

# DATAMINING

BY: JUNTA ZENIARJA, M.KOM

# Apa itu Data Mining?



# Mengapa Data Mining?

- ▶ Manusia dalam suatu organisasi, sadar atau tidak sadar telah memproduksi berbagai **data yang jumlahnya sangat besar**
  - ▶ Contoh data: bisnis, kedokteran, ekonomi, geografi, olahraga, ...
- ▶ Pada dasarnya, data adalah entitas yang **tidak memiliki arti**, meskipun **kemungkinan memiliki nilai** di dalamnya

# Apa itu Data Mining?

- ▶ Disiplin ilmu yang mempelajari **metode untuk mengekstr** pengetahuan atau **menemukan pola** dari suatu data
  1. **Data**: fakta yang terekam dan tidak membawa art
  2. **Pengetahuan**: pola, aturan atau model yang munc dari data
- ▶ Sehingga Data mining sering disebut **Knowledge Discov in Database (KDD)**
- ▶ Konsep Transformasi  
**Data→Informasi→Pengetahuan**



# Data

- ▶ Tidak membawa arti, merupakan kumpulan dari fakta-fakta tentang suatu kejadian.
- ▶ Suatu catatan terstruktur dari suatu transaksi.
- ▶ Merupakan materi penting dalam membentuk informasi.

# Pengetahuan

- ▶ Gabungan dari suatu **pengalaman, nilai, informasi kontekstual dan juga pandangan pakar** yang memberikan suatu framework untuk mengevaluasi dan menciptakan pengalaman baru dan informasi (*Thomas H. Davenport, Laurence Prusak*).
- ▶ Bisa berupa **solusi pemecahan suatu masalah, petunjuk suatu pekerjaan** dan ini bisa ditingkatkan nilainya, dipelajari dan juga bisa diajarkan kepada yang lain.

# ***Data - Informasi – Pengetahuan***

**Data** Kehadiran Pegawai

NIP	TGL	DATANG	PULANG
1103	02/12/2004	07:20	15:40
1142	02/12/2004	07:45	15:33
1156	02/12/2004	07:51	16:00
1173	02/12/2004	08:00	15:15
1180	02/12/2004	07:01	16:31
1183	02/12/2004	07:49	17:00

# Data - Informasi – Pengetahuan

## Informasi Akumulasi Bulanan Kehadiran Pegawai

NIP	Masuk	Alpa	Cuti	Sakit	Telat
1103	22				
1142	18	2		2	
1156	10	1	11		
1173	12	5			5
1180	10			12	



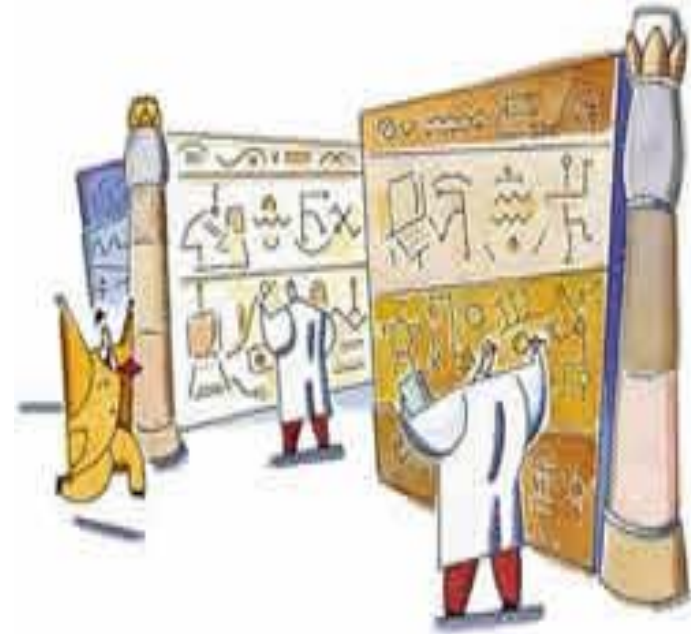
# Data - Informasi – Pengetahuan

Informasi Kondisi Kehadiran Mingguan Pegawai

	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

## *Data - Informasi – Pengetahuan*

- Pengetahuan tentang kebiasaan pegawai dalam jam datang/pulang kerja
- Pengetahuan tentang bagaimana teknik meningkatkan kehadiran pegawai → kebijakan



# Data - Informasi - Pengetahuan - Kebijakan

- ▶ **Kebijakan penataan jam kerja** karyawan khusus untuk hari senin dan jumat
- ▶ Peraturan jam kerja:
  - ▶ Hari Senin dimulai jam 10:00
  - ▶ Hari Jumat diakhiri jam 14:00
  - ▶ Sisa jam kerja dikompensasi ke hari lain:
    1. Senin pulang setelah maghrib, toh jalanan jakarta macet total di sore hari (**bayar hutang 2 jam**)
    2. Rabu dan kamis bayar hutang setengah jam di pagi hari dan setengah jam di sore hari (**bayar hutang 2 jam**)

