

DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

DATA MINING INTRODUCTION

Nizar Rabbi Radliya
nizar@email.unikom.ac.id



SILABUS

Data Mining Introduction

Preprocessing

Similarity

Association

UTS

Classification & Prediction

Clustering

UAS

REVIEW

Data → Informasi → Pengetahuan

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 Kehadiran Pegawai

REVIEW

Data → Informasi → Pengetahuan

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

Informasi Akumulasi Kehadiran Pegawai Per Bulan

REVIEW

Data → Informasi → Pengetahuan

	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

Pola Kehadiran Mingguan Pegawai

REVIEW

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**

DATA MINING ?

“We are drowning in data, but starving for knowledge”



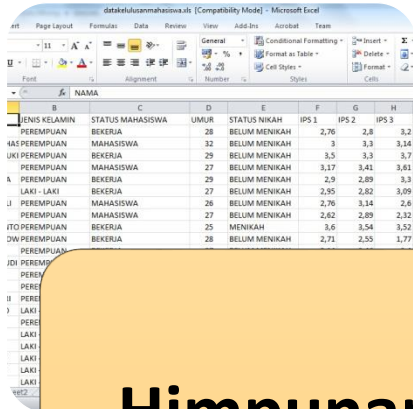
DATA MINING ?

Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan.

Darly Pregibon (2011) menyatakan bahwa data mining adalah campuran dari statistik, kecerdasan buatan, dan riset basis data.

Pramudiono (2006) mengartikan data mining sebagai serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

DATA MINING ?



	A	B	C	D	E	F	G	H
		NAMA						
1		JENIS KELAMIN	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2	IPS 3
2		PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,76	2,8	3,2
3		PEREMPUAN	MAHASISWA	32	BELUM MENIKAH	3	3,3	3,14
4		PEREMPUAN	BEKERJA	29	BELUM MENIKAH	3,5	3,3	3,7
5		PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3,17	3,41	3,61
6		PEREMPUAN	BEKERJA	29	BELUM MENIKAH	2,9	2,89	3,3
7		LAKI - LAKI	BEKERJA	27	BELUM MENIKAH	2,95	2,82	3,09
8		PEREMPUAN	MAHASISWA	28	BELUM MENIKAH	2,76	3,14	2,6
9		PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	2,62	2,89	2,32
10		LAKI - LAKI	BEKERJA	25	MENIKAH	3,6	3,54	3,52
11		PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,71	2,55	1,77

**Himpunan
Data**

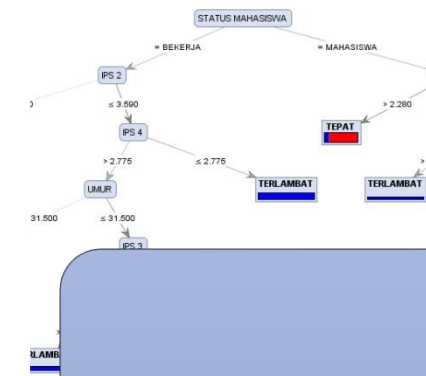


$$f(x) dx = \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} \cdot k\right)$$

$$= \left(-m \frac{1}{\omega} \sin(\omega t) \right) \left[l - \frac{r^2}{4l} + r \left(\cos(\omega t) + \frac{r}{4l} \cos(2\omega t) \right) \right]$$

$$= R_1 e^{\left(-\zeta + \sqrt{\zeta^2 - 1} \right) \omega t} - R_2 e^{\left(-\zeta - \sqrt{\zeta^2 - 1} \right) \omega t}$$

