ANALISIS DATA & *DATA MINING*

Oleh:
Lindung Parningotan Manik, PDEng.

Disampaikan pada:
Pelatihan Teknis Substantif JF ADI Batch 4

Kamis, 30 Juli 2026

Course Outline

1. Pengantar Analisis & Penambangan Data

1.1 Apa dan Mengapa Data Mining?
1.2 Peran Utama dan Metode Data Mining
1.3 Kategori dan Tipe Data
1.4 Analisis Data

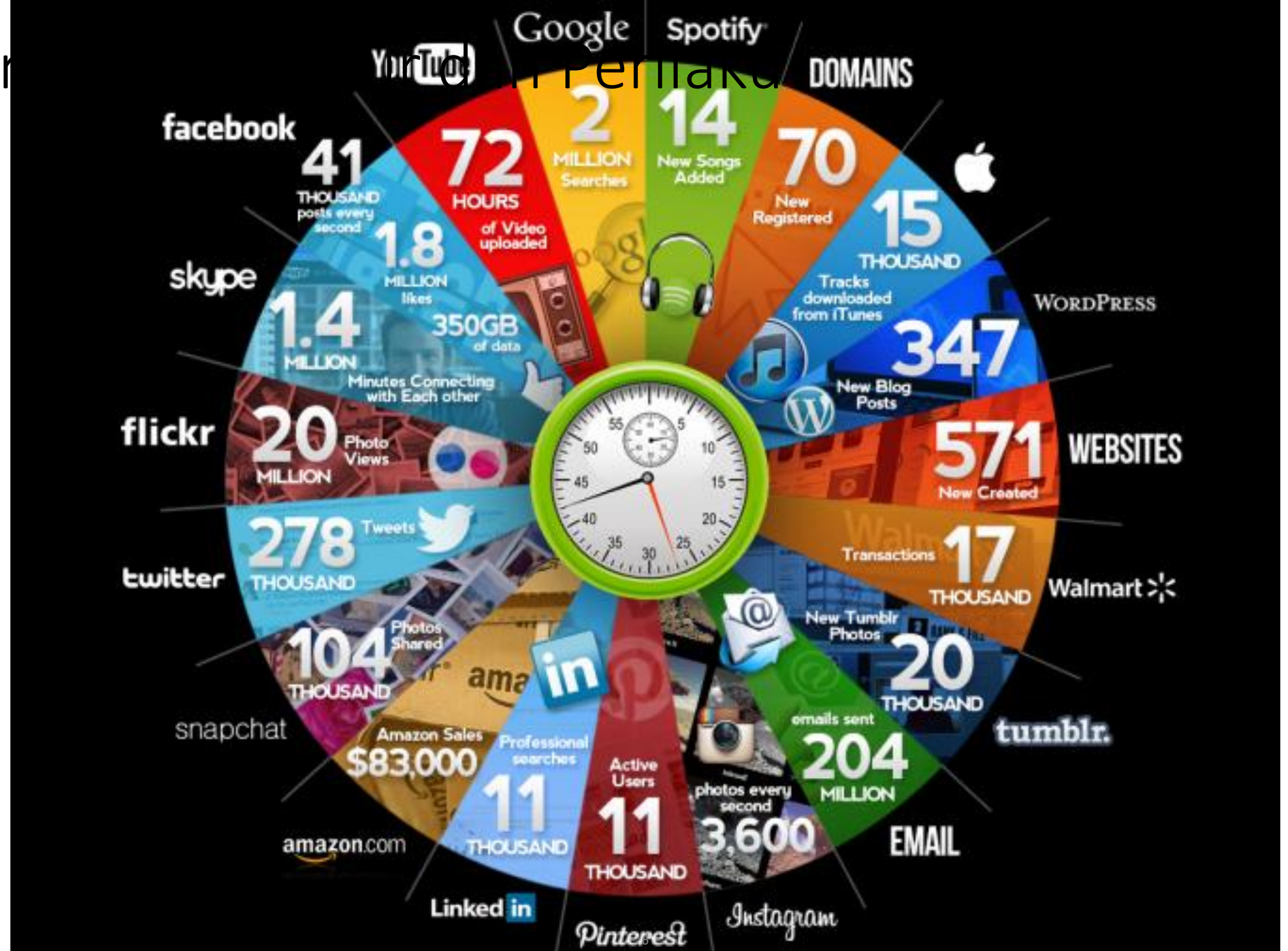
1.1 Apa dan Mengapa Data Mining?

Manusia Memproduksi Data

Manusia memproduksi beragam data yang jumlah dan ukurannya sangat besar

- Astronomi
- Bisnis
- Kedokteran
- Ekonomi
- Olahraga
- Cuaca
- Financial
- …

Per


Google
Spotify
YouTube
DOMAINS
facebook
41
THOUSAND
posts every second
72
HOURS
of Video uploaded
2
MILLION
Searches
14
New Songs Added
70
New Registered
15
THOUSAND
Tracks downloaded from iTunes
1.8
MILLION
likes
skype
1.4
MILLION
Minutes Connecting with Each other
350GB
of data
WordPress
347
New Blog Posts
571
New Created
WEBSITES
flickr
20
MILLION
Photo Views
Transactions
17
THOUSAND
Walmart
278
THOUSAND
Tweets
twitter
New Tumblr Photos
20
THOUSAND
104
THOUSAND
Photos Shared
snapchat
Amazon Sales
$83,000
Professional searches
11
THOUSAND
Active Users
11
THOUSAND
emails sent
204
MILLION
tumblr.
photos every second
3,600
amazon.com
EMAIL
Linked in
Pinterest
Instagram

Datangnya Tsunami Data

kilobyte (**kB**)	10^3
megabyte (**MB**)	10^6
gigabyte (**GB**)	10^9
terabyte (**TB**)	10^{12}
petabyte (**PB**)	10^{15}
exabyte (**EB**)	10^{18}
zettabyte (**ZB**)	10^{21}
yottabyte (**YB**)	10^{24}

- Mobile Electronics market
  - 7B smartphone subscriptions in 2015

- Web & Social Networks generates amount of data
  - Google processes 100 PB per day, 3 million servers
  - Facebook has 300 PB of user data per day
  - Youtube has 1000PB video storage

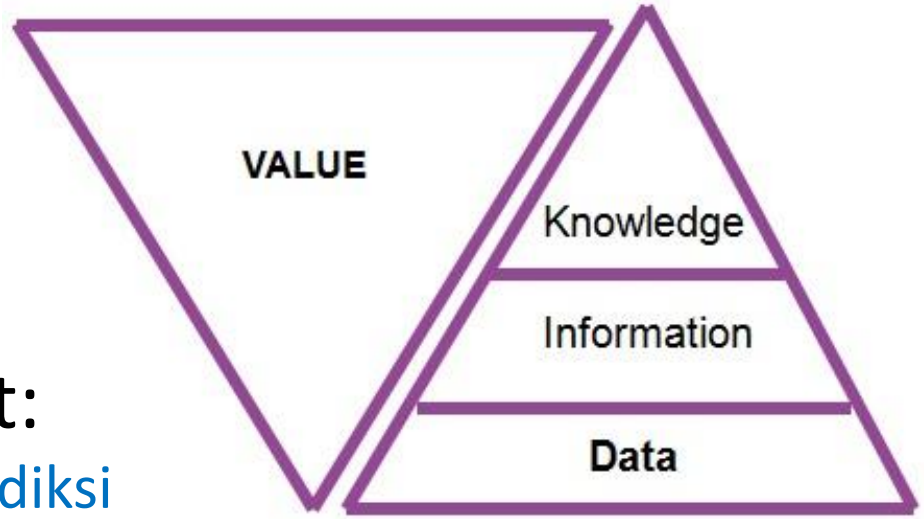
Kebanjiran Data tapi Miskin Pengetahuan

We are drowning in data,
but starving for knowledge!

(John Naisbitt, Megatrends, 1988)

Mengubah Data Menjadi Pengetahuan

- Data harus kita olah menjadi pengetahuan supaya bisa bermanfaat bagi manusia



- Dengan pengetahuan tersebut, manusia dapat:
  - Melakukan estimasi dan prediksi apa yang terjadi di depan
  - Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
  - Membantu pengambilan keputusan dan pembuatan kebijakan

Data - Informasi – Pengetahuan - Kebijakan

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

Data - Informasi – Pengetahuan - Kebijakan

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 Bulanan Kehadiran Pegawai

Data - Informasi – Pengetahuan - Kebijakan

	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

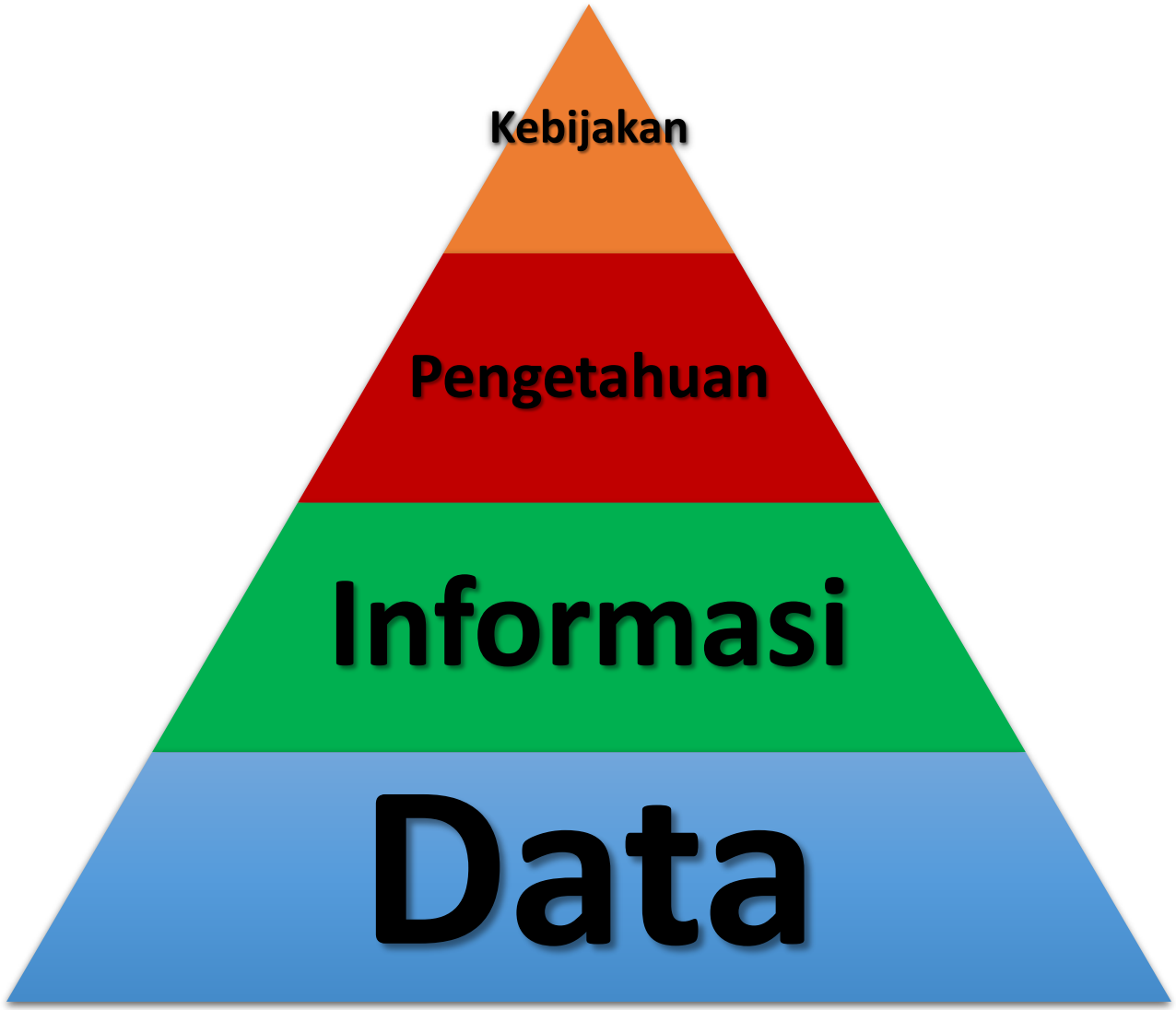
Pola Kebiasaan Kehadiran Mingguan Pegawai

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 - Informasi – Pengetahuan - Kebijakan



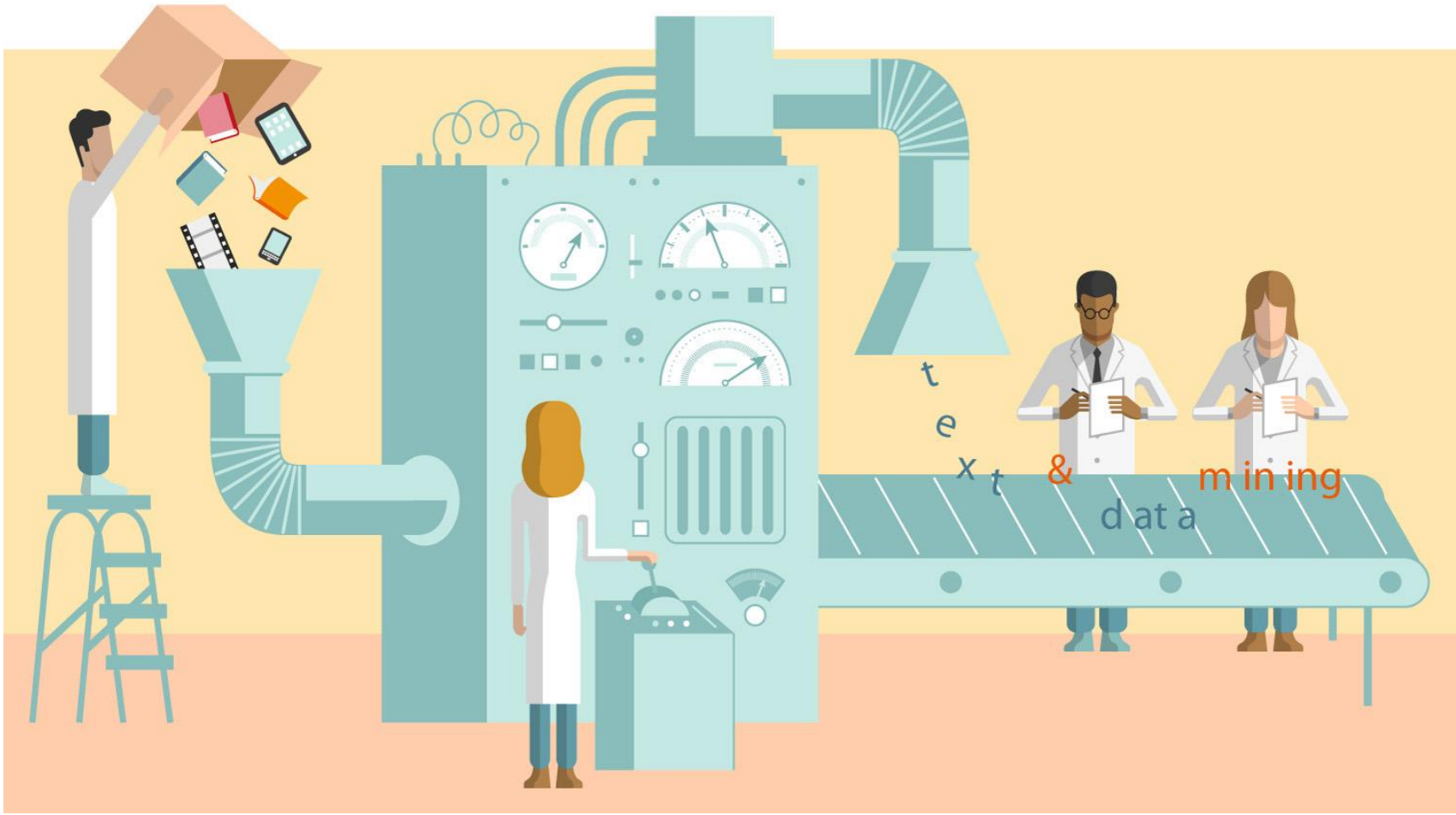
Kebijakan Penataan Jam Kerja Pegawai

Pola Kebiasaan Datang-Pulang Pegawai

Informasi Rekap Kehadiran Pegawai

Data Absensi Pegawai

Apa itu Data Mining?



Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

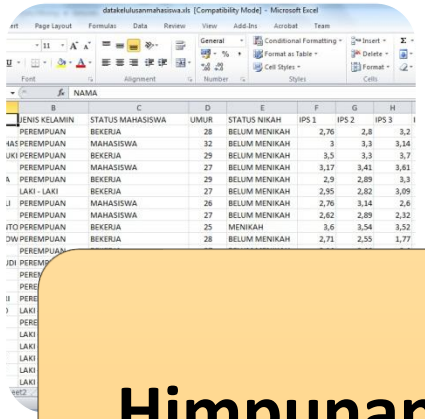
Terminologi dan Nama Lain Data Mining

- Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

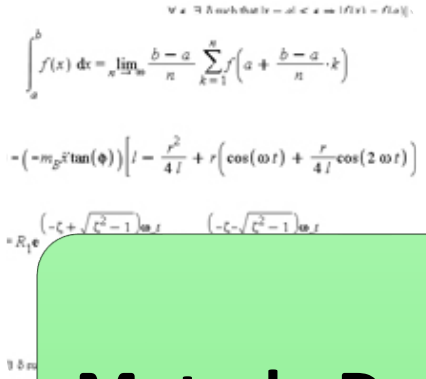
- Ekstraksi dari data ke pengetahuan:
  1. Data: fakta yang terekam dan tidak membawa arti
  2. Informasi: Rekap, rangkuman, penjelasan dan statistik dari data
  3. Pengetahuan: pola, rumus, aturan atau model yang muncul dari data

- Nama lain data mining:
  - Knowledge Discovery in Database (KDD)
  - Big data
  - Business intelligence
  - Knowledge extraction
  - Pattern analysis
  - Information harvesting

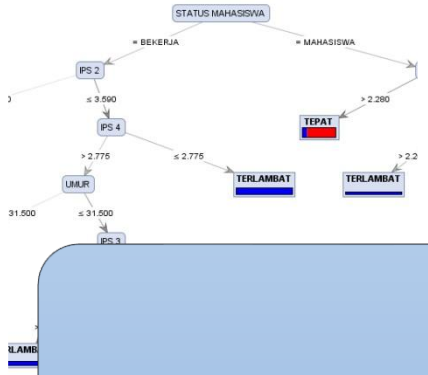
Konsep Proses Data Mining



Himpunan Data



Metode Data Mining



Pengetahuan

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)*

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data *(Han et al., 2011)*

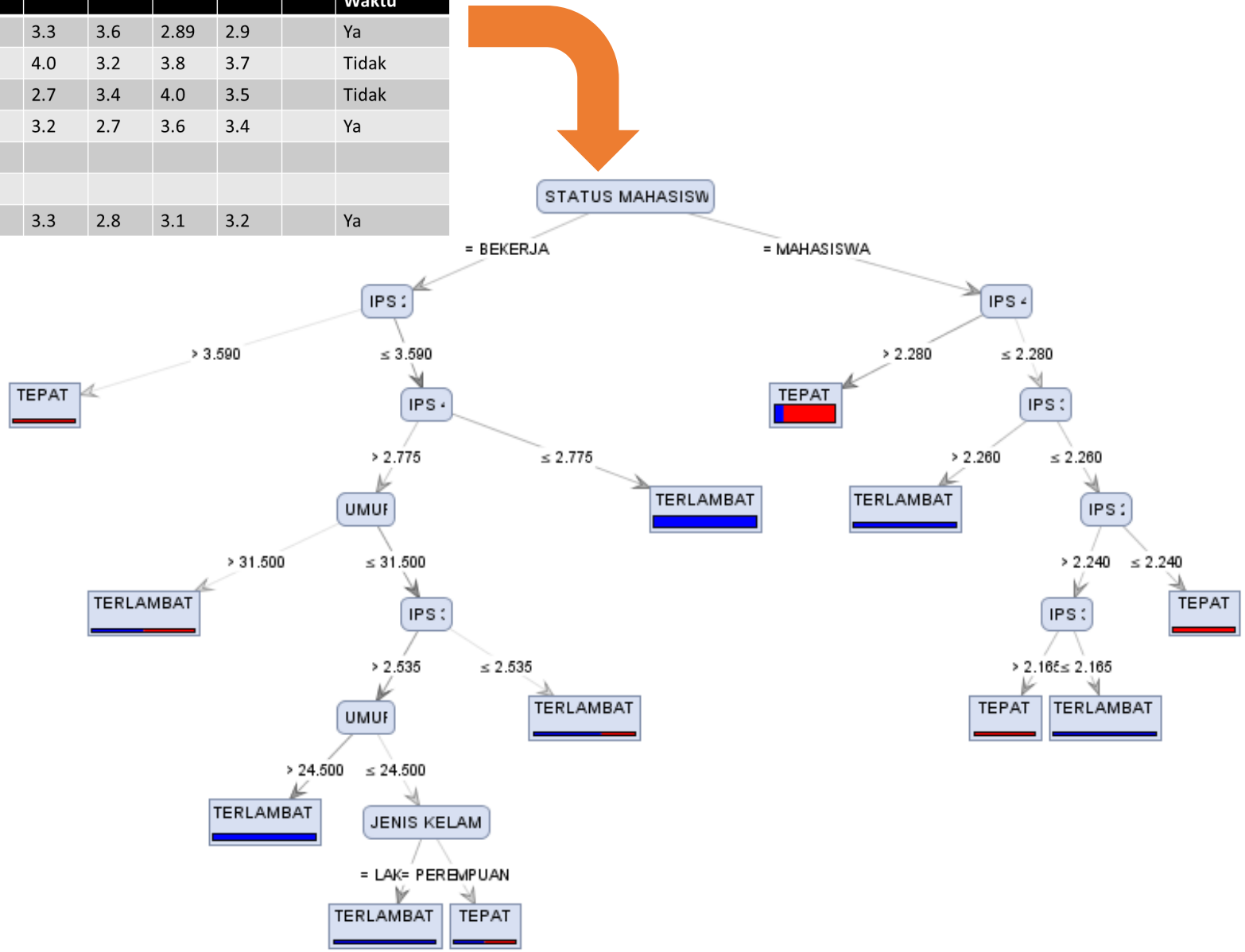
Contoh Data di Kampus

- Puluhan ribu data mahasiswa di kampus yang diambil dari sistem informasi akademik
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!
- Seperti apa pengetahuan itu? Rumus, Pola, Aturan

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

Prediksi 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



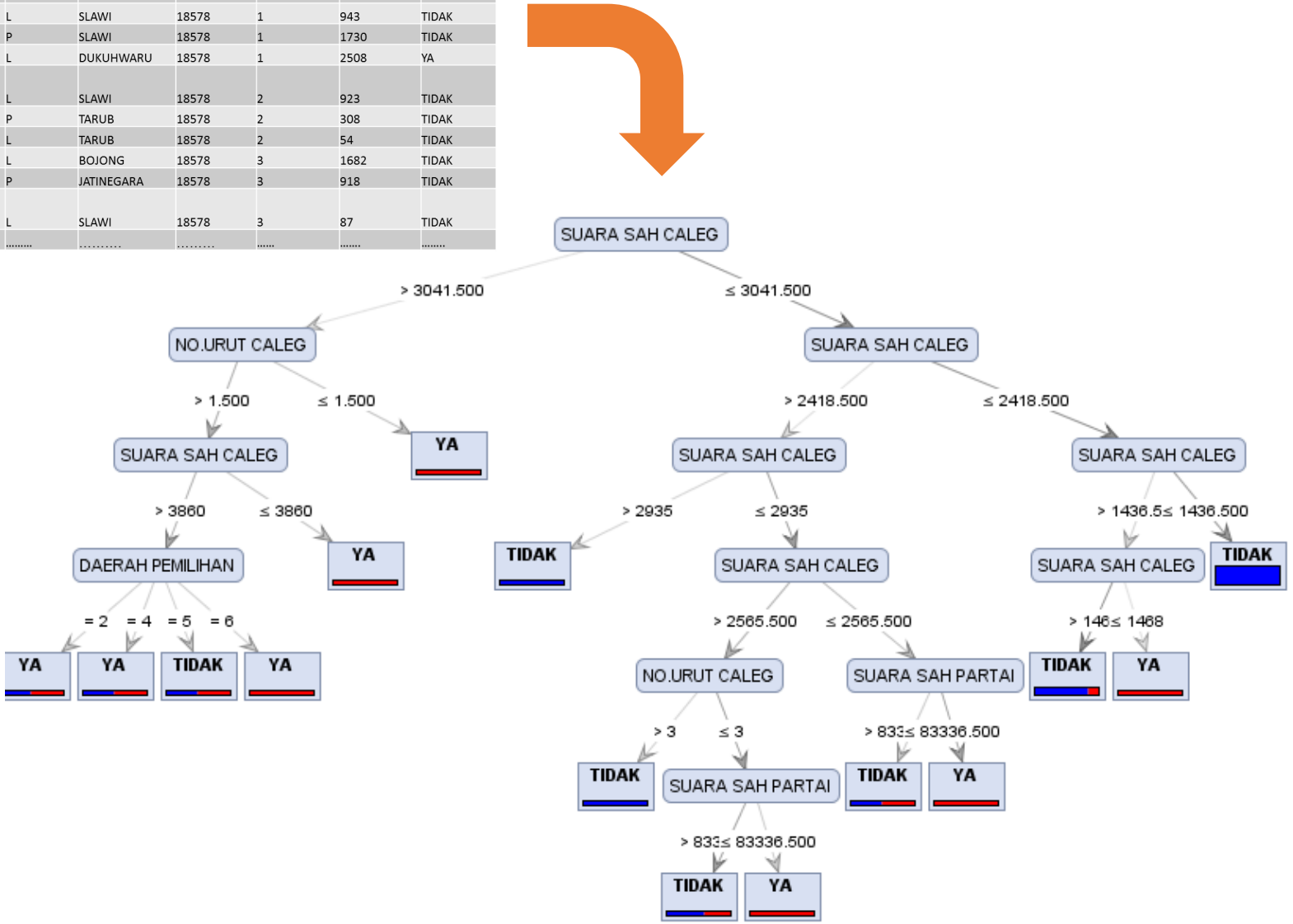
Contoh Data di Komisi Pemilihan Umum

- Puluhan ribu data calon anggota legislatif di KPU
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!

