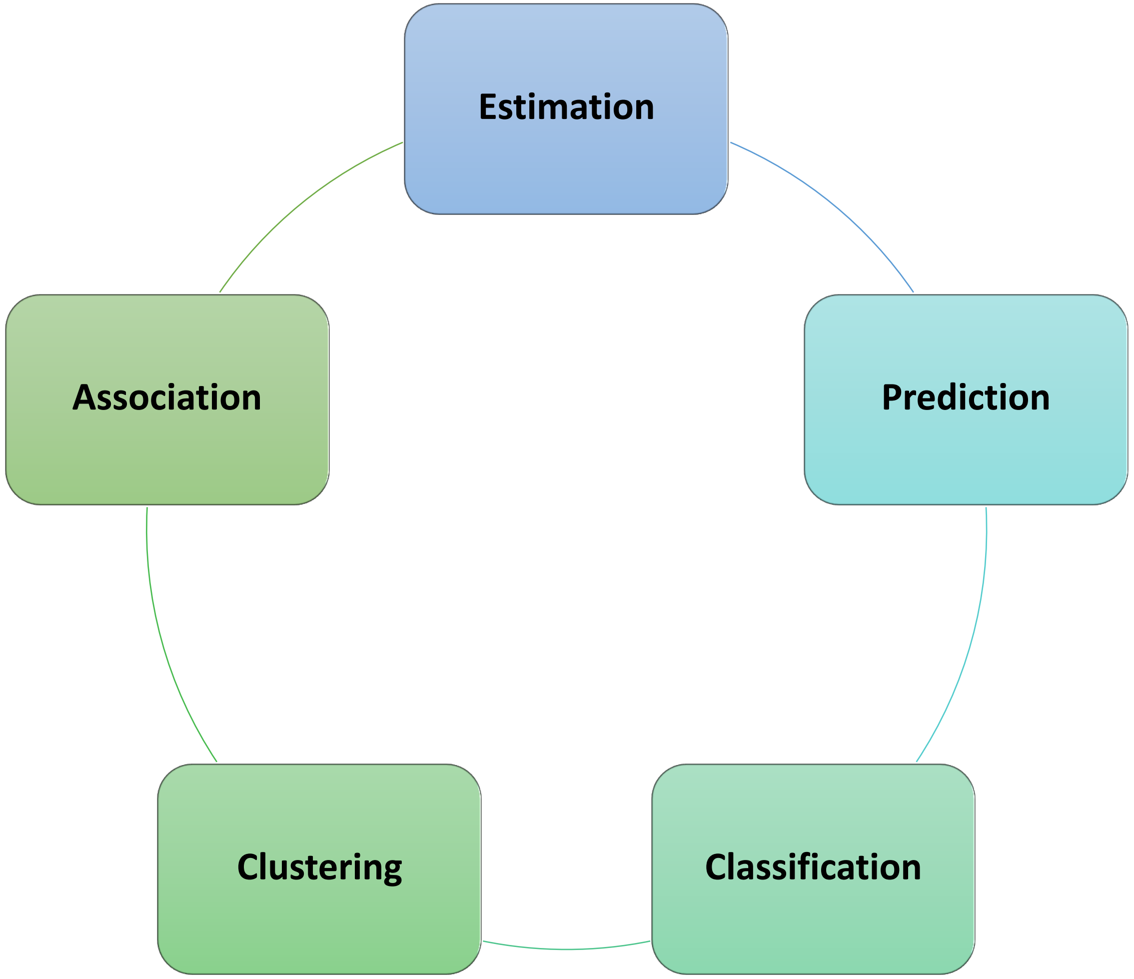
AssociationRules

```
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)
```



TEKNIK DATA MINING

1. Estimasi
2. Prediksi
3. Klasifikasi
4. Klastering
5. Asosiasi



ESTIMASI

- Algoritma estimasi mirip dengan algoritma klasifikasi, tapi variabel target adalah berupa bilangan numerik (kontinyu) dan bukan kategorikal (nominal atau diskrit)
- Estimasi nilai dari variable target ditentukan berdasarkan nilai dari variabel prediktor (atribut)
- Algoritma estimasi yang biasa digunakan adalah: Linear Regression, Neural Network, Support Vector Machine

