PROGRAM STUDI
TEKNIK INFORMATIKA – S1
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO

MATA KULIAH
DATA MINING

Technology vector created by sentavio - www.freepik.com

DATA MINING

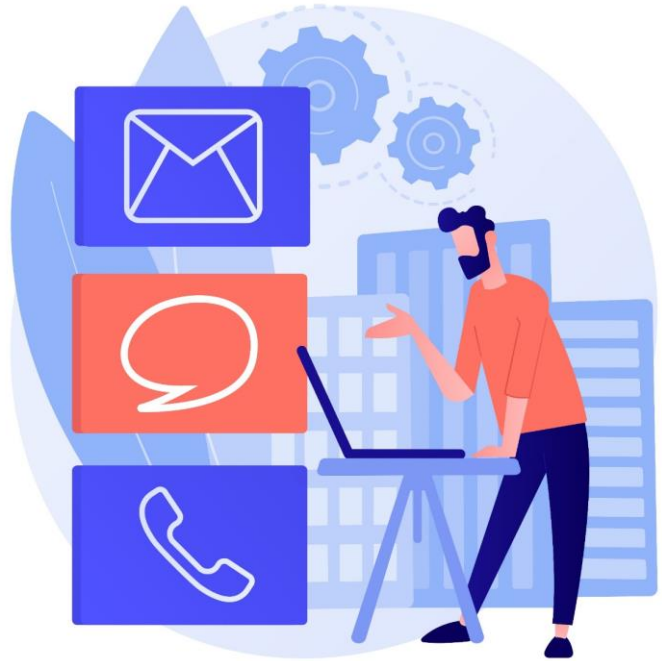
"Data untuk Data Mining"

TIM PENGAMPU DOSEN DATA MINING

2023

Kontak Dosen

- Junta Zeniarja, M.Kom
- Email: junta@dsn.dinus.ac.id
- Youtube : https://www.youtube.com/JuntaZeniarja
- Scholar : http://bit.do/JuntaScholar



Data untuk Data Mining

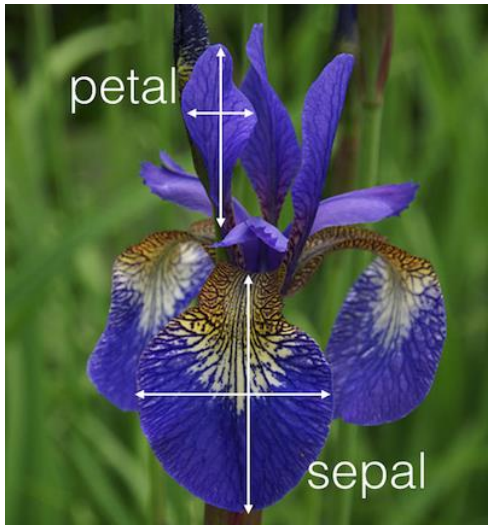
Pertemuan 2

Dataset (Koleksi Data)

Attribute/Feature/Dimension

Class/Label/Target

	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*

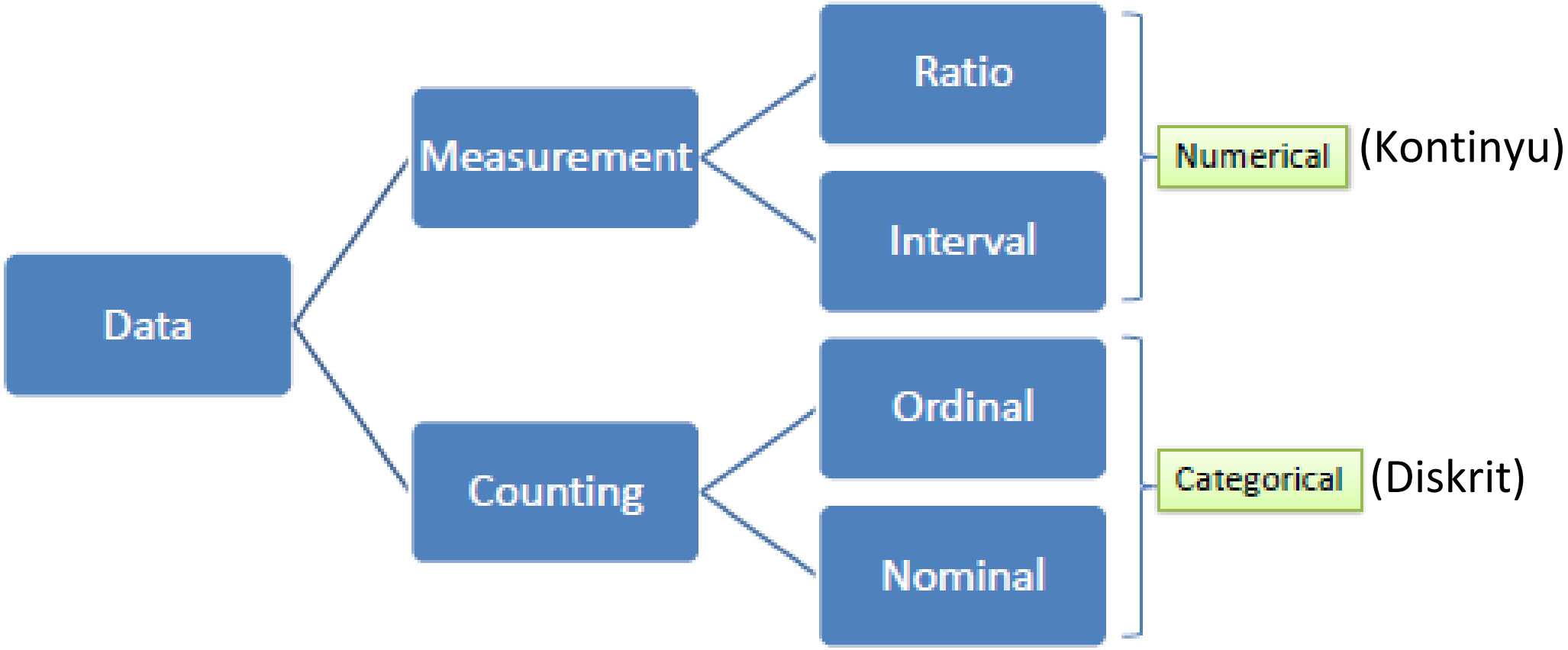


Record/ Object/ Sample/ Tuple/ Data

Nominal

Numerik

Tipe Data

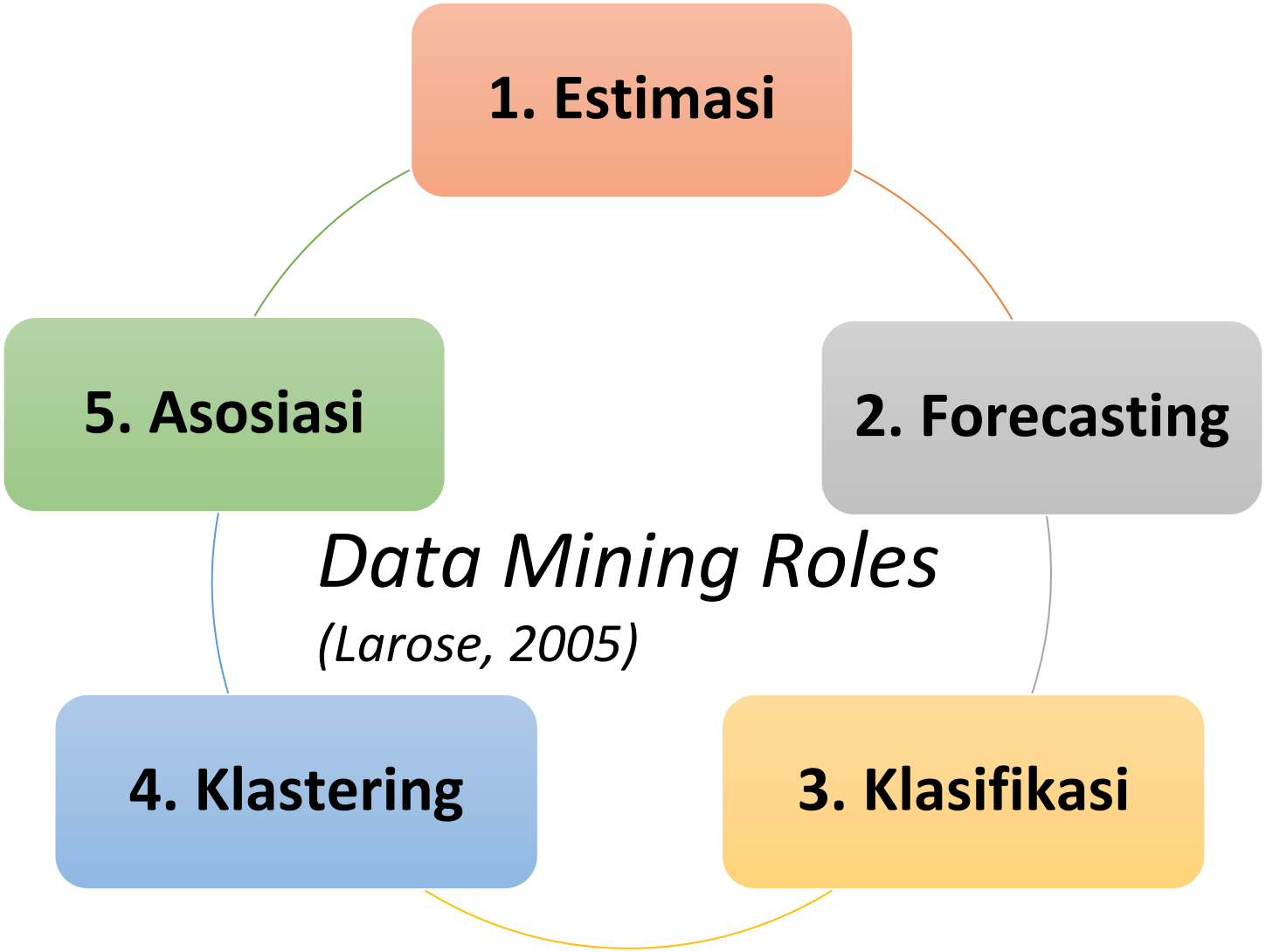

Data
Measurement
Counting
Ratio
Interval
Ordinal
Nominal
Numerical
Categorical
(Kontinyu)
(Diskrit)

Tipe Data	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Mempunyai titik nol yang absolut **(*, /)**	• Umur • Berat badan • Tinggi badan • Jumlah uang	geometric mean, harmonic mean, percent variation
Interval (Jarak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Tidak mempunyai titik nol yang absolut **(+, -)**	• Suhu 0°c-100°c, • Umur 20-30 tahun	mean, standard deviation, Pearson's correlation, t and F tests
Ordinal (Peringkat)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Tetapi diantara data tersebut terdapat hubungan atau berurutan **(<, >)**	• Tingkat kepuasan pelanggan (puas, sedang, tidak puas)	median, percentiles, rank correlation, run tests, sign tests
Nominal (Label)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Menunjukkan beberapa object yang berbeda **(=, ≠)**	• Kode pos • Jenis kelamin • Nomer id karyawan • Nama kota	mode, entropy, contingency correlation, χ^2 test

(Kontinyu)

(Diskrit)

Peran Utama Data Mining



1. Estimasi Waktu Pengiriman Pizza

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

Pembelajaran dengan
Metode Estimasi (*Regresi Linier*)

Waktu Tempuh (T) = 0.48P + 0.23TL + 0.5J

Pengetahuan

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

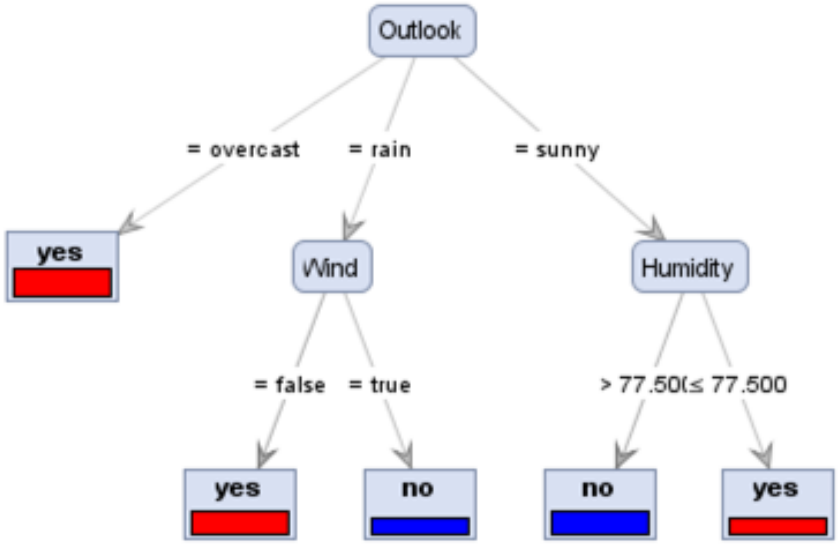
- Linear regression function

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

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)