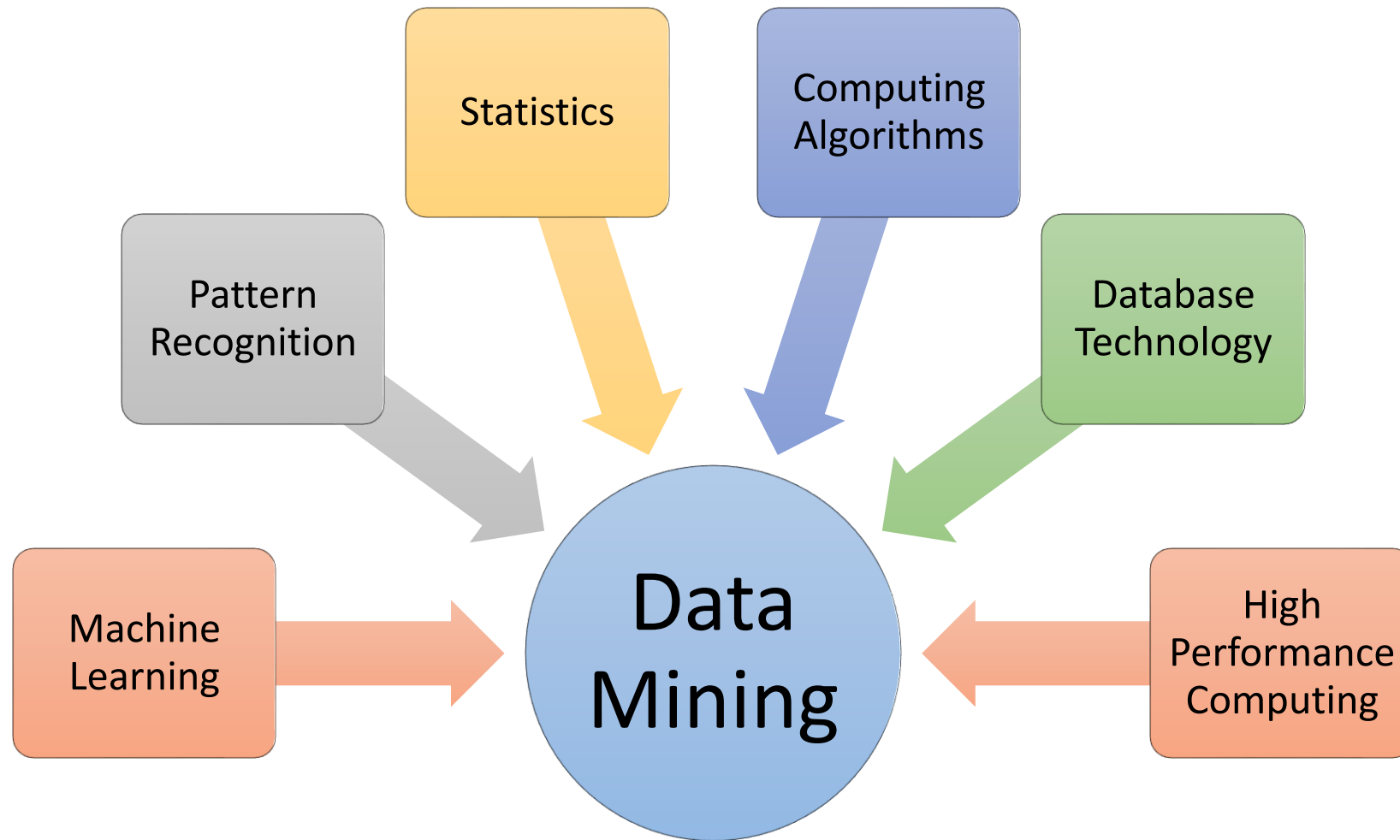
**Metode Data
Mining**



Pengetahuan

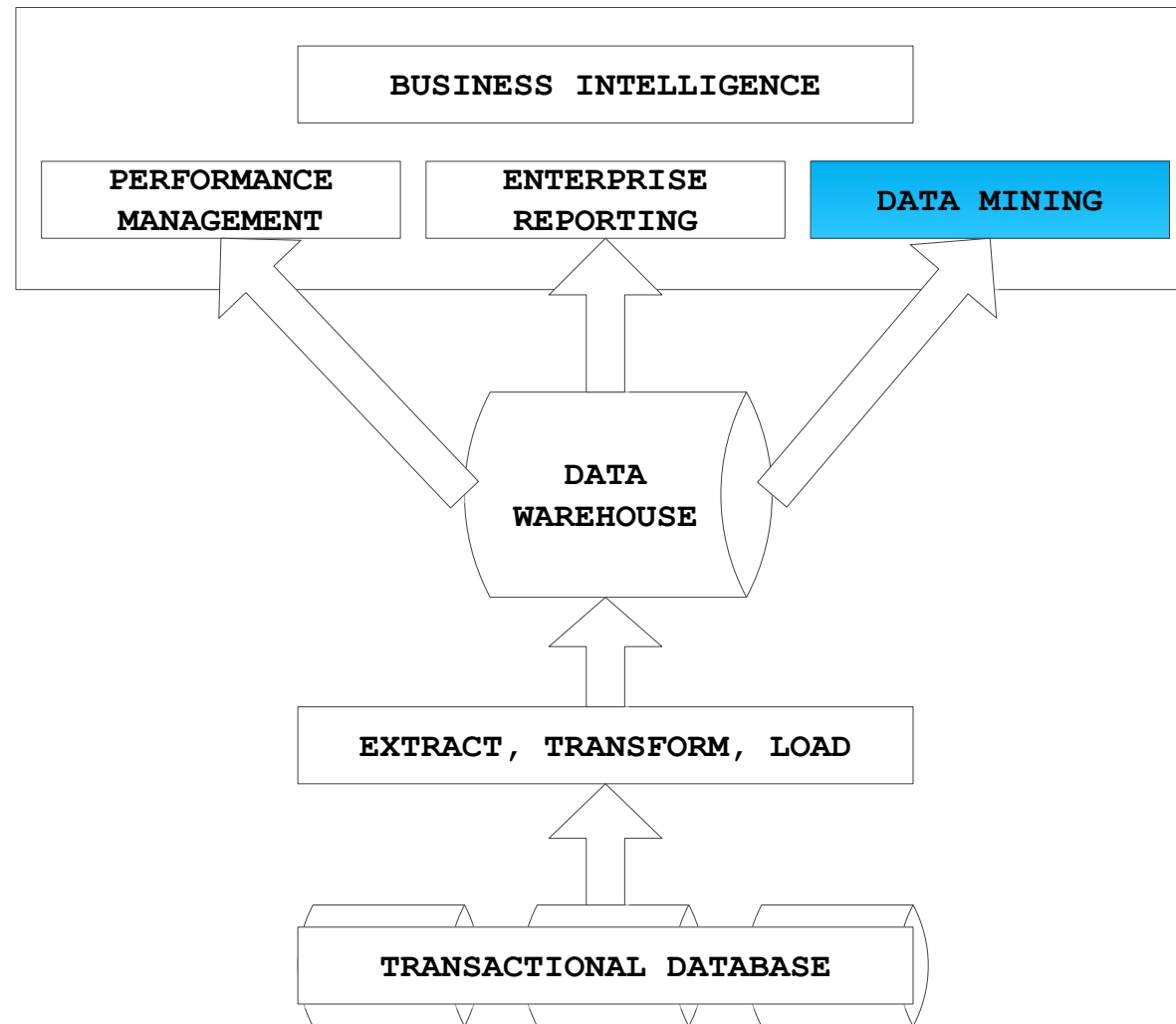
Proses ekstraksi dari DATA ke PENGETAHUAN (pola, rumus, aturan, model) dengan beberapa teknik dari kumpulan data besar.