# Definisi Data Mining

- ▶ Melakukan **ekstraksi** untuk mendapatkan **informasi penting** yang sifatnya **implisit** dan sebelumnya tidak diketahui, dari suatu data (*Witten et al., 2011*)
- ▶ Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk **menemukan keteraturan, pola dan hubungan** dalam set data berukuran besar (*Santosa, 2007*)

# Definisi Data Mining

- ▶ The analysis of (often large) observational data sets to find **unsuspected relationships** and to **summarize the data** in novel ways that are both understandable and useful to the data owner (*Han & Kamber, 2001*)
- ▶ The process of **discovering meaningful new correlations, patterns and trends** by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques (*Gartner Group*)

# Irisan Bidang Ilmu Data Mining

## 1. Statistik:

- ▶ Lebih bersifat teori
- ▶ Fokus ke pengujian hipotesis

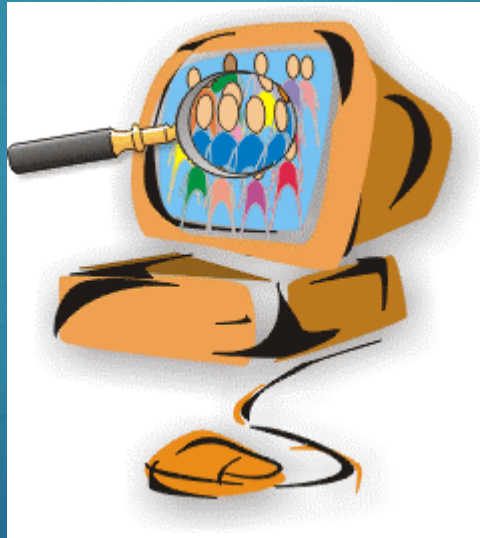
## 2. Machine Learning:

- ▶ Lebih bersifat heuristik
- ▶ Fokus pada perbaikan performansi dari suatu teknik learning

## 3. Data Mining:

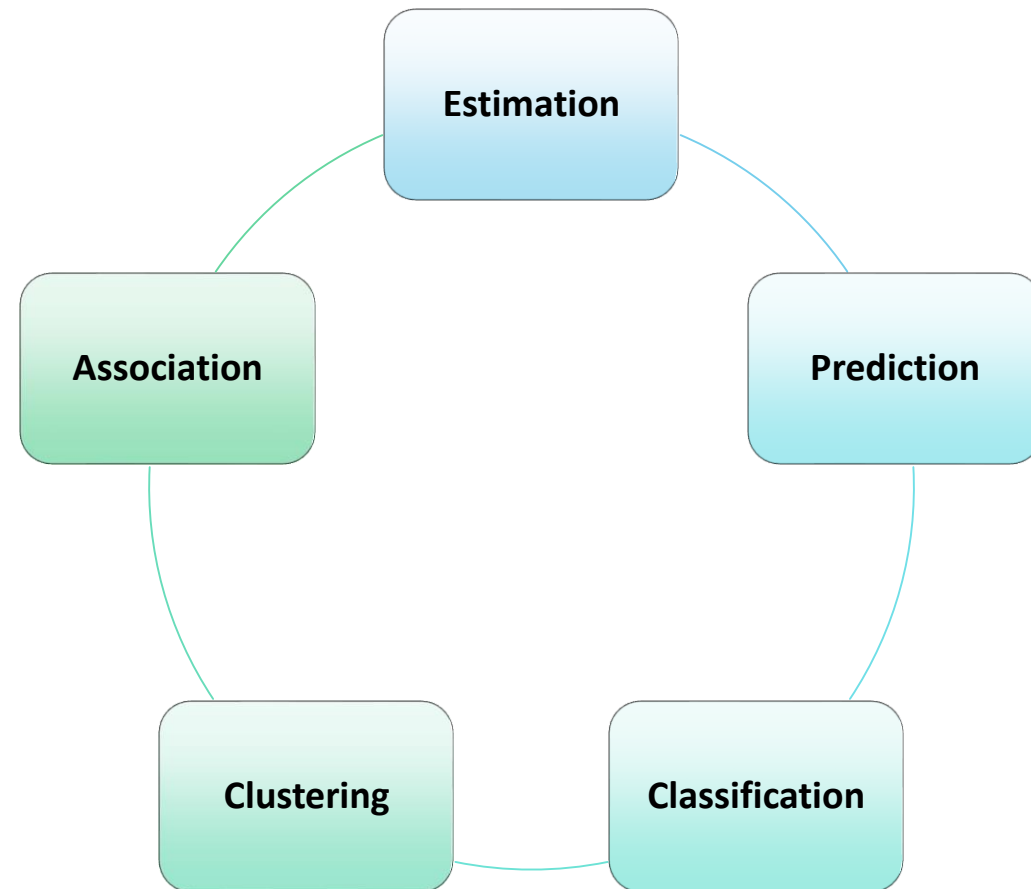
- ▶ Gabungan teori dan heuristik
- ▶ Fokus pada seluruh proses penemuan knowledge dan pola
- ▶ Termasuk data cleaning, learning dan visualisasi hasilnya