CONTOH: ESTIMASI PERFORMANSI CPU

Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Linear regression function

$$PRP = -55.9 + 0.0489\ MYCT + 0.0153\ MMIN + 0.0056\ MMAX + 0.6410\ CACH - 0.2700\ CHMIN + 1.480\ CHMAX$$

PREDIKSI

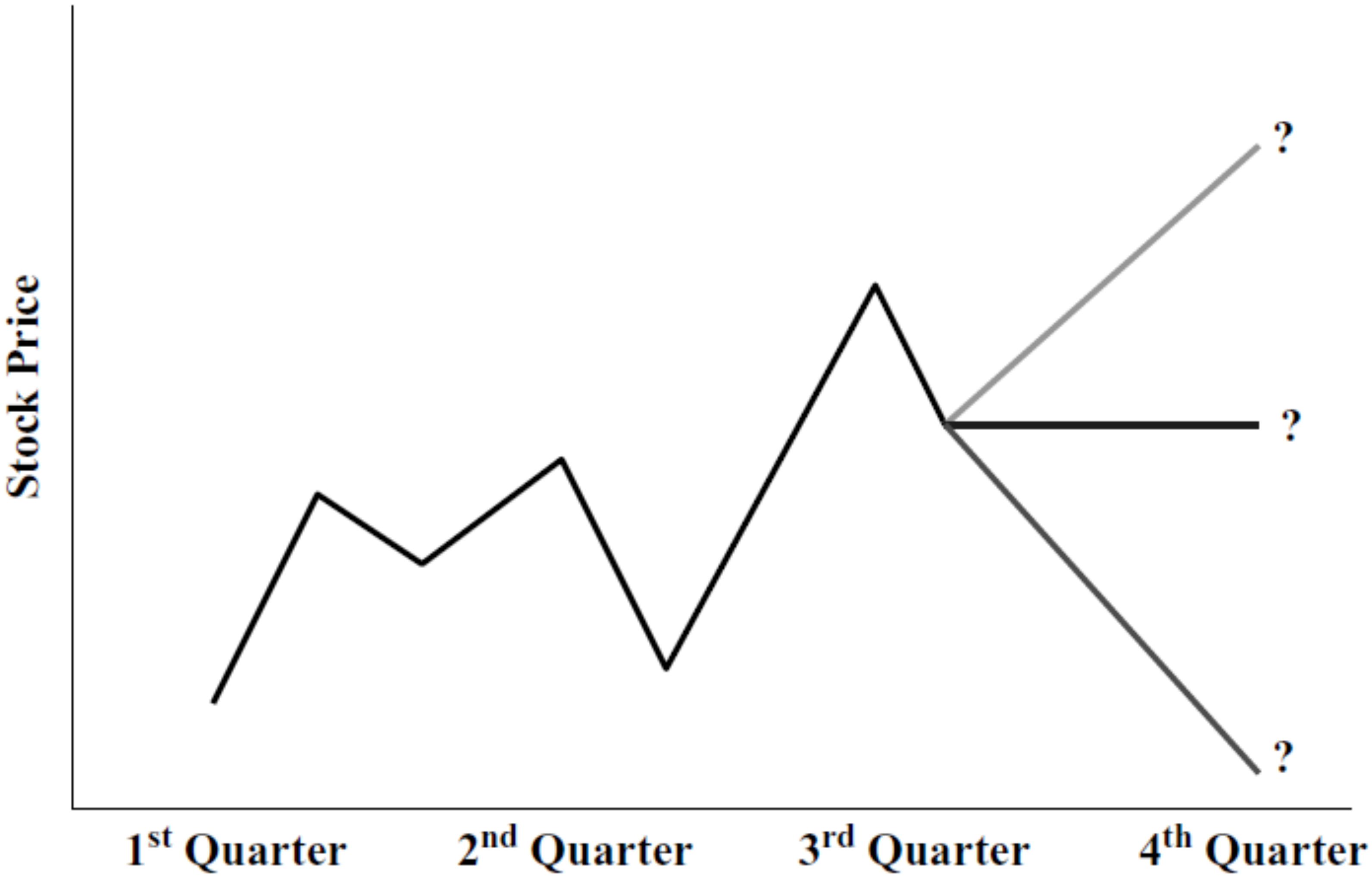
- Algoritma prediksi/forecasting sama dengan algoritma estimasi di mana label/target/class bertipe numerik, bedanya adalah data yang digunakan merupakan data rentet waktu (data time series)
- Istilah prediksi kadang digunakan juga untuk klasifikasi, tidak hanya untuk prediksi time series, karena sifatnya yang bisa menghasilkan class berdasarkan berbagai atribut yang kita sediakan
- Semua algoritma estimasi dapat digunakan untuk prediksi/forecasting