1. Korelasi dan Asosiasi

1. Rule (Aturan)
   IF ips3=2.8 THEN lulustepatwaktu

2. Cluster (Klaster)

2. Forecasting Harga Saham

Label Time Series

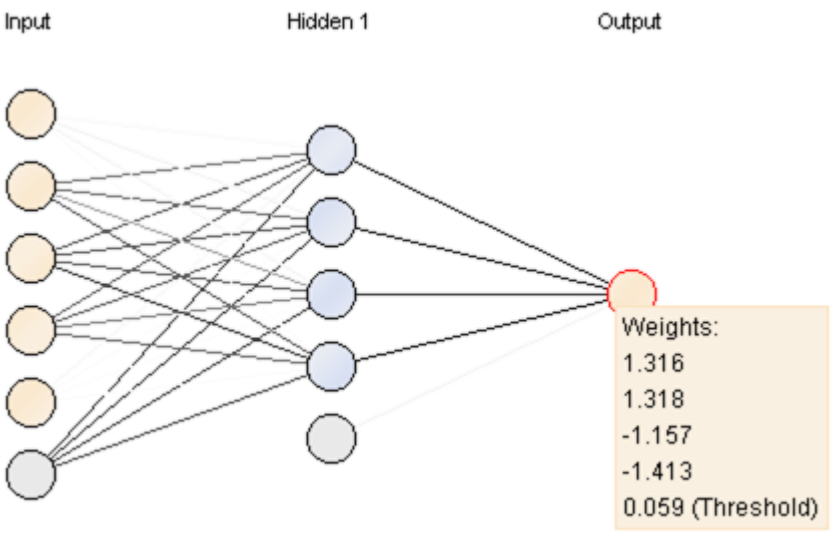
Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	277201000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Dataset harga saham dalam bentuk time series (rentet waktu)

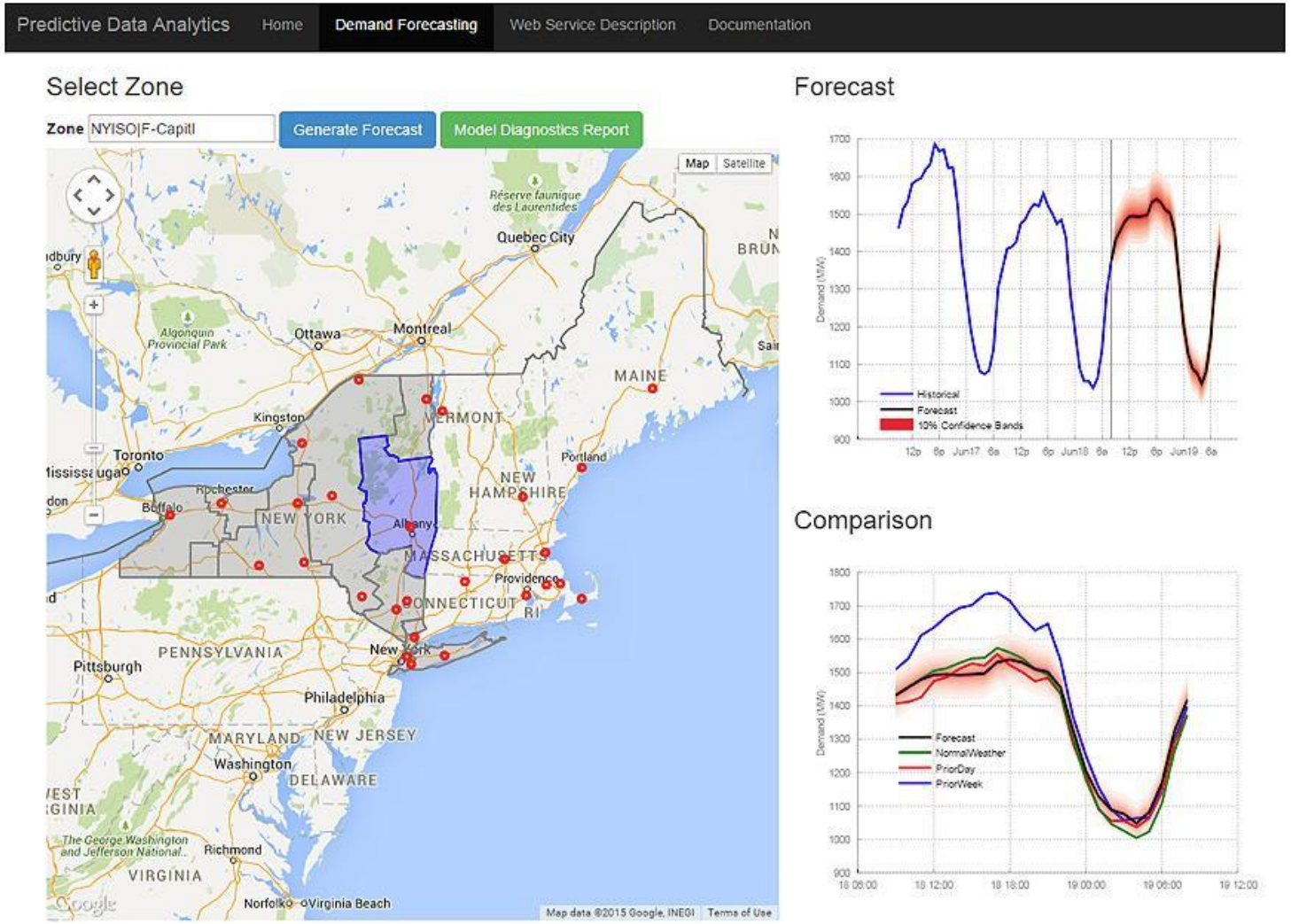
- Pembelajaran dengan
- Metode Forecasting (*Neural Network*)