# Peran Utama Data Mining



# Peran Utama Data Mining

1. Estimation
2. Prediction
3. Classification
4. Clustering
5. Association





# 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 Bangjo (B)	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

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

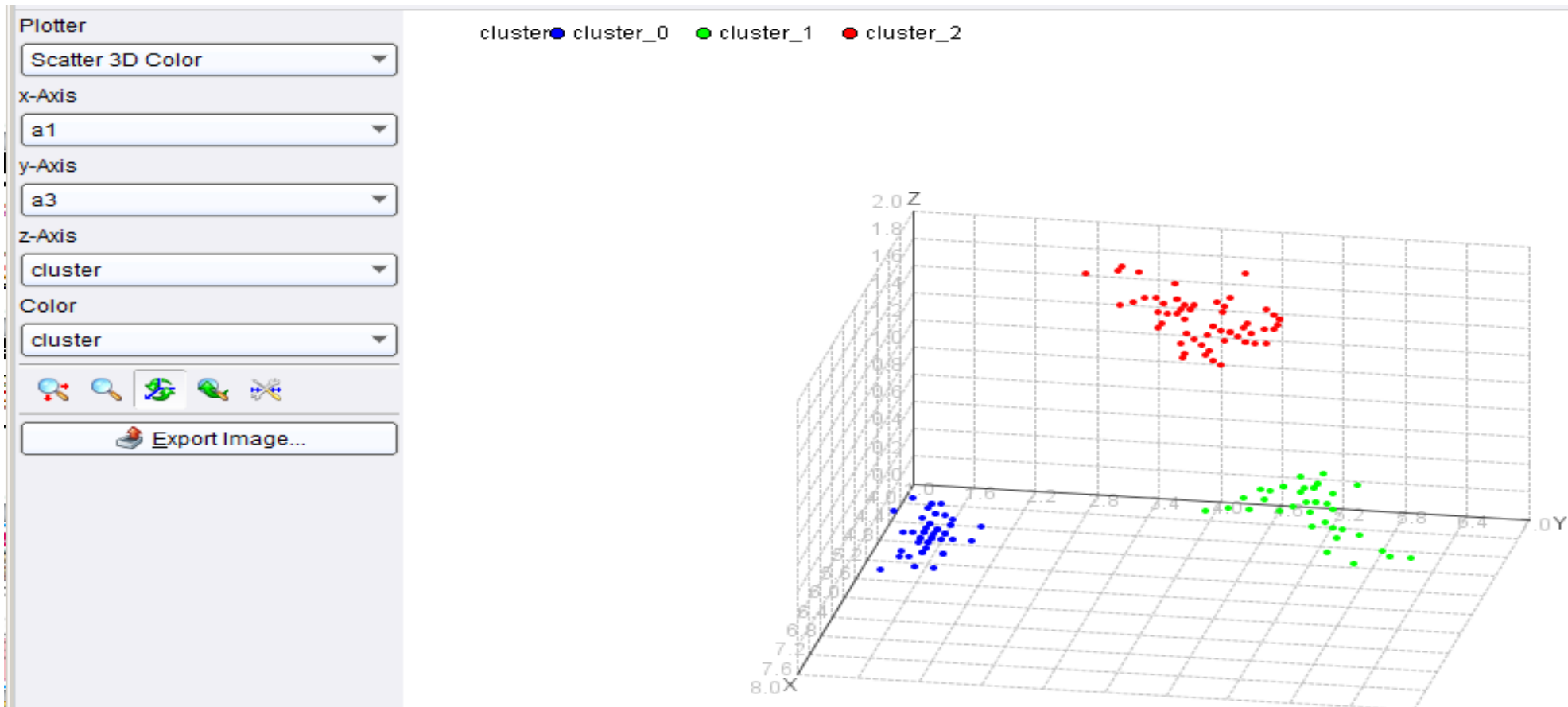
# Penentuan Kelulusan Mahasiswa

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



# 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

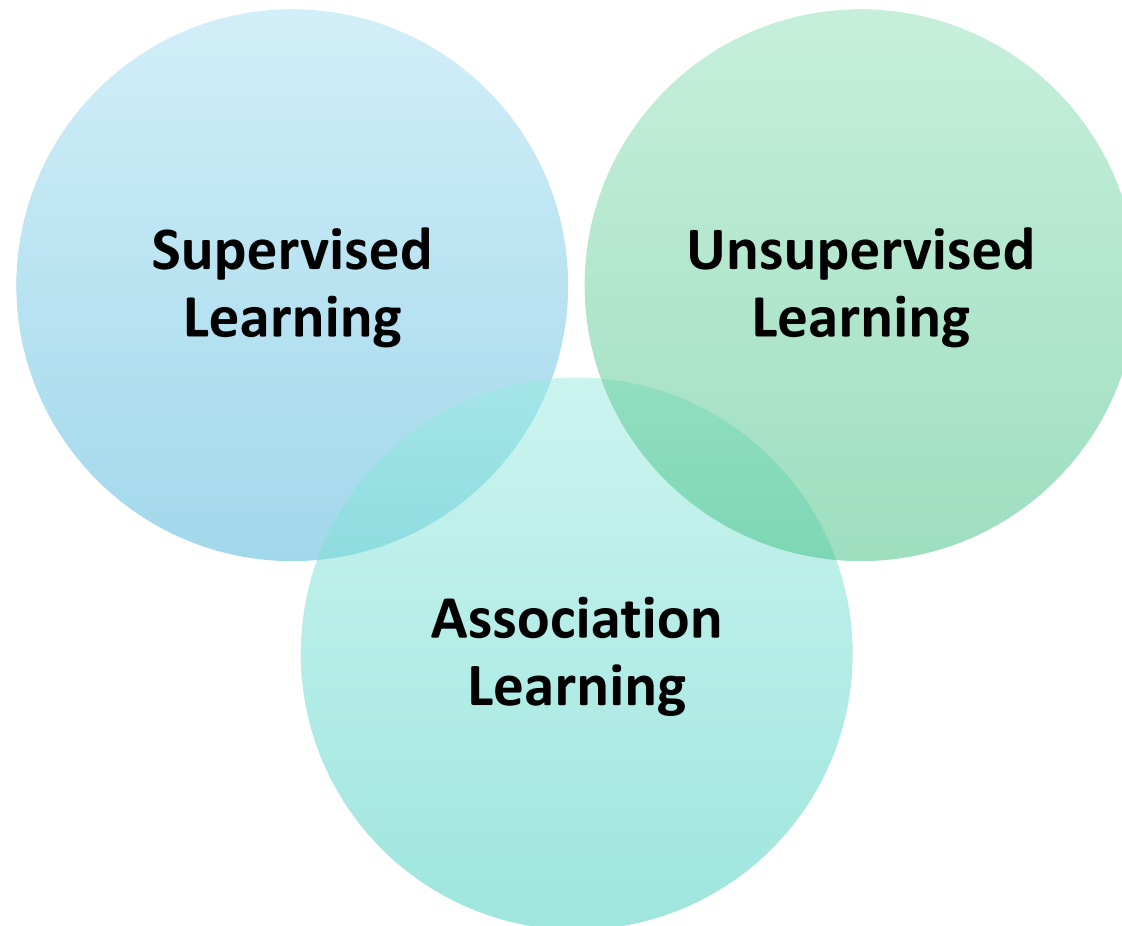
## 4. **Clustering** (Klastering):

- ▶ [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 (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*) 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