CONTOH: PREDIKSI HARGA SAHAM

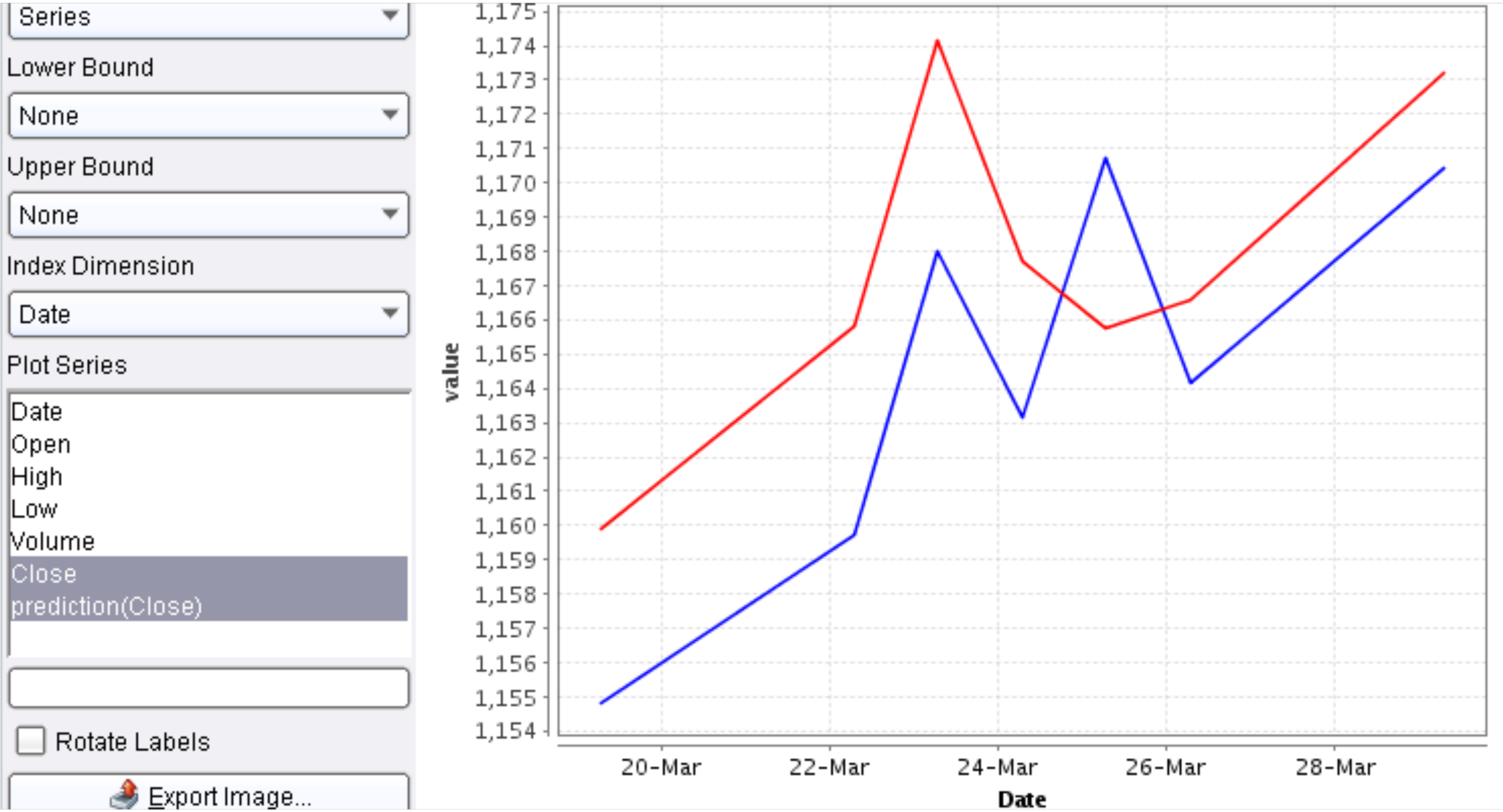
Dataset harga saham dalam bentuk time series (rentet waktu) harian

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	2232880000
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	1938100000
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	1891940000
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	1794650000
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	2595440000
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	2447310000
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	2512920000
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	2392630000
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	2117330000
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	2366380000
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	2502690000
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	2772010000
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	2419920000
14	1305.190	May 1, 2006	1310.610	1317.210	1303.460	2437040000
15	1313.210	May 2, 2006	1305.190	1313.660	1305.190	2403470000
16	1308.120	May 3, 2006	1313.210	1313.470	1303.920	2395230000
17	1312.250	May 4, 2006	1307.850	1315.140	1307.850	2431450000
18	1325.760	May 5, 2006	1312.250	1326.530	1312.250	2294760000
19	1324.660	May 8, 2006	1325.760	1326.700	1322.870	2151300000
20	1325.140	May 9, 2006	1324.660	1326.600	1322.480	2157290000
21	1322.850	May 10, 200	1324.570	1325.510	1317.440	2268550000
22	1305.920	May 11, 200	1322.630	1322.630	1303.450	2531520000
23	1291.240	May 12, 200	1305.880	1305.880	1290.380	2567970000
24	1294.500	May 15, 200	1291.190	1294.810	1284.510	2505660000

CONTOH: PREDIKSI HARGA SAHAM (PLOT)



CONTOH: PREDIKSI HARGA SAHAM (PLOT)



KLASIFIKASI

- Klasifikasi adalah algoritma yang menggunakan data dengan target/class/label berupa nilai kategorikal (nominal)
- Contoh, apabila target/class/label adalah pendapatan, maka bisa digunakan nilai nominal (kategorikal) sbb: pendapatan besar, menengah, kecil
- Contoh lain adalah rekomendasi contact lens, apakah menggunakan yang jenis soft, hard atau none
- Algoritma klasifikasi yang biasa digunakan adalah: Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, etc

CONTOH: REKOMENDASI MAIN GOLF

Input:

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Output (Rules):
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes