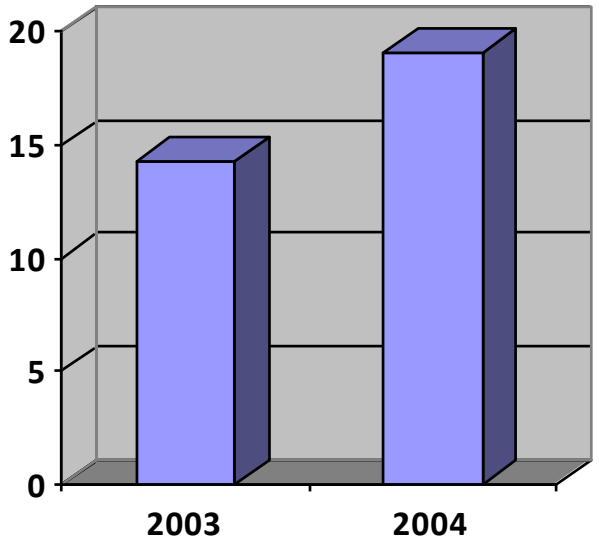
NAMA PARTAI POLITIK	NAMA CALON LEGESLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH CALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO,BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO,SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATI,SH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	DUKUHWARU	18578	1	2508	YA
GOLKAR	H.FAJAR SIGIT KUSUMAJAYA,SH	L	SLAWI	18578	2	923	TIDAK
GOLKAR	SUMIRAH	P	TARUB	18578	2	308	TIDAK
GOLKAR	DARYOTO	L	TARUB	18578	2	54	TIDAK
PKS	KHAPIP APRONI,S.Pdi	L	BOJONG	18578	3	1682	TIDAK
PKS	ENDANG SUCI RAHAYU	P	JATINEGARA	18578	3	918	TIDAK
PDI-P	KH.CHAFIDZ ISA MUFTI ,LC	L	SLAWI	18578	3	87	TIDAK
.....	.....	.....	.....	.....	.....	.....	.....

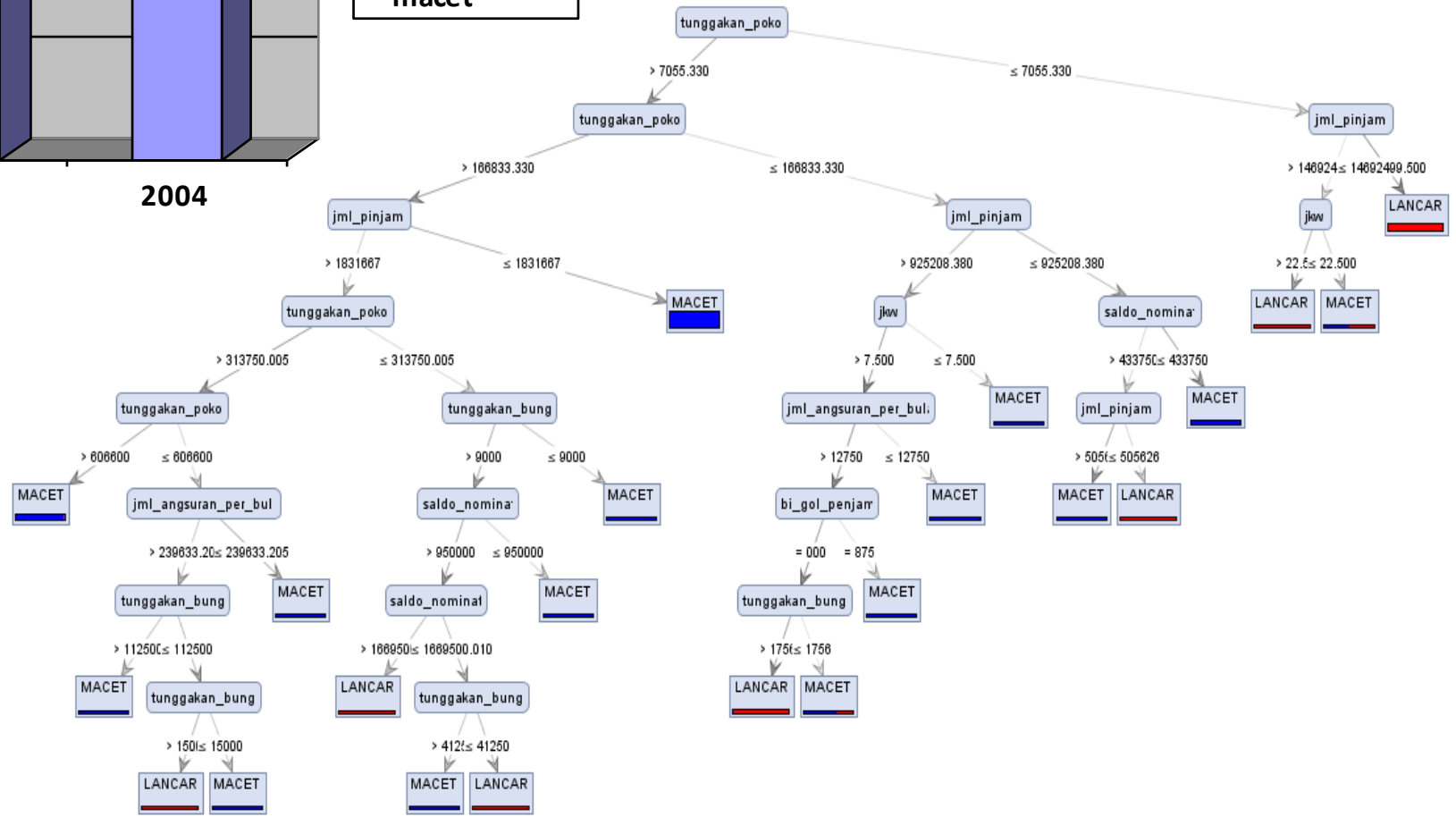
Prediksi Calon Legislatif DKI Jakarta

NAMA PARTAI POLITIK	NAMA CALON LEGESLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH CALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO,BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO,SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATI,SH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	DUKUHWARU	18578	1	2508	YA
GOLKAR	H.FAJAR SIGIT KUSUMAJAYA,SH	L	SLAWI	18578	2	923	TIDAK
GOLKAR	SUMIRAH	P	TARUB	18578	2	308	TIDAK
GOLKAR	DARYOTO	L	TARUB	18578	2	54	TIDAK
PKS	KHAPIP APRONI,S.Pdi	L	BOJONG	18578	3	1682	TIDAK
PKS	ENDANG SUCI RAHAYU	P	JATINEGARA	18578	3	918	TIDAK
PDI-P	KH.CHAFIDZ ISA MUFTI ,LC	L	SLAWI	18578	3	87	TIDAK
……	……	……	……	……	…..	…….	……..

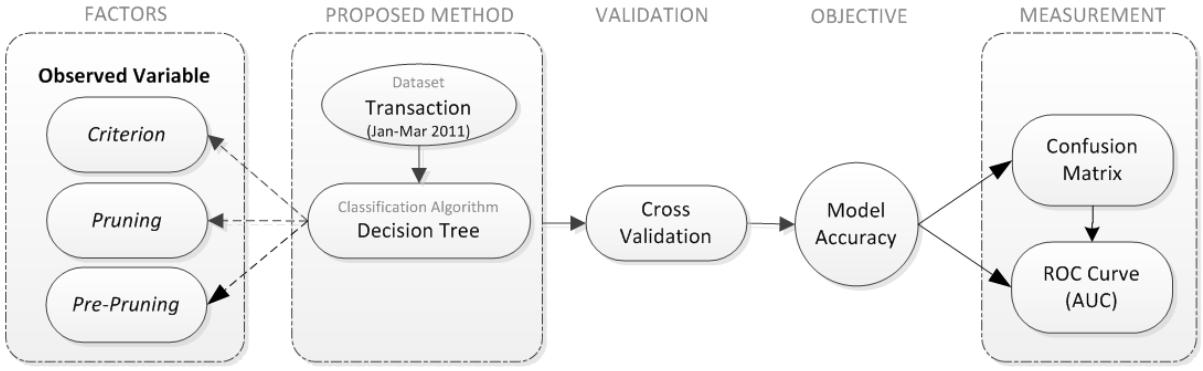


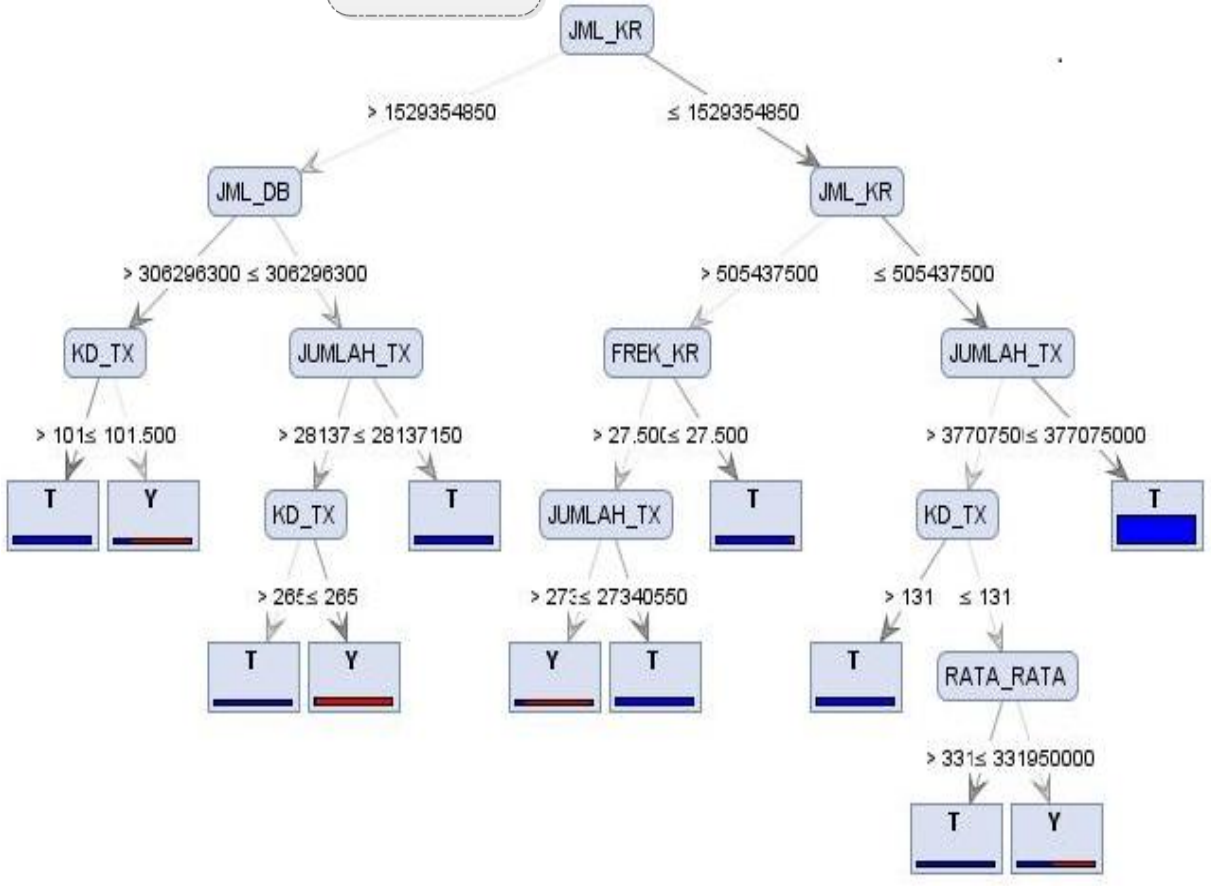
Penentuan Kelayakan Kredit





Deteksi Pencucian Uang

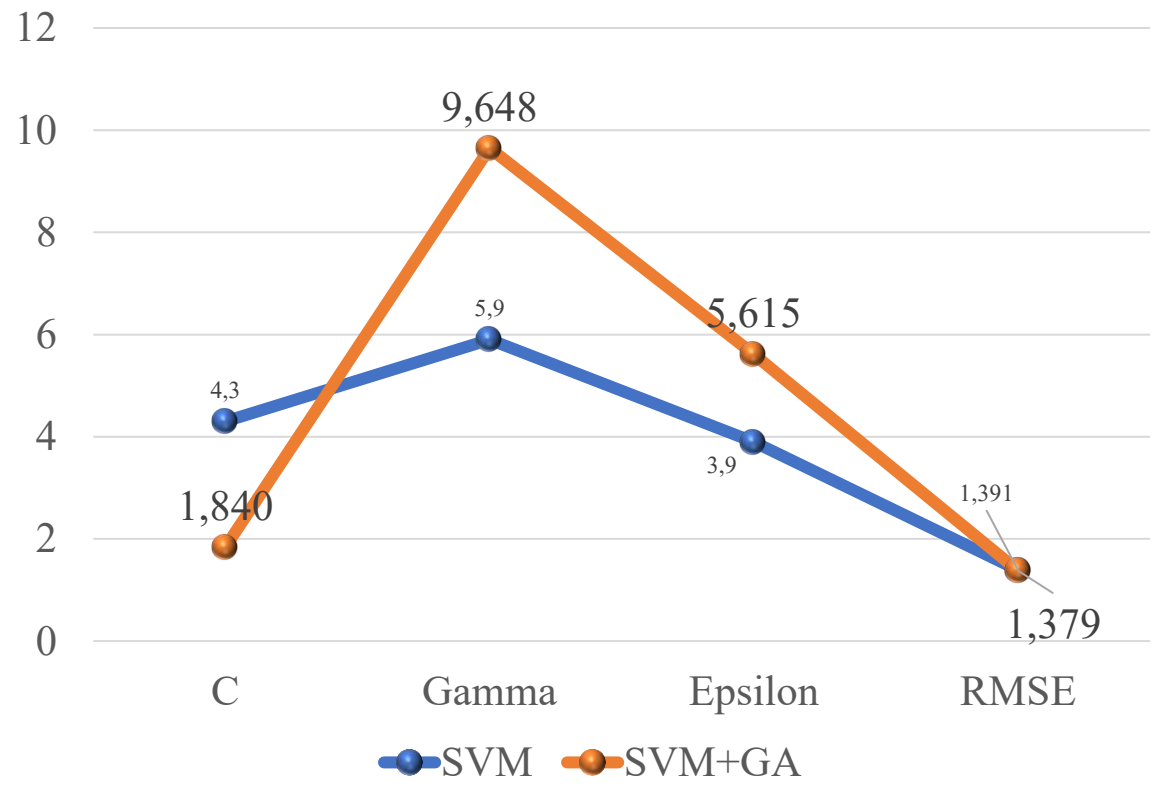




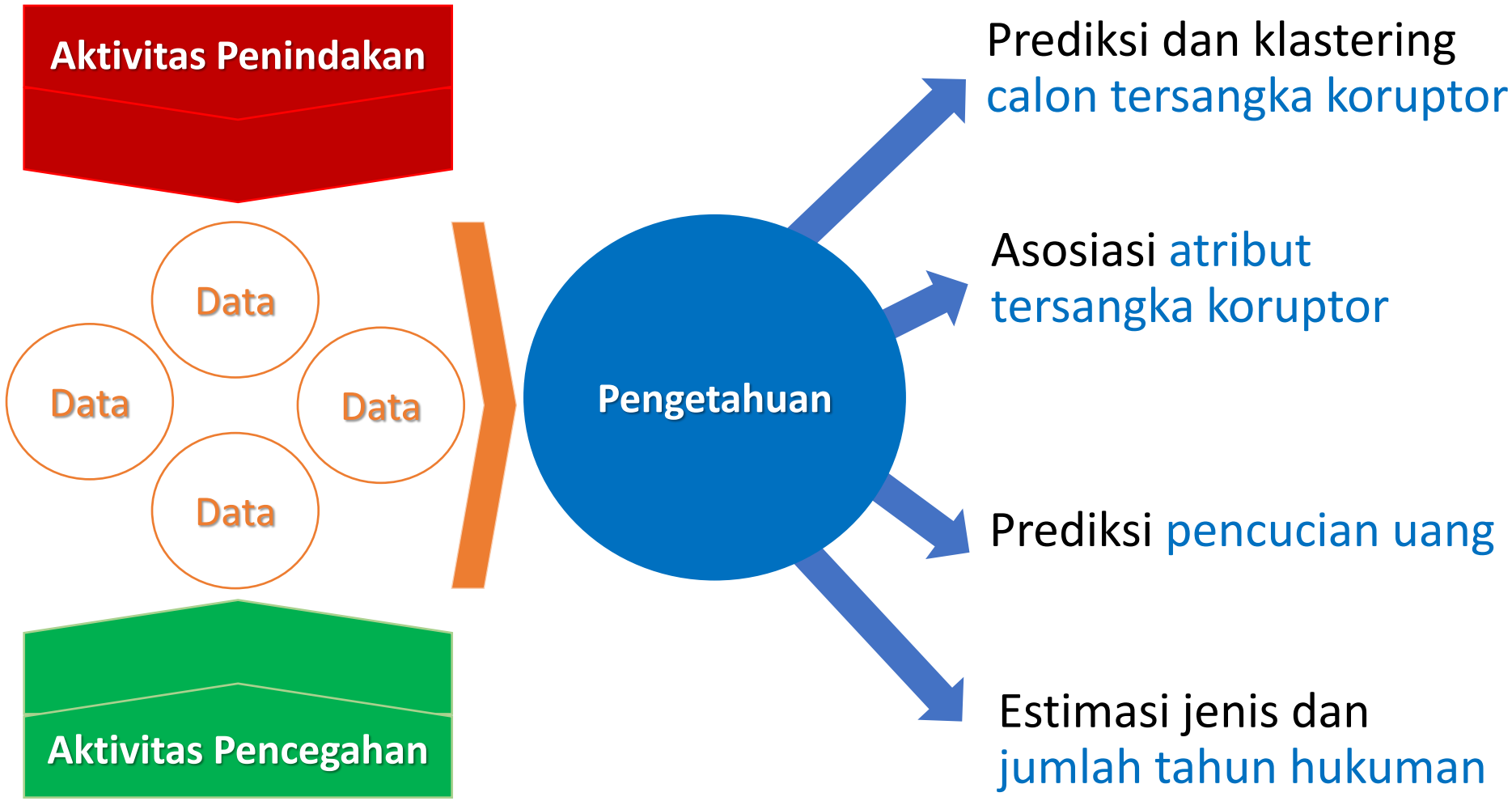
Prediksi Kebakaran Hutan

FFMC	DMC	DC	ISI	temp	RH	wind	rain	ln(area+1)
93.5	139.4	594.2	20.3	17.6	52	5.8	0	0
92.4	124.1	680.7	8.5	17.2	58	1.3	0	0
90.9	126.5	686.5	7	15.6	66	3.1	0	0
85.8	48.3	313.4	3.9	18	42	2.7	0	0.307485
91	129.5	692.6	7	21.7	38	2.2	0	0.357674
90.9	126.5	686.5	7	21.9	39	1.8	0	0.385262
95.5	99.9	513.3	13.2	23.3	31	4.5	0	0.438255

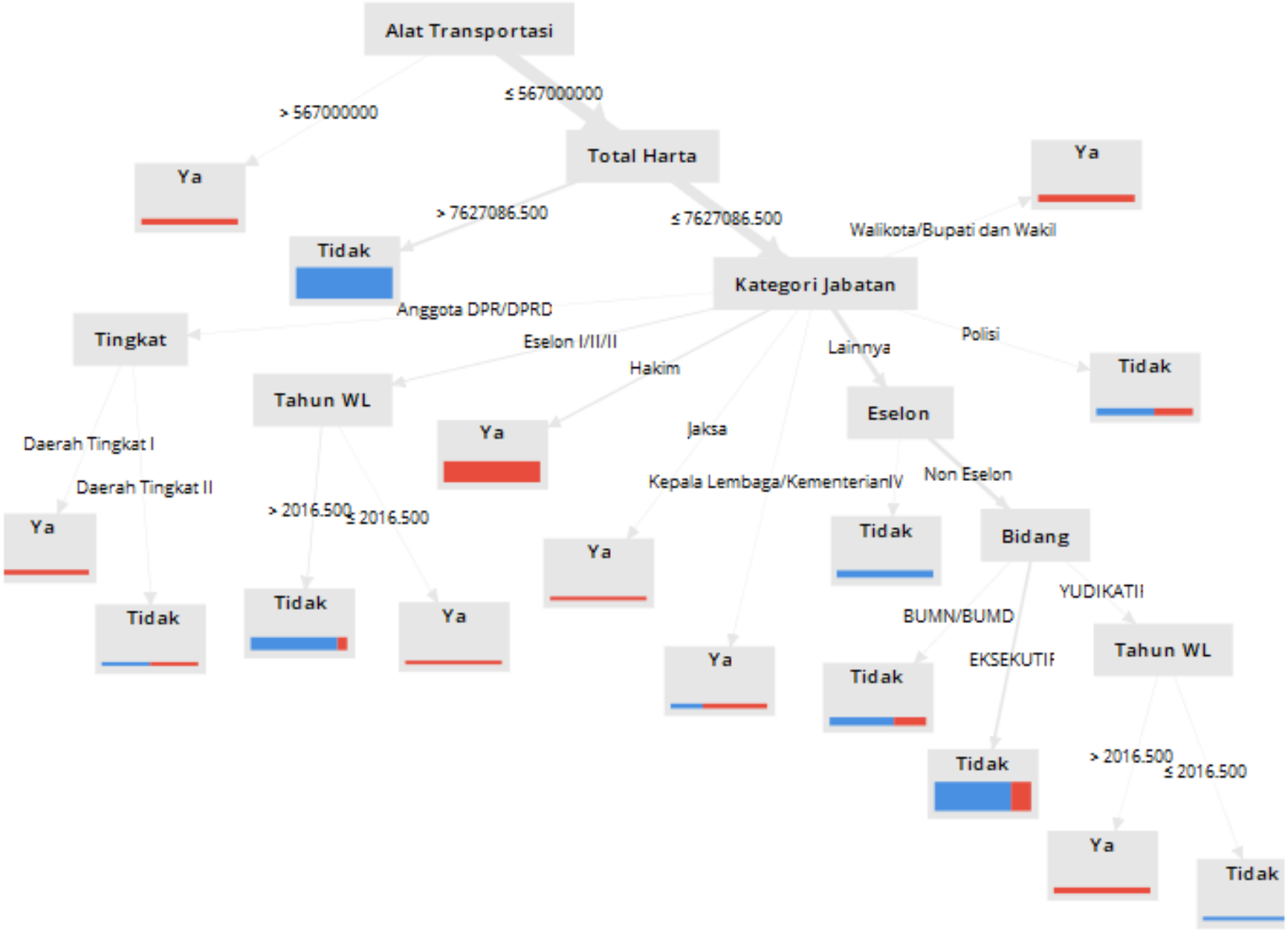
	SVM	SVM+GA
C	4.3	1,840
Gamma (γ)	5.9	9,648
Epsilon (ε)	3.9	5,615
RMSE	1.391	1.379