# Dataset Transaction

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

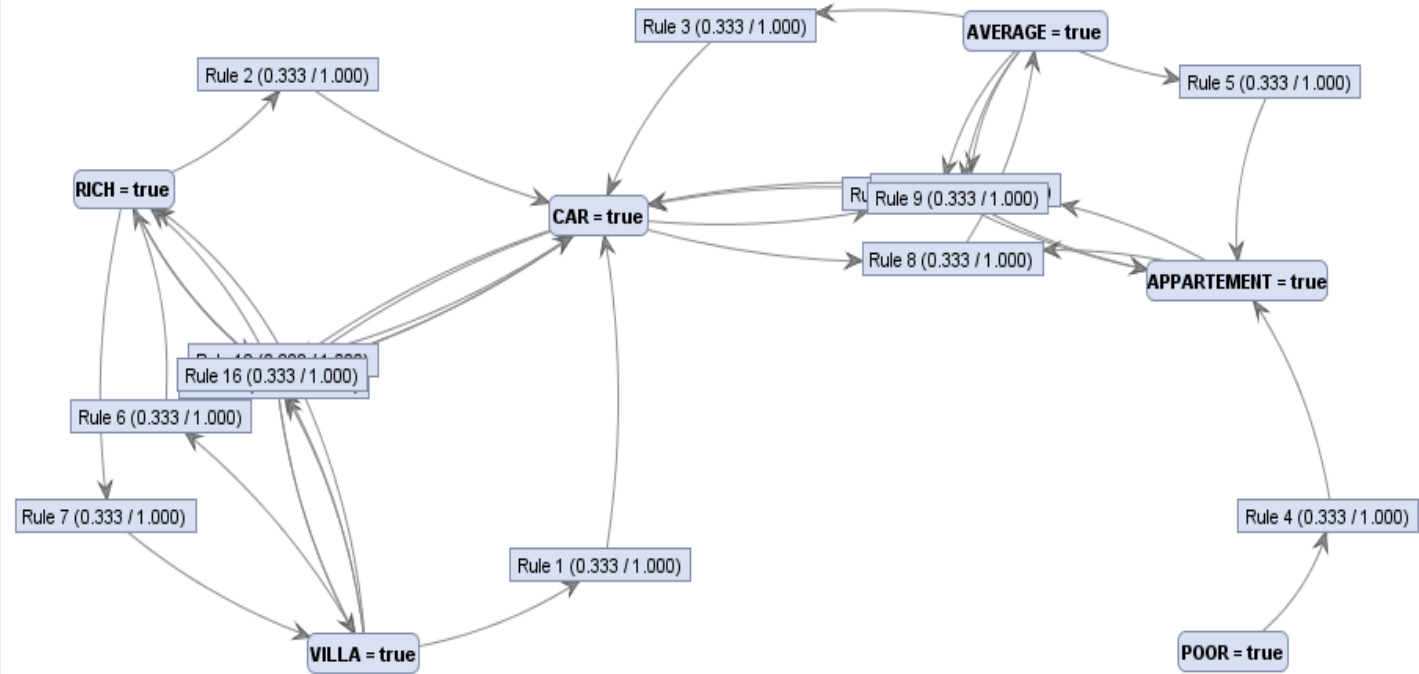
Row No.	CAR = true	APPARTEMENT = true	VILLA = true	POOR = true	AVERAGE = true	RICH = true
1	false	true	false	true	false	false
2	true	true	false	false	true	false
3	true	false	true	false	false	true