DATA MINING ?



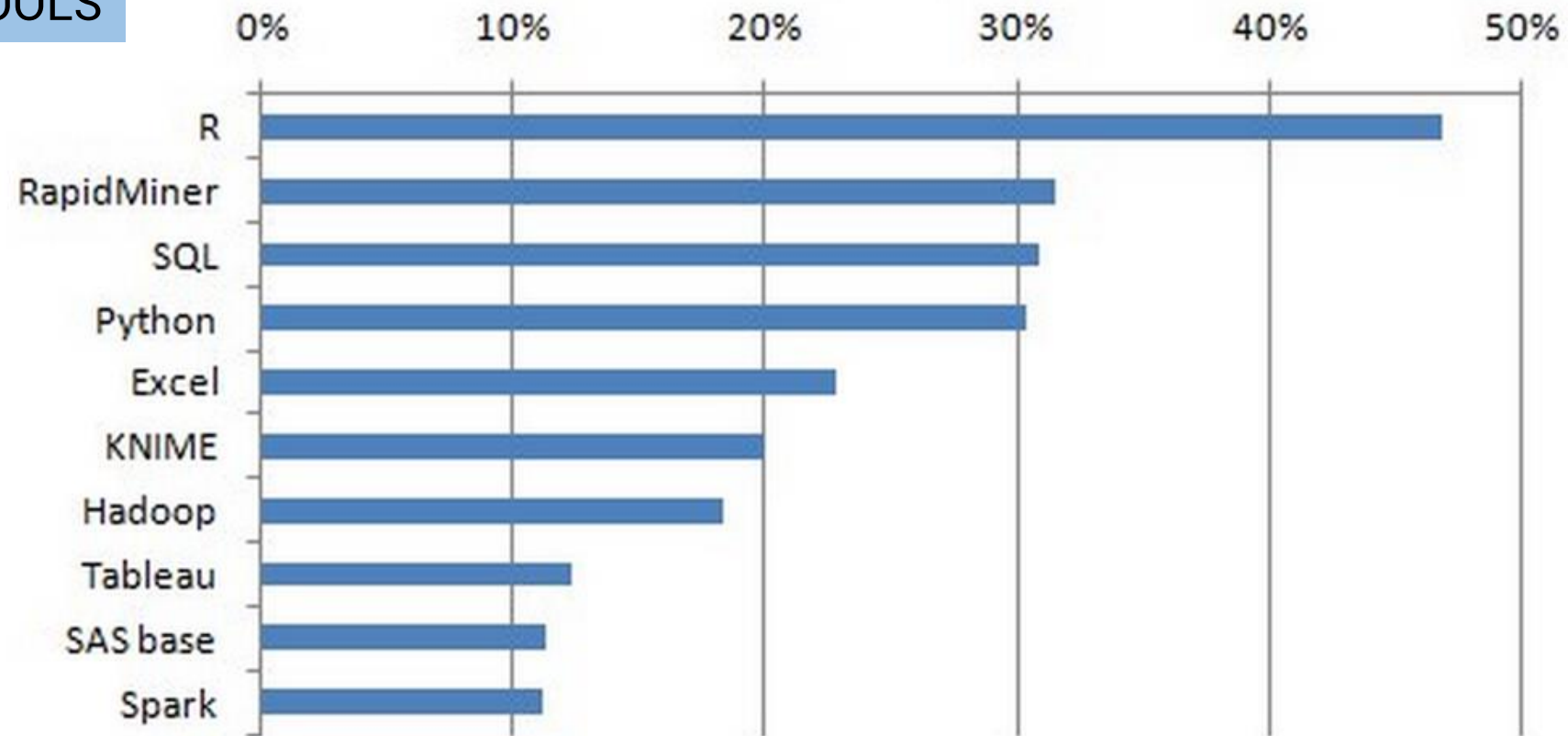
Hubungan dengan Bidang lainnya

DATA MINING ?