Profiling dan Prediksi Koruptor

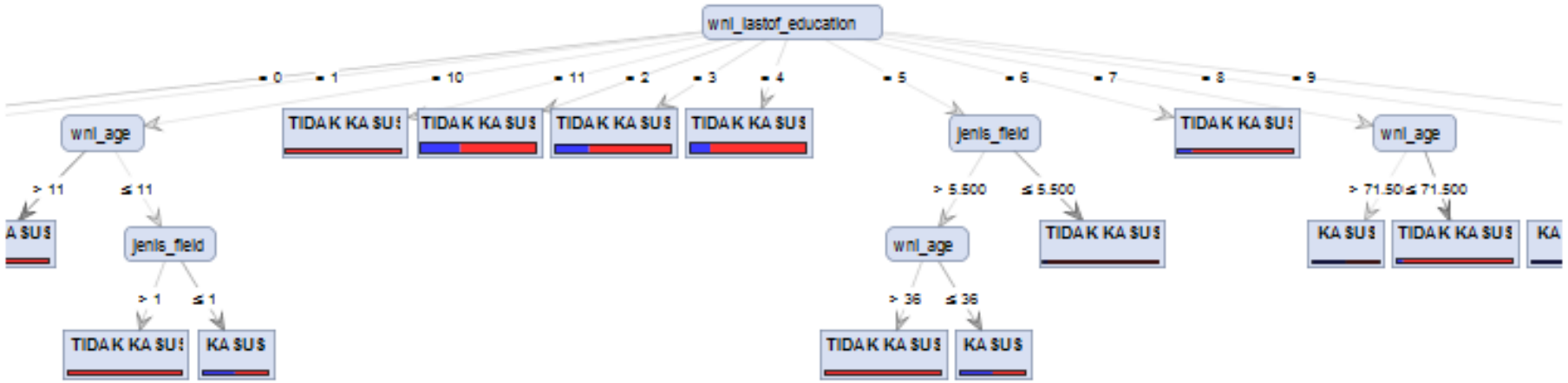


Pola Profil Tersangka Koruptor

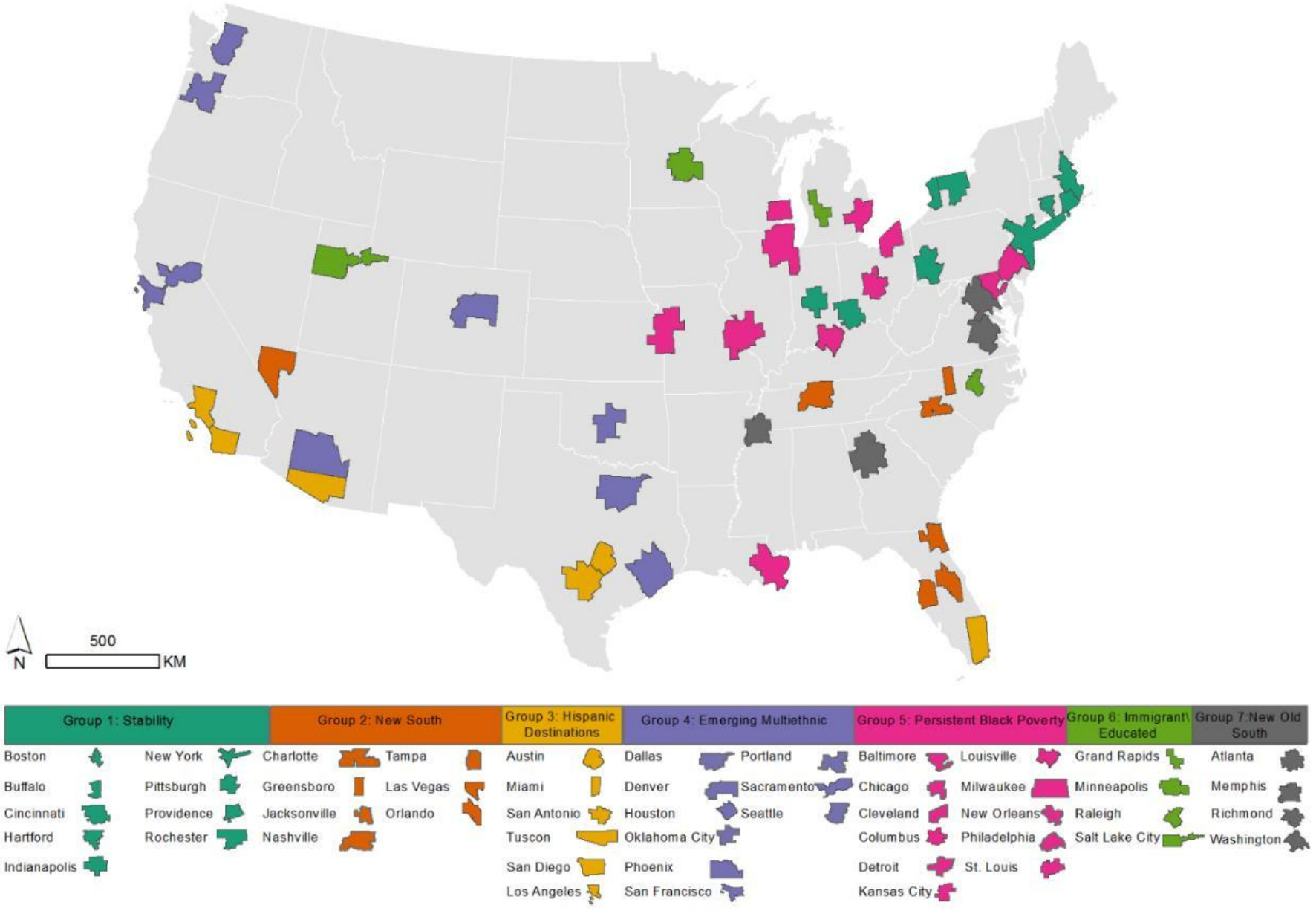


Profiling dan Deteksi Kasus TKI

Row No.	status_kasus	wni_age	wni_lastof_...	gender_name	wni_marital...	wni_local_p...	self_report_...	jenis_field
1	KASUS	-183	3	Perempuan	5	32	PEA	3
2	KASUS	-181	0	Perempuan	5	32	Yordania	6
3	KASUS	-4	0	Perempuan	0	36	RRC	6
4	KASUS	-3	0	Perempuan	0	33	Suriah	6
5	KASUS	-1	0	Laki-laki	0	12	Libya	2
6	KASUS	-1	0	Perempuan	0	32	Libanon	6
7	KASUS	0	0	Laki-laki	0	11	Jepang	2
8	KASUS	0	0	Laki-laki	0	11	Jepang	5
9	KASUS	0	0	Laki-laki	0	11	Libya	2
10	KASUS	0	0	Laki-laki	0	11	Malaysia	3
11	KASUS	0	0	Laki-laki	0	11	Malaysia	6
12	KASUS	0	0	Laki-laki	0	11	Yaman	2
13	KASUS	0	0	Laki-laki	0	12	Amerika Seri...	3



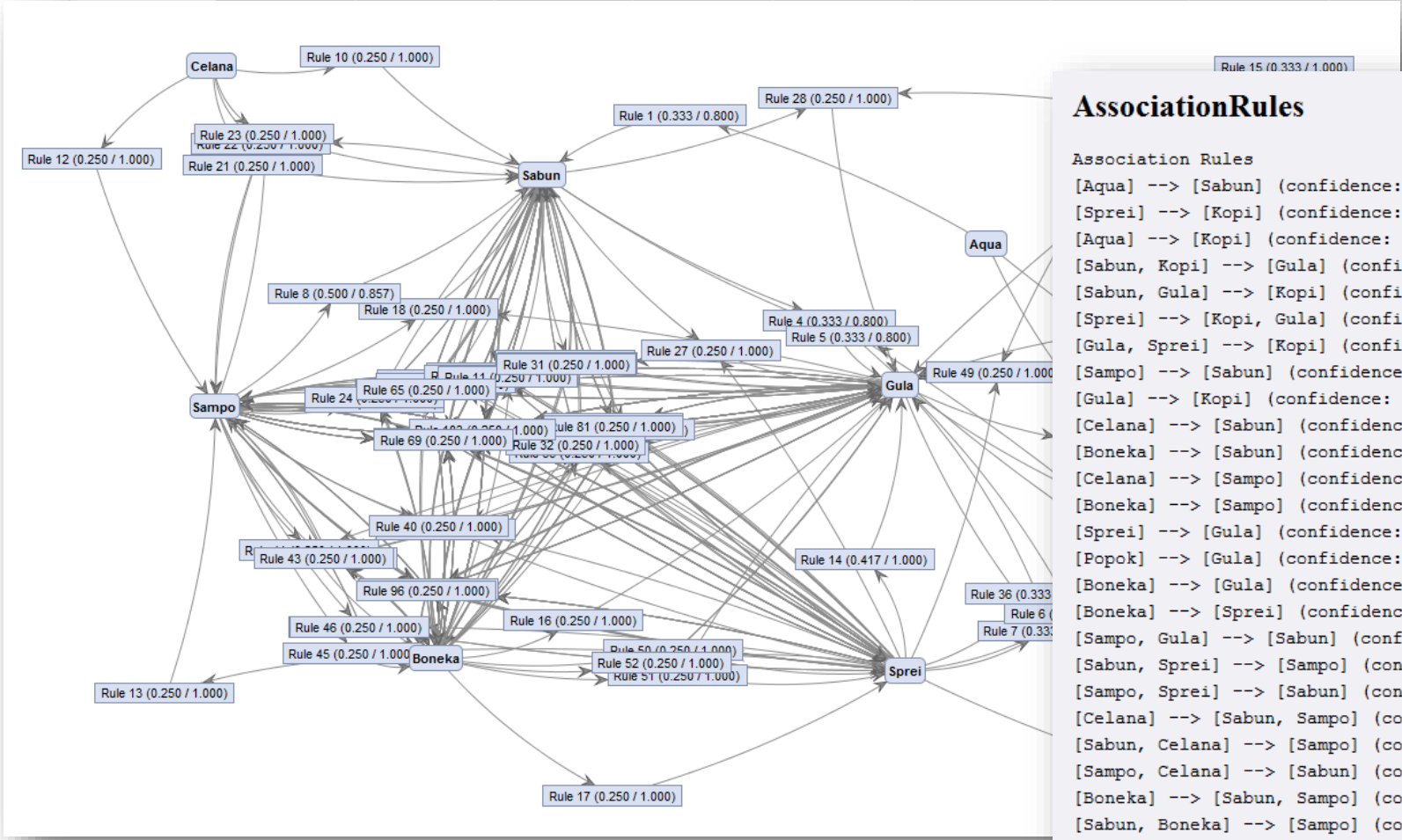
Klasterisasi Tingkat Kemiskinan



Pola Aturan Asosiasi dari Data Transaksi

ExampleSet (12 examples, 0 special attributes, 10 regular attributes)

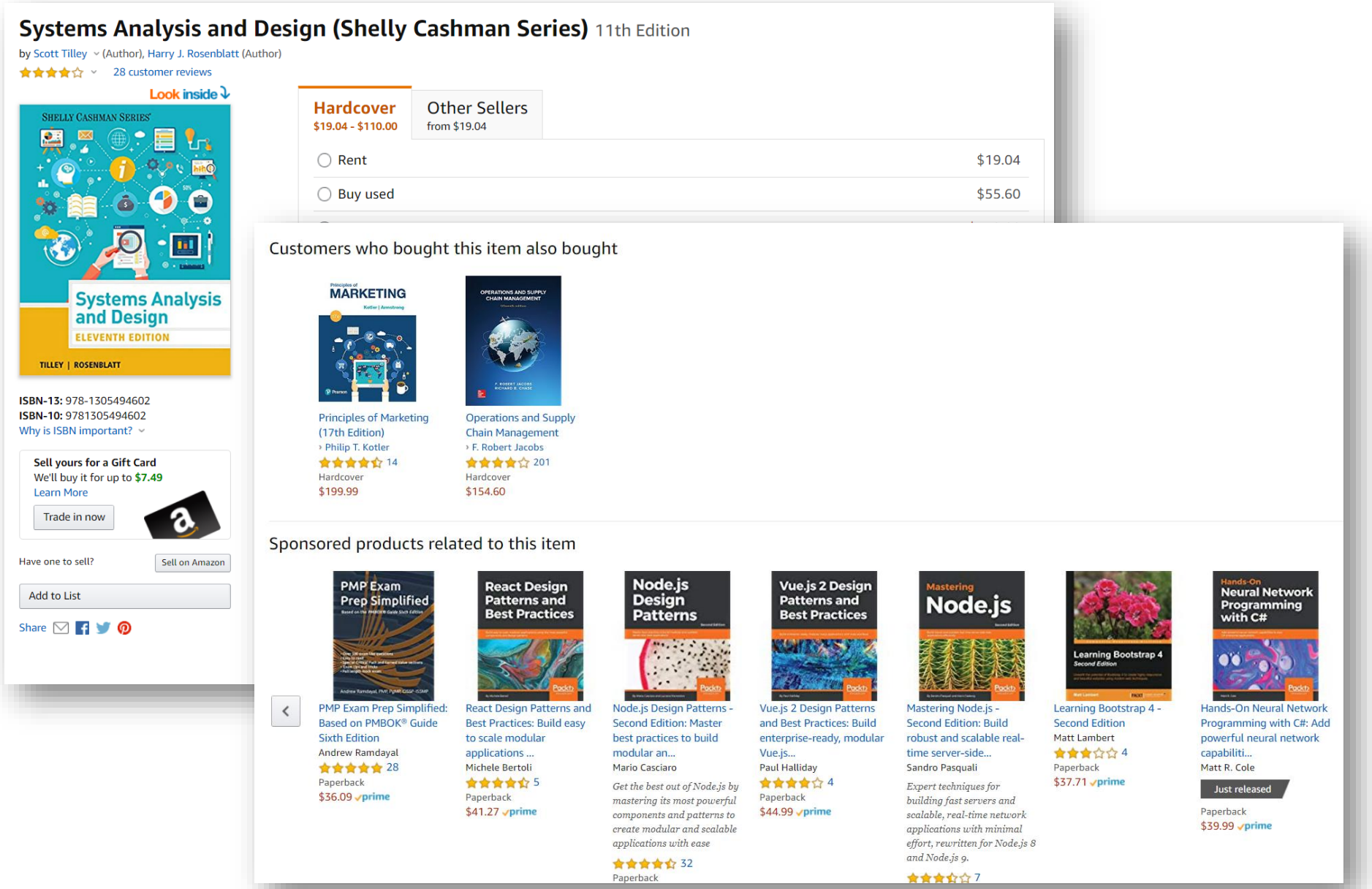
Row No.	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
										1.0