# Association Rules

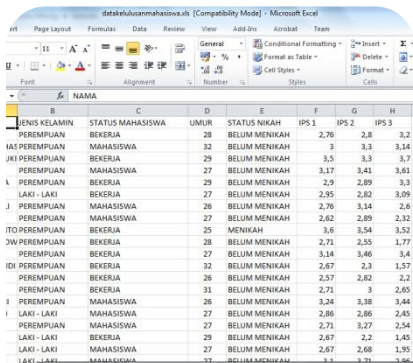
## Association Rules

### 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)
```



# Proses Utama pada Data Mining



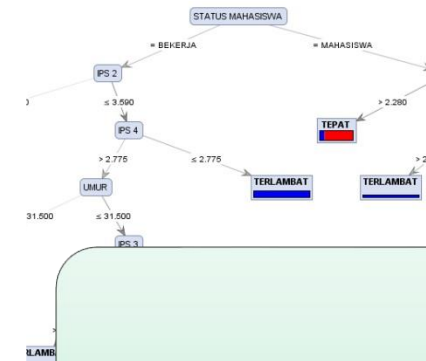
The screenshot shows a Microsoft Excel spreadsheet with the following data:

B	C	D	E	F	G	H
1	NAMA					
1	JENIS KELAMIN	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2
2	PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,76	2,8
3	PEREMPUAN	MAHASISWA	32	BELUM MENIKAH	3	3,3
4	PEREMPUAN	BEKERJA	29	BELUM MENIKAH	3,5	3,3
5	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3,17	3,41
6	PEREMPUAN	BEKERJA	29	BELUM MENIKAH	2,9	2,89
7	LAKI - LAKI	BEKERJA	27	BELUM MENIKAH	2,95	2,82
8	PEREMPUAN	MAHASISWA	26	BELUM MENIKAH	2,76	3,14
9	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	2,62	2,89
10	PEREMPUAN	BEKERJA	25	MENIKAH	3,6	3,54
11	PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,71	2,55
12	PEREMPUAN	BEKERJA	27	BELUM MENIKAH	3,14	3,46
13	PEREMPUAN	BEKERJA	32	BELUM MENIKAH	2,67	2,3
14	PEREMPUAN	BEKERJA	26	BELUM MENIKAH	2,57	2,82
15	PEREMPUAN	BEKERJA	31	BELUM MENIKAH	2,71	3
16	PEREMPUAN	MAHASISWA	26	BELUM MENIKAH	3,24	3,38
17	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2,86	2,86
18	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	2,71	3,27
19	LAKI - LAKI	BEKERJA	29	BELUM MENIKAH	2,67	2,2
20	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2,67	2,68
21	LAKI - LAKI	MAHASISWA	32	BELUM MENIKAH	3,3	3,31

**Input**  
(Data)

$$f(x) dx = \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} \cdot k\right)$$
$$= \left( -m_2 \tilde{x} \tan(\Phi) \right) \left[ l = \frac{r^2}{4l} + r \left( \cos(\omega l) + \frac{r}{4l} \cos(2\omega l) \right) \right]$$
$$= R_1 e^{\left( -\zeta + \sqrt{\zeta^2 - 1} \right) \omega l} \left( -\zeta + \sqrt{\zeta^2 - 1} \right) \omega l$$

**Metode**  
(Algoritma  
Data Mining)



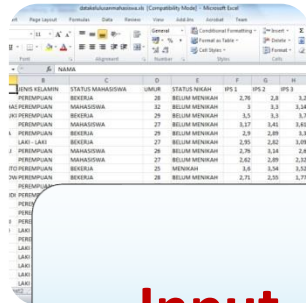
**Output**  
(Pola/Model)