Hubungan dengan Bidang lainnya

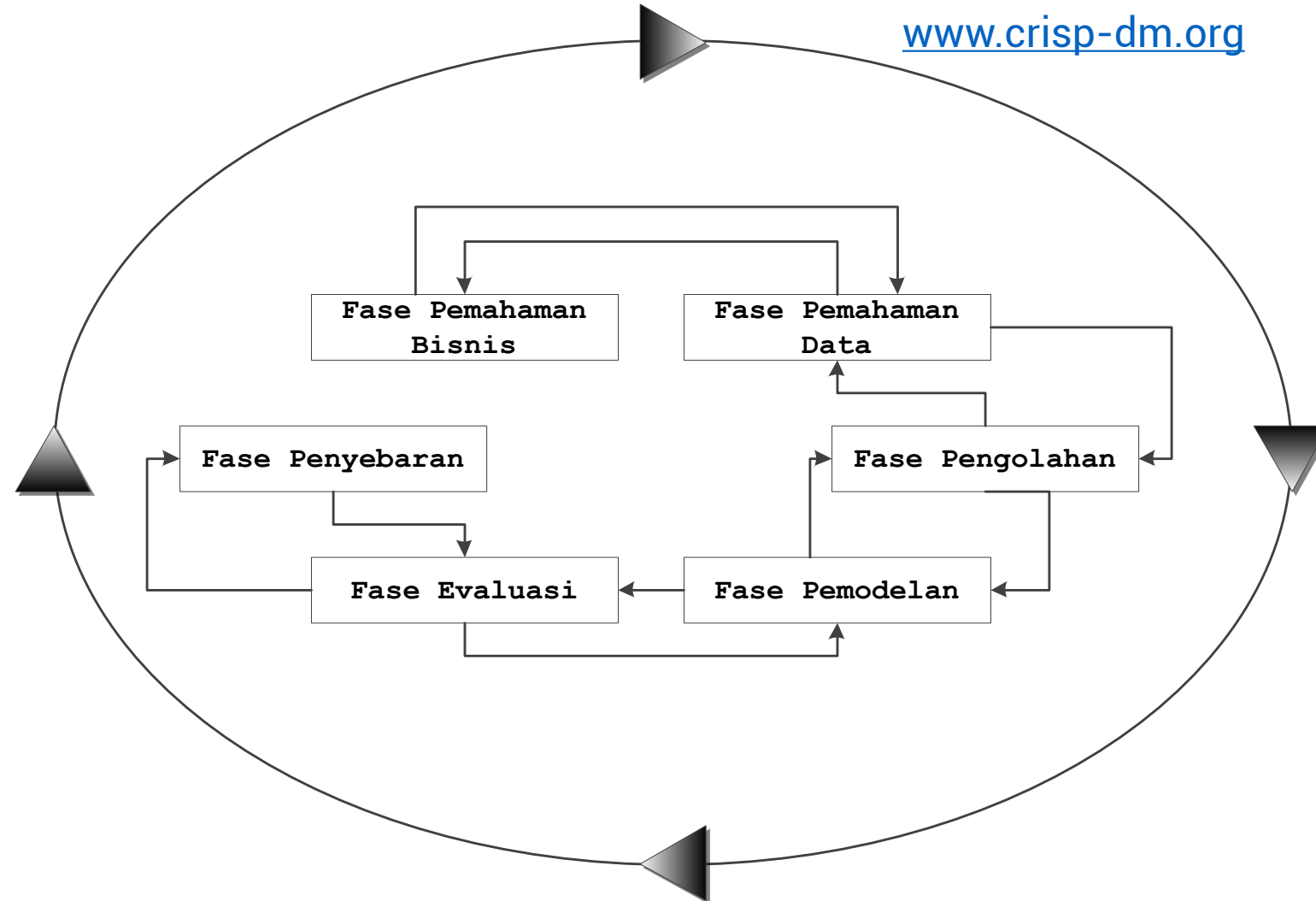
DATA MINING TOOLS



<http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html>

Penggunaan Software Data Mining

www.crisp-dm.org

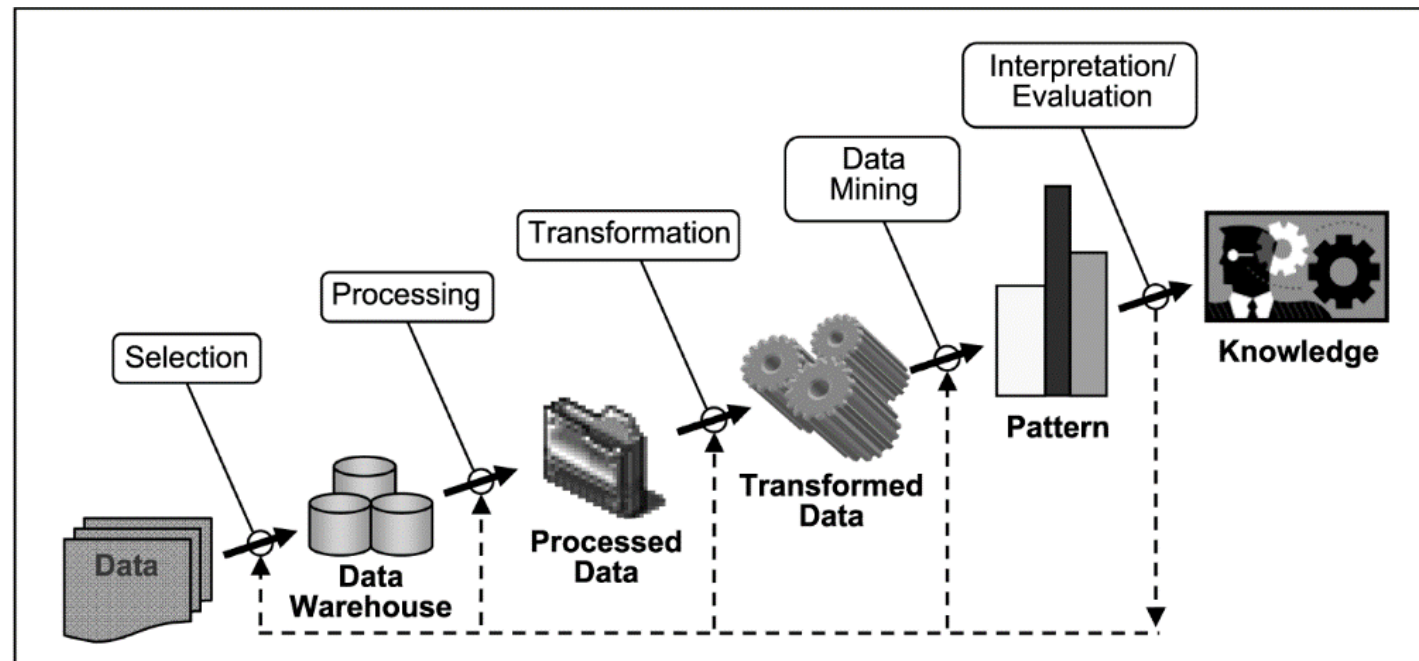


DATA MINING PROCESSES

Tiga langkah utama dalam proses data mining (Gonunescu, 2011)

