

MACHINE LEARNING DAN DATA MINING

NUGROHO D.S.

APA ITU MACHINE LEARNING?

suatu area dalam artificial intelligence yang berhubungan dengan pengembangan teknik-teknik yang bisa diprogramkan dan belajar dari data masa lalu. (Santoso, 2007)

Bagaimana membuat komputer yang dapat belajar dari lingkungan sekitar sehingga memiliki “pengetahuan” yang berkembang

PERBEDAAN DM DAN ML

DM (data mining)

- **fokus pada memanfaatkan program untuk membantu manusia belajar dari data yang sangat besar.**

- **ML (machine learning)**

- **Fokus pada perbaikan performansi dari suatu teknik learning**

PERAN UTAMA DATA MINING

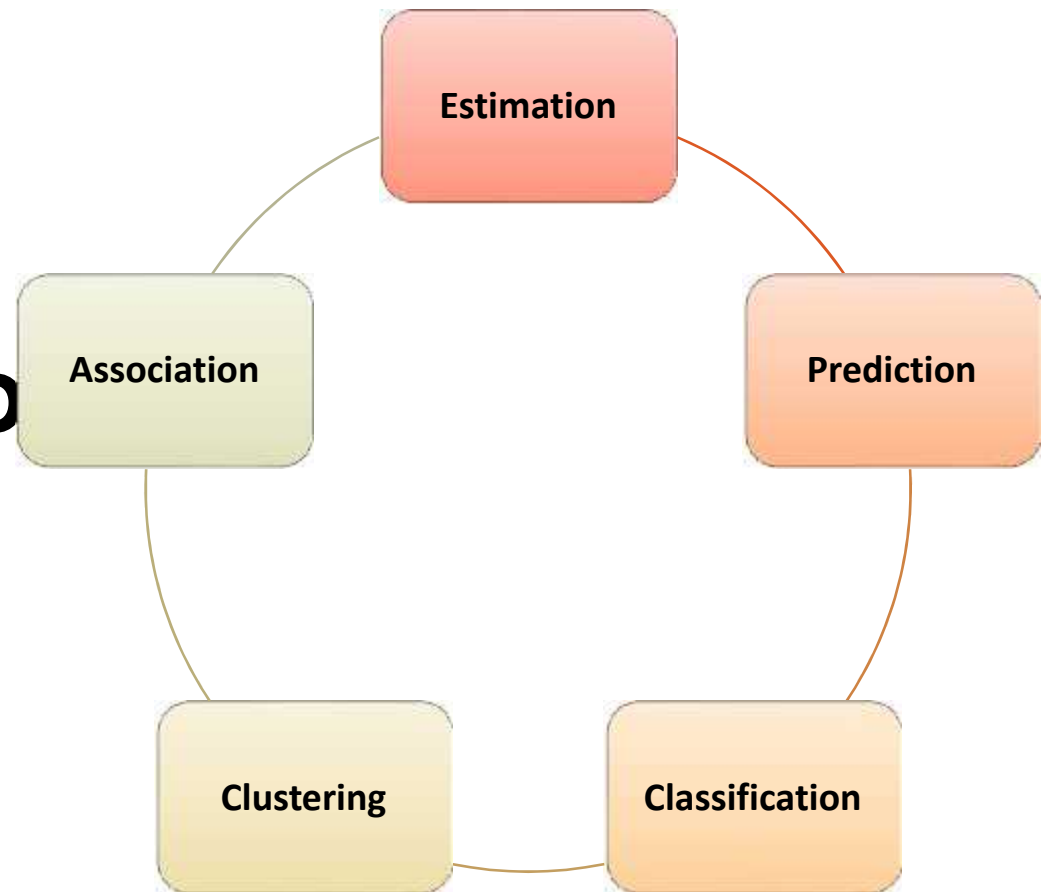
1. Estimation

2. Prediction

3. Classification

4. Clustering

5. Association



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

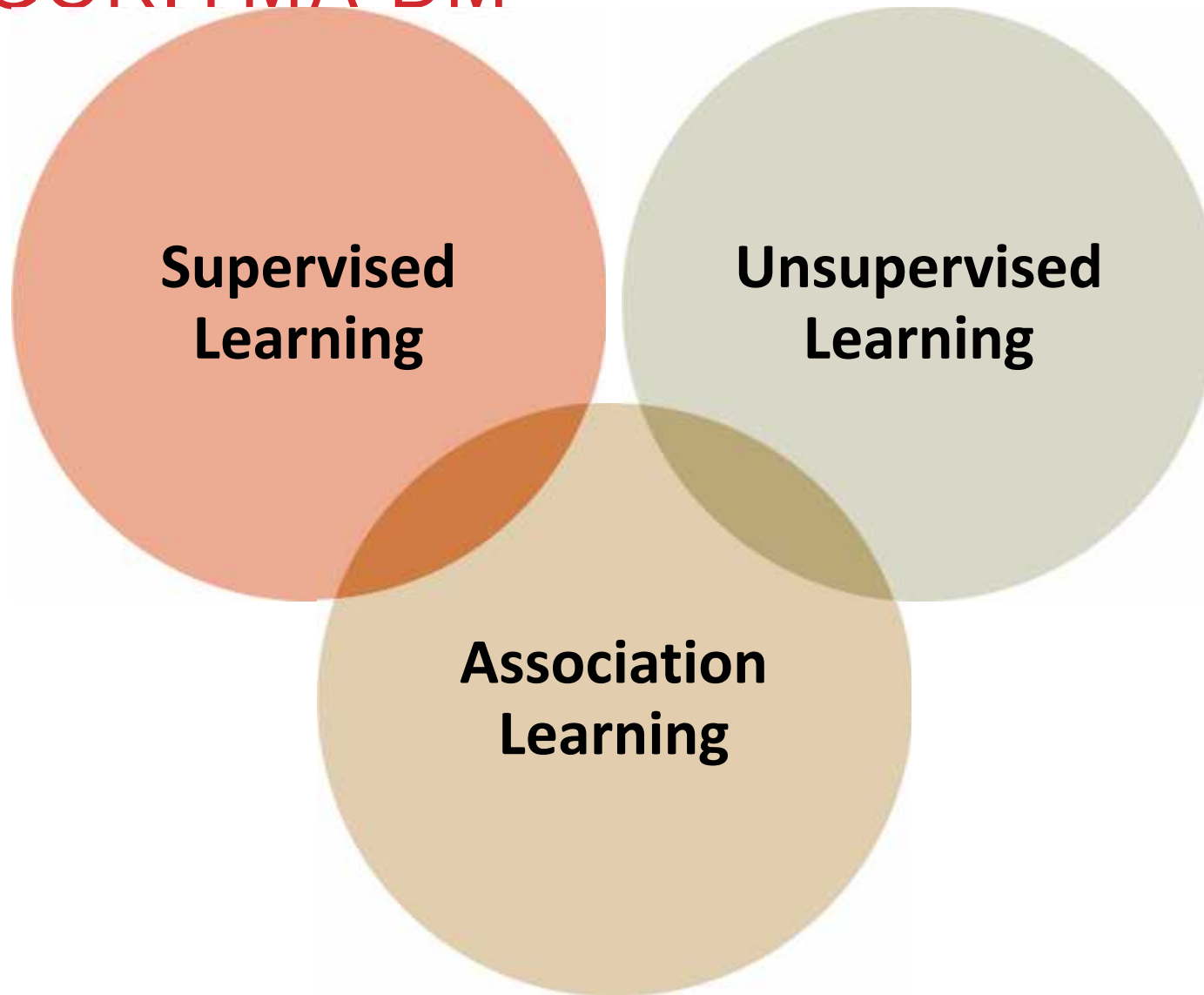
4. Clustering (Klastering):

- **K-Means**, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc

5. Association (Asosiasi):

- FP-Growth, **A Priori**, etc

METODE LEARNING PADA ALGORITMA DM



METODE LEARNING PADA ALGORITMA DM


1. **Supervised** Learning :

- 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> |
| ... | | | | | |

ALGORITMA 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**

ALGORITMA 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

| 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, 2006 | 1324.570 | 1325.510 | 1317.440 | 2268550000 |
| 22 | 1305.920 | May 11, 2006 | 1322.630 | 1322.630 | 1303.450 | 2531520000 |
| 23 | 1291.240 | May 12, 2006 | 1305.880 | 1305.880 | 1290.380 | 2567970000 |
| 24 | 1294.500 | May 15, 2006 | 1291.190 | 1294.810 | 1284.510 | 2505660000 |

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

ALGORITMA 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 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 |

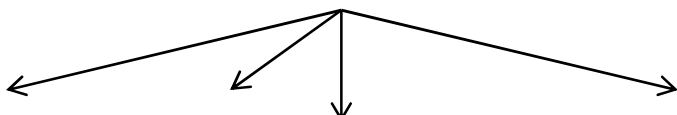
METODE LEARNING PADA ALGORITMA DM

2. **Unsupervised** Learning:

- 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 |
| ... | | | | |

ALGORITMA KLAUSTERING

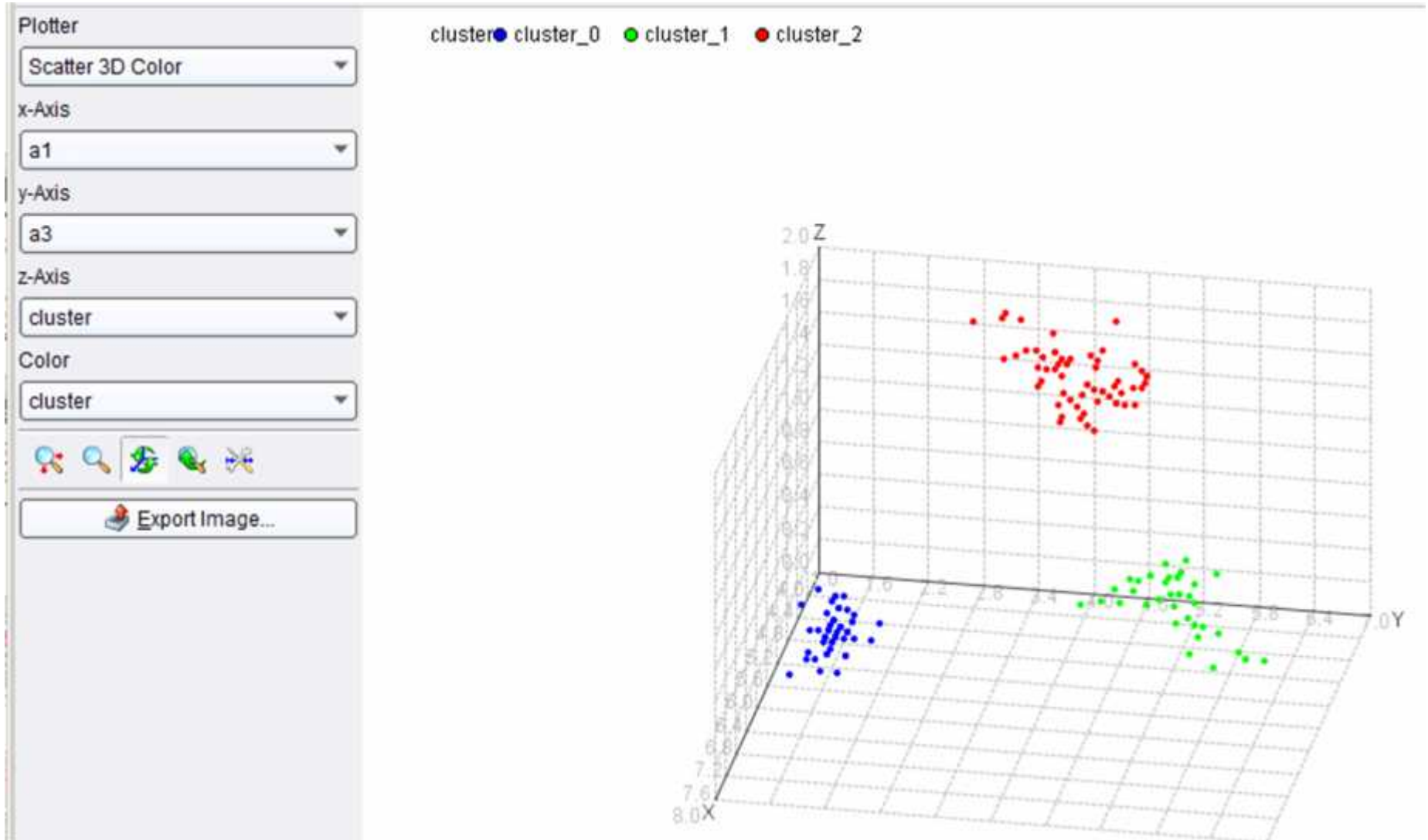
Klastering adalah **pengelompokkan data**, hasil observasi dan kasus ke dalam **class yang mirip**

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 BUNGA IRIS (PLOT)



CONTOH: KLASTERING BUNGA IRIS (TABLE)

| ExampleSet (150 examples, 3 special attributes, 4 regular attributes) | | | | | | | View |
|---|-------|-------------|-----------|-------|-------|-------|-------|
| 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

METODE LEARNING PADA ALGORITMA DM

3. **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 **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)
```