2. Forecasting Harga Saham [2]

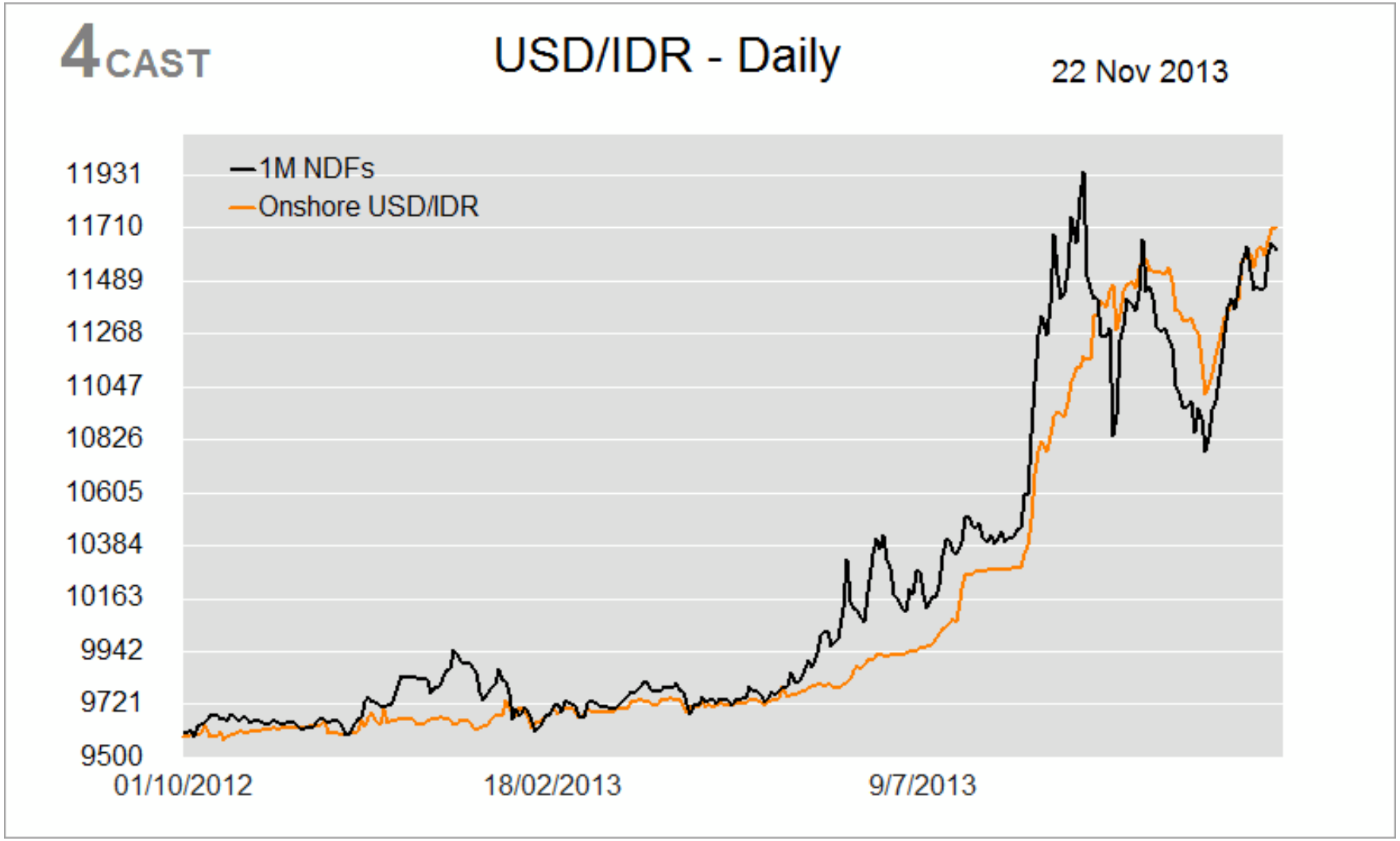
Pengetahuan berupa Rumus Neural Network



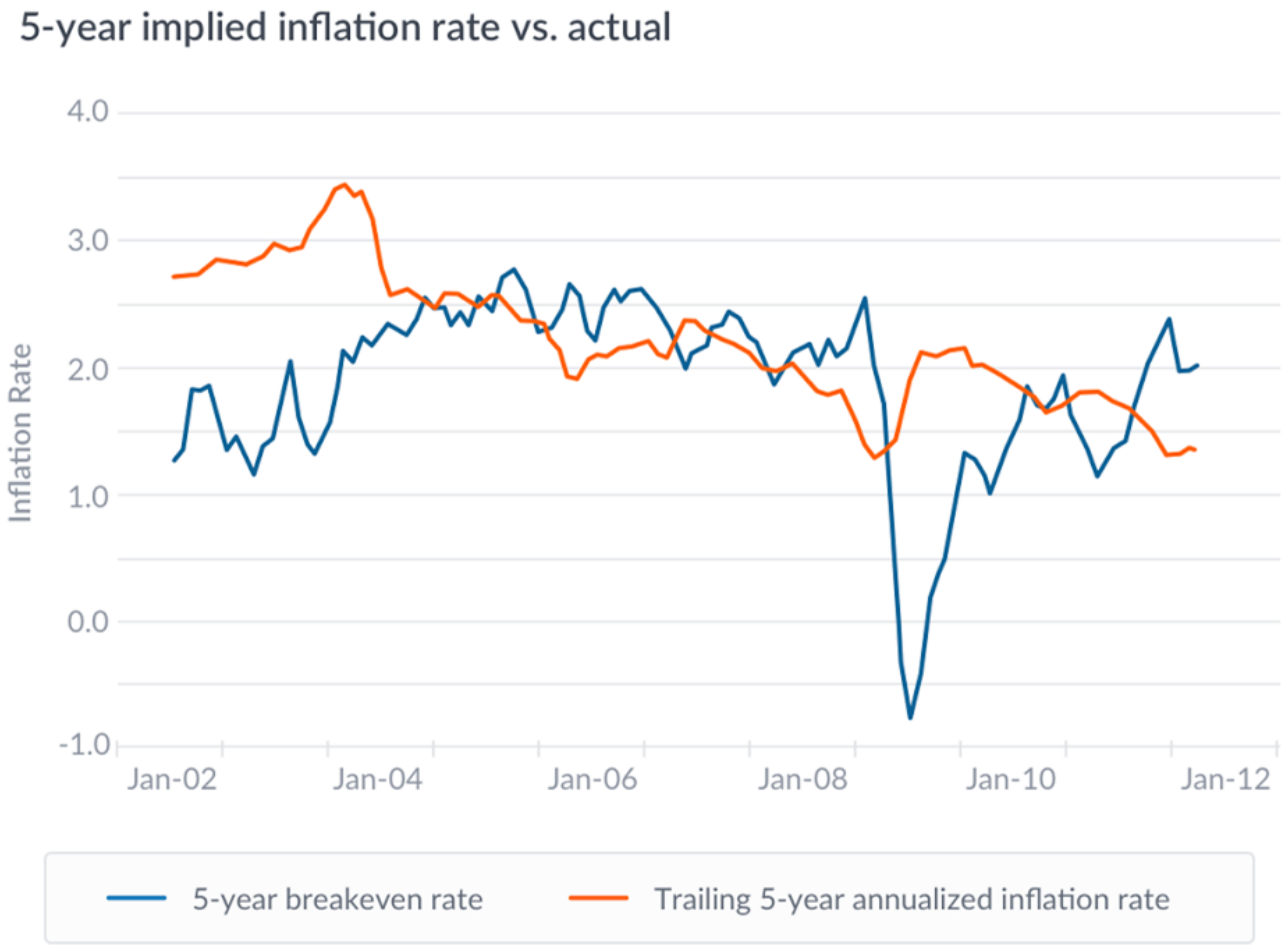
Contoh : Forecasting Cuaca



Contoh : Exchange Rate Forecasting



Contoh : Inflation Rate Forecasting



*Source: Bloomberg

3. Klasifikasi Kelulusan Mahasiswa

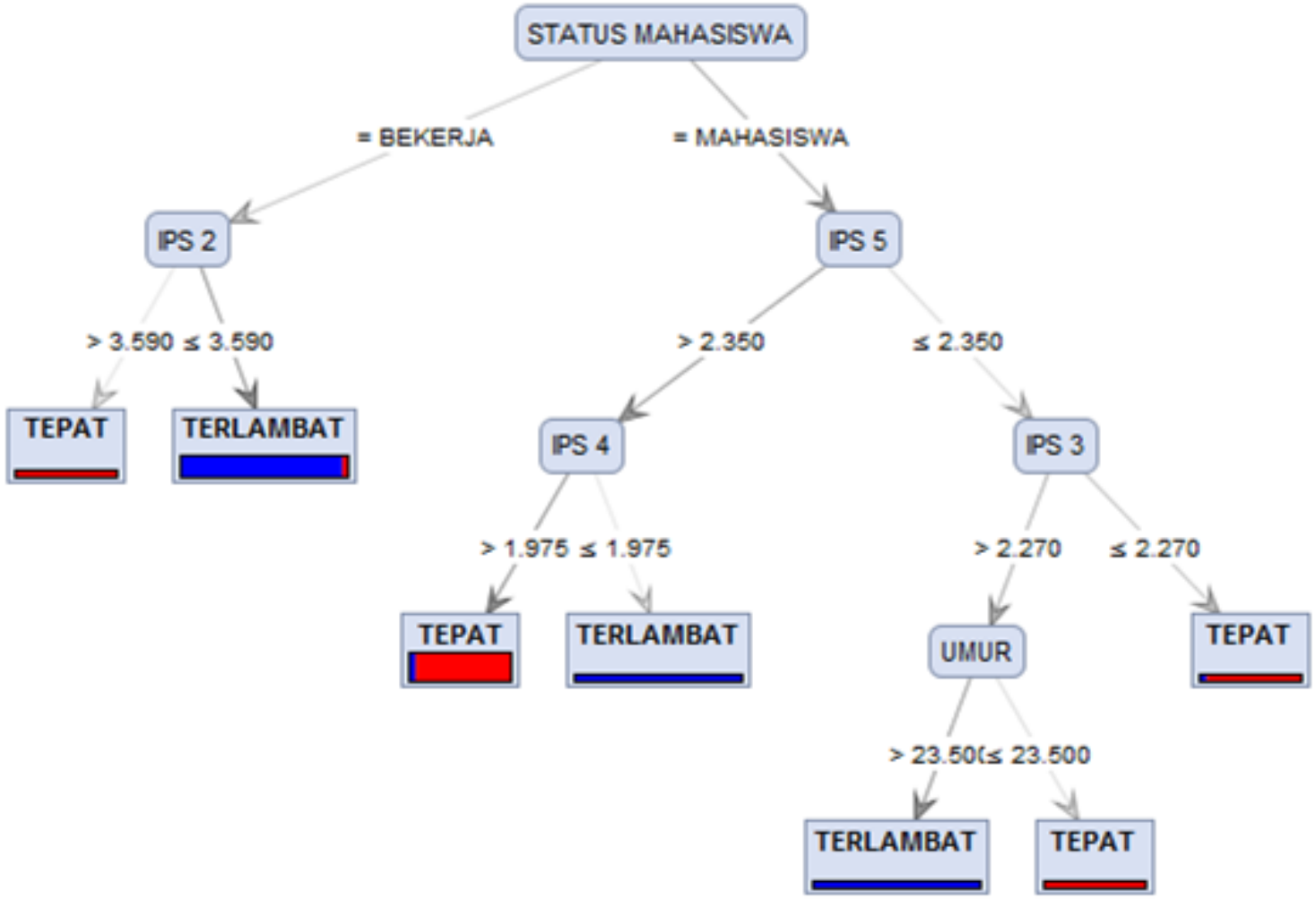
Label

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

- Pembelajaran dengan
- Metode Klasifikasi (*C4.5*)

3. Klasifikasi Kelulusan Mahasiswa [2]

- Pengetahuan Berupa Pohon Keputusan



Contoh: Rekomendasi Main Golf

- **Input**:

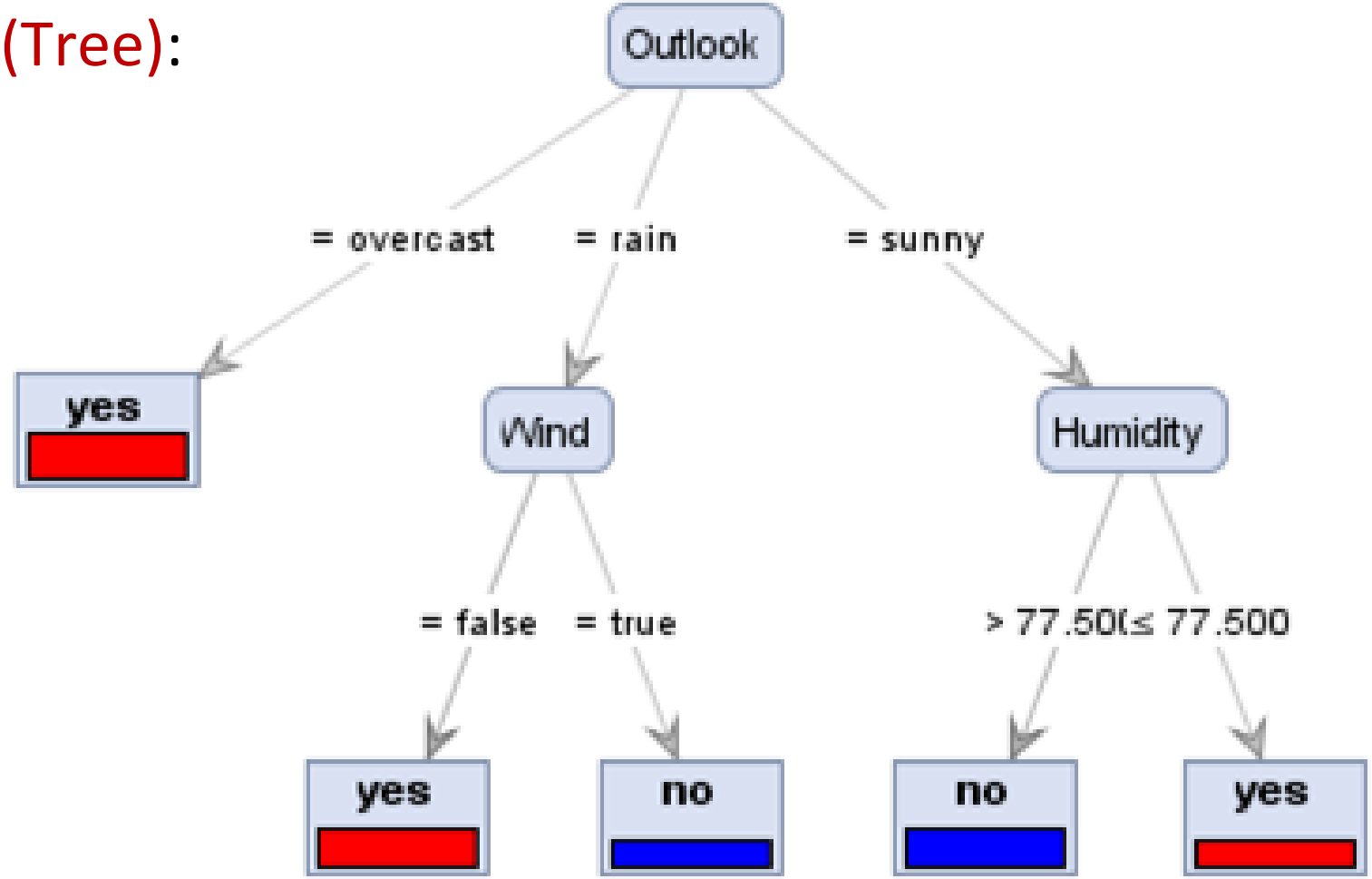
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	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 [2]

- Output (Tree):



Contoh: Rekomendasi Contact Lens

- Input:

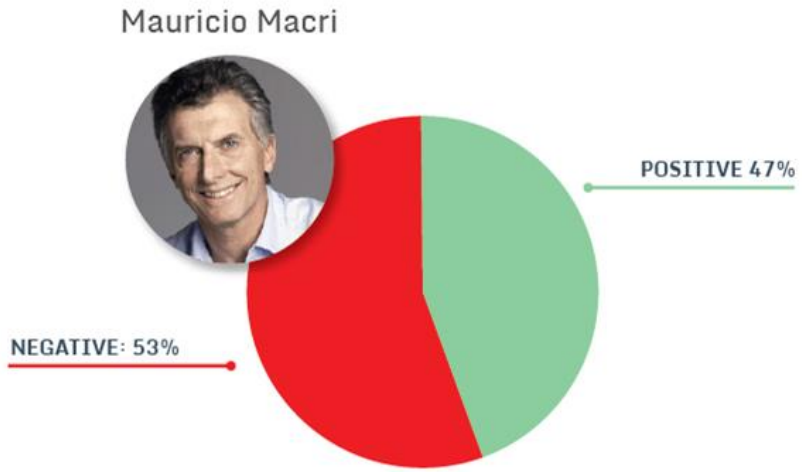
Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young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

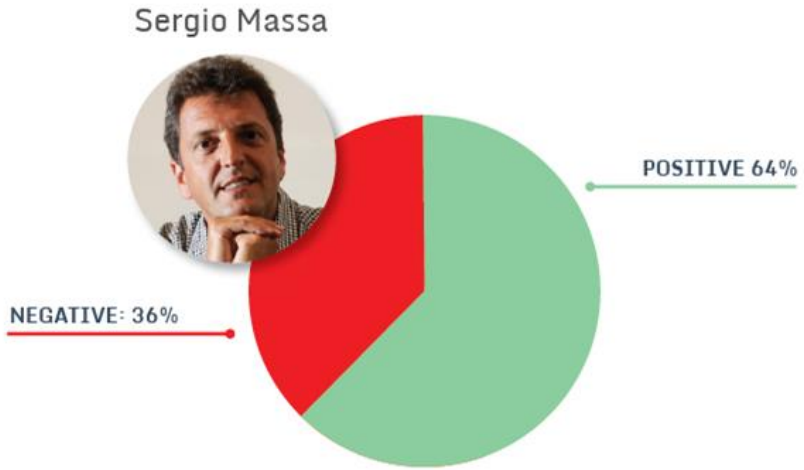
Contoh: Rekomendasi Contact Lens [2]

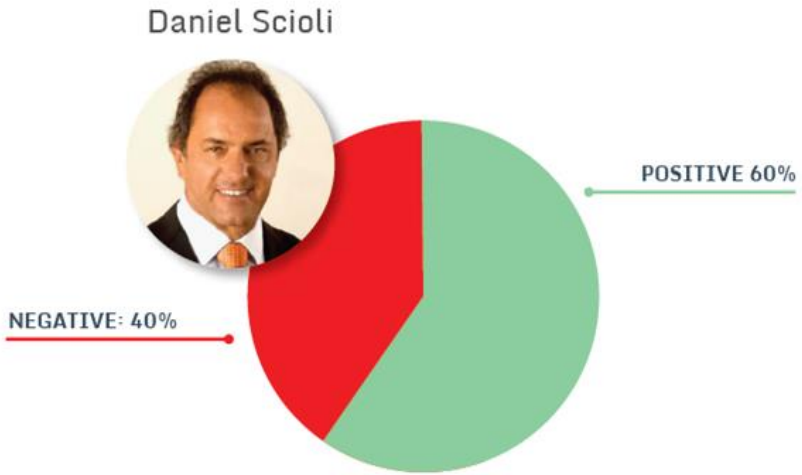
- Output:

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

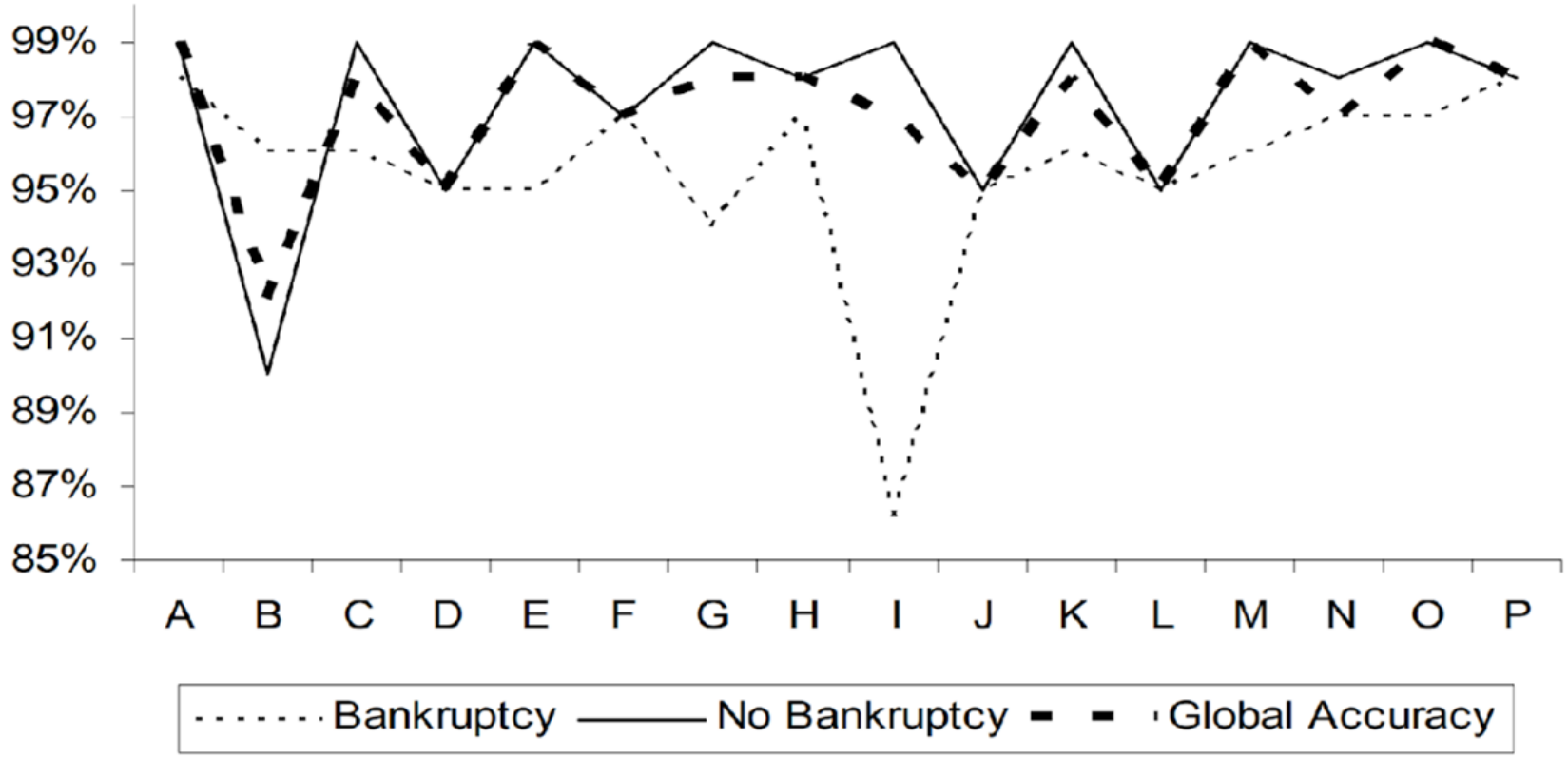
Contoh : Klasifikasi Sentiment Analysis







Contoh : Bankruptcy Prediction



4. Klastering Bunga Iris

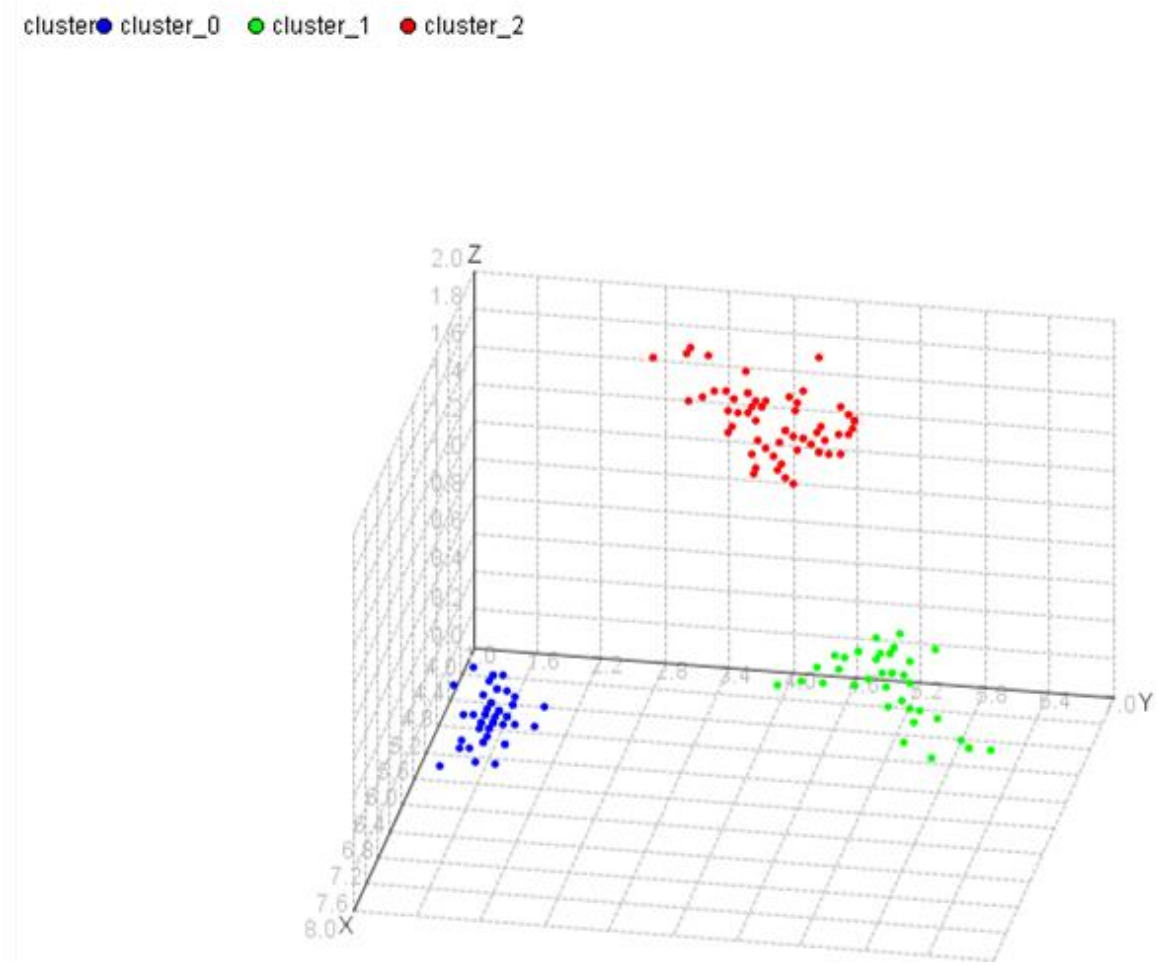
Dataset Tanpa Label

Row No.	id	a1	a2	a3	a4
1	id_1	5.100	3.500	1.400	0.200
2	id_2	4.900	3	1.400	0.200
3	id_3	4.700	3.200	1.300	0.200
4	id_4	4.600	3.100	1.500	0.200
5	id_5	5	3.600	1.400	0.200
6	id_6	5.400	3.900	1.700	0.400
7	id_7	4.600	3.400	1.400	0.300
8	id_8	5	3.400	1.500	0.200
9	id_9	4.400	2.900	1.400	0.200
10	id_10	4.900	3.100	1.500	0.100
11	id_11	5.400	3.700	1.500	0.200

- Pembelajaran dengan
- Metode Klastering (*K-Means*)