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)
```

Pola Aturan Asosiasi di Amazon.com



From Stupid Apps to Smart Apps

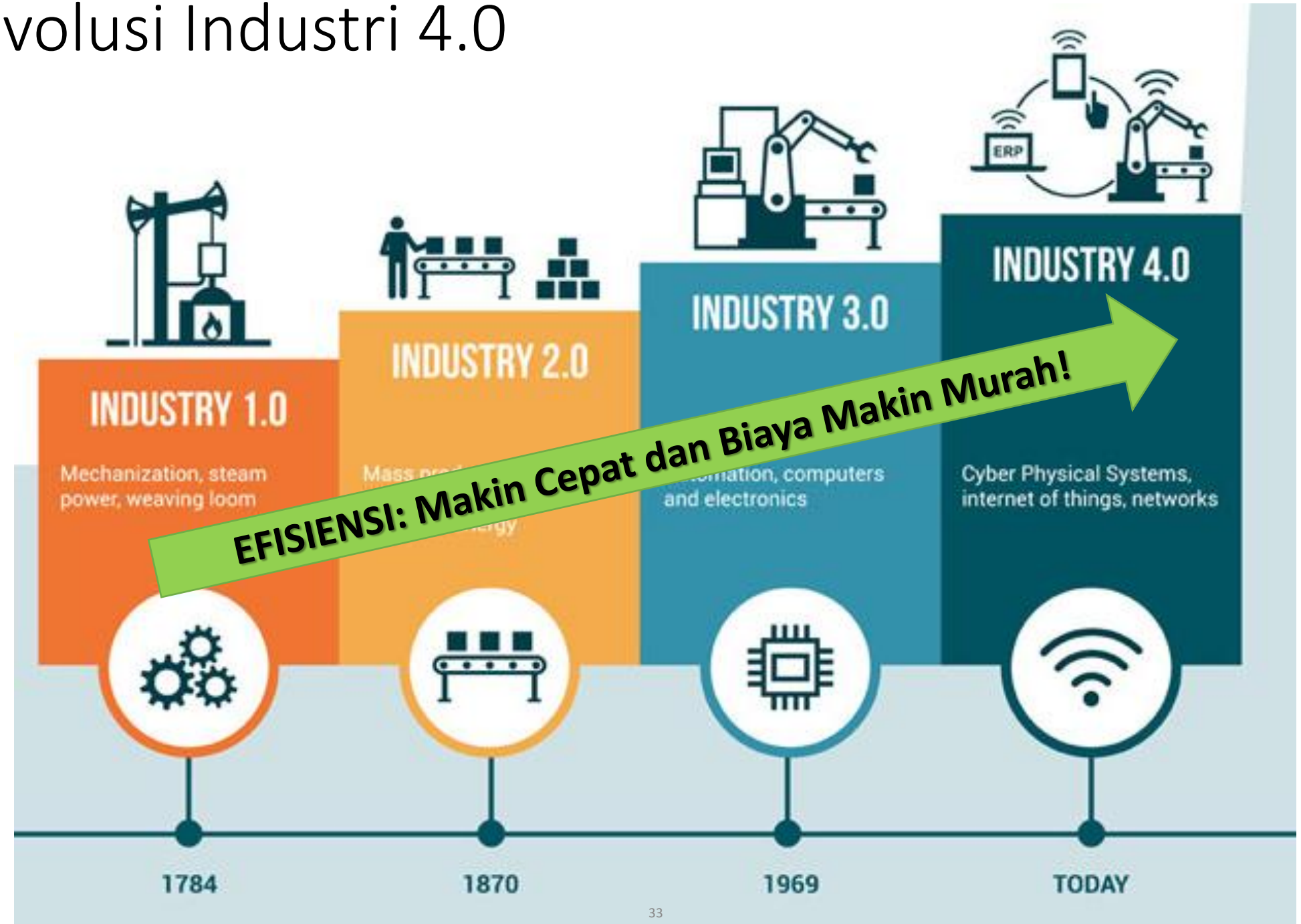
Stupid Applications

- Sistem Informasi Akademik
- Sistem Pencatatan Pemilu
- Sistem Laporan Kekayaan Pejabat
- Sistem Pencatatan Kredit

Smart Applications

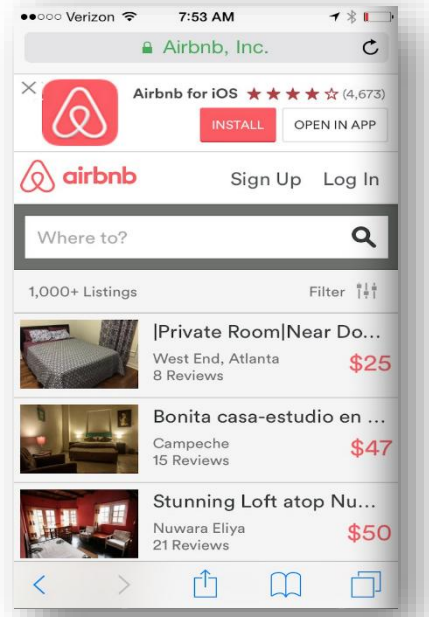
- Sistem Prediksi Kelulusan Mahasiswa
- Sistem Prediksi Hasil Pemilu
- Sistem Prediksi Koruptor
- Sistem Penentu Kelayakan Kredit

Revolusi Industri 4.0

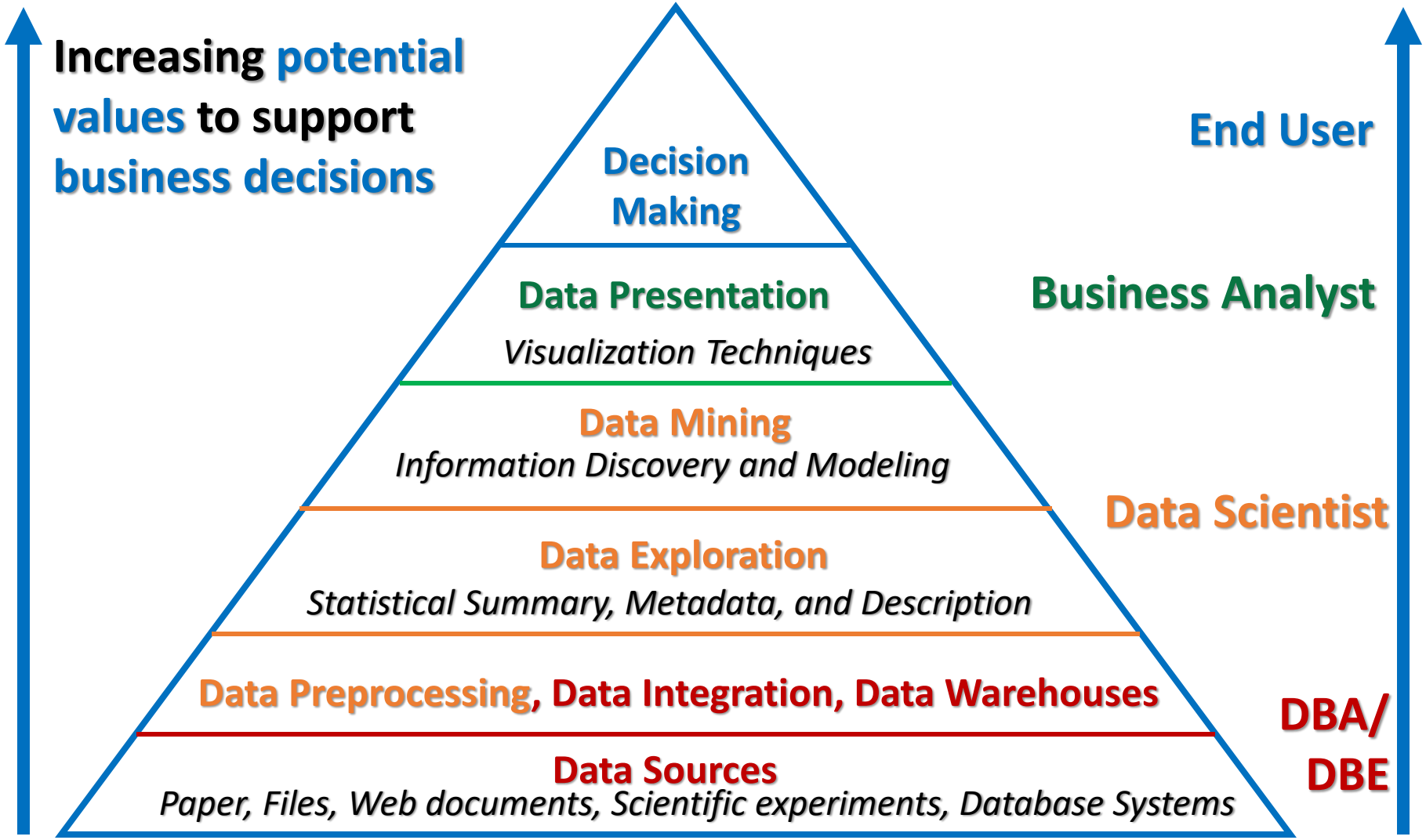


Perusahaan Pengolah Pengetahuan

- Uber - the world's largest taxi company, owns no vehicles
- Google - world's largest media/advertising company, creates no content
- Alibaba - the most valuable retailer, has no inventory
- Airbnb - the world's largest accommodation provider, owns no real estate
- Gojek - perusahaan angkutan umum, tanpa memiliki kendaraan



Data Mining Tasks and Roles in General

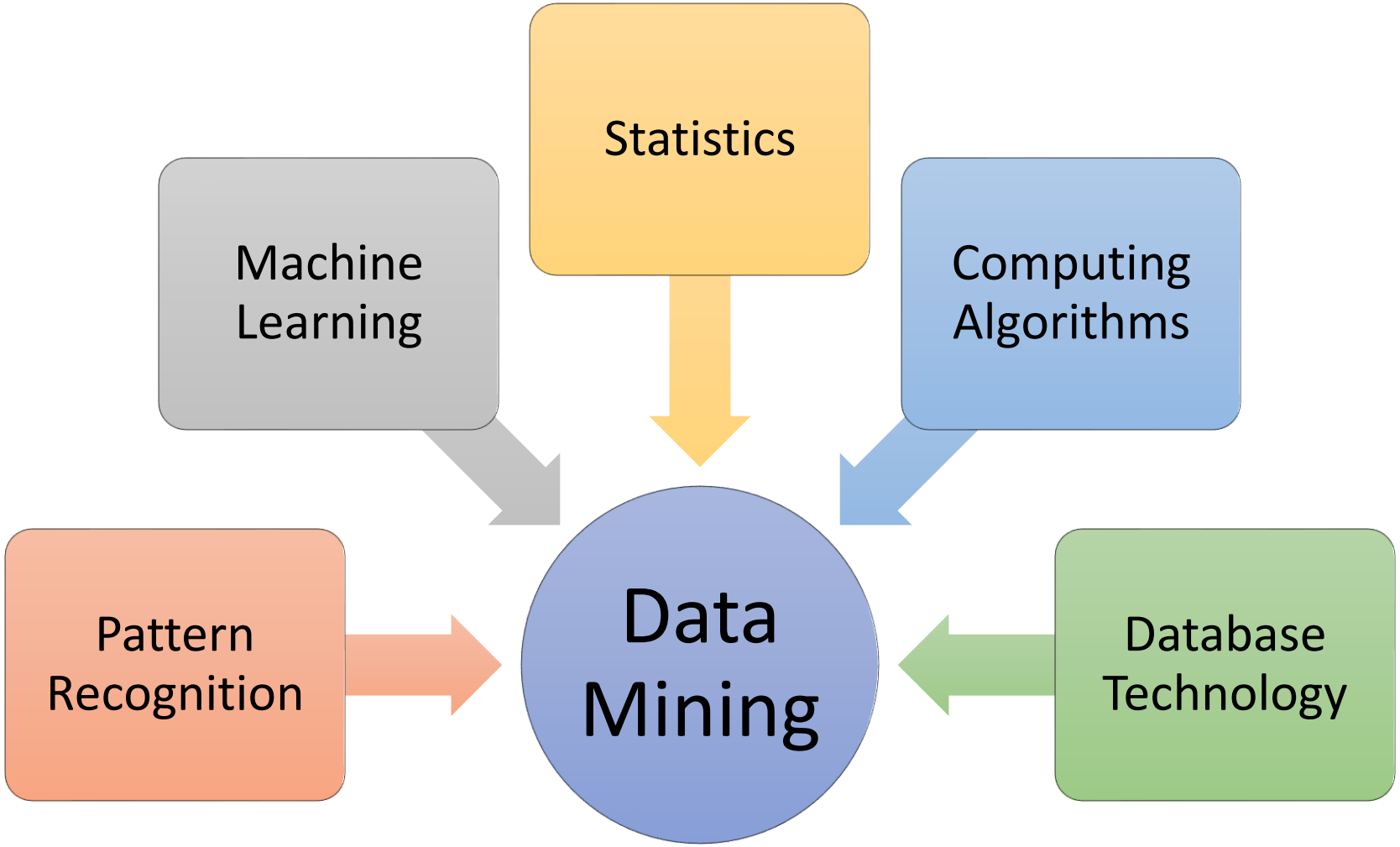


Data Mining Tasks and Roles in Product Development

- Menganalisis data untuk dijadikan pola pengetahuan (model/rule/formula/tree)
- Pola pengetahuan (model/rule/formula/tree) dimasukkan ke dalam sistem (software)
- Sistem (software) menjadi cerdas dan bermanfaat signifikan dalam meningkatkan value dan benefit dari perusahaan/organisasi

- Dimana peran data scientist dalam perusahaan pengembang teknologi (startup business atau GAFAM)?
  - Data Scientist? Software Engineer? Researcher? IT Infrastructure Engineer?

Hubungan Data Mining dan Bidang Lain

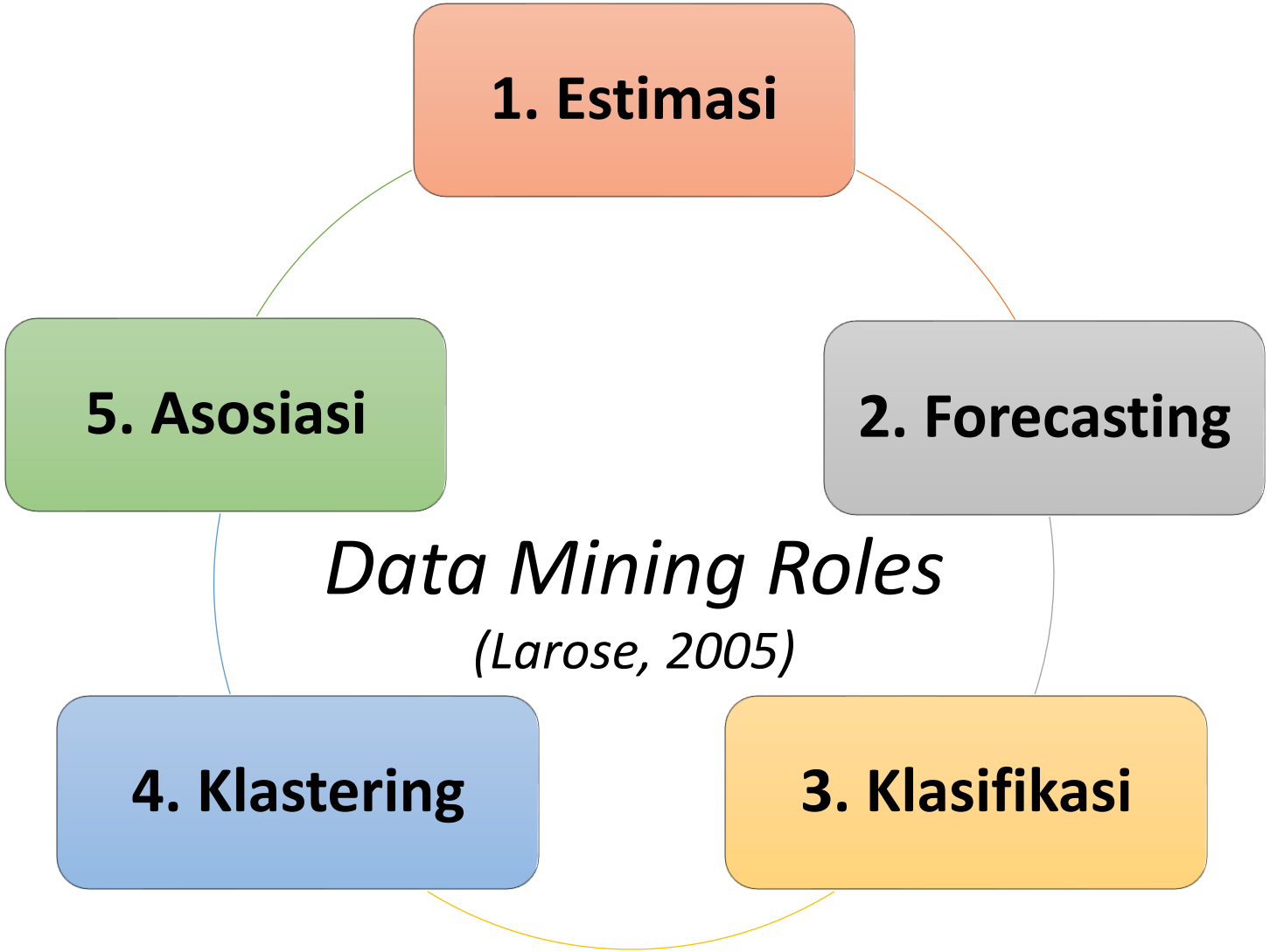


Masalah-Masalah di Data Mining

1. Tremendous amount of data
   - Algorithms must be highly scalable to handle such as tera-bytes of data
2. High-dimensionality of data
   - Micro-array may have tens of thousands of dimensions
3. High complexity of data
   - Data streams and sensor data
   - Time-series data, temporal data, sequence data
   - Structure data, graphs, social networks and multi-linked data
   - Heterogeneous databases and legacy databases
   - Spatial, spatiotemporal, multimedia, text and Web data
   - Software programs, scientific simulations
4. New and sophisticated applications

1.2 Peran Utama dan Metode Data Mining

Peran Utama Data Mining



1. Estimasi Waktu Pengiriman Pizza

Label

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

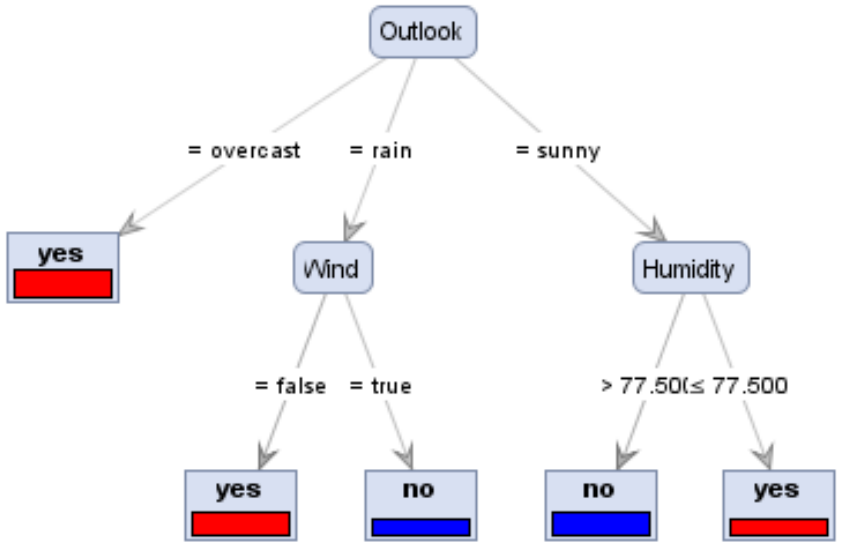
- Linear regression function

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)



3. Korelasi dan Asosiasi

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

5. Cluster (Klaster)

2. Forecasting Harga Saham

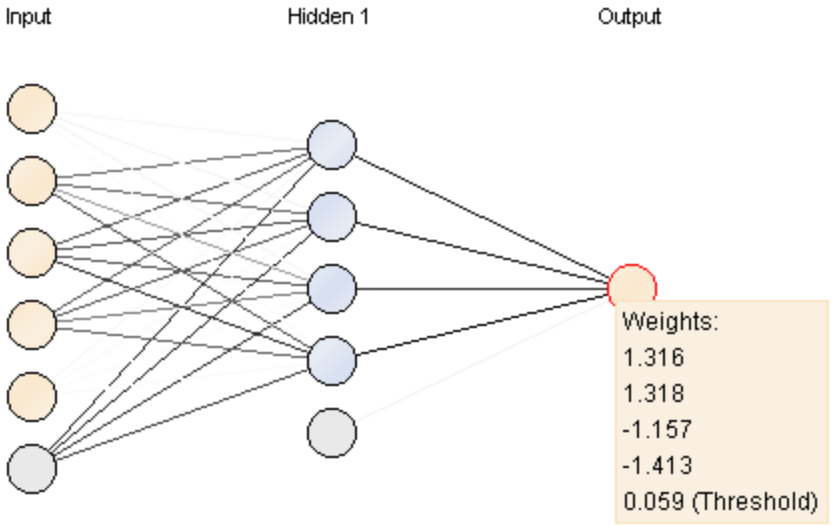
Label Time Series

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)

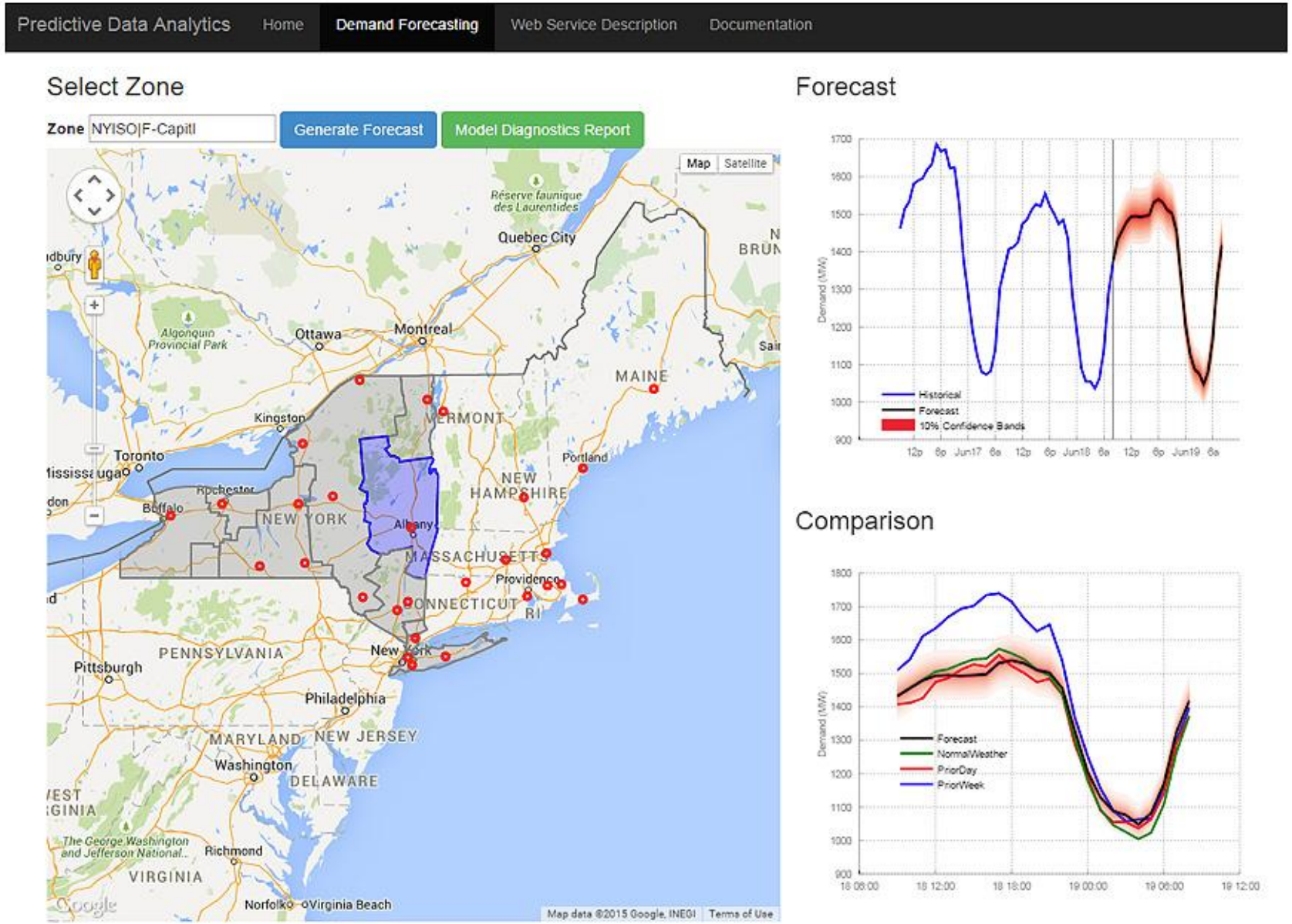
Pembelajaran dengan Metode Forecasting (*Neural Network*)

Pengetahuan berupa Rumus Neural Network

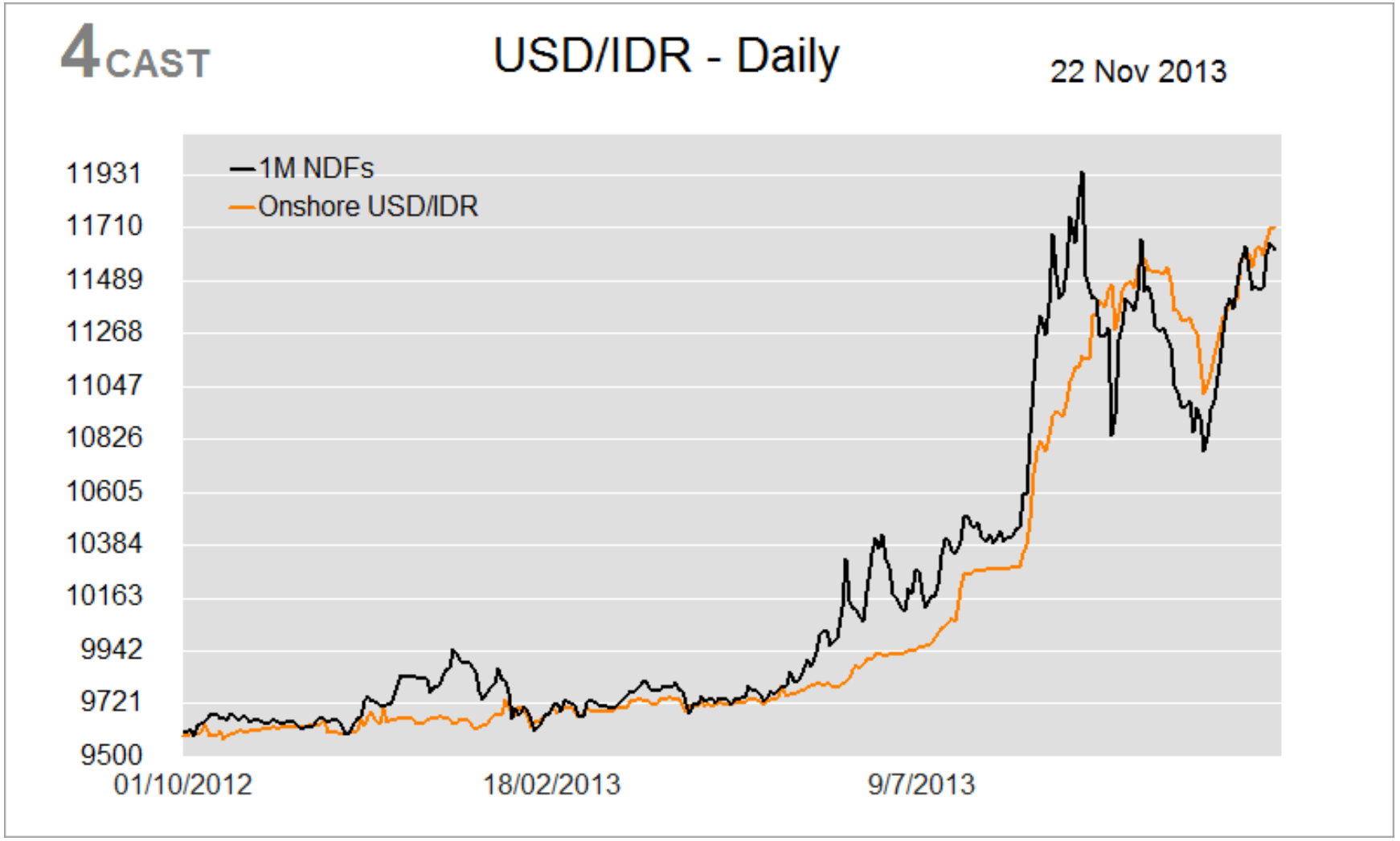


Prediction Plot

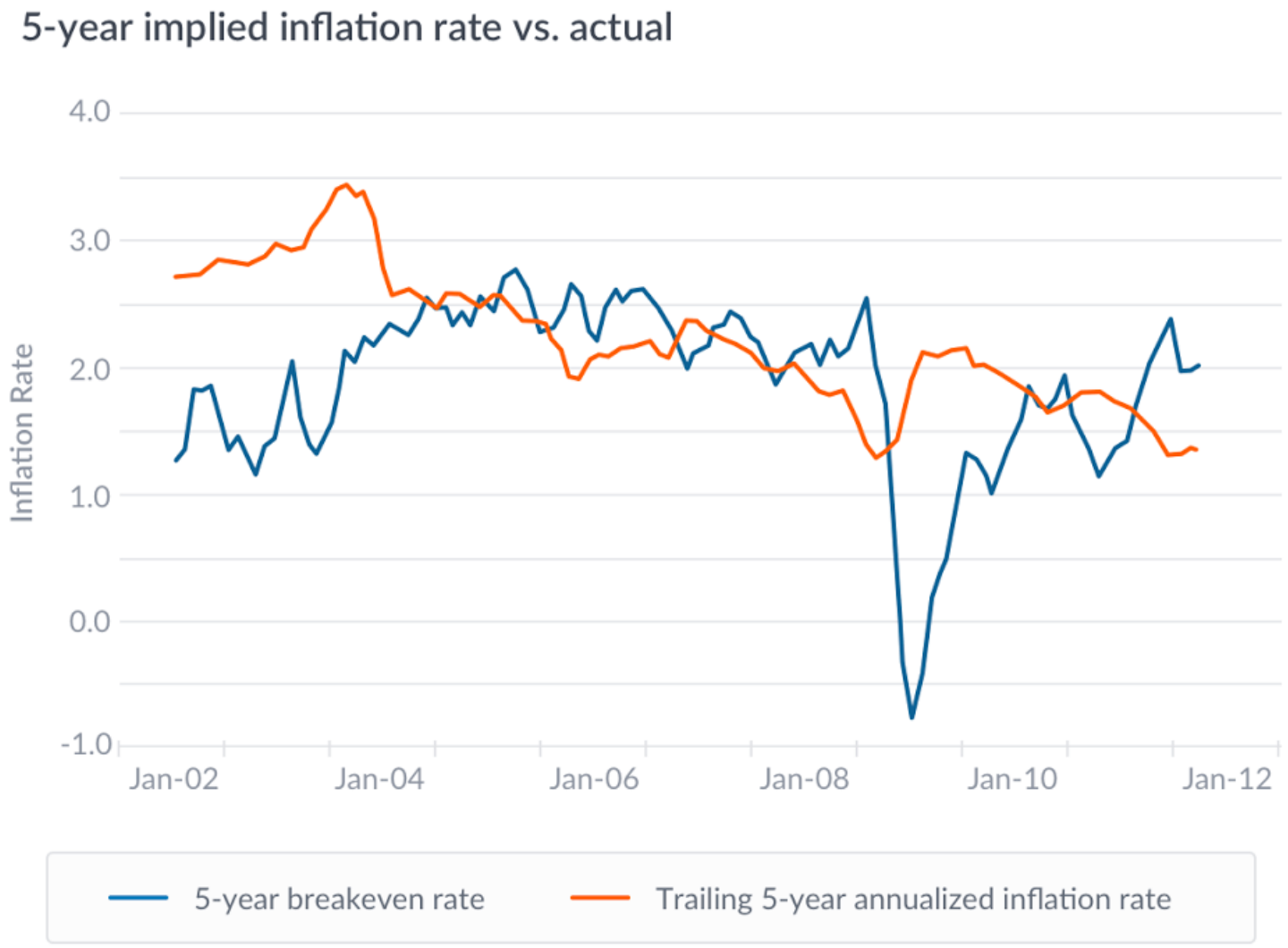
Forecasting Cuaca



Exchange Rate Forecasting



Inflation Rate Forecasting



*Source: Bloomberg

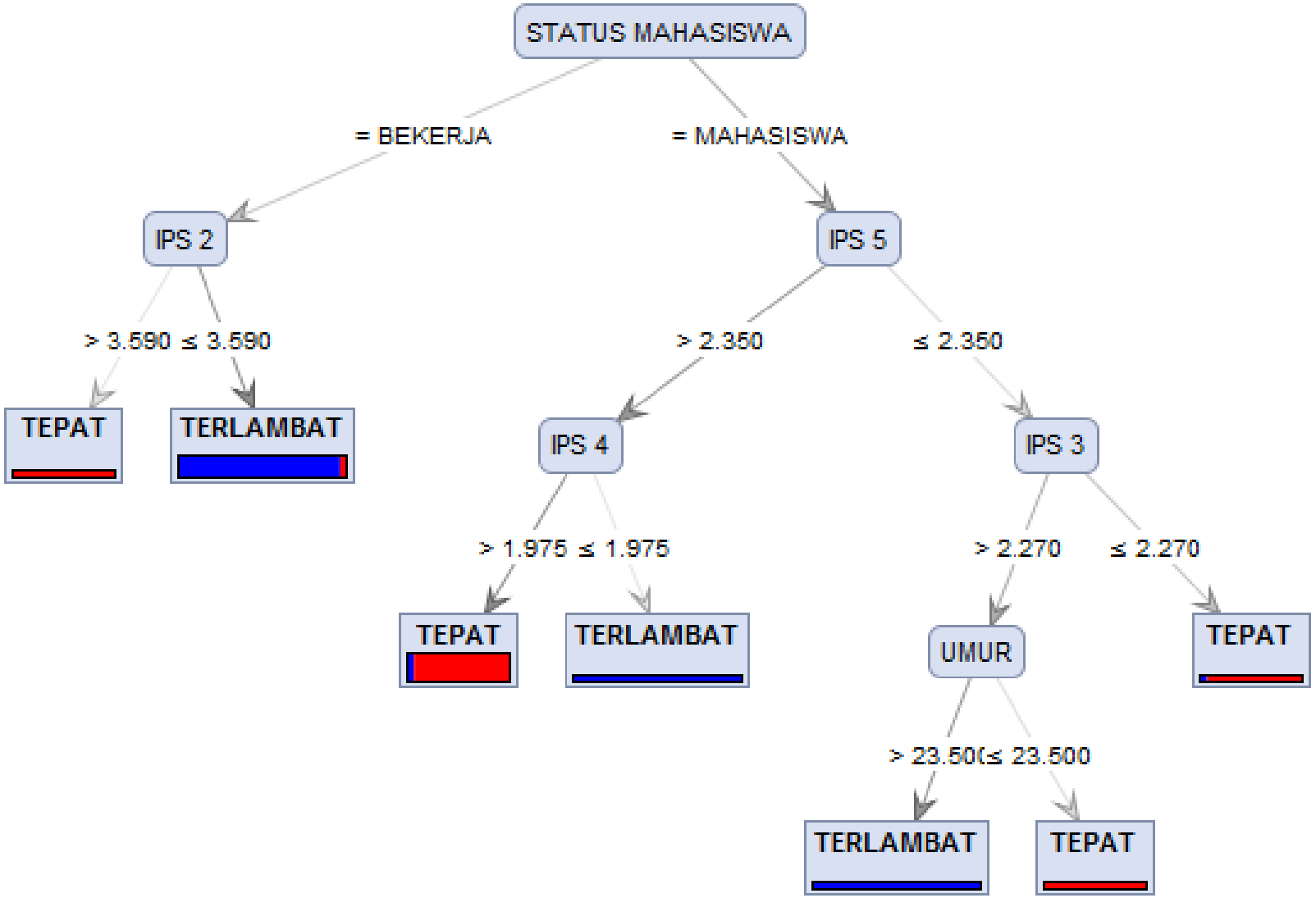
3. Klasifikasi Kelulusan Mahasiswa

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*)

Pengetahuan Berupa Pohon Keputusan



Contoh: Rekomendasi Main Golf

- **Input**:

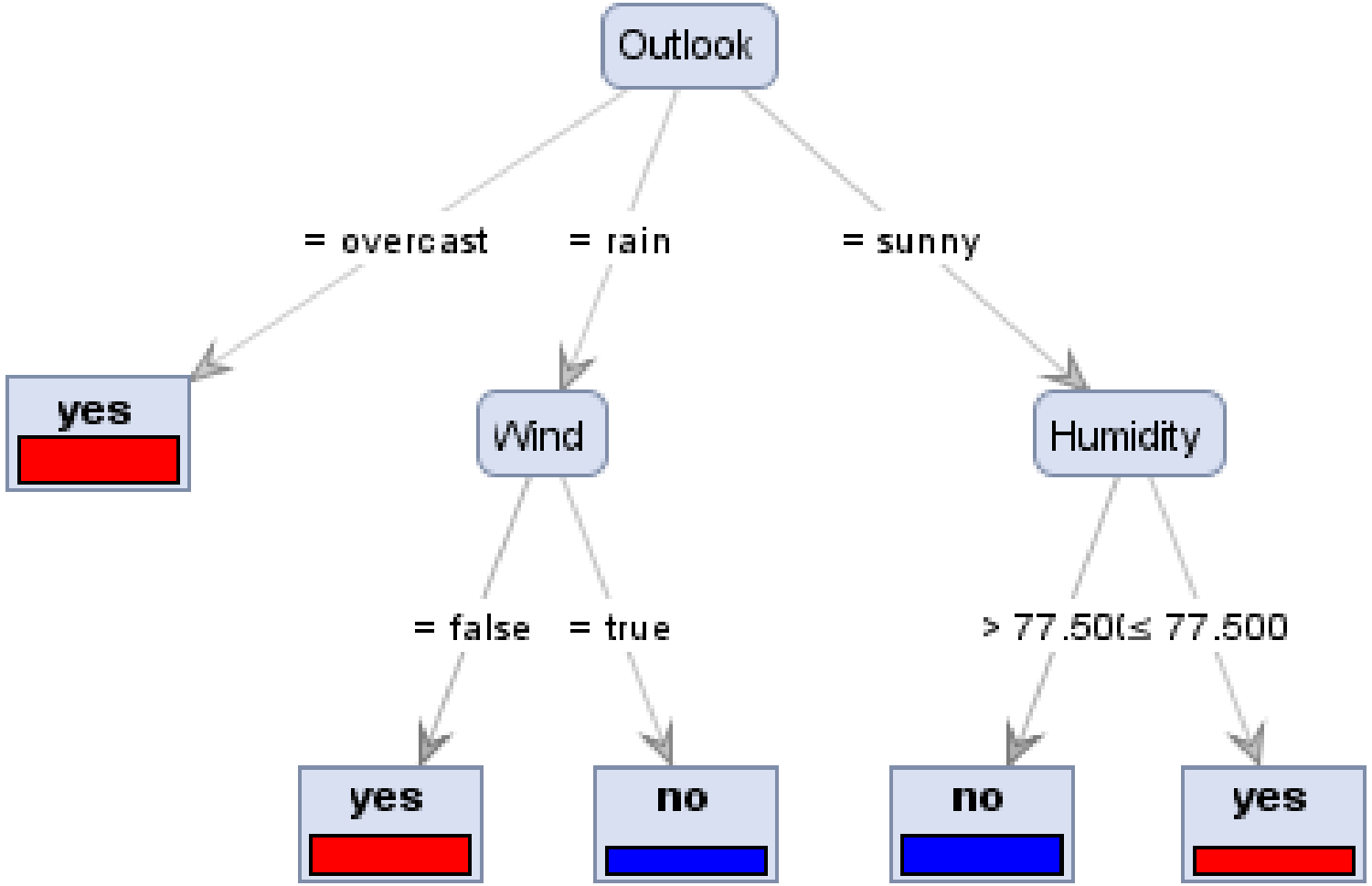
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false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

- Output (Tree):



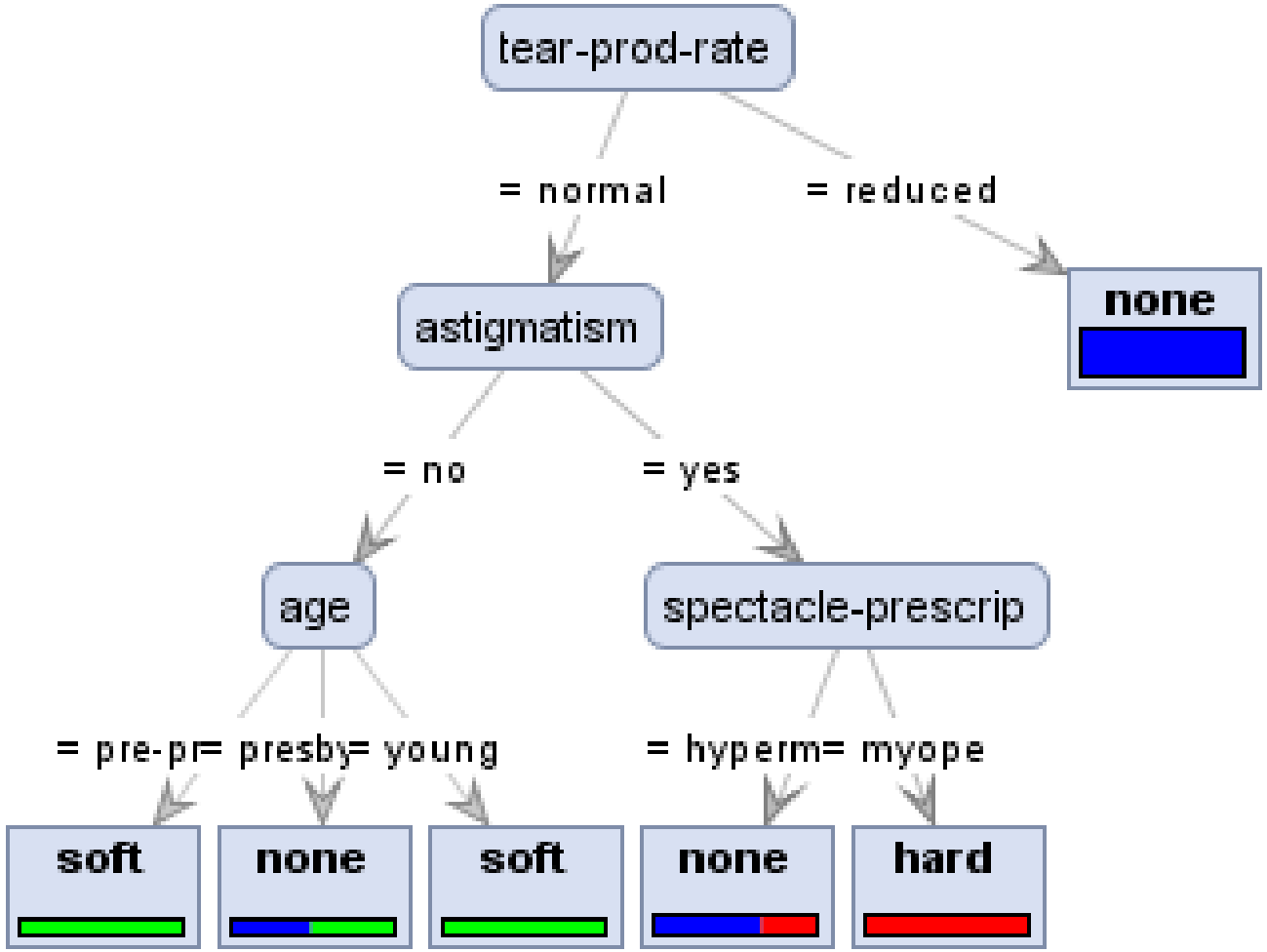
Contoh: Rekomendasi Contact Lens

- 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):



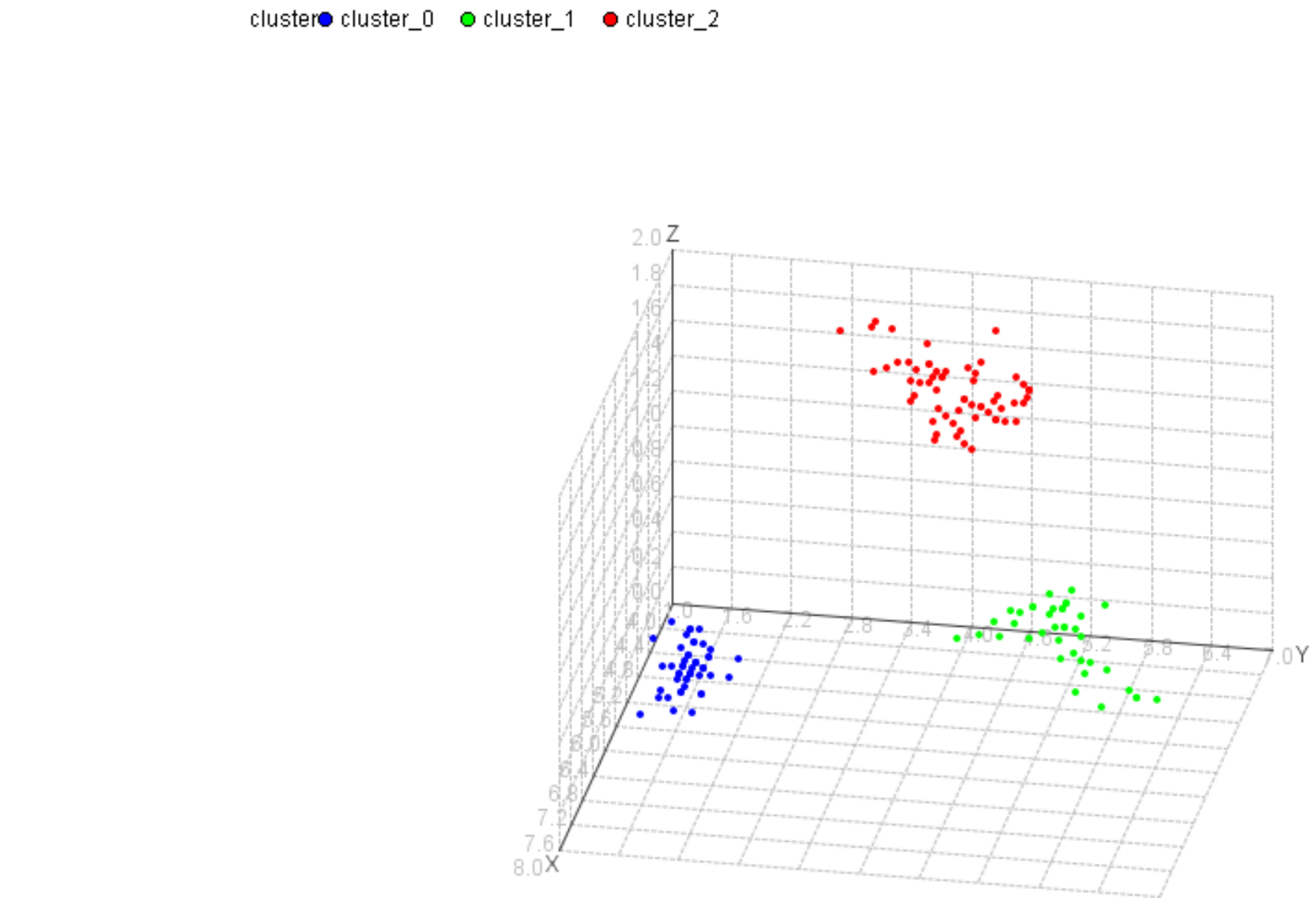
4. Klastering Bunga Iris

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

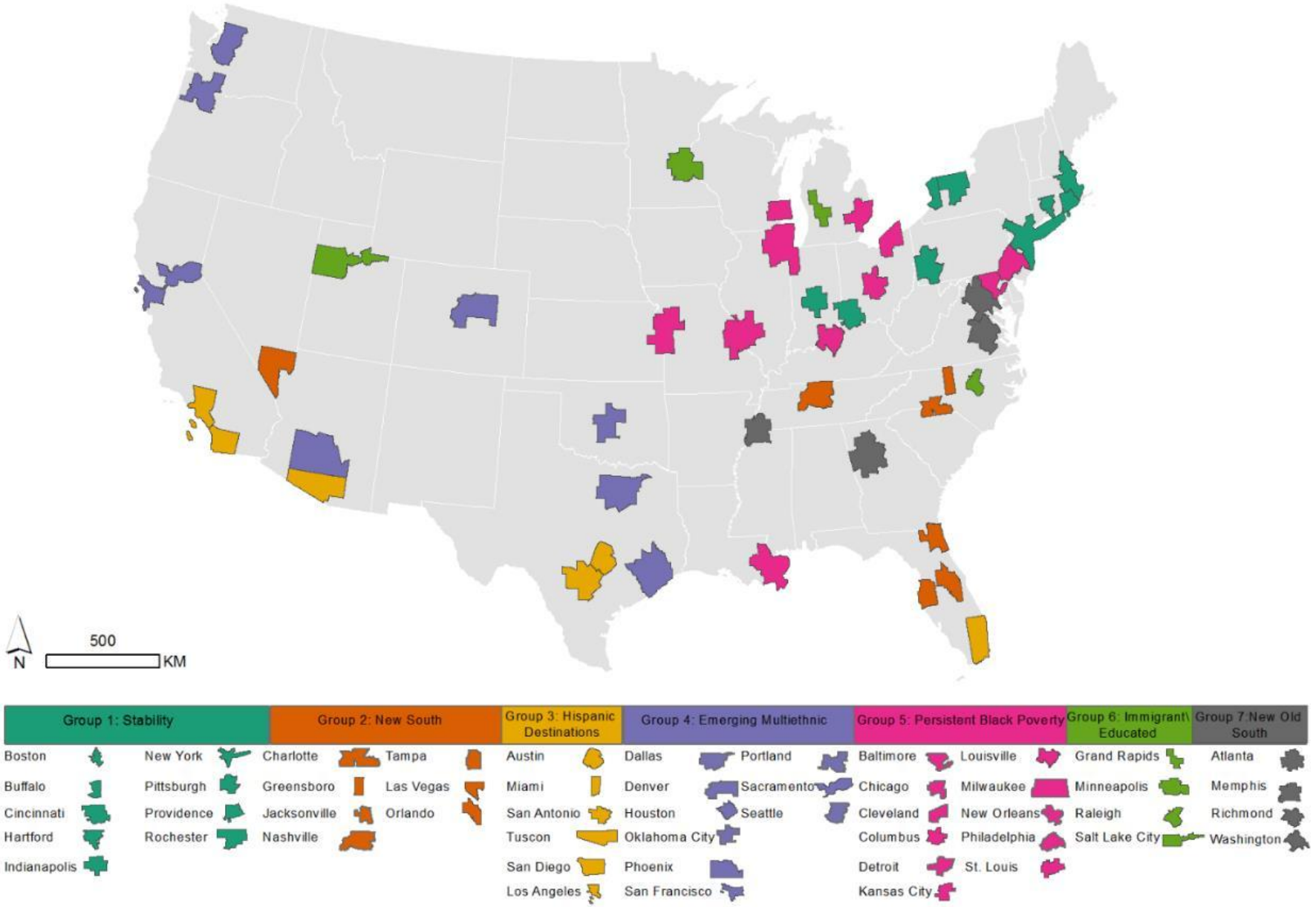
Pembelajaran dengan
Metode Klastering (*K-Means*)

Pengetahuan (Model) Berupa Klaster



Klastering Jenis Pelanggan

Poverty Rate Clustering



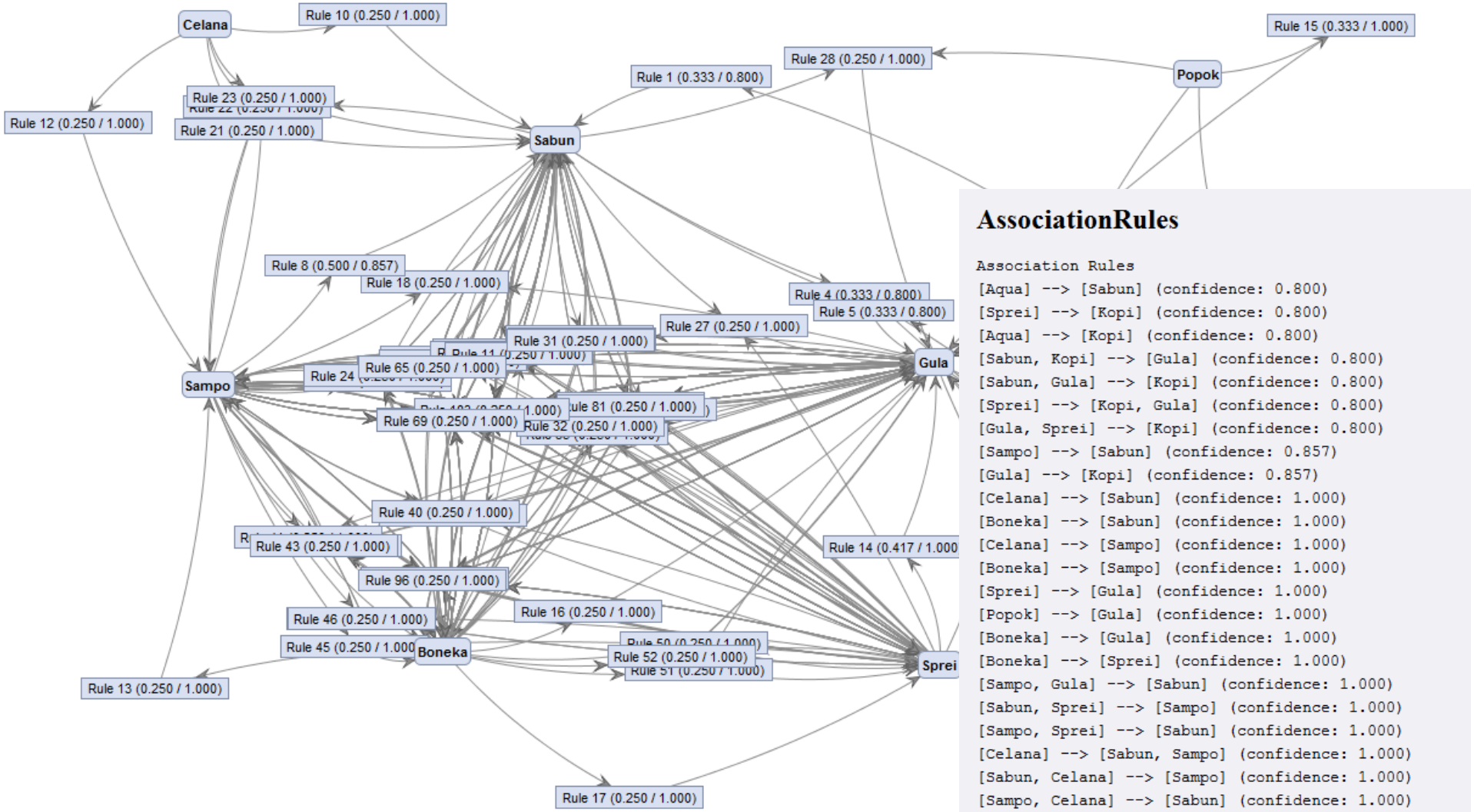
5. Aturan Asosiasi Pembelian Barang

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



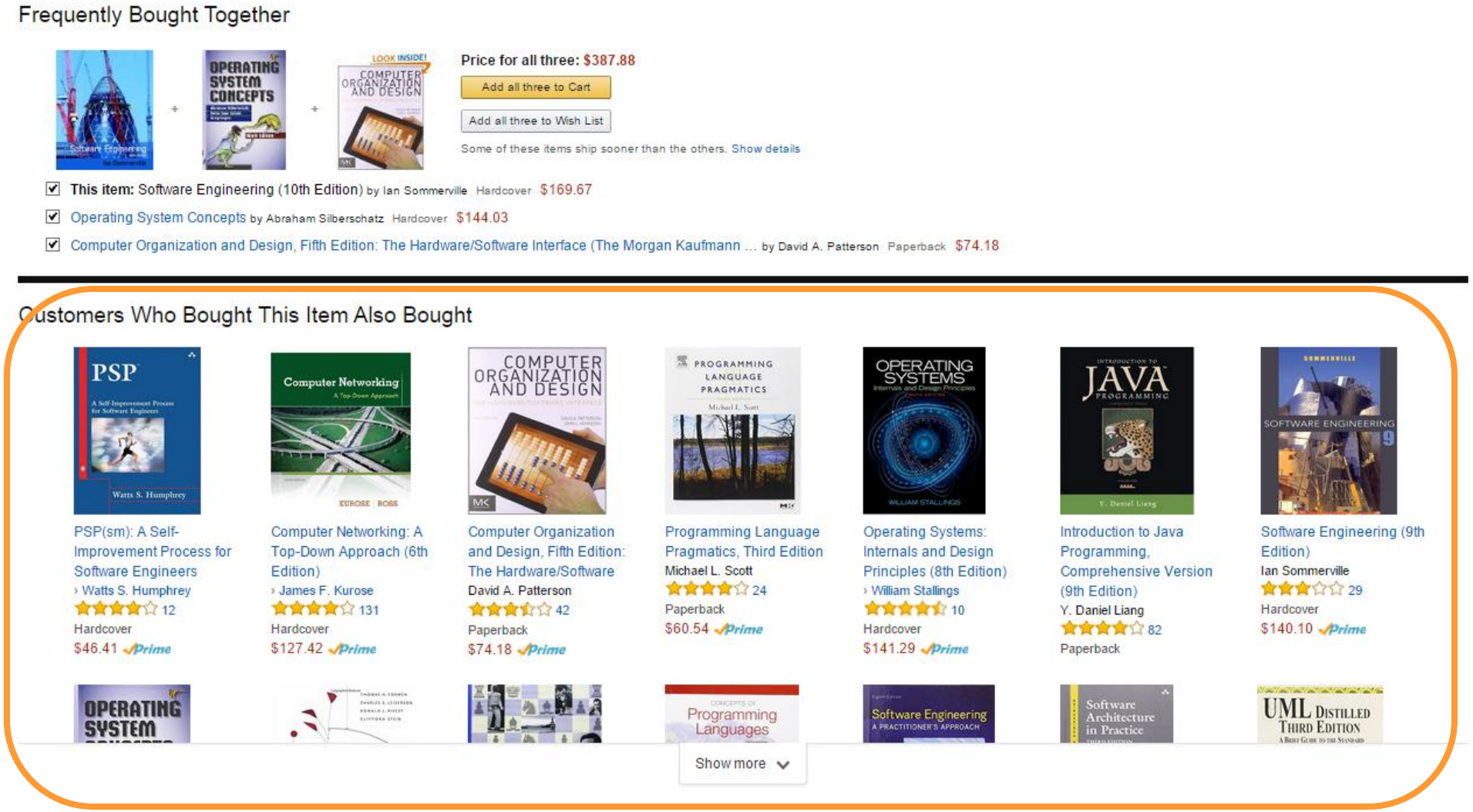
AssociationRules

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)

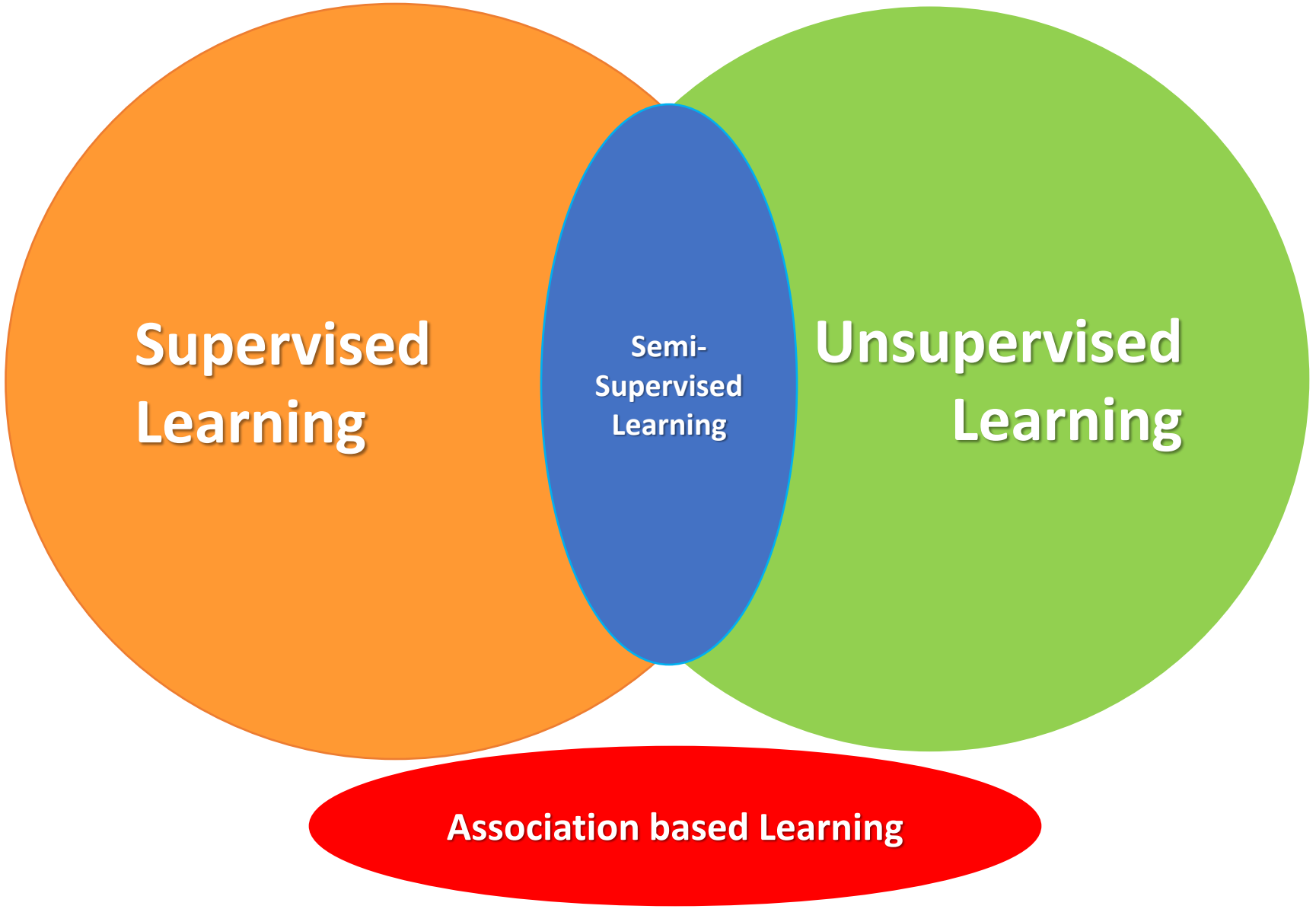
Contoh Aturan Asosiasi