CONTOH: REKOMENDASI MAIN GOLF

- **Input (Atribut Nominal dan Numerik):**

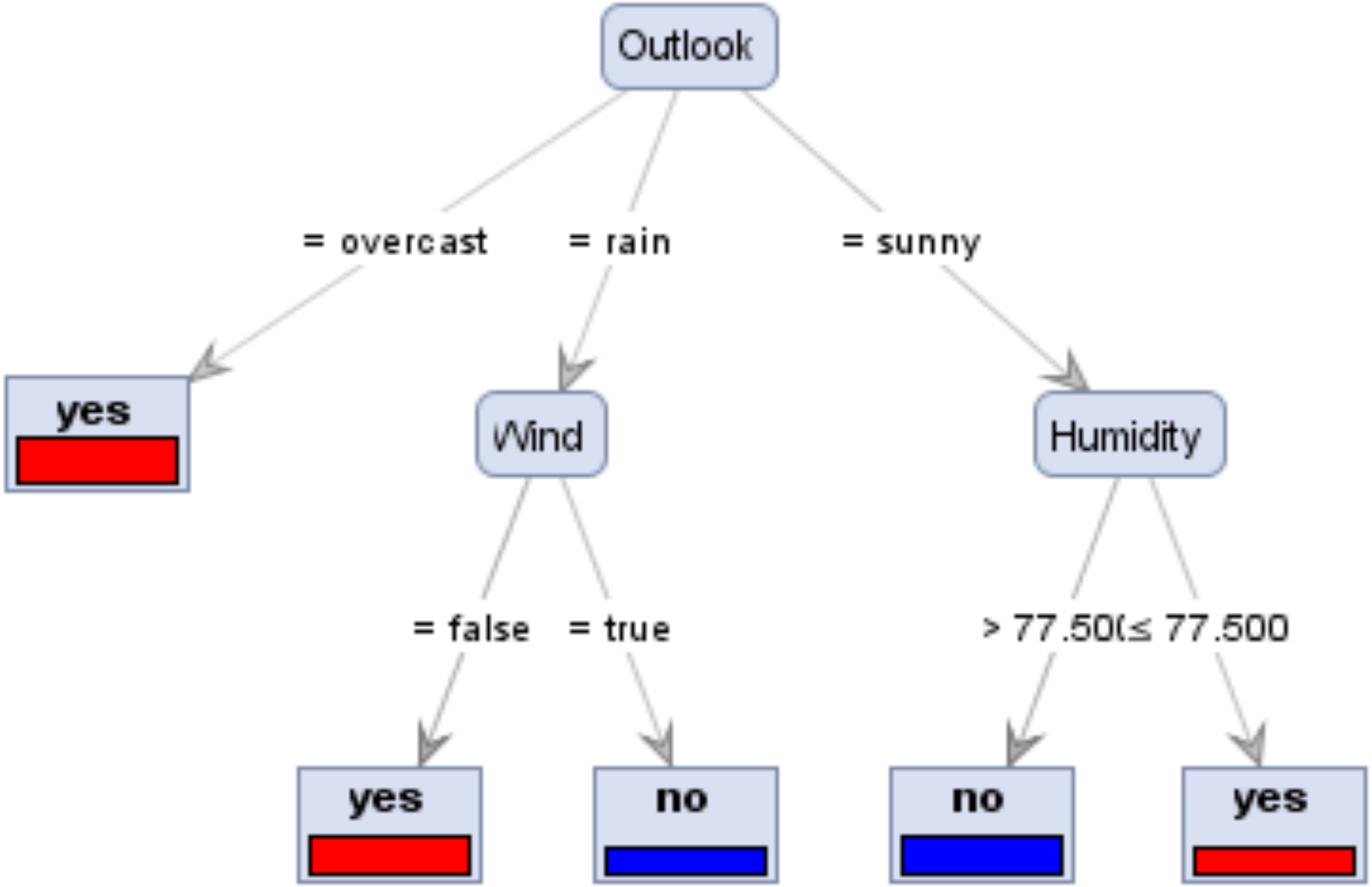
Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Output (Rules):

If outlook = sunny and humidity = high then play = no
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes

CONTOH: REKOMENDASI MAIN GOLF

- **Output (Tree):**



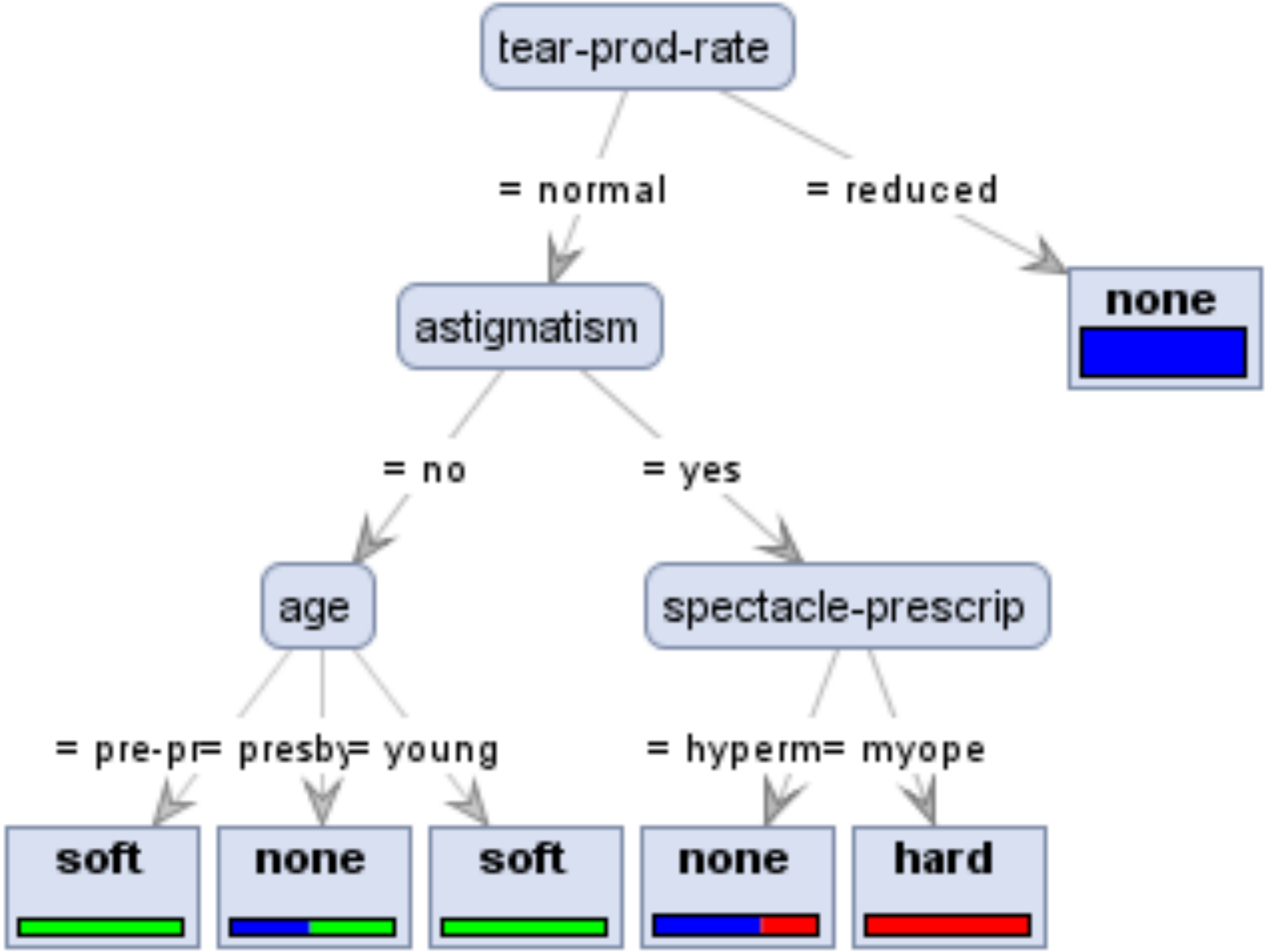
CONTOH: REKOMENDASI CONTACT LENS

- **Input**:

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft

CONTOH: REKOMENDASI CONTACT LENS

- Output/Model (Tree):



CONTOH: PENENTUAN JENIS BUNGA IRIS

- **Input**:

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

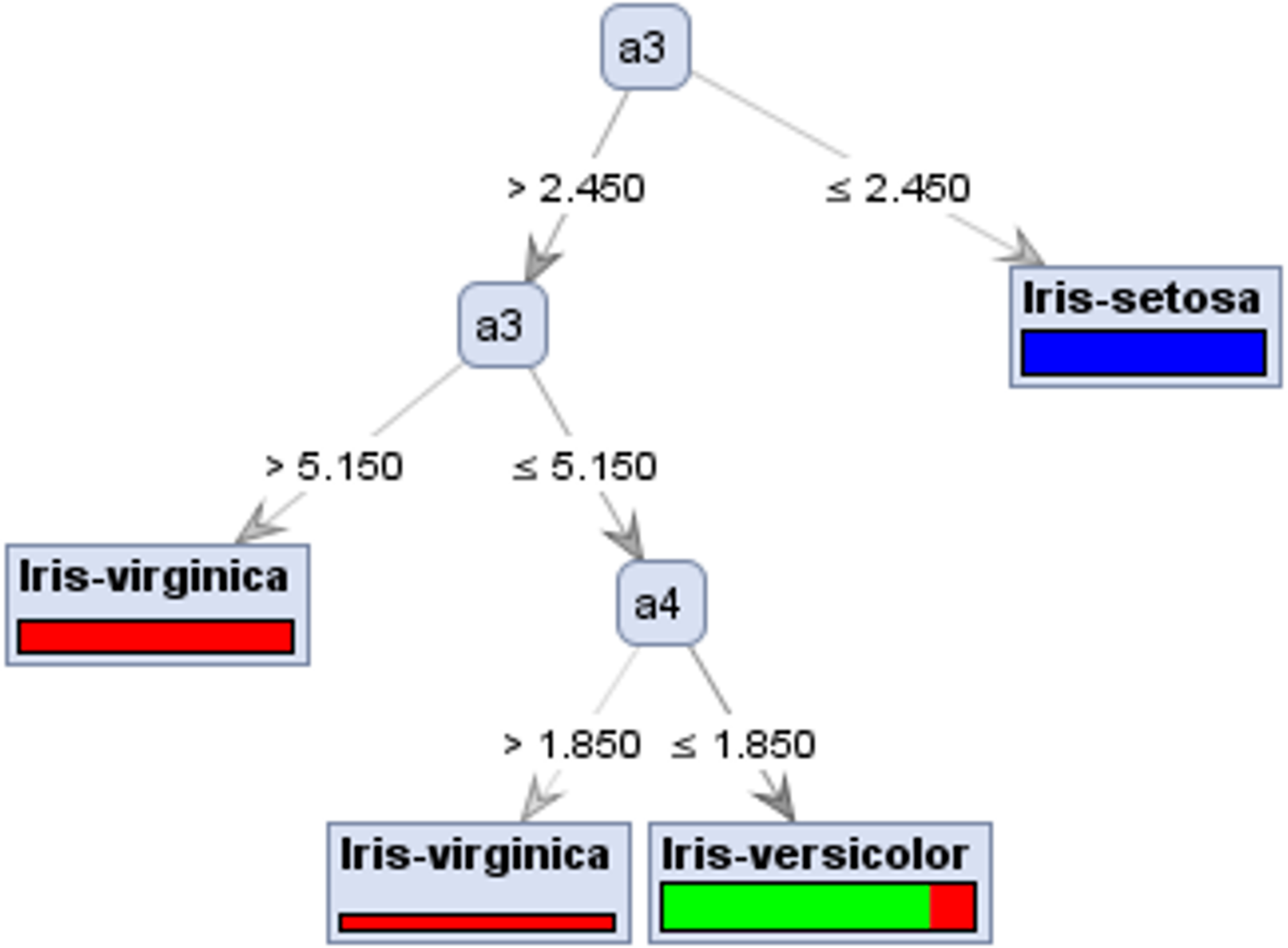
CONTOH: PENENTUAN JENIS BUNGA IRIS

- **Output** (Rules):

```
If petal-length < 2.45 then Iris-setosa
If sepal-width < 2.10 then Iris-versicolor
If sepal-width < 2.45 and petal-length < 4.55 then Iris-versicolor
If sepal-width < 2.95 and petal-width < 1.35 then Iris-versicolor
If petal-length ≥ 2.45 and petal-length < 4.45 then Iris-versicolor
If sepal-length ≥ 5.85 and petal-length < 4.75 then Iris-versicolor
If sepal-width < 2.55 and petal-length < 4.95 and
   petal-width < 1.55 then Iris-versicolor
If petal-length ≥ 2.45 and petal-length < 4.95 and
   petal-width < 1.55 then Iris-versicolor
If sepal-length ≥ 6.55 and petal-length < 5.05 then Iris-versicolor
If sepal-width < 2.75 and petal-width < 1.65 and
   sepal-length < 6.05 then Iris-versicolor
If sepal-length ≥ 5.85 and sepal-length < 5.95 and
   petal-length < 4.85 then Iris-versicolor
If petal-length ≥ 5.15 then Iris-virginica
If petal-width ≥ 1.85 then Iris-virginica
If petal-width ≥ 1.75 and sepal-width < 3.05 then Iris-virginica
If petal-length ≥ 4.95 and petal-width < 1.55 then Iris-virginica
```

CONTOH: PENENTUAN JENIS BUNGA IRIS

- **Output** (Tree):



KLASTERING

- Klastering adalah pengelompokkan data, hasil observasi dan kasus ke dalam class yang mirip
- Suatu klaster (cluster) adalah koleksi data yang mirip antara satu dengan yang lain, dan memiliki perbedaan bila dibandingkan dengan data dari klaster lain
- Perbedaan utama algoritma klastering dengan klasifikasi adalah klastering tidak memiliki target/class/label, jadi termasuk *unsupervised learning*
- Klastering sering digunakan sebagai tahap awal dalam proses data mining, dengan hasil klaster yang terbentuk akan menjadi input dari algoritma berikutnya yang digunakan

CONTOH: KLASTERING JENIS GAYA HIDUP

- Claritas, Inc. provide a demographic profile of each of the geographic areas in the country, as defined by zip code. One of the clustering mechanisms they use is the PRIZM segmentation system, which describes every U.S. zip code area in terms of distinct lifestyle types (66 segments). Just go to the company's Web site, enter a particular zip code, and you are shown the most common PRIZM clusters for that zip code.
- What do these clusters mean? For illustration, let's look up the clusters for zip code 90210, Beverly Hills, California. The resulting clusters for zip code 90210 are:
  - *Cluster 01: Blue Blood Estates*
  - *Cluster 10: Bohemian Mix*
  - *Cluster 02: Winner's Circle*
  - *Cluster 07: Money and Brains*
  - *Cluster 08: Young Literati*

Nielsen PRIZM - Understanding Social and Lifestage G

- What is Nielsen PRIZM?
- Features and Benefits
- Lifestyle Segmentation
- Urbanization Classes
- Social Groups
- Lifestage Classes
- Lifestage Groups
- Summary

Features and Benefits

and market to them with tailored messages and products designed just for them. Captured by catchy names, images and behavior snapshots that bring the segments to life for marketers, PRIZM segments are memorable and summarize complex consumer profiles in a way that is intuitive and easy to communicate.

For example, PRIZM Segment number 16 is known as *Bohemian Mix*. We can describe both the demographic traits, as well as the lifestyle characteristics of the households in this segment. You can review these segment descriptors in the image at right.