4. Klastering Bunga Iris [2]

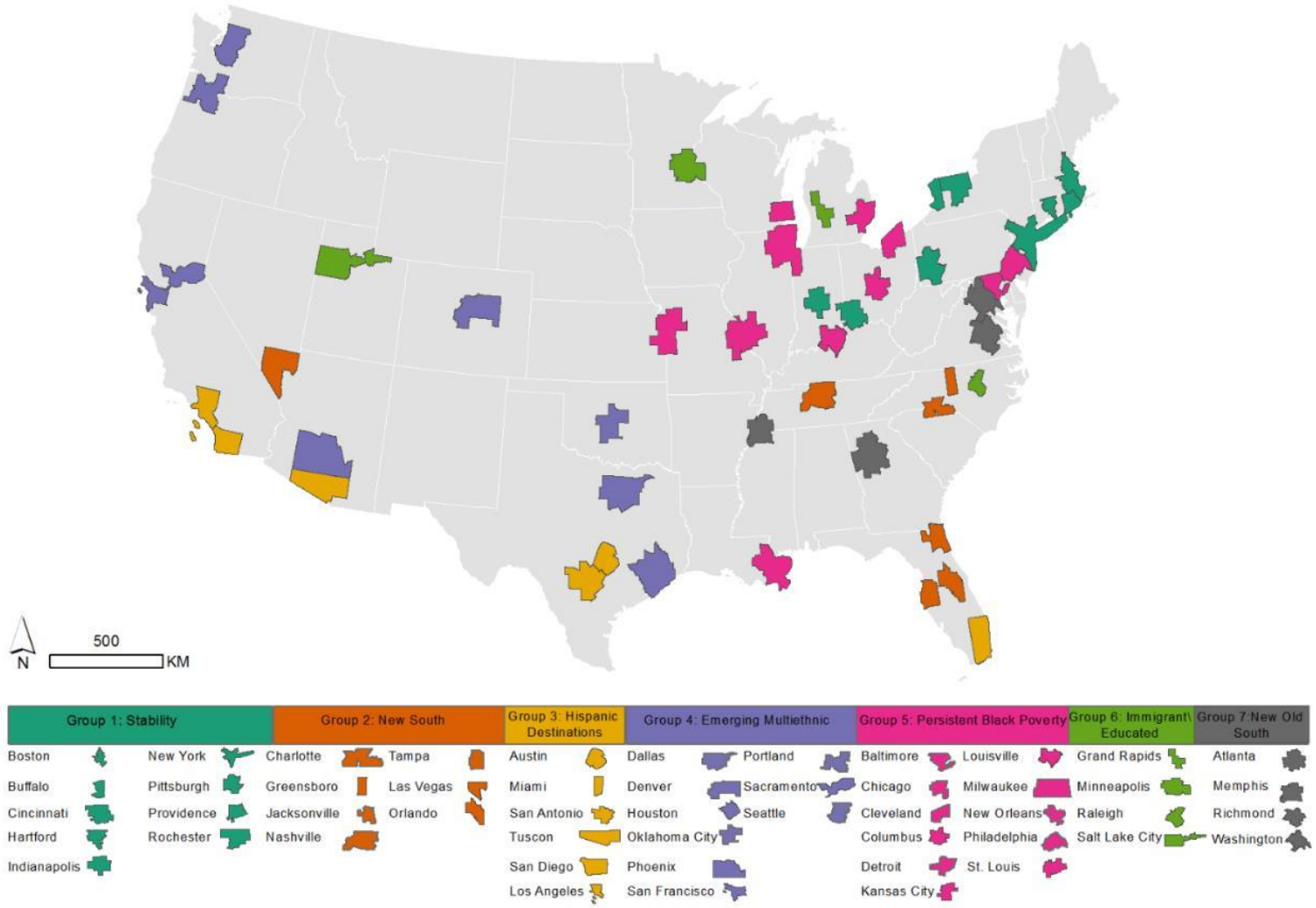
- Pengetahuan (Model) Berupa Klaster



Contoh : Klastering Jenis Pelanggan

Contoh : Klastering Sentimen Warga

Contoh : Poverty Rate Clustering



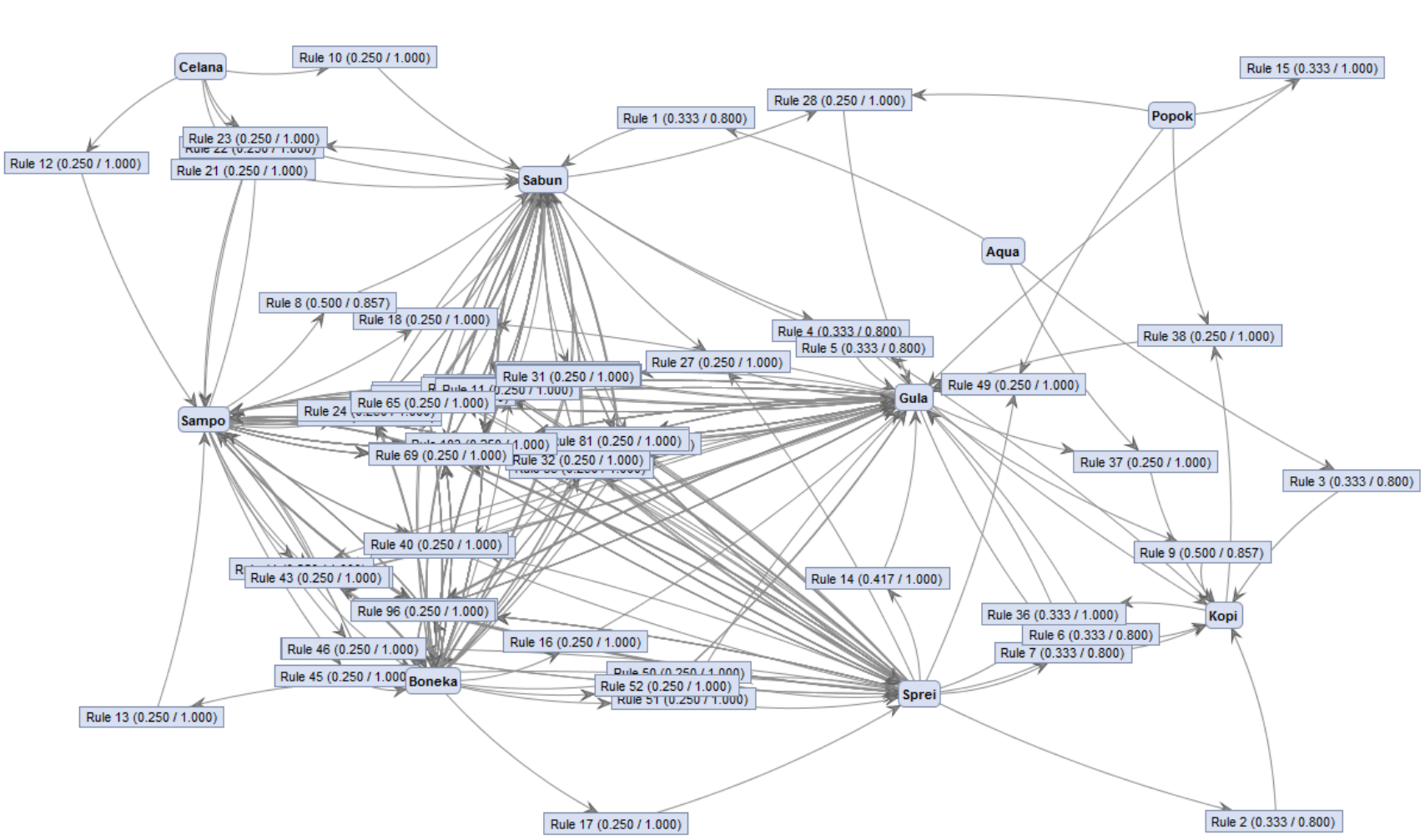
5. Aturan Asosiasi Pembelian Barang

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

Pembelajaran dengan
Metode Asosiasi (*FP-Growth*)