- 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



Kategorisasi Algoritma Data Mining



1. Supervised Learning

- Pembelajaran dengan guru, data set memiliki target/label/class
- Sebagian besar algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
- Algoritma melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variable prediktor

Dataset dengan Class

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*

Nominal

Numerik

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 tanpa Class

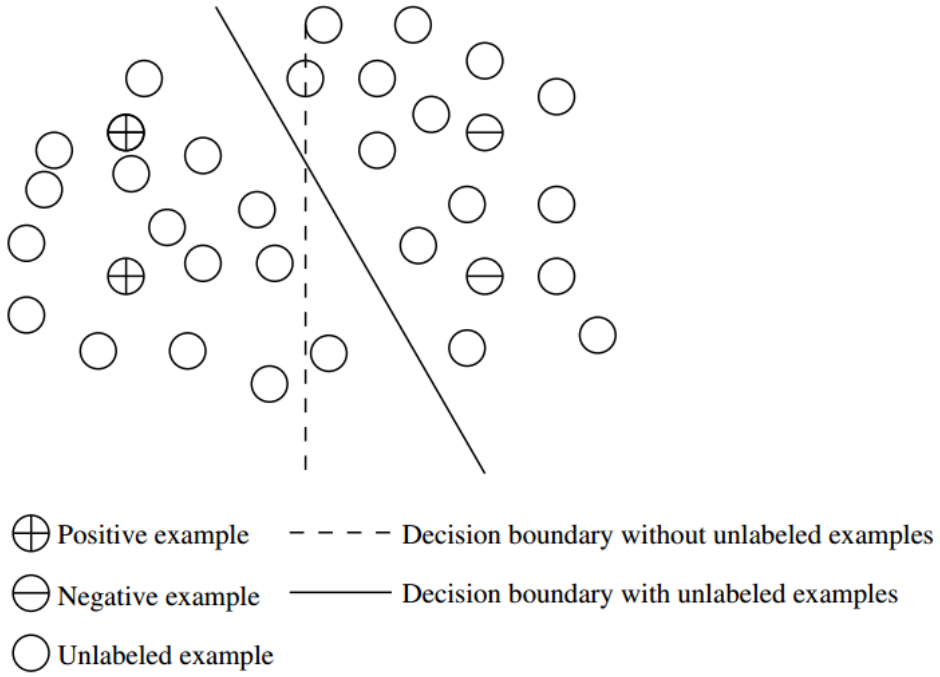
Attribute/Feature/Dimension

	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

3. Semi-Supervised Learning

- Semi-supervised learning adalah metode data minir yang menggunakan data dengan label dan tidak berlabel sekaligus dalam proses pembelajarannya

- Data yang memiliki kelas digunakan untuk membentuk model (pengetahuan), data tanpa label digunakan untuk membuat batasan antara kelas



Contoh Penerapan Data Mining

- Penentuan kelayakan kredit pemilihan rumah di bank
- Penentuan pasokan listrik PLN untuk wilayah Jakarta
- Prediksi profile tersangka koruptor dari data pengadilan
- Perkiraan harga saham dan tingkat inflasi
- Analisis pola belanja pelanggan
- Memisahkan minyak mentah dan gas alam
- Penentuan pola pelanggan yang loyal pada perusahaan operator telepon
- Deteksi pencucian uang dari transaksi perbankan
- Deteksi serangan (*intrusion*) pada suatu jaringan

1.3 Kategori dan Tipe Data

Kategorisasi Data

Dataset adalah input utama dalam data mining: baris = objek, kolom = atribut.

Dataset: tabel dua dimensi

ID	Kota	Waktu	Temperatur
1	Bandung	19:45	17°C
2	Surabaya	19:45	34°C

Baris: objects / samples / instances / records

Kolom: attributes / dimensions / features

1 Univariate

Distribusi data melibatkan satu atribut saja.

2 Bivariate

Distribusi data melibatkan dua atribut.

n Multivariate

Distribusi data melibatkan banyak atribut.

Intinya: Sebelum memilih metode data mining, pahami dulu bentuk dataset dan atributnya.

Data Terstruktur dan Tidak Terstruktur

Data tidak terstruktur perlu diolah agar siap menjadi dataset tabular.

T Terstruktur

Format standar, struktur jelas, urutan tetap, mudah diakses. Umumnya tersimpan sebagai tabel pada basis data relasional.

ID	Kota	Waktu	Temp.
1	Bandung	19:45	17°C
2	Surabaya	19:45	34°C

→

U Tidak Terstruktur

Tidak memiliki format baku. Contohnya teks, video, gambar, audio, atau unggahan media sosial.

139492 Malam ini, Bandung dingin banget.
154239 Buset, panas sekali Surabaya malam ini.

→

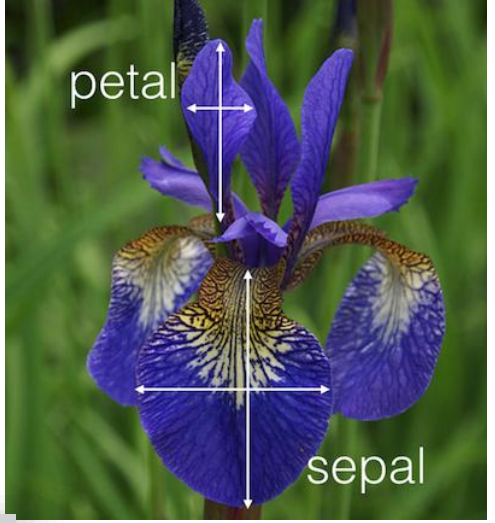
C Konversi

Temu kembali informasi atau NLP dapat membantu mengekstrak atribut dari teks bebas menjadi tabel.

ID	Kota	Waktu	Temp.
139492	Bandung	19:45	Dingin
154239	Surabaya	19:45	Panas

! **Intinya: Data mining biasanya membutuhkan input berbentuk dataset: data tidak terstruktur perlu diekstraksi dulu menjadi atribut.**

Dataset (Himpunan Data)



Attribute/Feature/Dimension

Class/Label/Target

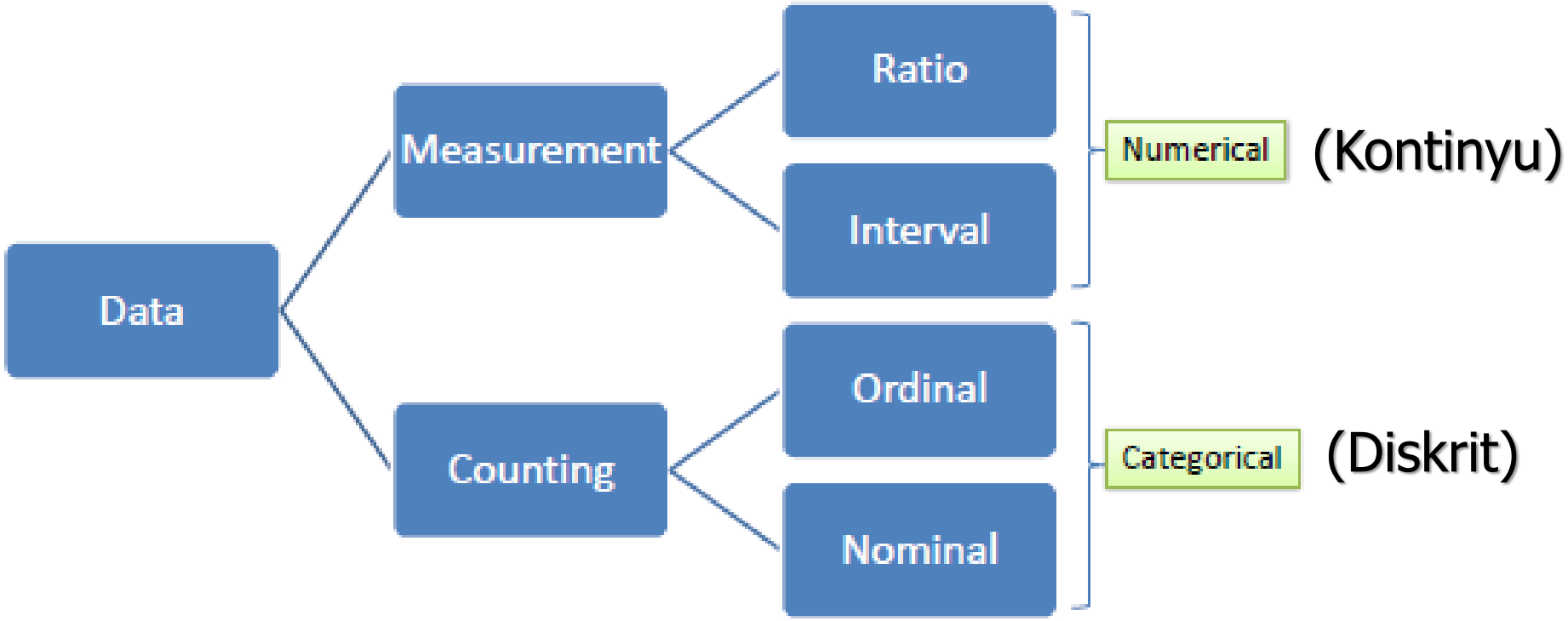
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
…					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris 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

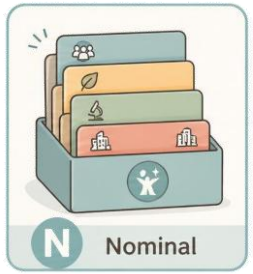
Nominal

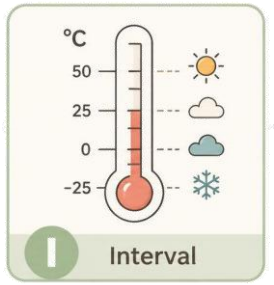
Numerik

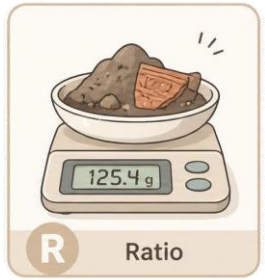
Tipe Data


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

Jenis Data dan Skala Pengukuran







Tujuan	membedakan kategori	mengurutkan kategori	membandingkan selisih	menghitung
Contoh	nama unit kerja	akreditasi lab	suhu	dana riset
Bentuk data	Biro… Direktorat… Pusat Riset…	A B C	20 30 40	100 juta 200 juta
	Data kategorik		Data numerik	

Jenis data menentukan cara analisis dan visualisasi.

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

1.4 Analisis Data

Kenapa Analisis Data Penting?

Data belum otomatis menjadi informasi

Analisis data membantu kita untuk:

- **Melihat pola**
- **Membandingkan kondisi**
- **Menemukan masalah prioritas**
- **Menjawab pertanyaan berbasis data**

Analisis yang kurang tepat dapat menghasilkan jawaban yang keliru

Apa Itu Analisis Data?

Analisis data adalah proses mengubah data menjadi informasi yang berguna untuk menjawab pertanyaan

Contoh:
Data publikasi tahunan
→ dianalisis