# Output/Pola/Model/Knowledge

1. Formula/**Function** (Rumus atau Fungsi Regresi)
  - ▶  $\text{WAKTU TEMPUH} = 0.48 + 0.6 \text{ JARAK} + 0.34 \text{ LAMPU} + 0.2 \text{ PESANAN}$
2. Decision **Tree** (Pohon Keputusan)
3. **Rule** (Aturan)
  - ▶ IF  $\text{ips3}=2.8$  THEN  $\text{lulustepatwaktu}$
4. **Cluster** (Klaster)



# Input – Metode – Output – Evaluation

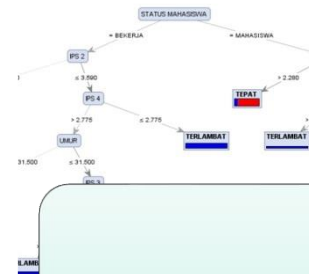


ID	STATUS MAHASISWA	UMUR	STATUS KAWIN	IP1	IP2	IP3
1	MAHASISWA	20	BELUM MENIKAH	2.26	2.2	3.2
2	MAHASISWA	32	BELUM MENIKAH	3	3.3	3.34
3	MAHASISWA	29	BELUM MENIKAH	3.5	3.5	3.7
4	MAHASISWA	27	BELUM MENIKAH	3.57	3.45	3.45
5	MAHASISWA	29	BELUM MENIKAH	2.8	2.89	3.3
6	MAHASISWA	27	BELUM MENIKAH	2.95	2.82	3.89
7	MAHASISWA	26	BELUM MENIKAH	2.76	3.36	3.4
8	MAHASISWA	27	BELUM MENIKAH	3.82	3.89	3.32
9	MAHASISWA	25	MAHISWA	3.6	3.56	3.52
10	MAHASISWA	28	BELUM MENIKAH	2.75	2.55	3.77

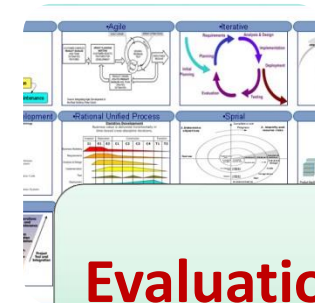
**Input**  
(Data)

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} \cdot k\right)$$
$$= \left( -m_2 \tan(\phi) \right) \left[ l - \frac{r^2}{4l} + r \left( \cos(\omega r) + \frac{r}{4l} \cos(2\omega r) \right) \right]$$
$$= \left( -\zeta + \sqrt{\zeta^2 - 1} \right) \log x - \left( -\zeta - \sqrt{\zeta^2 - 1} \right) \log x$$

**Metode**  
(Algoritma  
Data Mining)



**Output**  
(Pola/Model)



**Evaluation**  
(Akurasi, AUC,  
RMSE, etc)



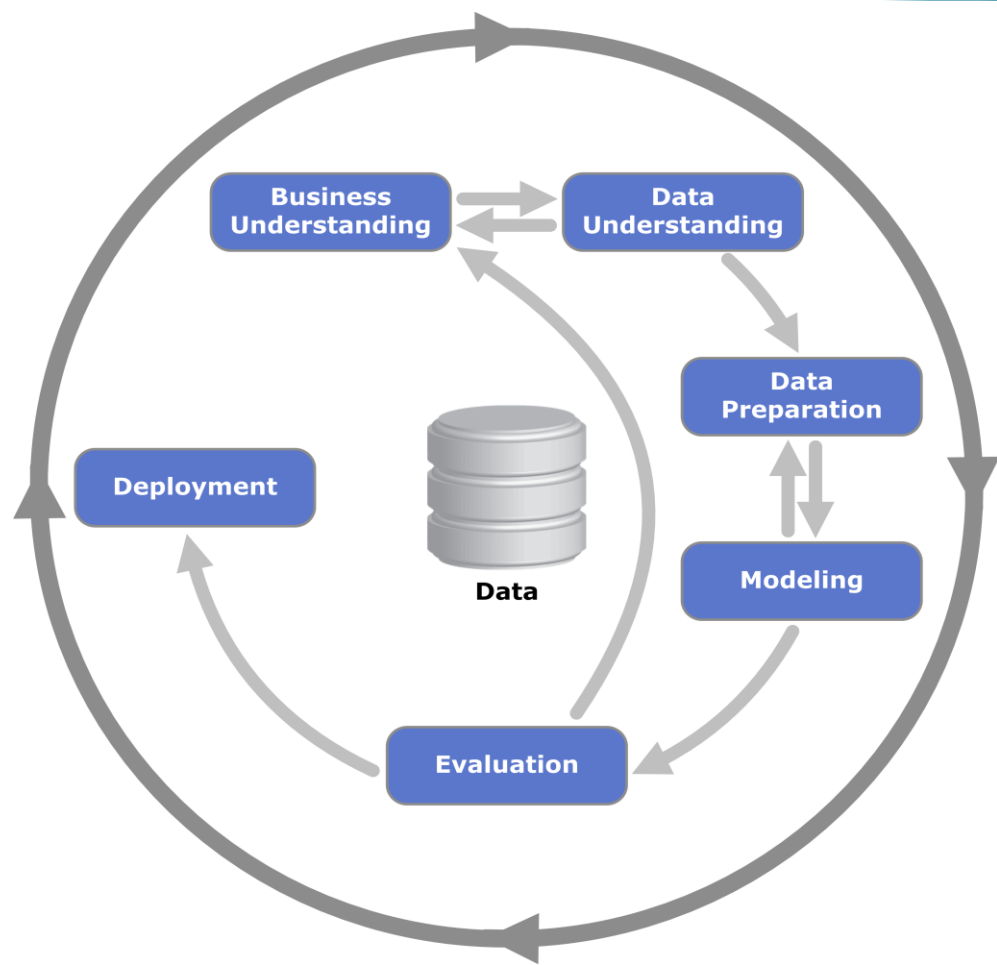
# CRISP-DM

*(Cross Industry Standard Process for Data Mining)*

# Pengenalan CRISP-DM

- ▶ **CRISP-DM** (Cross Industry Standard Process for Data Mining) merupakan model proses data mining yang menggambarkan urutan atau langkah pada penelitian Data Mining yang digunakan para ahli untuk mengatasi masalah. Jajak pendapat yang dilakukan pada tahun 2002, 2004, dan 2007 menunjukkan bahwa CRISP-DM termasuk metodologi terkemuka yang digunakan oleh para ahli Data Mining.