Pengetahuan berupa Aturan Asosiasi



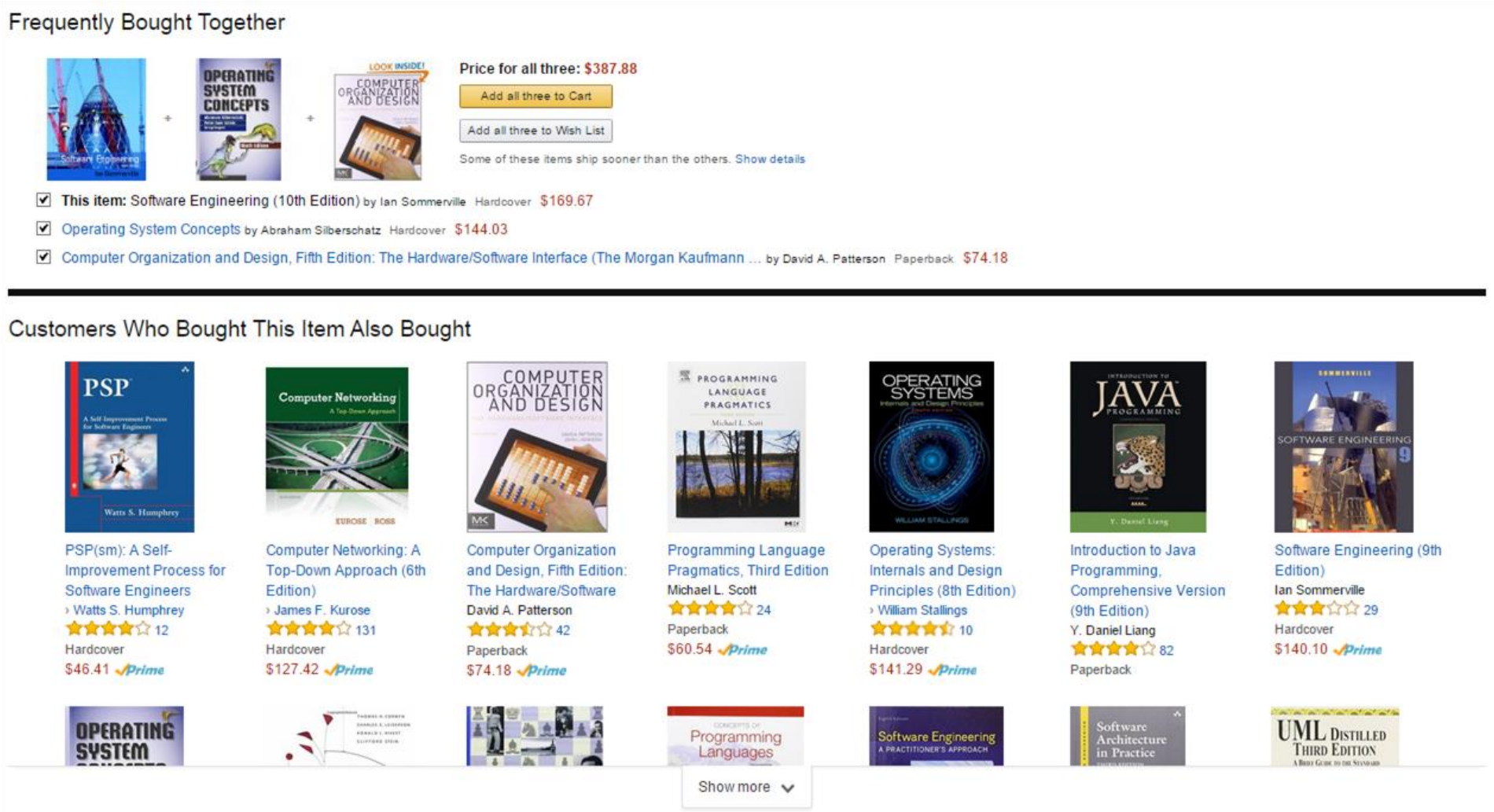
AssociationRules

```
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 Aturan Asosiasi

- Algoritma *association rule* (aturan asosiasi) adalah algoritma yang menemukan atribut yang "muncul bersamaan"
- 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

Aturan Asosiasi di Amazon.com



Metode Data Mining

1. Estimation (Estimasi):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc
2. Forecasting (Prediksi/Peramalan):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc
3. Classification (Klasifikasi):
   Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, Adaptative Credal C4.5), Naive Bayes (NB), K-Nearest Neighbor (kNN), Linear Discriminant Analysis (LDA), Logistic Regression (LogR), etc
4. Clustering (Klastering):
   K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means (FCM), etc
5. Association (Asosiasi):
   FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

Pola / Model / Knowledge / Output

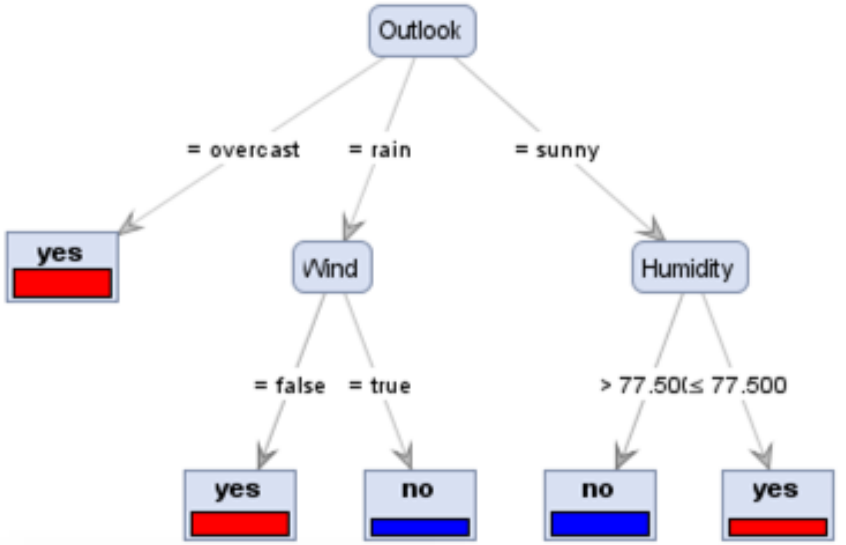
1. Formula/Function (Rumus atau Fungsi Regresi) :

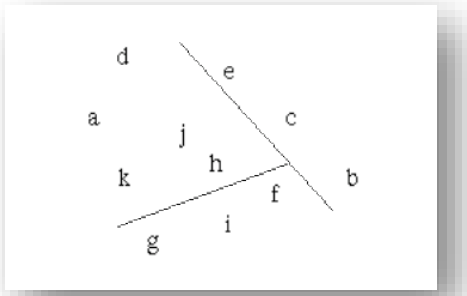
   WAKTU TEMPUH = 0.48 + 0.6 JARAK + 0.34 LAMPU + 0.2 PESANAN

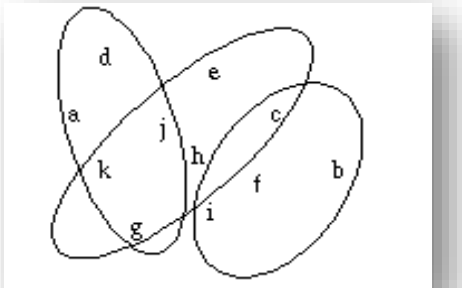
2. Decision Tree (Pohon Keputusan).
3. Tingkat Korelasi.
4. Rule (Aturan):

   IF ips3=2.8 THEN lulustepatwaktu

5. Cluster (Klaster).





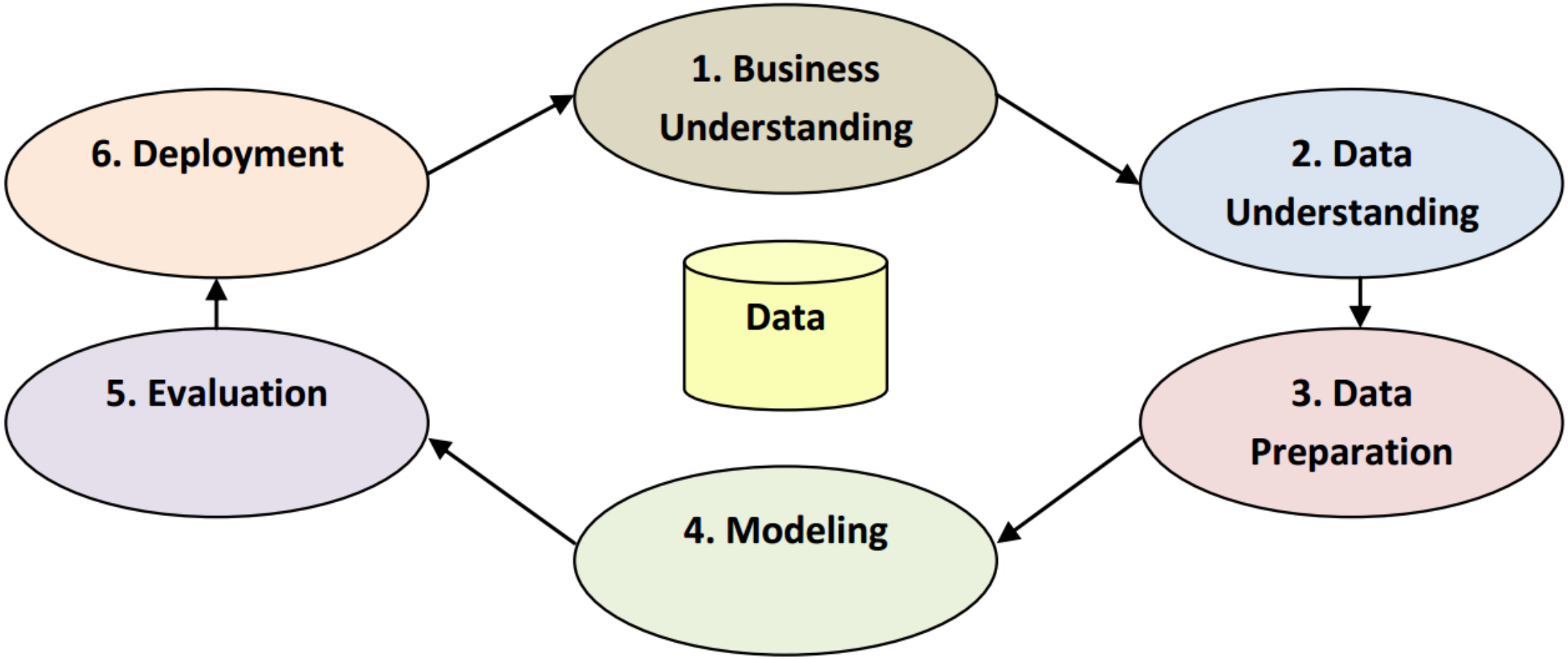


Proses Data Mining dengan CRISP-DM

Cross Industry Standart Process for Data Mining (CRISP – DM)

- Dunia industri yang beragam bidangnya memerlukan proses yang standard yang mampu mendukung penggunaan data mining untuk menyelesaikan masalah bisnis.
- Proses tersebut harus dapat digunakan di lintas industry (cross-industry) dan netral secara bisnis, tool dan aplikasi yang digunakan, serta mampu menangani strategi pemecahan masalah bisnis dengan menggunakan data mining.
- Pada tahun 1996, lahirlah salah satu standard proses di dunia data mining yang kemudian disebut dengan: the *Cross-Industry Standard Process for Data Mining* (CRISP–DM) *(Chapman, 2000).*

Cross Industry Standart Process for Data Mining (CRISP – DM) [2]



1. Pemahaman Proses Bisnis (*Business Understanding*)

- Nyatakan tujuan dan persyaratan proyek dengan jelas dalam kaitannya dengan bisnis atau unit penelitian secara keseluruhan.
- Terjemahkan tujuan dan batasan ini ke dalam rumusan definisi masalah Data Mining.
- Persiapkan strategi awal untuk mencapai tujuan ini.
- Merancang apa yang akan Anda bangun.

2. Pemahaman Data (*Data Understanding*)

- Kumpulkan datanya.
- Gunakan analisis data eksplorasi untuk membiasakan diri dengan data dan menemukan wawasan awal.
- Evaluasi kualitas data.
- Jika diinginkan, pilih subset menarik yang mungkin berisi pola yang dapat ditindaklanjuti.

3. Persiapan Data (*Data Preparation*)

- Persiapkan dari data mentah awal kumpulan data akhir yang akan digunakan untuk semua tahap selanjutnya.
- Pilih kasus dan variabel yang ingin Anda analisis dan yang sesuai untuk analisis Anda.
- Melakukan pembersihan data, integrasi, reduksi dan transformasi, sehingga siap untuk alat pemodelan.

4. Pemodelan (*Modeling*)

- Pilih dan terapkan teknik pemodelan yang sesuai.
- Kalibrasi pengaturan model untuk mengoptimalkan hasil.
- Ingatlah bahwa seringkali, beberapa teknik berbeda dapat digunakan untuk masalah data mining yang sama.
- Jika perlu, putar kembali ke tahap persiapan data untuk menyesuaikan bentuk data dengan persyaratan khusus dari teknik penambangan data tertentu.

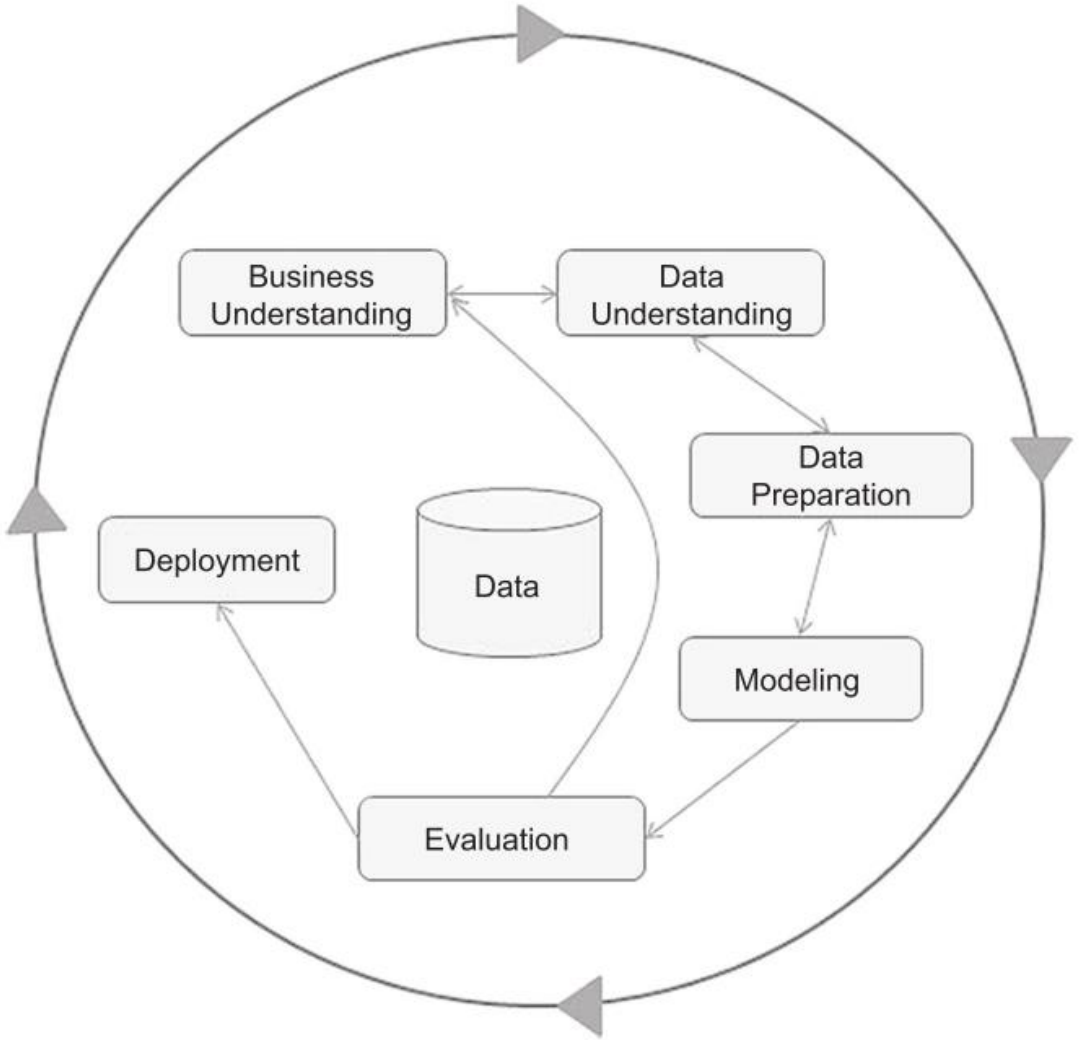
5. Evaluasi (*Evaluation*)

- Mengevaluasi kualitas dan efektivitas satu atau lebih model yang dikirimkan dalam fase pemodelan sebelum menerapkannya untuk digunakan di lapangan.
- Tentukan apakah model benar-benar mencapai tujuan yang ditetapkan.
- Tentukan apakah beberapa aspek penting dari masalah bisnis atau penelitian belum diperhitungkan secara memadai.
- Ambil keputusan tentang penggunaan hasil Data Mining.

6. Penerapan (*Deployment*)

- Manfaatkan model yang dibuat:
  - pembuatan model tidak menandakan penyelesaian proyek.
- Contoh penerapan sederhana:
  - Buat laporan.
- Contoh penerapan yang lebih kompleks:
  - Menerapkan proses Data Mining paralel di departemen lain.
- Untuk bisnis, pelanggan sering melakukan penerapan berdasarkan model Anda.

CRISP – DM : Detail Flow

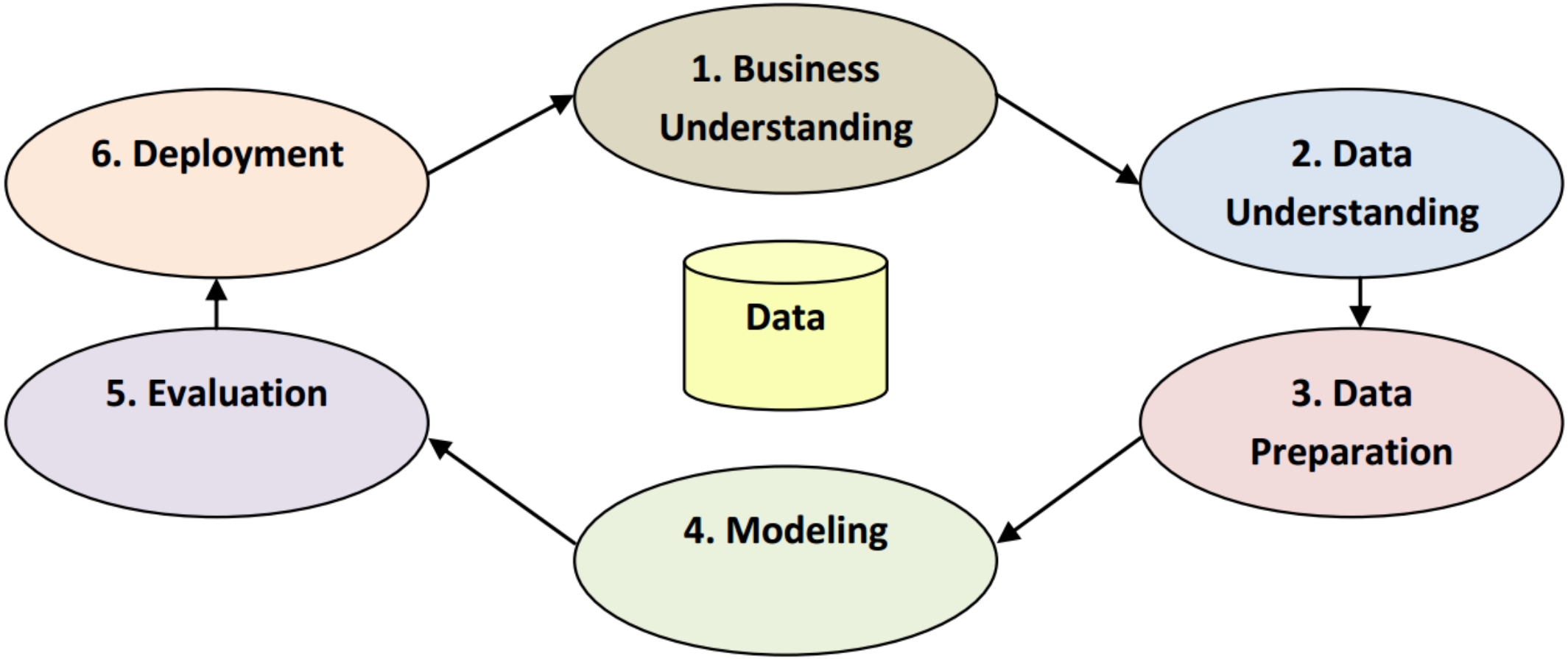


Studi Kasus dengan CRISP-DM

Kelulusan Mahasiswa di Universitas Suka Suka

Dataset: *datakelulusanmahasiswa.xls*

Cross Industry Standart Process for Data Mining (CRISP – DM)



1. Pemahaman Proses Bisnis (*Business Understanding*)

Problems:

- Andi adalah Rektor di Universitas Suka Suka.
- Universitas Suka Suka memiliki masalah besar karena rasio kelulusan mahasiswa tiap angkatan sangat rendah.
- Andi ingin memahami dan membuat pola dari profile mahasiswa yang bisa lulus tepat waktu dan yang tidak lulus tepat waktu.
- Dengan pola tersebut, Andi bisa melakukan konseling, terapi, dan memberi peringatan dini kepada mahasiswa kemungkinan tidak lulus tepat waktu untuk memperbaiki diri, sehingga akhirnya bisa lulus tepat waktu.

Objective:

- Menemukan pola dari mahasiswa yang lulus tepat waktu dan tidak.

2. Pemahaman Data (*Data Understanding*)

- Untuk menyelesaikan masalah, Andi mengambil data dari Sistem Informasi Akademik di Universitasnya.
- Data-data dikumpulkan dari data profil mahasiswa dan indeks prestasi semester mahasiswa, dengan atribut seperti di bawah:
    1. NAMA.
    2. JENIS KELAMIN: Laki-Laki atau Perempuan.
    3. STATUS MAHASISWA: Mahasiswa atau Bekerja.
    4. UMUR.
    5. STATUS NIKAH: Menikah atau Belum Menikah.
    6. IPS 1: Indeks Prestasi Semester 1.
    7. IPS 2: Indeks Prestasi Semester 2.
    8. IPS 3: Indeks Prestasi Semester 3.
    9. IPS 4: Indeks Prestasi Semester 4.
    10. IPS 5: Indeks Prestasi Semester 5.
    11. IPS 6: Indeks Prestasi Semester 6.
    12. IPS 7: Indeks Prestasi Semester 7.
    13. IPS 8: Indeks Prestasi Semester 8.
    14. IPK: Indeks Prestasi Kumulatif.
    15. STATUS KELULUSAN: Terlambat atau Tepat.

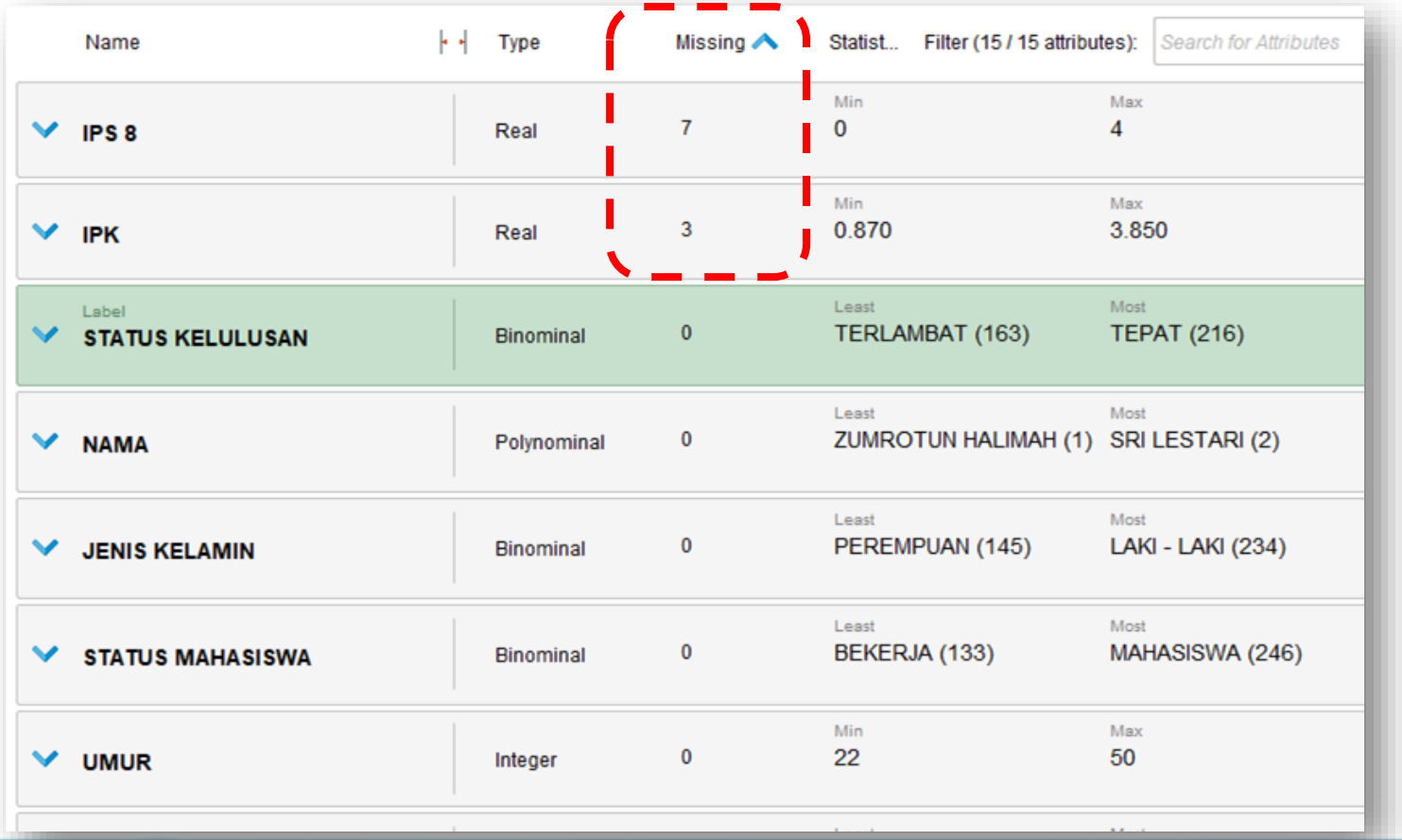
3. Persiapan Data (*Data Preparation*)

- Data set: datakelulusanmahasiswa.xls

Row No.	STATUS KEL...	NAMA	JENIS KELA...	STATUS MA...	UMUR	STATUS NIK...	IPS 1	IPS 2
1	TERLAMBAT	ANIK WIDAYA...	PEREMPUAN	BEKERJA	28	BELUM MENI...	2.760	2.800
2	TERLAMBAT	DWI HESTYN...	PEREMPUAN	MAHASISWA	32	BELUM MENI...	3	3.300
3	TERLAMBAT	MURYA ARIE...	PEREMPUAN	BEKERJA	29	BELUM MENI...	3.500	3.300
4	TERLAMBAT	NANIK SUSA...	PEREMPUAN	MAHASISWA	27	BELUM MENI...	3.170	3.410
5	TERLAMBAT	RIFKA ISTIQF...	PEREMPUAN	BEKERJA	29	BELUM MENI...	2.900	2.890
6	TERLAMBAT	SUHARYONO	LAKI - LAKI	BEKERJA	27	BELUM MENI...	2.950	2.820
7	TEPAT	FARIKHATUN...	PEREMPUAN	MAHASISWA	26	BELUM MENI...	2.760	3.140
8	TEPAT	FIFI SUNALISA	PEREMPUAN	MAHASISWA	27	BELUM MENI...	2.620	2.890
9	TERLAMBAT	HENDRIK M...	PEREMPUAN	BEKERJA	25	MENIKAH	3.600	3.540
10	TERLAMBAT	IMAM AGUNG...	PEREMPUAN	BEKERJA	28	BELUM MENI...	2.710	2.550
11	TERLAMBAT	IMAM SANTO...	PEREMPUAN	BEKERJA	27	BELUM MENI...	3.140	3.460
12	TERLAMBAT	IRFAN EKO ...	PEREMPUAN	BEKERJA	32	BELUM MENI...	2.670	2.300
13	TERLAMBAT	IWAN HAMBALI	PEREMPUAN	BEKERJA	26	BELUM MENI...	2.570	2.820
14	TERLAMBAT	M SYAIFULLAH	PEREMPUAN	BEKERJA	31	BELUM MENI...	2.710	3

3. Persiapan Data (*Data Preparation*) [2]

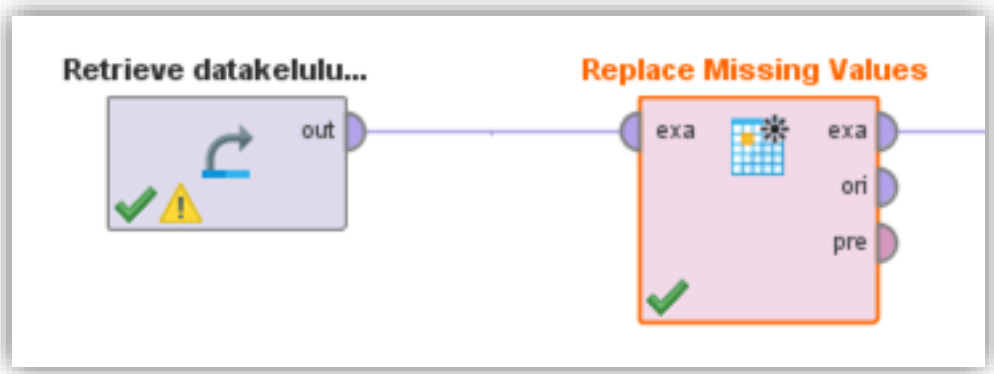
- Terdapat 379 data mahasiswa dengan 15 atribut.
- Missing Value sebayak 10 data, dan tidak terdapat data noise.



Name	Type	Missing	Statist...	Filter (15 / 15 attributes): Search for Attributes
IPS 8	Real	7	Min 0	Max 4
IPK	Real	3	Min 0.870	Max 3.850
Label STATUS KELULUSAN	Binominal	0	Least TERLAMBAT (163)	Most TEPAT (216)
NAMA	Polynominal	0	Least ZUMROTUN HALIMAH (1)	Most SRI LESTARI (2)
JENIS KELAMIN	Binominal	0	Least PEREMPUAN (145)	Most LAKI - LAKI (234)
STATUS MAHASISWA	Binominal	0	Least BEKERJA (133)	Most MAHASISWA (246)
UMUR	Integer	0	Min 22	Max 50

3. Persiapan Data (*Data Preparation*) [3]

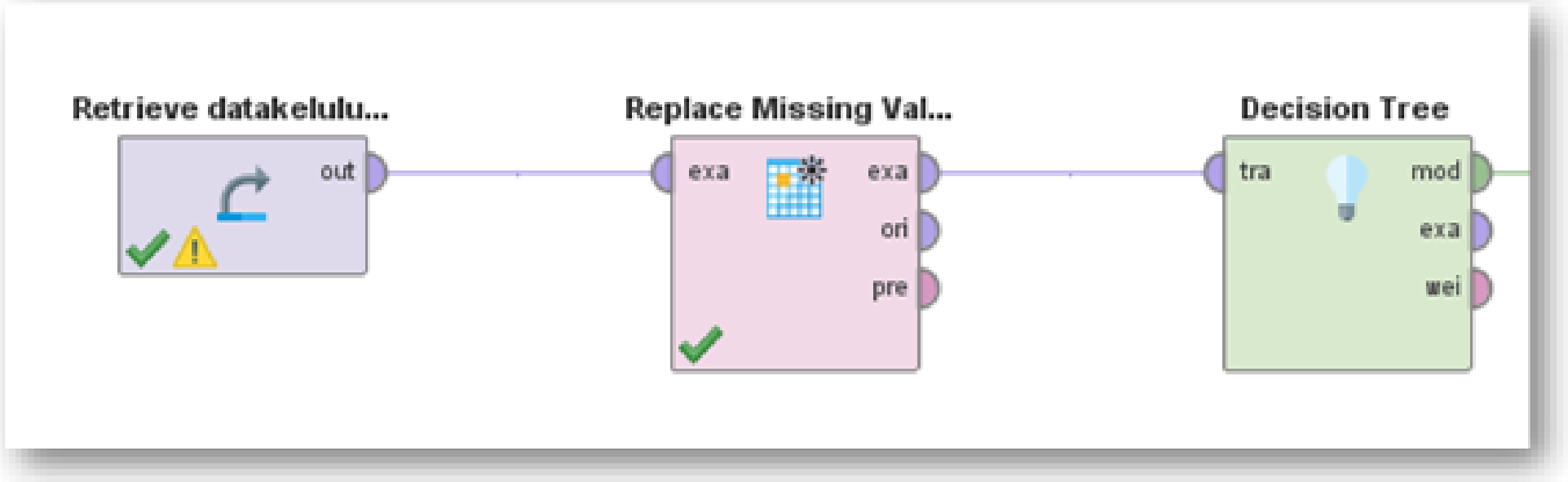
- Missing Value dipecahkan dengan menambahkan data dengan nilai rata-rata.
- Hasilnya adalah data bersih tanpa missing value.



Name	Type	Missing	Statist...	Filter (15 / 15 attributes):
Label STATUS KELULUSAN	Binominal	0	Least TERLAMBAT (163)	Most TEPA
NAMA	Polynominal	0	Least ZUMROTUN HALIMAH (1)	Most SRI LE
JENIS KELAMIN	Binominal	0	Least PEREMPUAN (145)	Most LAKI -
STATUS MAHASISWA	Binominal	0	Least BEKERJA (133)	Most MAHA
UMUR	Integer	0	Min 22	Max 50
STATUS NIKAH	Binominal	0	Least MENIKAH (8)	Most BELU
IPS 1	Real	0	Min 0.330	Max 3.790

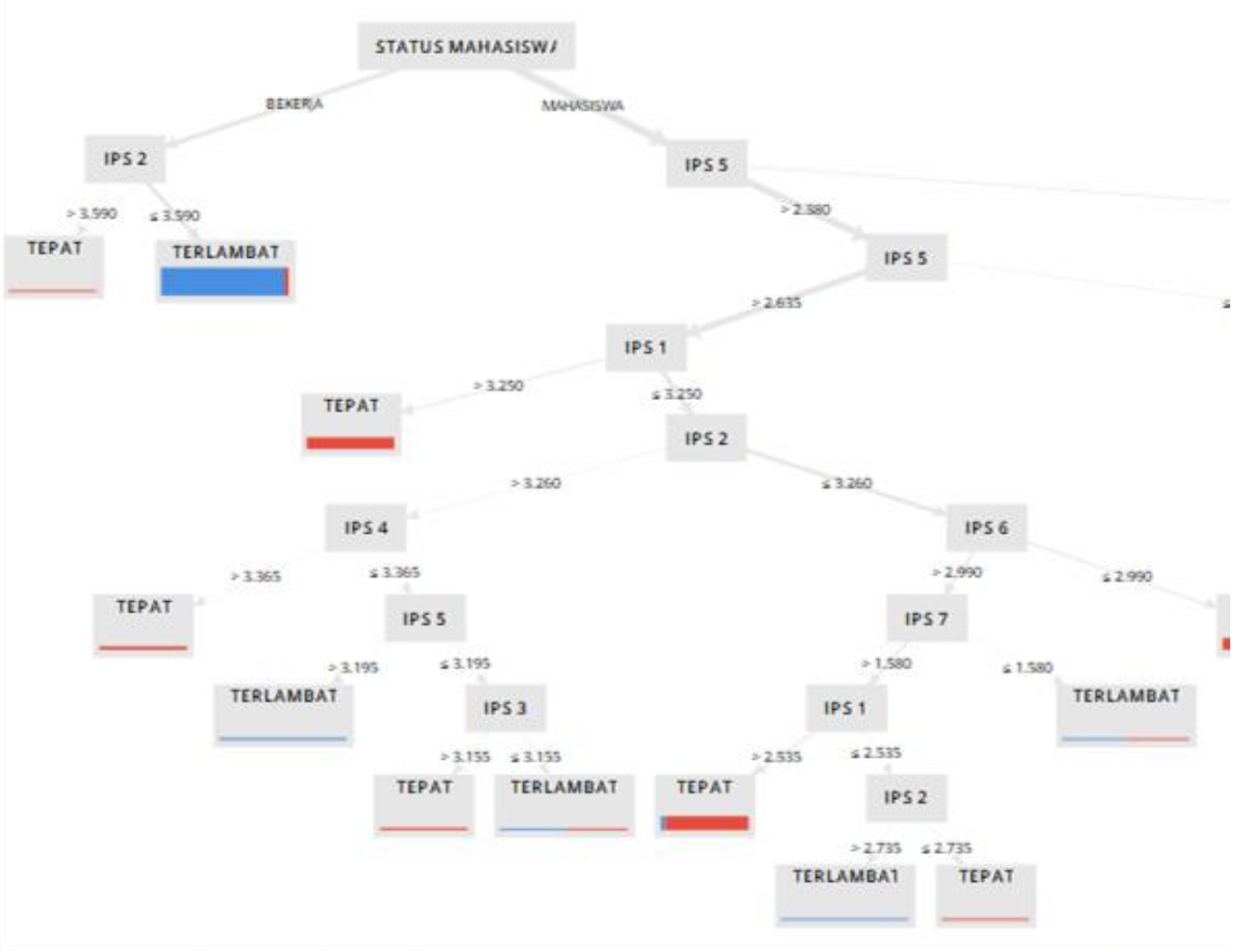
4. Pemodelan (*Modeling*)

- Modelkan dataset dengan Decision Tree.
- Pola yang dihasilkan bisa berbentuk *Tree* atau *If-Then*.



4. Pemodelan (*Modeling*) [2]

- Hasil pola dari data berupa berupa decision tree (pohon keputusan).



5. Evaluasi (*Evaluation*)

- Hasil pola dari data berupa berupa peraturan if-then.