→ diperoleh informasi:
tren publikasi, unit dengan kontribusi tertinggi, dan bidang riset yang berkembang.

Analisis Statistik

1. Statistik Deskriptif
   - Nilai mean (rata-rata), standar deviasi, varians, data maksimal, data minimal, dsb

2. Statistik Inferensi
   - Perkiraan dan estimasi
   - Pengujian Hipotesis

Jenis Analisis Data

Deskriptif:
Apa yang terjadi?
Berlaku untuk yang dianalisis saja

Inferensia:
Apa kesimpulan umumnya? Menggunakan sampel untuk menarik kesimpulan tentang populasi.

Deskriptif menggambarkan data yang ada.
Inferensia menarik kesimpulan lebih luas tentang populasi.

Analisis Deskriptif

Digunakan untuk menjawab:

- **Berapa jumlahnya?**
- **Mana yang paling tinggi/rendah?**
- **Kategori apa yang paling sering muncul?**
- **Bagaimana sebarannya?**
- **Bagaimana trennya?**
- **Apakah ada nilai ekstrem?**

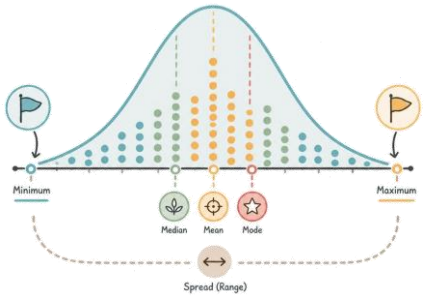
Contoh :

- **Unit mana yang memiliki jumlah publikasi tertinggi?**
- **Kategori apa yang paling banyak muncul?**
- **Bagaimana tren permintaan layanan per bulan?**

Statistik Ringkas

Ukuran	Makna praktis	Catatan
Frekuensi	jumlah kemunculan	cocok untuk kategorik
Persentase	proporsi dari total	perlu total pembanding
Mean	rata-rata	dipengaruhi nilai ekstrem
Median	nilai tengah	cocok untuk data miring/ada pencilan
Modus	paling sering muncul	cocok untuk data kategorik
Minimum-maksimum	rentang terendah - tertinggi	dapat dipengaruhi nilai ekstrem
Standar deviasi	tingkat keragaman data	perlu dibaca bersama dg rata-rata

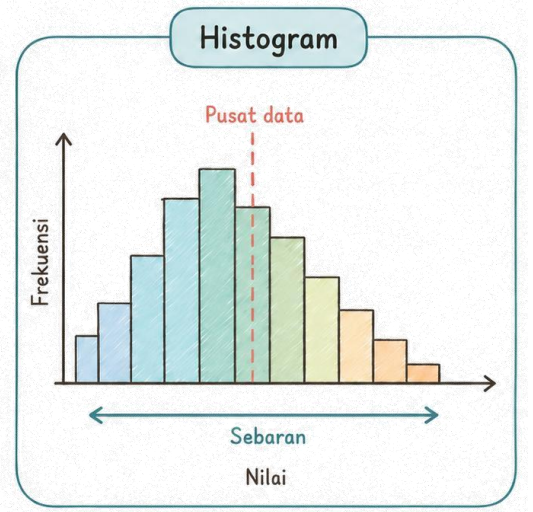
Pilih ukuran sesuai karakter data

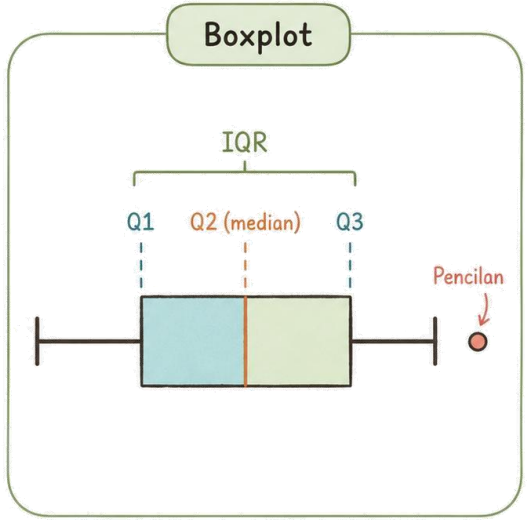


Membaca Distribusi Data Numerik

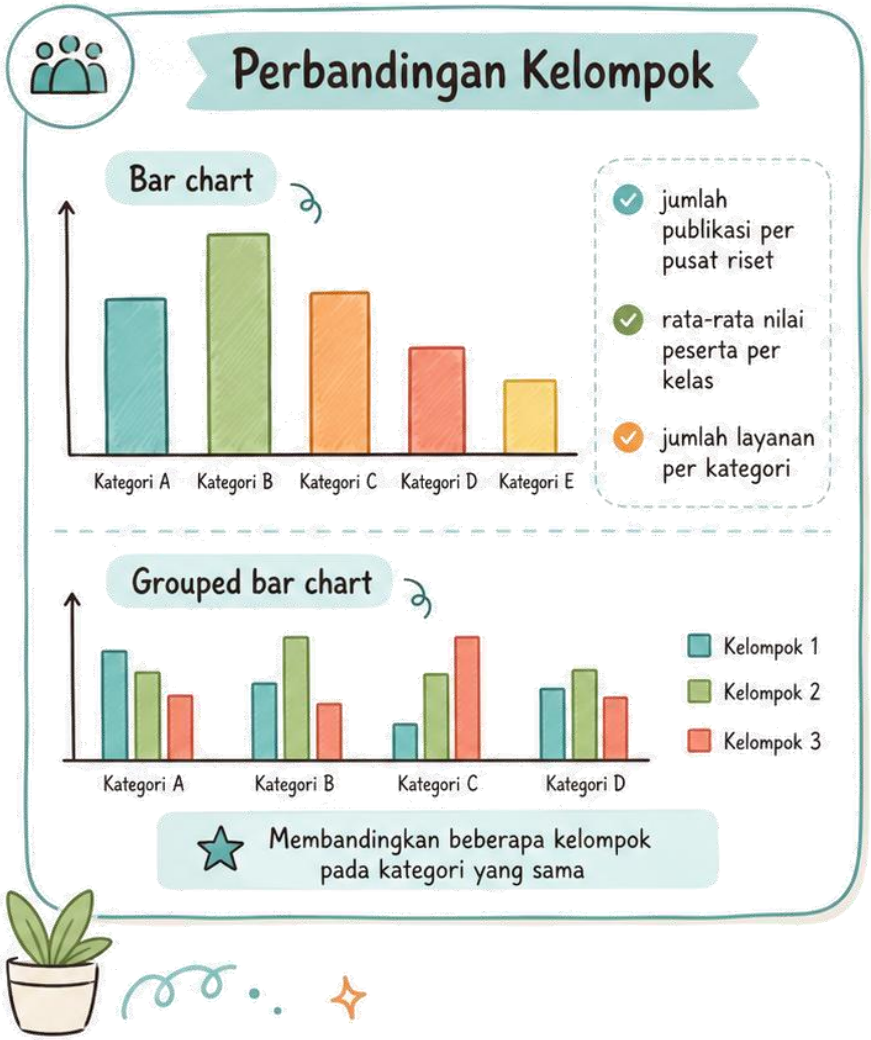
Untuk data numerik, perhatikan:

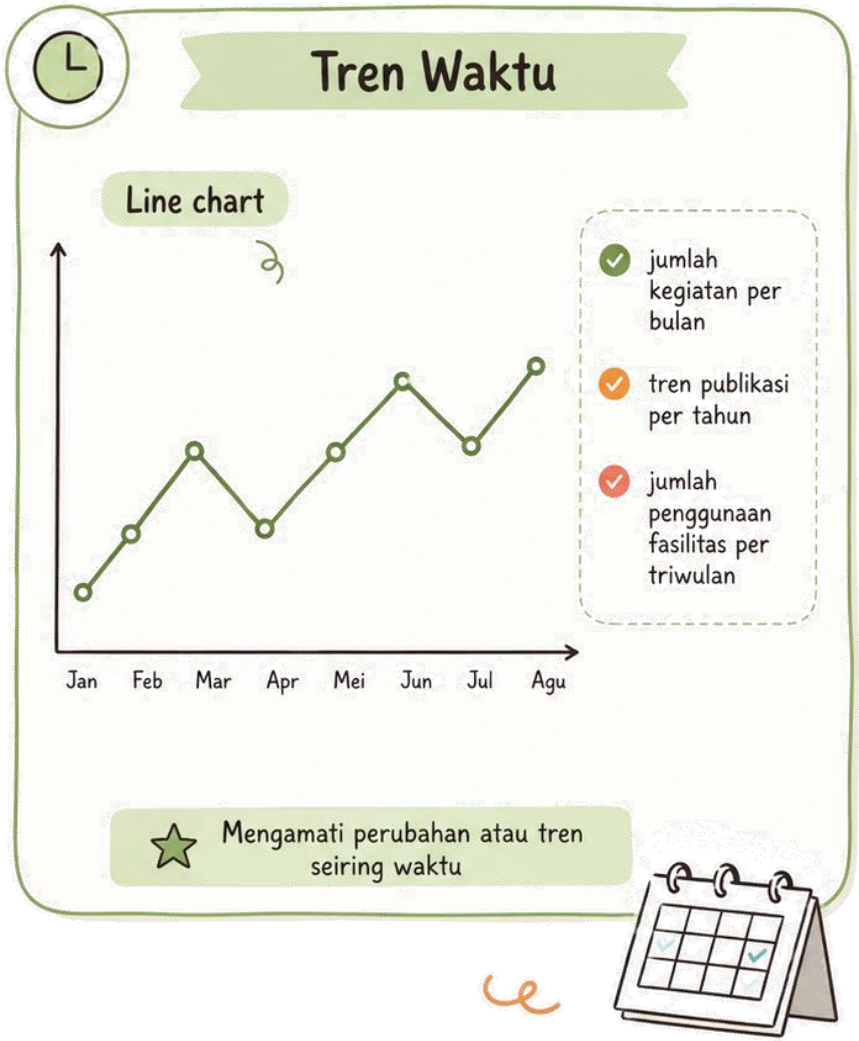
- Pusat data: mean atau median.
- Sebaran data: standar deviasi, rentang, IQR
- Bentuk distribusi: simetris atau miring.
- Pencilan: nilai yang sangat berbeda.





Deskriptif untuk Perbandingan dan Tren



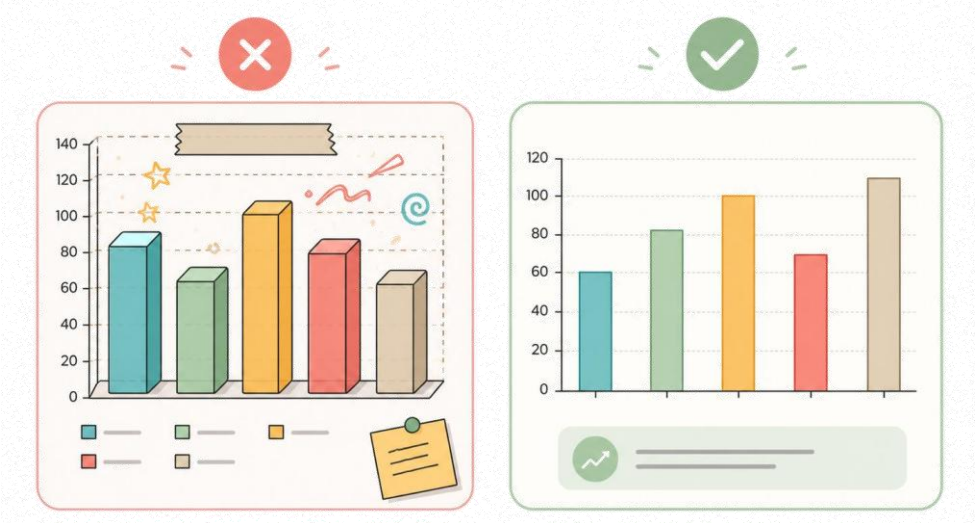


Memilih Grafik yang Tepat

Pilih grafik sesuai tujuan:

Tujuan	Grafik yang cocok
Membandingkan kategori	Bar chart
Melihat tren waktu	Line chart
Melihat proporsi sederhana	Pie chart
Melihat distribusi data numerik	Histogram / boxplot
Melihat hubungan dua variabel numerik	Scatter plot

Grafik yang Efektif

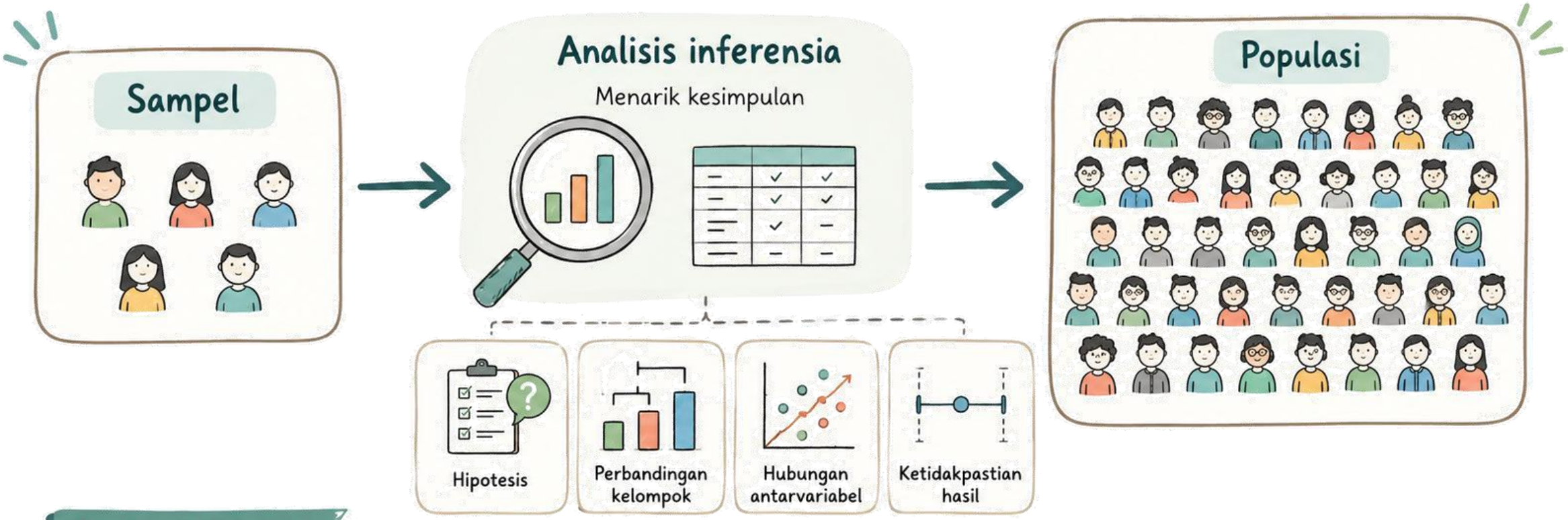


Grafik yang baik harus

- Memiliki judul yang jelas
- Memiliki label sumbu/kategori
- Menggunakan skala yang wajar
  - Tidak menyesatkan
  - Fokus pada pesan utama

Grafik adalah alat komunikasi, bukan hiasan.

Analisis Inferensia



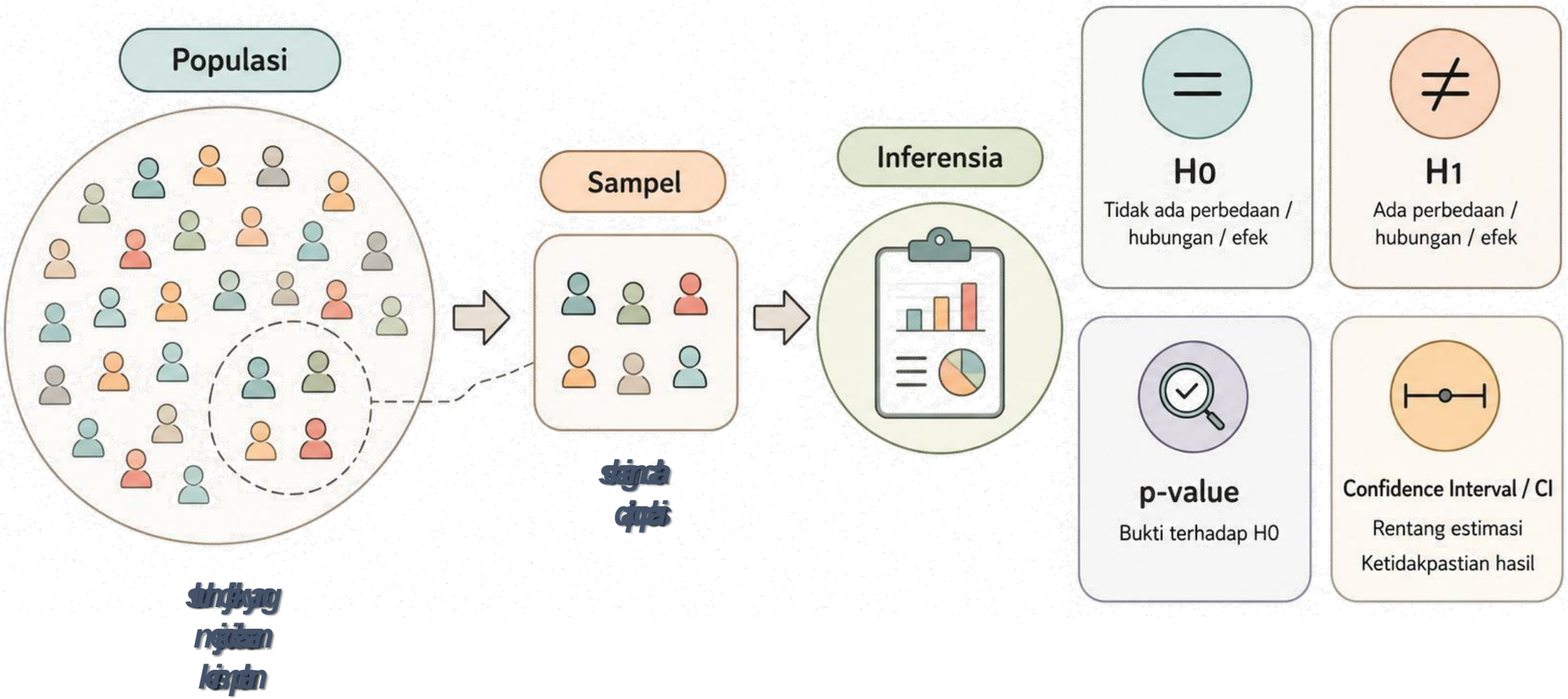
Fungsi analisis inferensia

- Menarik kesimpulan dari sampel ke populasi
- Menguji dugaan atau hipotesis
- Membandingkan kelompok secara statistik
- Melihat hubungan antarvariabel
- Mengukur ketidakpastian hasil

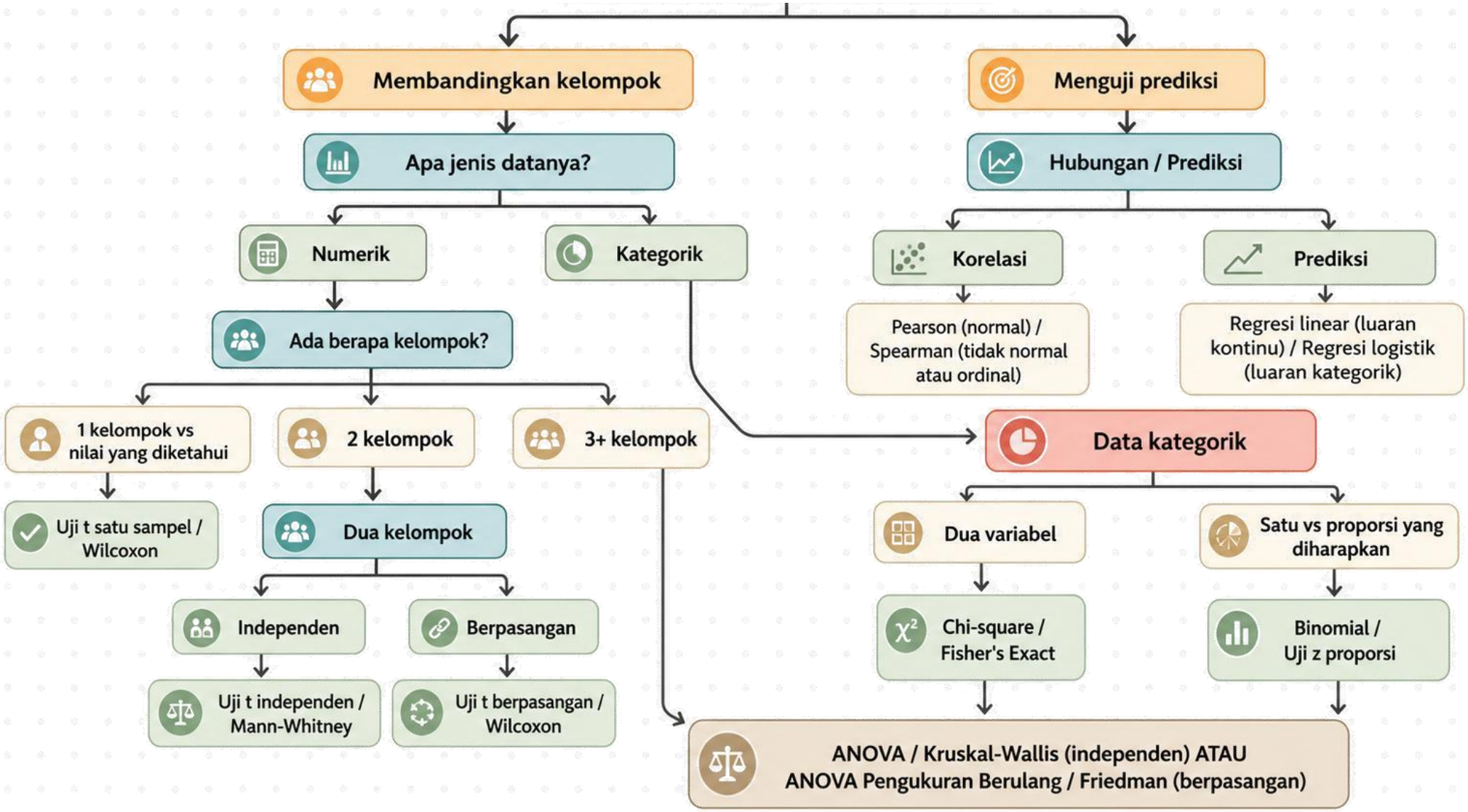
Contoh pertanyaan

- Apakah rata-rata nilai dua kelas berbeda?
- Apakah status akreditasi berhubungan dengan jenis fasilitas?
- Apakah dana riset berkaitan dengan jumlah publikasi?

Istilah Dasar Inferensia



Analisis Inferensia



Statistik Inferensi (Pengujian Hipotesis)

Penggunaan	Parametrik	Non Parametrik
Dua sampel saling berhubungan (*Two Dependent samples*)	**T Test** Z Test	Sign test Wilcoxon Signed-Rank Mc Nemar Change test
Dua sampel tidak berhubungan (*Two Independent samples*)	T Test Z Test	Mann-Whitney U test Moses Extreme reactions Chi-Square test Kolmogorov-Smirnov test Walt-Wolfowitz runs
Beberapa sampel berhubungan (*Several Dependent Samples*)		**Friedman test** Kendall W test Cochran's Q
Beberapa sampel tidak Berhubungan (*Several Independent Samples*)	Anova test (F test)	Kruskal-Wallis test Chi-Square test Median test

Metode Parametrik

- Metode parametrik dapat dilakukan jika beberapa persyaratan dipenuhi, yaitu:
  - Sampel yang dianalisis haruslah berasal dari populasi yang berdistribusi normal
  - Jumlah data cukup banyak
  - Jenis data yang dianalisis adalah biasanya interval atau rasio

Metode Non Parametrik

- Metode ini dapat dipergunakan secara lebih luas, karena tidak mengharuskan datanya berdistribusi normal
  - Dapat dipakai untuk data nominal dan ordinal sehingga sangat berguna bagi para peneliti sosial untuk meneliti perilaku konsumen, sikap manusia, dsb