1. Ekplorasi/pemrosesan awal data
2. Membangun model dan melakukan validasi terhadapnya
3. Penerapan

Peran Data Mining dalam Knowledge Discovery in Database (KDD)



DATA MINING TECHNIQUES

Classification,
Clustering,
Association,
Anomaly,
Prediction,
Estimation
Regression,
Sequential Pattern,
Deviation Detection
DII

C4.5, Nearest Neighbor, A Priori, Fuzzy C-Means,
Bayesian Classification, C4.5, K-Means, SVM, EM,
PageRank, AdaBoost, kNN, CART, dll

CLASSIFICATION

Kasifikasi (classification) digunakan untuk pembuatan model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan variabel yang baru.

Algoritma:

Decision Tree Induction (C4.5)

Nearest-Neighbor

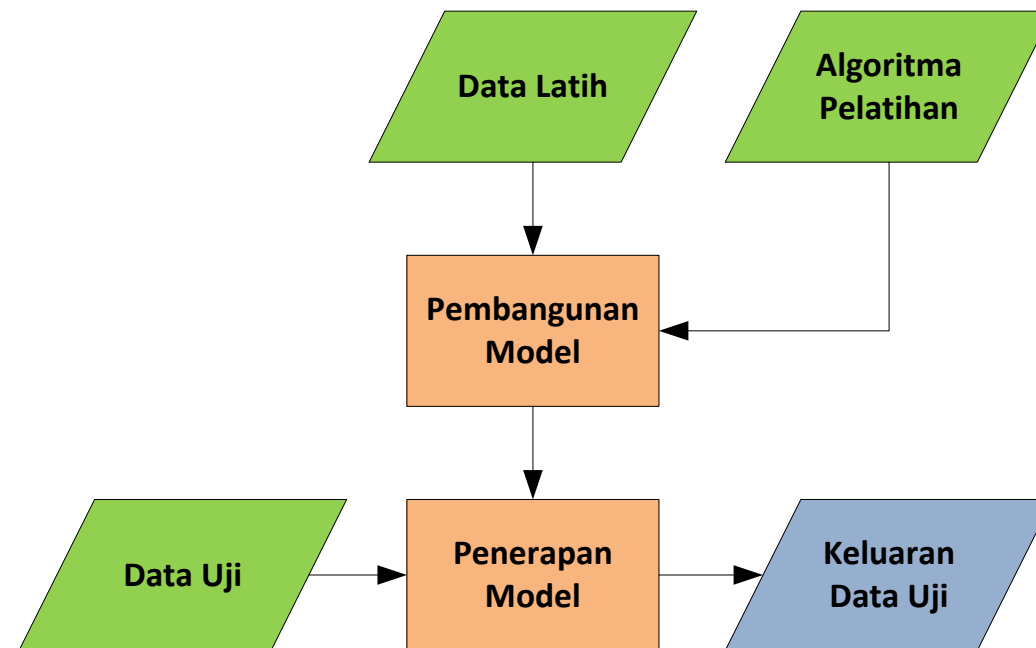
Bayesian Classification

Neural Network

Model Evaluation and Selection