```
STATUS MAHASISWA = BEKERJA
|   IPS 2 > 3.590: TEPAT {TERLAMBAT=0, TEPAT=2}
|   IPS 2 ≤ 3.590: TERLAMBAT {TERLAMBAT=127, TEPAT=4}
STATUS MAHASISWA = MAHASISWA
|   IPS 5 > 2.380
|   |   IPS 5 > 2.635
|   |   |   IPS 1 > 3.250: TEPAT {TERLAMBAT=0, TEPAT=50}
|   |   |   IPS 1 ≤ 3.250
|   |   |   |   IPS 2 > 3.260
|   |   |   |   |   IPS 4 > 3.365: TEPAT {TERLAMBAT=0, TEPAT=10}
|   |   |   |   |   IPS 4 ≤ 3.365
|   |   |   |   |   |   IPS 5 > 3.195: TERLAMBAT {TERLAMBAT=4, TEPAT=0}
|   |   |   |   |   |   IPS 5 ≤ 3.195
|   |   |   |   |   |   |   IPS 3 > 3.155: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   |   |   |   IPS 3 ≤ 3.155: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   IPS 2 ≤ 3.260
|   |   |   |   |   IPS 6 > 2.990
|   |   |   |   |   |   IPS 7 > 1.580
|   |   |   |   |   |   |   IPS 1 > 2.535: TEPAT {TERLAMBAT=3, TEPAT=58}
|   |   |   |   |   |   |   IPS 1 ≤ 2.535
|   |   |   |   |   |   |   |   IPS 2 > 2.735: TERLAMBAT {TERLAMBAT=2, TEPAT=0}
|   |   |   |   |   |   |   |   IPS 2 ≤ 2.735: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   |   IPS 7 ≤ 1.580: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   |   IPS 6 ≤ 2.990: TEPAT {TERLAMBAT=0, TEPAT=51}
|   |   IPS 5 ≤ 2.635
|   |   |   IPS 3 > 2.480
|   |   |   |   IPS 1 > 2.920: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   IPS 1 ≤ 2.920
|   |   |   |   |   IPS 3 > 3.075: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   IPS 3 ≤ 3.075: TERLAMBAT {TERLAMBAT=6, TEPAT=0}
|   |   |   IPS 3 ≤ 2.480: TEPAT {TERLAMBAT=0, TEPAT=11}
```

- Atribut atau faktor yang paling berpengaruh adalah Status Mahasiswa, IPS2, IPS5, IPS1.
- Atribut atau faktor yang tidak berpengaruh adalah Nama, Jenis Kelamin, Umur, IPS6, IPS7, IPS8.

5. Evaluasi (*Evaluation*)

- Hasil pola dari data berupa berupa peraturan if-then.