  - Cenderung lebih sederhana dibandingkan dengan metode parametrik

- Selain keuntungannya, berikut kelemahan metode non parametrik:
  - Tidak adanya sistematika yang jelas seperti metode parametrik
  - Terlalu sederhana sehingga sering meragukan
  - Memakai tabel-tabel yang lebih bervariasi dibandingkan dengan tabel-tabel standar pada metode parametrik

Interpretasi Statistik

- Ho = tidak ada perbedaan signifikan
- Ha = ada perbedaan signifikan

alpha=0.05

Bila $p < 0.05$, maka Ho ditolak

- Contoh: kasus p=0.03, maka dapat ditarik kesimpulan?

Interpretasi

2. Proses Data Mining

2.1 Proses dan Tools Data Mining

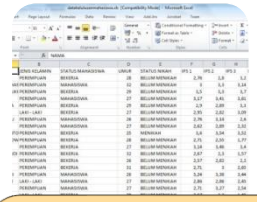
2.2 Penerapan Proses Data Mining

2.3 Evaluasi Model Data Mining & Analisis Data

2.4 Proses Data Mining berbasis CRISP-DM

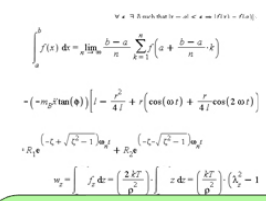
2.1 Proses dan Tools Data Mining

Proses Data Mining



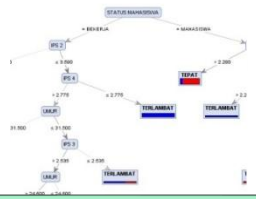
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



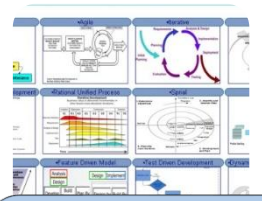
2. Metode Data Mining

(**Pilih Metode** Sesuai Karakter Data)



3. Pengetahuan

(**Pahami Model** dan Pengetahuan yg Sesuai)



4. Evaluation

(**Analisis Model** dan **Kinerja** Metode)

DATA PREPROCESSING

Data Cleaning
Data Integration
Data Reduction
Data Transformation

MODELING

Estimation
Prediction
Classification
Clustering
Association

MODEL

Formula
Tree
Cluster
Rule
Correlation

KINERJA

Akurasi
Tingkat Error
Jumlah Cluster

MODEL

Atribute/Faktor
Korelasi
Bobot

1. Himpunan Data (Dataset)

- Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi
- Jenis dataset ada dua: Private dan Public
- Private Dataset: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
  - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- Public Dataset: data set dapat diambil dari repositori pubik yang disepakati oleh para peneliti data mining
  - UCI Repository (*http://www.ics.uci.edu/~mlearn/MLRepository.html*)
  - ACM KDD Cup (*http://www.sigkdd.org/kddcup/*)
  - PredictionIO (*http://docs.prediction.io/datacollection/sample/*)
- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: comparable, repeatable dan verifiable

Public Data Set (UCI Repository)

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search
Repository Web
Google

View ALL Data Sets

Browse Through: **360** Data Sets

Table View List View

Default Task
- Classification (262)
- Regression (63)
- Clustering (54)
- Other (52)

Attribute Type
- Categorical (37)
- Numerical (213)
- Mixed (56)

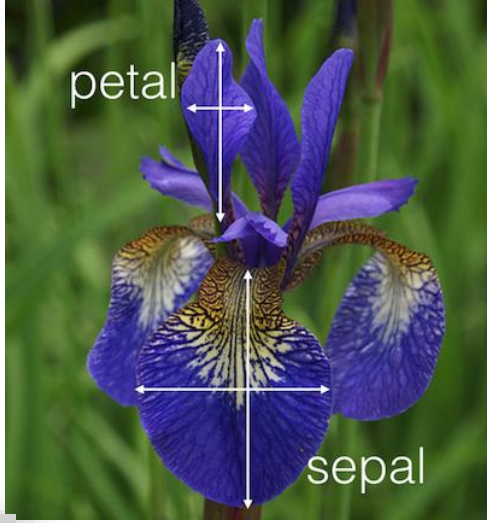
Data Type
- Multivariate (281)
- Univariate (16)
- Sequential (36)
- Time-Series (65)
- Text (32)
- Domain-Theory (22)
- Other (21)

Area
- Life Sciences (82)
- Physical Sciences (43)
- CS / Engineering (111)
- Social Sciences (23)
- Business (21)
- Game (10)
- Other (67)

 # Attributes
- Less than 10 (86)
- 10 to 100 (162)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987

Dataset (Himpunan 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

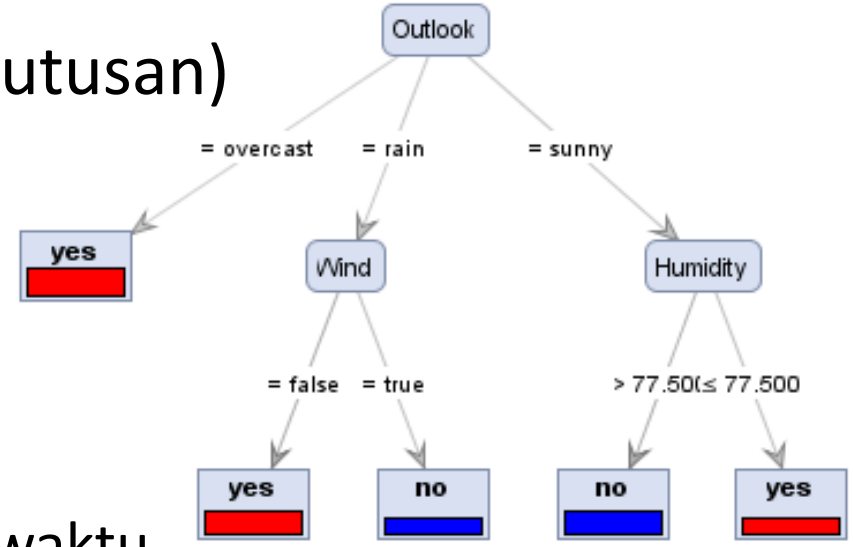
2. Metode Data Mining

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

3. Pengetahuan (Pola/Model)

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

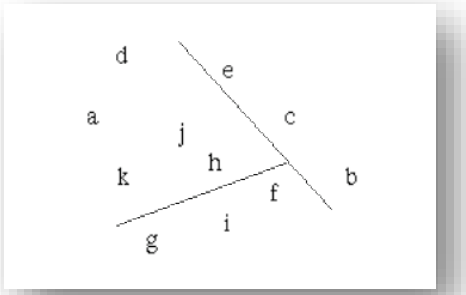
2. Decision Tree (Pohon Keputusan)

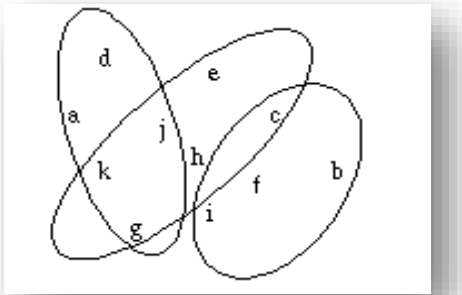


3. Tingkat Korelasi

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

5. Cluster (Klaster)





4. Evaluasi (Akurasi, Error, etc)

1. Estimation:
   Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
   Confusion Matrix: Accuracy
   ROC Curve: Area Under Curve (AUC)
4. Clustering:
   Internal Evaluation: Davies–Bouldin index, Dunn index,
   External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:
   Lift Charts: Lift Ratio
   Precision and Recall (F-measure)

Kriteria Evaluasi dan Validasi Model

1. Akurasi
   - Ukuran dari seberapa baik model mengkorelasikan antara hasil dengan atribut dalam data yang telah disediakan
   - Terdapat berbagai model akurasi, tetapi semua model akurasi tergantung pada data yang digunakan
2. Kehandalan
   - Ukuran di mana model data mining diterapkan pada dataset yang berbeda