Bohemian Mix

16

Y2 Young Achievers

Upper-Mid Middle Age Family Mix

<55

Renters

White-Collar, Mix

College Graduate

White, Black, Asian, Hispanic

Eat at Au Bon Pain

Buy Spanish/Latin music

Read *The Economist*

Watch soccer

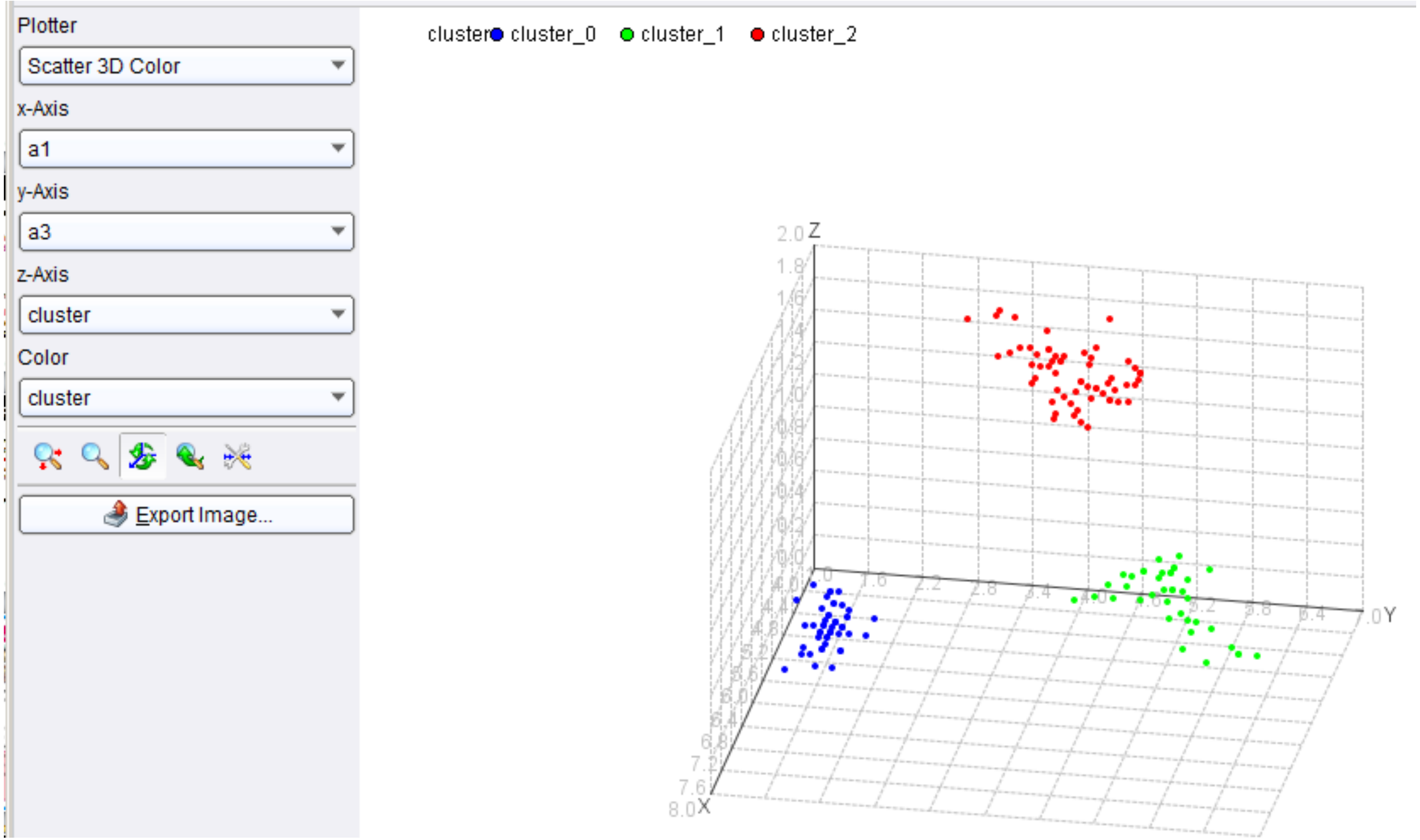
Audi A4

CONTOH: 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

CONTOH: KLASTERING BUNGA IRIS (PLOT)



CONTOH: KLASTERING BUNGA IRIS (TABLE)

ExampleSet (150 examples, 3 special attributes, 4 regular attributes)

Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_0	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	cluster_0	4.900	3	1.400	0.200
3	id_3	Iris-setosa	cluster_0	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	cluster_0	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	cluster_0	5	3.600	1.400	0.200
6	id_6	Iris-setosa	cluster_0	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	cluster_0	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	cluster_0	5	3.400	1.500	0.200
9	id_9	Iris-setosa	cluster_0	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	cluster_0	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	cluster_0	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	cluster_0	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	cluster_0	4.800	3	1.400	0.100
14	id_14	Iris-setosa	cluster_0	4.300	3	1.100	0.100
15	id_15	Iris-setosa	cluster_0	5.800	4	1.200	0.200
16	id_16	Iris-setosa	cluster_0	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	cluster_0	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	cluster_0	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	cluster_0	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	cluster_0	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	cluster_0	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	cluster_0	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	cluster_0	4.600	3.600	1	0.200
24	id_24	Iris-setosa	cluster_0	5.100	3.300	1.700	0.500

Cluster Model

```
Cluster 0: 50 items
Cluster 1: 39 items
Cluster 2: 61 items
Total number of items: 150
```

ALGORITMA ASOSIASI

- Algoritma *association rule* (aturan asosiasi) adalah algoritma yang menemukan atribut yang "muncul bersamaan"
- Dalam dunia bisnis, sering disebut dengan *affinity analysis* atau *market basket analysis*
- Algoritma asosiasi akan mencari aturan yang menghitung hubungan diantara dua atau lebih atribut
- Algoritma association rules berangkat dari pola "If *antecedent*, then *consequent*," bersamaan dengan pengukuran support (*coverage*) dan confidence (*accuration*) yang terasosiasi dalam aturan