```
STATUS MAHASISWA = BEKERJA
|   IPS 2 > 3.590: TEPAT {TERLAMBAT=0, TEPAT=2}
|   IPS 2 ≤ 3.590: TERLAMBAT {TERLAMBAT=127, TEPAT=4}
STATUS MAHASISWA = MAHASISWA
|   IPS 5 > 2.380
|   |   IPS 5 > 2.635
|   |   |   IPS 1 > 3.250: TEPAT {TERLAMBAT=0, TEPAT=50}
|   |   |   IPS 1 ≤ 3.250
|   |   |   |   IPS 2 > 3.260
|   |   |   |   |   IPS 4 > 3.365: TEPAT {TERLAMBAT=0, TEPAT=10}
|   |   |   |   |   IPS 4 ≤ 3.365
|   |   |   |   |   |   IPS 5 > 3.195: TERLAMBAT {TERLAMBAT=4, TEPAT=0}
|   |   |   |   |   |   IPS 5 ≤ 3.195
|   |   |   |   |   |   |   IPS 3 > 3.155: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   |   |   |   IPS 3 ≤ 3.155: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   IPS 2 ≤ 3.260
|   |   |   |   |   IPS 6 > 2.990
|   |   |   |   |   |   IPS 7 > 1.580
|   |   |   |   |   |   |   IPS 1 > 2.535: TEPAT {TERLAMBAT=3, TEPAT=58}
|   |   |   |   |   |   |   IPS 1 ≤ 2.535
|   |   |   |   |   |   |   |   IPS 2 > 2.735: TERLAMBAT {TERLAMBAT=2, TEPAT=0}
|   |   |   |   |   |   |   |   IPS 2 ≤ 2.735: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   |   IPS 7 ≤ 1.580: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   |   IPS 6 ≤ 2.990: TEPAT {TERLAMBAT=0, TEPAT=51}
|   |   IPS 5 ≤ 2.635
|   |   |   IPS 3 > 2.480
|   |   |   |   IPS 1 > 2.920: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   IPS 1 ≤ 2.920
|   |   |   |   |   IPS 3 > 3.075: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   IPS 3 ≤ 3.075: TERLAMBAT {TERLAMBAT=6, TEPAT=0}
|   |   |   IPS 3 ≤ 2.480: TEPAT {TERLAMBAT=0, TEPAT=11}
```

- Atribut atau faktor yang paling berpengaruh adalah Status Mahasiswa, IPS2, IPS5, IPS1.
- Atribut atau faktor yang tidak berpengaruh adalah Nama, Jenis Kelamin, Umur, IPS6, IPS7, IPS8.

6. Penerapan (*Deployment*)

- Andi membuat program peningkatan disiplin dan pendampingan ke mahasiswa di semester awal (1-2) dan semester 5, karena faktor yang paling menentukan kelulusan mahasiswa ada di dua semester itu.
- Andi membuat peraturan melarang mahasiswa bekerja paruh waktu di semester awal perkuliahan, karena beresiko tinggi di kelulusan tepat waktu.
- Andi membuat program kerja paruh waktu di dalam kampus, sehingga banyak pekerjaan kampus yang bisa intens ditangani, sambil mendidik mahasiswa supaya memiliki pengalaman kerja. Dan yang paling penting mahasiswa tidak meninggalkan kuliah karena pekerjaan.
- Andi memasukkan pola dan model yang terbentuk ke dalam Sistem Informasi Akademik, secara berkala diupdate setiap semester.
- Sistem Informasi Akademik dibuat cerdas, sehingga bisa mengirimkan email Analisis Pola Kelulusan secara otomatis ke mahasiswa sesuai profilnya.

Latihan Soal (Kuis)

- Carilah data yang dapat Anda gunakan untuk ***Proses Data Mining***, kemudian gunakan ***CRISP – DM*** untuk menyelesaikan masalah tersebut.
- Jelaskan setiap langkahnya seperti ***Studi Kasus Data Kelulusan Mahasiswa*** diatas menggunakan ***proses CRISP-DM*** dimulai dari ***Business Understanding***, ***Data Understanding***, ***Data Preparation***, ***Modeling***, ***Evaluation*** dan ***Deployment***.

Referensi

1. Jiawei Han, Micheline Kamber, Jian Pei, Data mining : concepts and techniques – 3rd ed, Elsevier, 2012.
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 4th Edition, *Elsevier*, 2017.
3. Budi Santosa, Ardian Umam, Data Mining dan Big Data Analytics, Penebar Media Pustaka, 2018.
4. Max Bramer, Principles of Data Mining – Undergraduate Topics in Computer Science – 4th ed, Springer, 2020.
5. Romi Satrio Wahono, *Lecture Notes – Data Mining*, diakses 3 Maret 2021, <https://romisatriawahono.net/dm/>.
6. Sumber gambar: www.freepik.com.

THANKS

ANY QUESTIONS?