- ▶ Pada tahun 2009, CRISP-DM disebutkan sebagai standar **de facto** untuk mengembangkan project Data Mining atau KDD (*Knowledge Discovery in Database*).

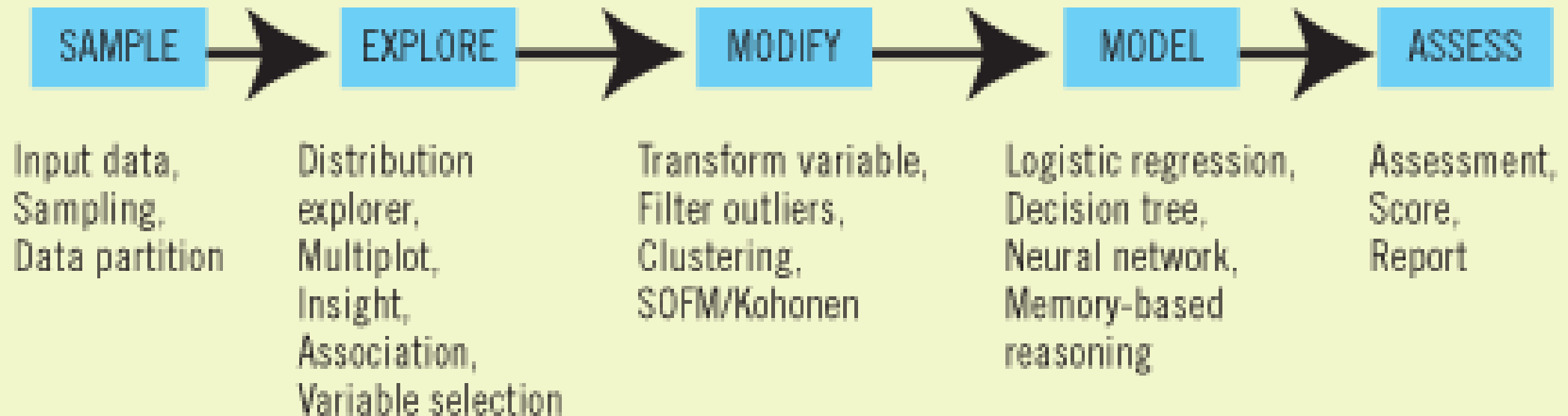
# C R I S P D M



# SEMMA –DM

Figure 1: SEMMA Methodology Diagram

Supported by SAS Enterprise Mining Environment



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p><b>Determine Business Objectives</b>  <i>Background</i>  <i>Business Objectives</i>  <i>Business Success Criteria</i></p> <p><b>Assess Situation</b>  <i>Inventory of Resources</i>  <i>Requirements, Assumptions, and Constraints</i>  <i>Risks and Contingencies</i>  <i>Terminology</i>  <i>Costs and Benefits</i></p> <p><b>Determine Data Mining Goals</b>  <i>Data Mining Goals</i>  <i>Data Mining Success Criteria</i></p> <p><b>Produce Project Plan</b>  <i>Project Plan</i>  <i>Initial Assessment of Tools and Techniques</i></p>	<p><b>Collect Initial Data</b>  <i>Initial Data Collection Report</i></p> <p><b>Describe Data</b>  <i>Data Description Report</i></p> <p><b>Explore Data</b>  <i>Data Exploration Report</i></p> <p><b>Verify Data Quality</b>  <i>Data Quality Report</i></p>	<p><b>Select Data</b>  <i>Rationale for Inclusion/Exclusion</i></p> <p><b>Clean Data</b>  <i>Data Cleaning Report</i></p> <p><b>Construct Data</b>  <i>Derived Attributes</i>  <i>Generated Records</i></p> <p><b>Integrate Data</b>  <i>Merged Data</i></p> <p><b>Format Data</b>  <i>Reformatted Data</i>    <i>Dataset</i>  <i>Dataset Description</i></p>	<p><b>Select Modeling Techniques</b>  <i>Modeling Technique</i>  <i>Modeling Assumptions</i></p> <p><b>Generate Test Design</b>  <i>Test Design</i></p> <p><b>Build Model</b>  <i>Parameter Settings</i>  <i>Models</i>  <i>Model Descriptions</i></p> <p><b>Assess Model</b>  <i>Model Assessment</i>  <i>Revised Parameter Settings</i></p>	<p><b>Evaluate Results</b>  <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i>  <i>Approved Models</i></p> <p><b>Review Process</b>  <i>Review of