Techniques to Improve Classification Accuracy: Ensemble Methods

dll



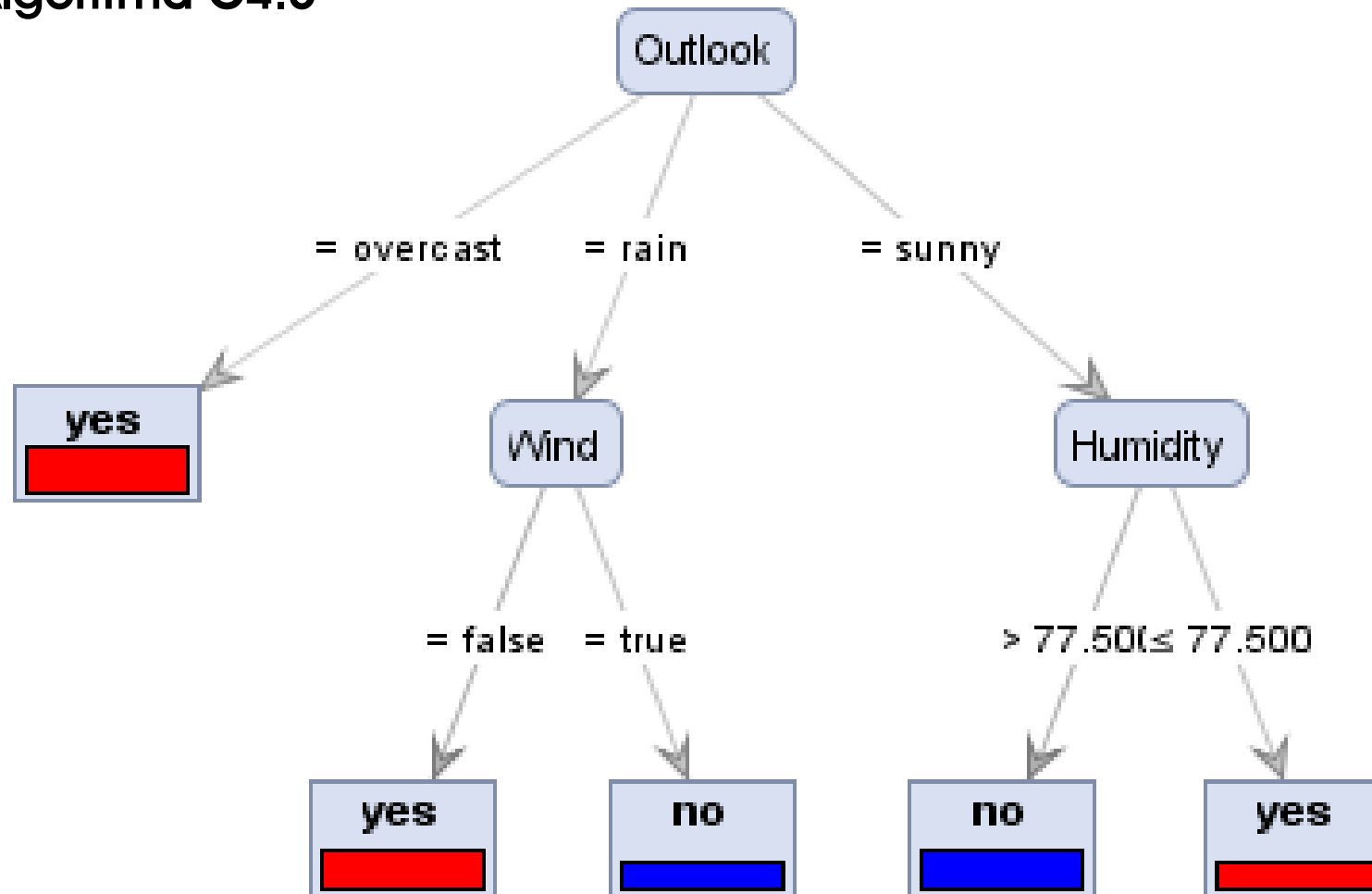
CLASSIFICATION

Data Keputusan Bermain Tenis (data set)

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

CLASSIFICATION

Pohon Keputusan Bermain Tenis (model) Algoritma C4.5



CLASSIFICATION

Seleksi Kondisi untuk Rekomendasi Bermain Tenis (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

CLUSTERING

Penklusteran (clustering) digunakan untuk melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan karakteristik masing-masing data pada kelompok-kelompok yang ada.

Algoritma:

Cluster Analysis: Basic Concepts

Partitioning Methods

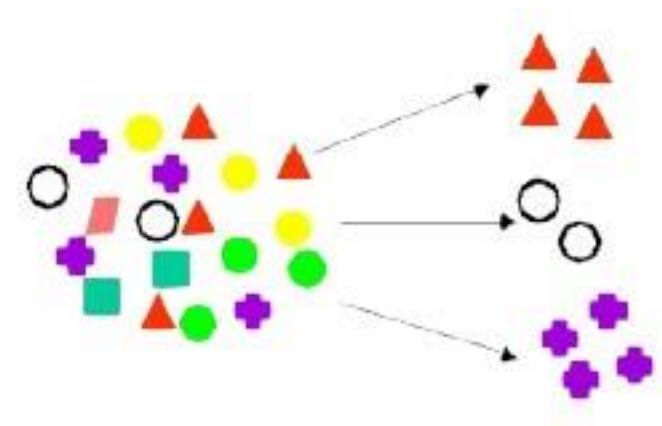
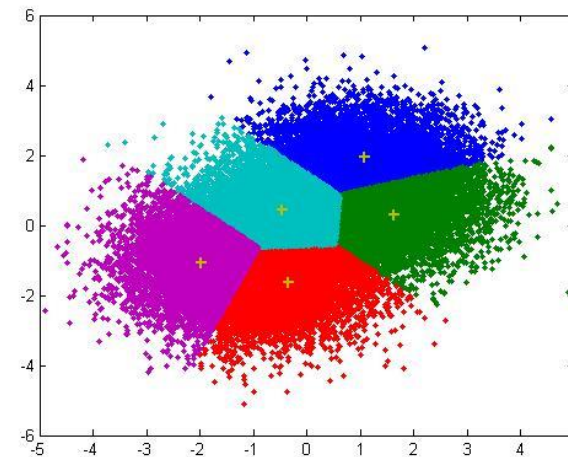
Hierarchical Methods