ALGORITMA ASOSIASI

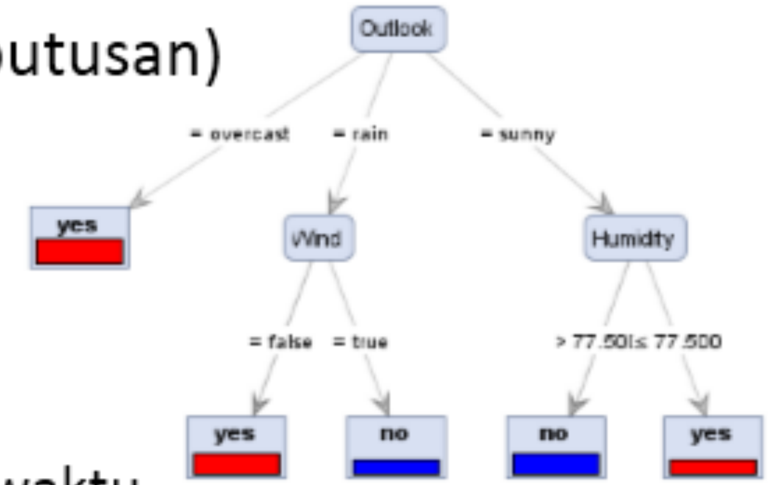
- Contoh, pada hari kamis malam, 1000 pelanggan telah melakukan belanja di supermaket ABC, dimana:
  - 200 orang membeli Sabun Mandi
  - dari 200 orang yang membeli sabun mandi, 50 orangnya membeli Fanta
- Jadi, association rule menjadi, “Jika membeli sabun mandi, maka membeli Fanta”, dengan nilai support = 200/1000 = 20% dan nilai confidence = 50/200 = 25%
- Algoritma association rule diantaranya adalah: A priori algorithm, FP-Growth algorithm, GRI algorithm

ALGORITMA DATA MINING (DM)

1. **Estimation** (Estimasi):
   Linear Regression, Neural Network, Support Vector Machine, etc
2. **Prediction/Forecasting** (Prediksi/Peramalan):
   Linear Regression, Neural Network, Support Vector Machine, etc
3. **Classification** (Klasifikasi):
   Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, etc
4. **Clustering** (Klastering):
   K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc
5. **Association** (Asosiasi):
   FP-Growth, A Priori, etc

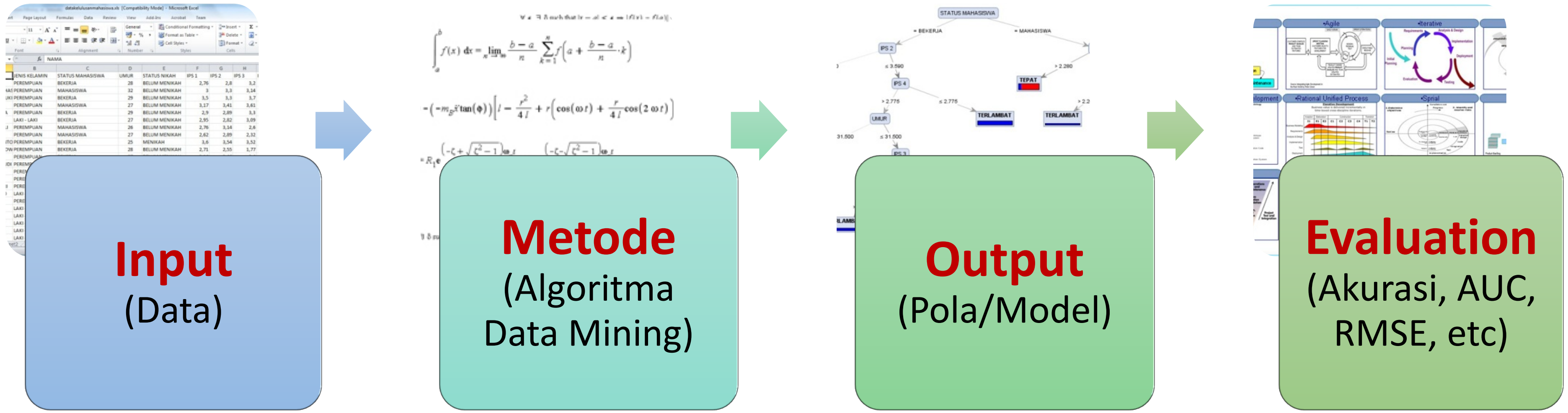
OUTPUT/POLA/MODEL/KNOWLEDGE

1. Formula/Function (Rumus atau Fungsi Regresi)
   - WAKTU TEMPUH = 0.48 + 0.6 JARAK + 0.34 LAMPU + 0.2 PESANAN
2. Decision Tree (Pohon Keputusan)



3. Rule (Aturan)
   - IF ips3=2.8 THEN lulustepatwaktu
4. Cluster (Klaster)

INPUT – METODE – OUTPUT – EVALUATION



CONTOH PENERAPAN DATA MINING

- Penentuan kelayakan aplikasi peminjaman uang di bank
- Penentuan pasokan listrik PLN untuk wilayah Jakarta
- Diagnosis pola kesalahan mesin
- Perkiraan harga saham dan tingkat inflasi
- Analisis pola belanja pelanggan
- Pemilihan program TV otomatis
- Penentuan pola pelanggan yang loyal pada perusahaan operator telepon
- Deteksi pencucian uang dari transaksi perbankan
- Deteksi serangan (intrusion) pada suatu jaringan

Latihan Soal (Kuis)

1. Sebutkan 5 peran utama data mining!
2. algoritma apa saja yang dapat digunakan untuk 5 peran utama data mining di atas?
3. Jelaskan perbedaan estimasi dan prediksi!
4. Jelaskan perbedaan estimasi dan klasifikasi!
5. Jelaskan perbedaan klasifikasi dan klastering!
6. Jelaskan perbedaan klastering dan prediksi!
7. Jelaskan perbedaan supervised dan unsupervised learning!
8. Sebutkan tahapan utama proses data mining!