Process</i></p> <p><b>Determine Next Steps</b>  <i>List of Possible Actions</i>  <i>Decision</i></p>	<p><b>Plan Deployment</b>  <i>Deployment Plan</i></p> <p><b>Plan Monitoring and Maintenance</b>  <i>Monitoring and Maintenance Plan</i></p> <p><b>Produce Final Report</b>  <i>Final Report</i>  <i>Final Presentation</i></p> <p><b>Review Project Experience</b>  <i>Documentation</i></p>

# Keterangan

## 1. Pemahaman Bisnis(*Business Understanding*)

- ▶ Merupakan tahap awal yaitu pemahaman penelitian, penentuan tujuan dan rumusan masalah *data mining*.

## 2. Pemahaman Data(*Data Understanding*)

- ▶ Dalam tahap ini dilakukan pengumpulan data, mengenali lebih lanjut data yang akan digunakan.

## 3. Pengolahan Data(*Data Preparation*)

- ▶ Tahap ini adalah pekerjaan berat yang perlu dilaksanakan secara intensif. Memilih kasus atau variable yang ingin dianalisis, melakukan perubahan pada beberapa variable jika diperlukan sehingga data siap untuk dimodelkan.

## 4. Pemodelan(*Modeling*)

- ▶ Memilih teknik pemodelan yang sesuai dan sesuaikan aturan model untuk hasil yang maksimal. Dapat kembali ke tahap pengolahan untuk menjadikan data ke dalam bentuk yang sesuai dengan model tertentu.

## 5. Evaluasi (*Evaluation*)

- ▶ Mengevaluasi satu atau model yang digunakan dan menetapkan apakah terdapat model yang memenuhi tujuan pada tahap awal. Kemudian menentukan apakah ada permasalahan yang tidak dapat tertangani dengan baik serta mengambil keputusan hasil penelitian.

## 6. Penyebaran (*Deployment*)

- ▶ Menggunakan model yang dihasilkan seperti pembuatan laporan atau penerapan proses *data mining* pada departemen lain.

# Referensi

1. Ian H. Witten, Frank Eibe, Mark A. Hall, **Data mining: Practical Machine Learning Tools and Techniques 3rd Edition**, *Elsevier*, 2011
2. Santosa Budi, **Teknik Pemanfaatan Data Untuk Keperluan Bisnis**, *Graha Ilmu*, 2007
3. [www.ilmukomputer.com](http://www.ilmukomputer.com).





Thank You!