Density-Based Methods

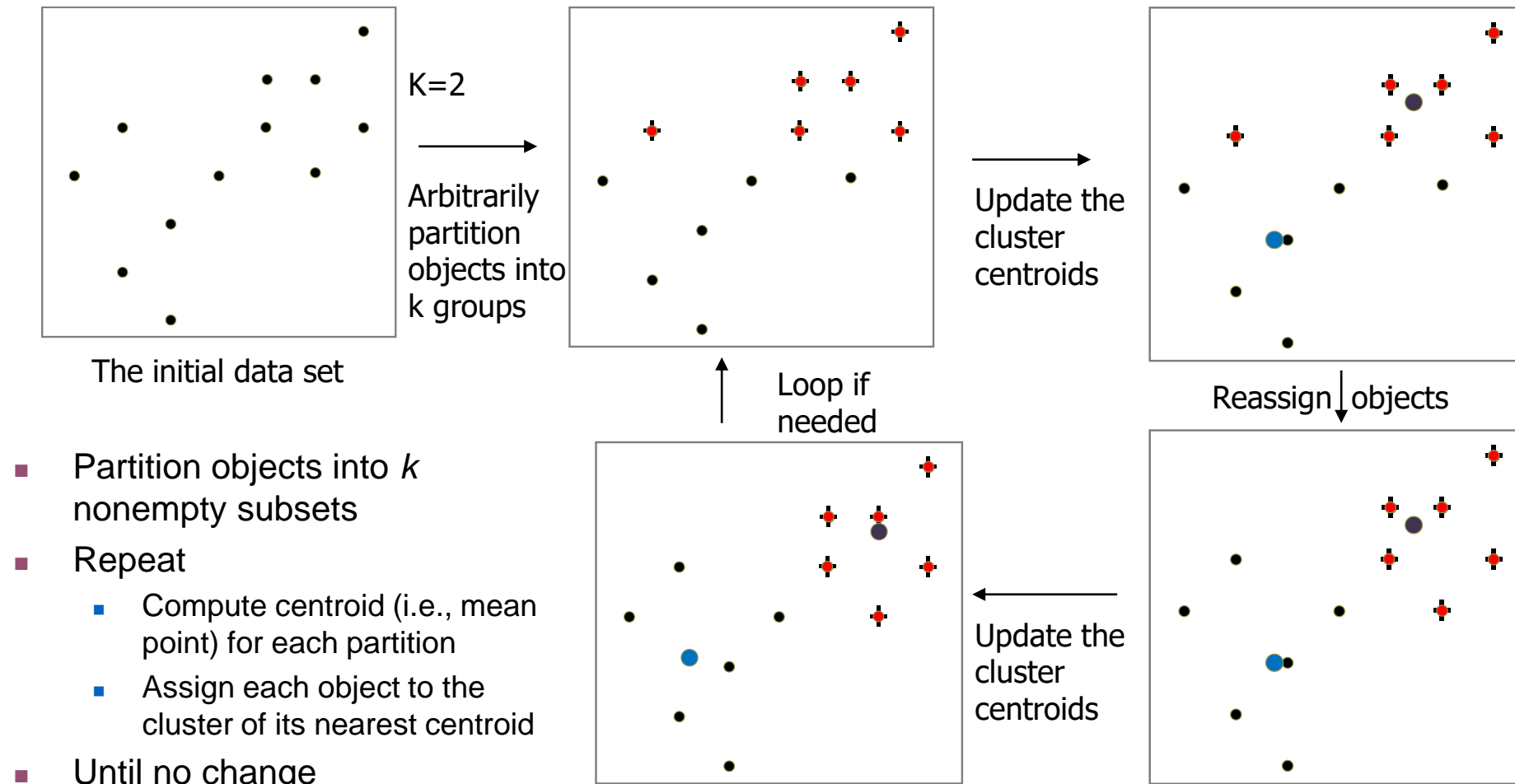
Grid-Based Methods

Evaluation of Clustering

dll



CLUSTERING



CLUSTERING

Menentukan Strategi Marketing Universitas

Himpunan Data yang digunakan ...

1. Data Mahasiswa yang Telah Lulus

No.	Name	Jurusan	Kota asal	IPK
-----	------	---------	-----------	-----

2. Data Kota Asal Mahasiswa yang Telah Lulus (transformasi data)

Wilayah	Frekuensi	Inisial
DKI Jakarta	84	1
Jawa Barat	82	2

3. Data Jurusan Mahasiswa yang Telah Lulus (transformasi data)

Major	Singkatan	Frekuensi	Inisial
Accounting	ACC	46	1
Management, concentration in International Business	IB	37	2

CLUSTERING

Menentukan Strategi Marketing Universitas

Pemodelan menggunakan algoritma K-Means ...

1. Menentukan jumlah cluster = 3
2. Menentukan nilai centroid dari setiap cluster

Titik Pusat awal	Nama	Jurusan	Kota asal	IPK
Cluster 1	DALLY TEGUH SESARIO	9	3	2.94
Cluster 2	HERVINA JULIANA	1	1	3.18
Cluster 3	PASCAL MUHAMMADI	1	2	3.15

3. Petakan setiap data pada centroid cluster (cari yang terdekat)

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \dots + (X_{ki} - X_{kj})^2}$$

dimana:

$D(i,j)$ = Jarak data ke i ke pusat cluster j

X_{ki} = Data ke i pada atribut data ke k

X_{kj} = Titik pusat ke j pada atribut ke k

CLUSTERING

Menentukan Strategi Marketing Universitas

Pemodelan menggunakan algoritma K-Means ...

Hasil pemetaan setiap data ke setiap cluster

No	Nama	Jurusan	Kota asal	IPK	Jarak Ke			Jarak terdekat ke cluster
					C 1	C 2	C 3	
1	ADE SUPRYAN STEFANUS	14	1	3.16	5.390	13.000	13.038	1
2	ADELINA GANARDI PUTRI HARDI	1	5	3.22	8.251	4.000	3.001	3
3	ADELINE DEWITA	4	2	3.29	5.111	3.164	3.003	3
4	ADIPUTRA	2	1	2.83	7.281	1.059	1.450	2
5	AFRIESKA LAURA TRISYANA	3	1	3.15	6.328	2.000	2.236	2

- Hitung kembali pusat cluster yang baru berdasarkan rata-rata anggota yang ada pada cluster tersebut
- Setelah didapatkan centroid yang baru dari setiap cluster, lakukan kembali dari langkah ketiga hingga centroid dari setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain

CLUSTERING

Menentukan Strategi Marketing Universitas

Hasil Analisis Clustering

<p>Cluster 1 terdiri dari 70 orang, yang berasal dari jurusan</p> <p>IT = 19 orang MKT = 15 orang VCD = 12 orang HTM = 9 orang EE = 6 orang BA = 4 orang IR = 2 orang MGT = 1 orang IS = 1 orang HRM = 1 orang</p> <p>Dan berasal dari Wilayah:</p> <p>DKI Jakarta = 30 orang Jawa Barat = 20 orang Sumatera Utara = 12 orang Sulawesi = 2 orang Jawa Timur = 2 orang Sumatera Selatan = 2 orang Bali = 1 orang Kalimantan = 1 orang</p> <p>Dengan rata-rata nilai IPK 3.2</p>	<p>Cluster 2 terdiri dari 132 orang, yang berasal dari aktifis</p> <p>ACC = 39 orang IB = 30 orang BF = 22 orang PR = 21 orang IE = 20 orang</p> <p>Dan berasal dari Wilayah:</p> <p>Jawa Barat = 62 orang DKI Jakarta = 54 orang Sumatera Utara = 16 orang</p> <p>Dengan rata-rata nilai IPK 3.25</p>	<p>Cluster 3 terdiri dari 41 orang, yang berasal dari jurusan:</p> <p>PR = 14 orang ACC = 7 orang IB = 7 orang BF = 6 orang E-3 = 3 orang MKT = 3 orang IT = 1 orang</p> <p>Dan berasal dari Wilayah:</p> <p>Sulawesi = 12 orang. Jawa Timur = 11 orang Sumatera Selatan = 11 orang Bali = 7 orang</p> <p>Dengan rata-rata nilai IPK 3.31</p>
--	--	--

CLUSTERING

Menentukan Strategi Marketing Universitas **Strategi Promosi yang Dihasilkan**

1. Promosi Dengan Mengirim Tim Marketing yang Sesuai dengan Jurusan yang Paling Banyak Diminati.
2. Promosi Pada Kota Berdasarkan Tingkat Akademik dari Calon Mahasiswa.

ASSOCIATION

Asosiasi (association) digunakan untuk menemukan pola yang mendeteksi kumpulan atribut-atribut yang muncul bersamaan dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut.

Biasa disebut dengan **affinity analysis** atau **market basket analysis**.

Algoritma:

A Priori

FP-Growth

GRI

dll

*Customers who bought this item ...
also bought ...*



ASSOCIATION

Data Transaksi (Format Tabular)

Algoritma A Priori

ExampleSet (12 examples, 0 special attributes, 10 regular attributes)										
Row No.	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
3	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
6	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
8	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
9	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
10	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
11	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0

ASSOCIATION

AssociationRules

Association Rules

```
[Aqua] --> [Sabun] (confidence: 0.800)
[Sprei] --> [Kopi] (confidence: 0.800)
[Aqua] --> [Kopi] (confidence: 0.800)
[Sabun, Kopi] --> [Gula] (confidence: 0.800)
[Sabun, Gula] --> [Kopi] (confidence: 0.800)
[Sprei] --> [Kopi, Gula] (confidence: 0.800)
[Gula, Sprei] --> [Kopi] (confidence: 0.800)
[Sampo] --> [Sabun] (confidence: 0.857)
[Gula] --> [Kopi] (confidence: 0.857)
[Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sampo] (confidence: 1.000)
[Boneka] --> [Sampo] (confidence: 1.000)
[Sprei] --> [Gula] (confidence: 1.000)
[Popok] --> [Gula] (confidence: 1.000)
[Boneka] --> [Gula] (confidence: 1.000)
[Boneka] --> [Sprei] (confidence: 1.000)
[Sampo, Gula] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Sampo] (confidence: 1.000)
[Sampo, Sprei] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Celana] --> [Sampo] (confidence: 1.000)
[Sampo, Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Boneka] --> [Sampo] (confidence: 1.000)
[Sampo, Boneka] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Gula] (confidence: 1.000)
[Sabun, Popok] --> [Gula] (confidence: 1.000)
[Boneka] --> [Sabun, Gula] (confidence: 1.000)
[Sabun, Boneka] --> [Gula] (confidence: 1.000)
[Gula, Boneka] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Boneka] (confidence: 1.000)
[Boneka] --> [Sabun, Sprei] (confidence: 1.000)
[Sabun, Boneka] --> [Sprei] (confidence: 1.000)
[Sprei, Boneka] --> [Sabun] (confidence: 1.000)
```

Contoh, pada hari kamis malam, 1000 pelanggan telah melakukan belanja di supermarket ABC, dimana: 200 orang membeli **Teh**, dan dari 200 orang yang membeli **Teh**, 50 orangnya membeli **Gula**.

Jadi, association rule menjadi, “**Jika membeli Teh, maka membeli Gula**”, dengan nilai **support** =

$200/1000 \times 100\% = 20\%$ dan nilai **confidence** =

$50/200 \times 100\% = 25\%$

ESTIMATION

Estimasi Waktu Pengiriman Pizza

Regresi Linier

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	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

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

Pengetahuan (Rumus)

NEXT

PREPROCESSING