   - Model data mining dapat diandalkan jika menghasilkan pola umum yang sama terlepas dari data testing yang disediakan
3. Kegunaan
   - Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna

Keseimbangan diantaranya ketiganya diperlukan karena belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna

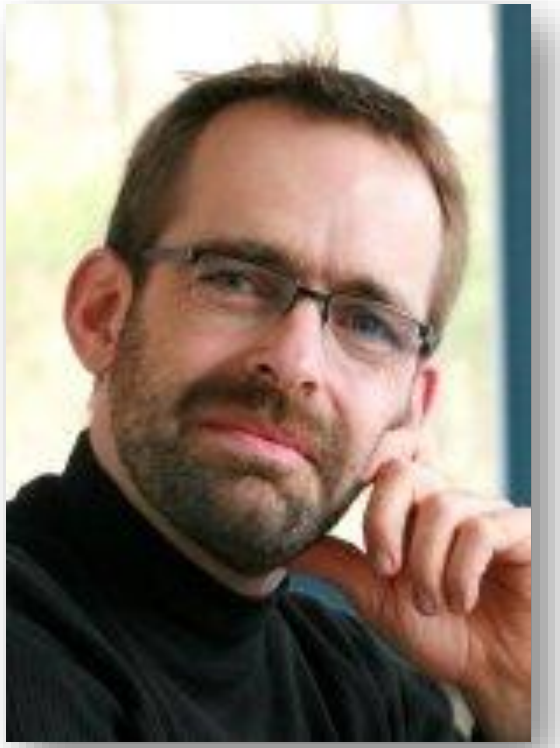
Magic Quadrant for Data Science Platform

(Gartner, 2017)

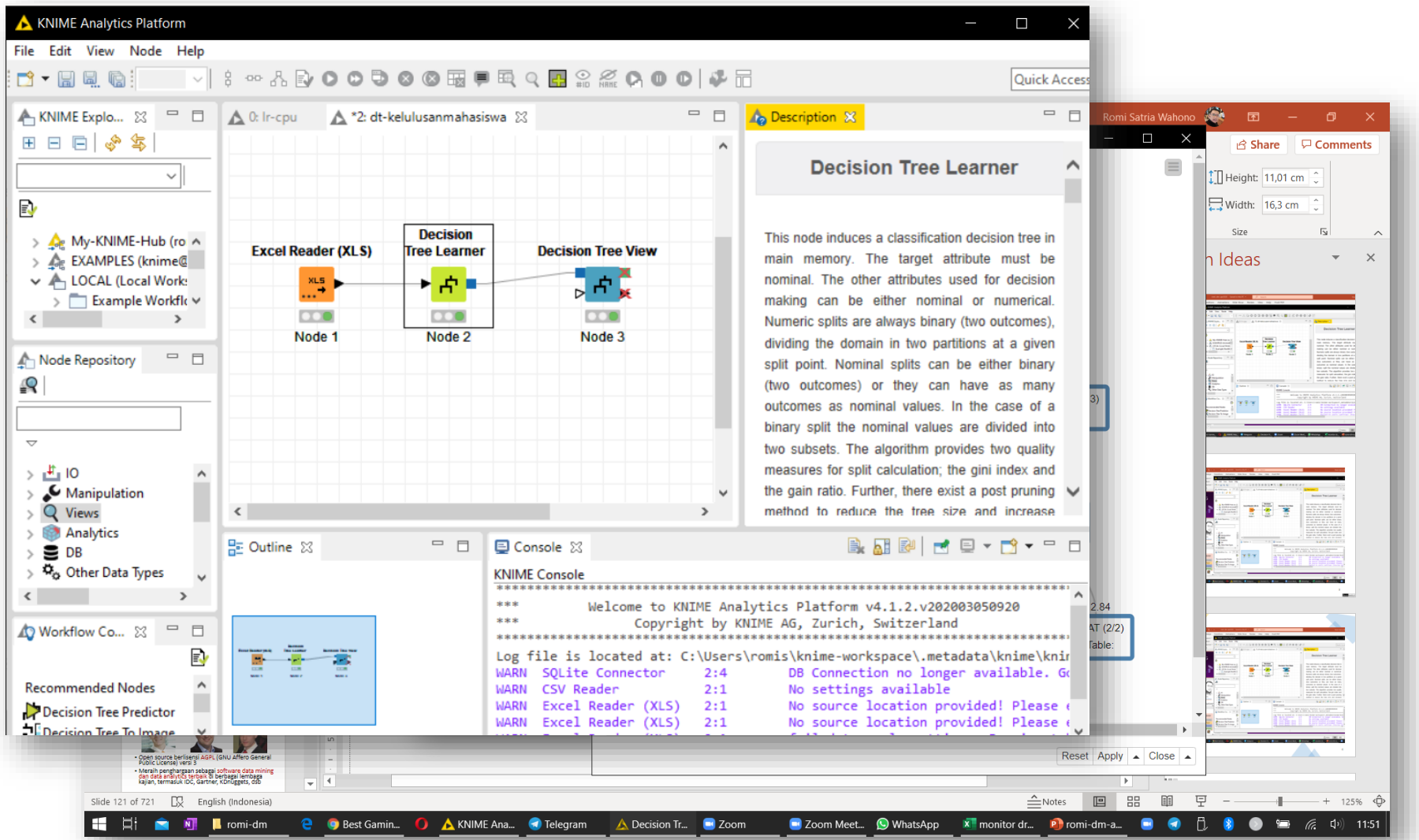
Magic Quadrant for Data Science Platform *(Gartner, 2018)*

KNIME

- KNIME (Konstanz Information Miner) adalah platform data mining untuk analisis, pelaporan, dan integrasi data yang termasuk perangkat lunak bebas dan sumber terbuka
- KNIME mulai dikembangkan tahun 2004 oleh tim pengembang perangkat lunak dari Universitas Konstanz, yang dipimpin oleh Michael Berthold, yang awalnya digunakan untuk penelitian di industri farmasi
- Mulai banyak digunakan orang sejak tahun 2006, dan setelah itu berkembang pesat sehingga tahun 2017 masuk ke Magic Quadrant for Data Science Platform (Gartner Group)



KNIME


KNIME Analytics Platform
File Edit View Node Help
Quick Access
KNIME Explo...
My-KNIME-Hub (ro
EXAMPLES (knime@
LOCAL (Local Works
Example Workfl
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Workflow Co...
Recommended Nodes
Decision Tree Predictor
0: lr-cpu
*2: dt-kelulusanmahasiswa
Excel Reader (XLS)
Decision Tree Learner
Decision Tree View
Node 1
Node 2
Node 3
Description
Decision Tree Learner
This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gini index and the gain ratio. Further, there exist a post pruning method to reduce the tree size and increase
Outline
Console
KNIME Console
Welcome to KNIME Analytics Platform v4.1.2.v202003050920
Copyright by KNIME AG, Zurich, Switzerland
Log file is located at: C:\Users\romis\knime-workspace\.metadata\knime\knir
WARN SQLite Connector 2:4 DB Connection no longer available.
WARN CSV Reader 2:1 No settings available
WARN Excel Reader (XLS) 2:1 No source location provided! Please
WARN Excel Reader (XLS) 2:1 No source location provided! Please
Romi Satria Wahono
Share
Comments
Height: 11,01 cm
Width: 16,3 cm
Size
Ideas
Reset
Apply
Close
Slide 121 of 721
English (Indonesia)
Notes
125%
romi-dm
Best Gamin...
KNIME Ana...
Telegram
Decision Tr...
Zoom
Zoom Meet...
WhatsApp
monitor dr...
romi-dm-a...
11:51

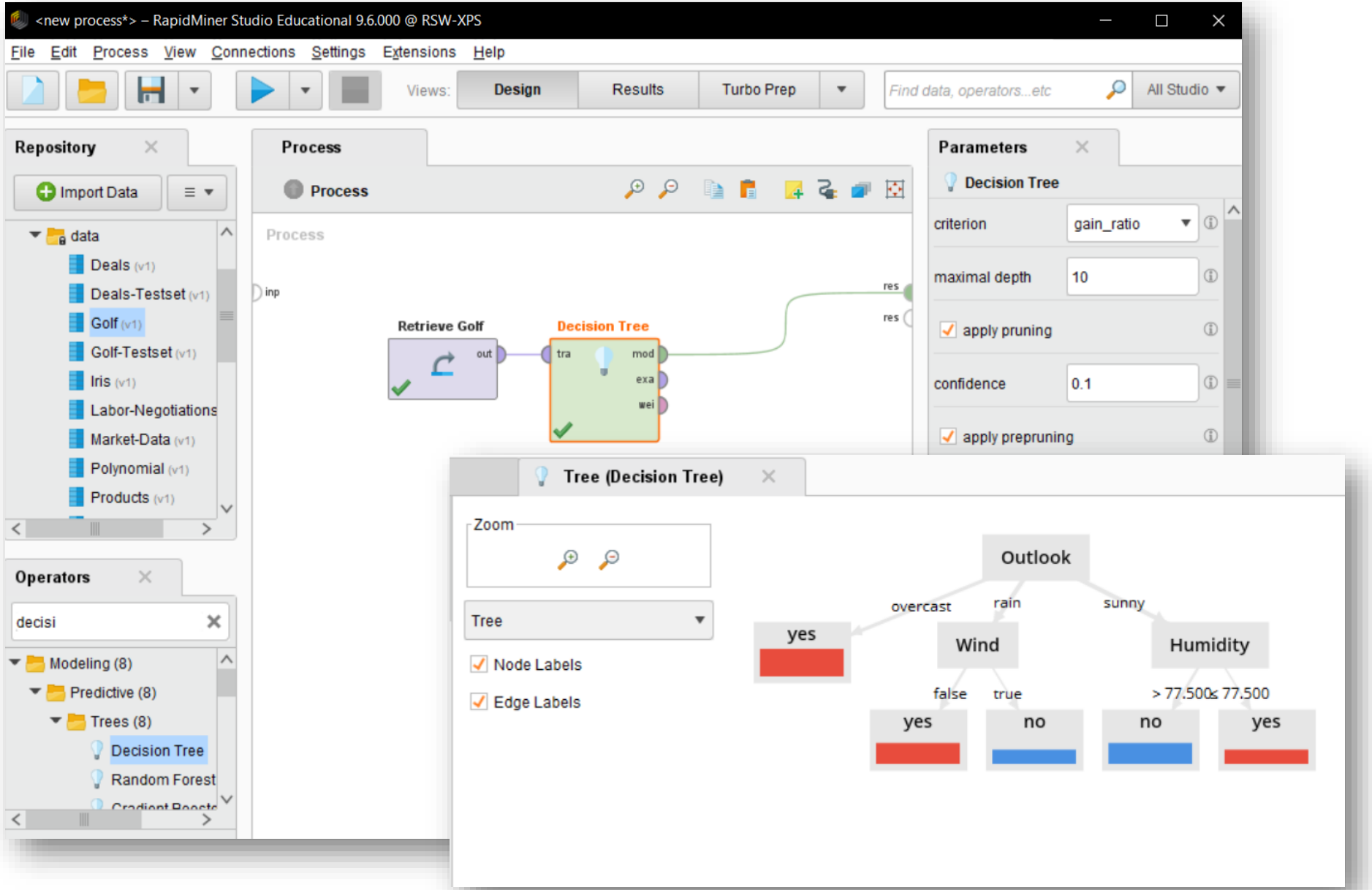
Rapidminer

- Pengembangan dimulai pada 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund, ditulis dalam bahasa Java



- Open source berlisensi AGPL (GNU Affero General Public License) versi 3
- Meraih penghargaan sebagai software data mining dan data analytics terbaik di berbagai lembaga kajian, termasuk IDC, Gartner, KDnuggets, dsb

Rapidminer


<new process*> – RapidMiner Studio Educational 9.6.000 @ RSW-XPS
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep
Find data, operators...etc
All Studio
Repository
Import Data
data
Deals (v1)
Deals-Testset (v1)
Golf (v1)
Golf-Testset (v1)
Iris (v1)
Labor-Negotiations
Market-Data (v1)
Polynomial (v1)
Products (v1)
Operators
decisi
Modeling (8)
Predictive (8)
Trees (8)
Decision Tree
Random Forest
Process
Process
Process
inp
Retrieve Golf
out
Decision Tree
tra
mod
exa
wei
res
res
Parameters
Decision Tree
criterion
gain_ratio
maximal depth
10
apply pruning
confidence
0.1
apply prepruning
Tree (Decision Tree)
Zoom
Tree
Node Labels
Edge Labels
Outlook
overcast
rain
sunny
yes
Wind
Humidity
false
true
> 77.500
≤ 77.500
yes
no
no
yes

Role Atribut Pada Rapidminer

1. Atribut: karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi
   - ID, atribut biasa

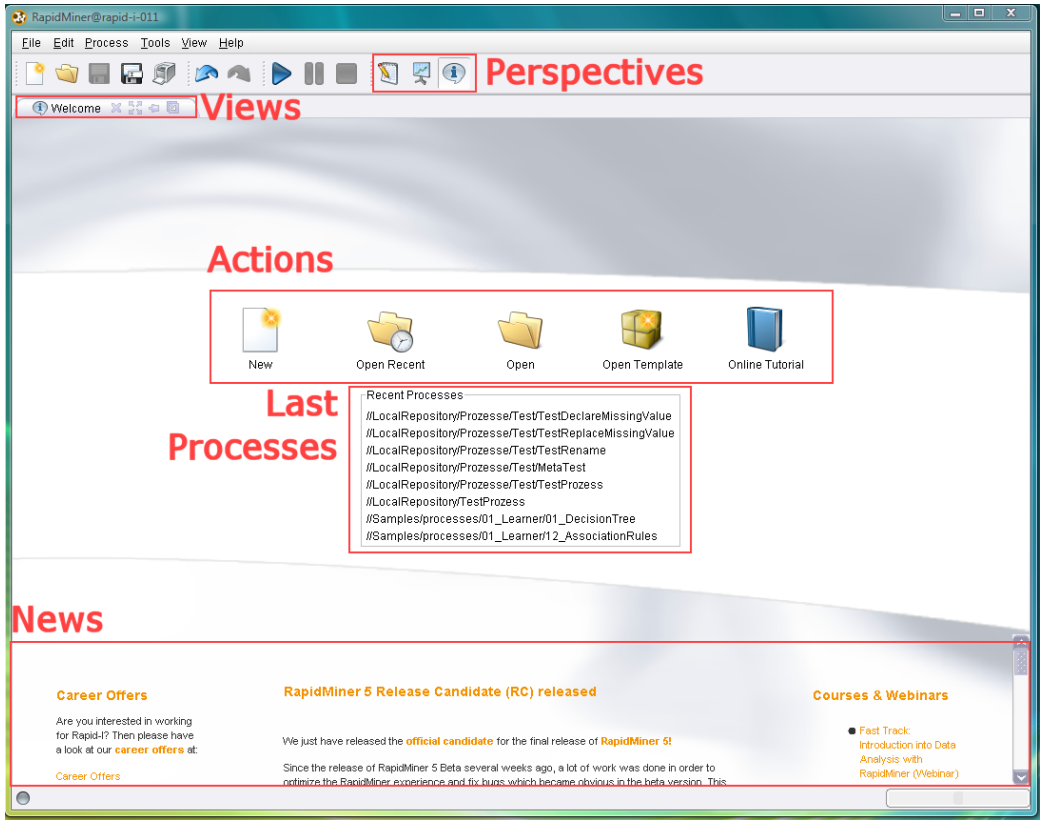
2. Atribut target: atribut yang menjadi tujuan untuk diisi oleh proses data mining
   - Label, cluster, weight

Tipe Nilai Atribut pada Rapidminer

1. nominal: nilai secara kategori
2. binominal: nominal dua nilai
3. polynominal: nominal lebih dari dua nilai
4. numeric: nilai numerik secara umum
5. integer: bilangan bulat
6. real: bilangan nyata
7. text: teks bebas tanpa struktur
8. date_time: tanggal dan waktu
9. date: hanya tanggal
10. time: hanya waktu

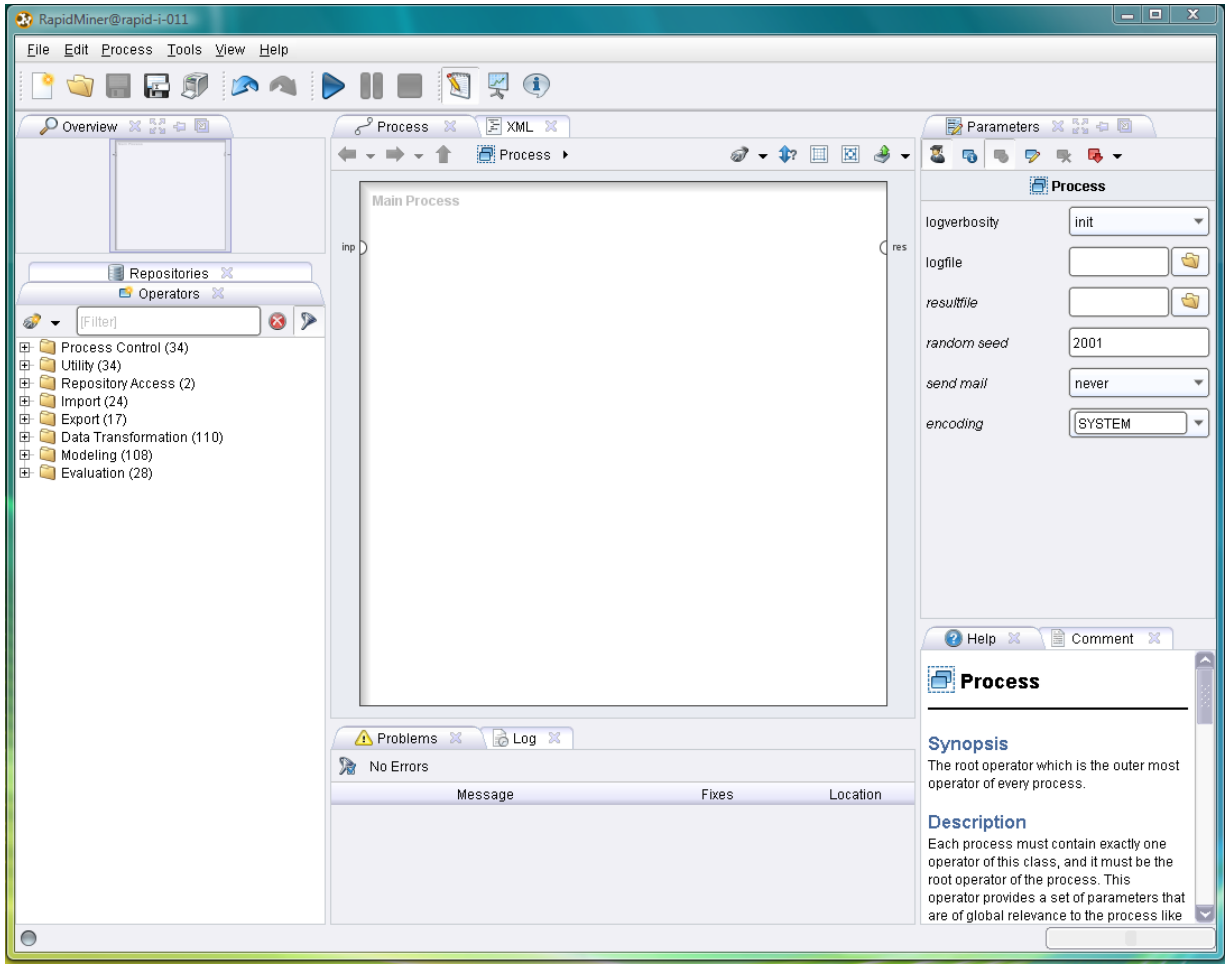
Perspektif dan View

1. Perspektif Selamat Datang (Welcome perspective)
2. Perspektif Desain (Design perspective)
3. Perspektif Hasil (Result perspective)

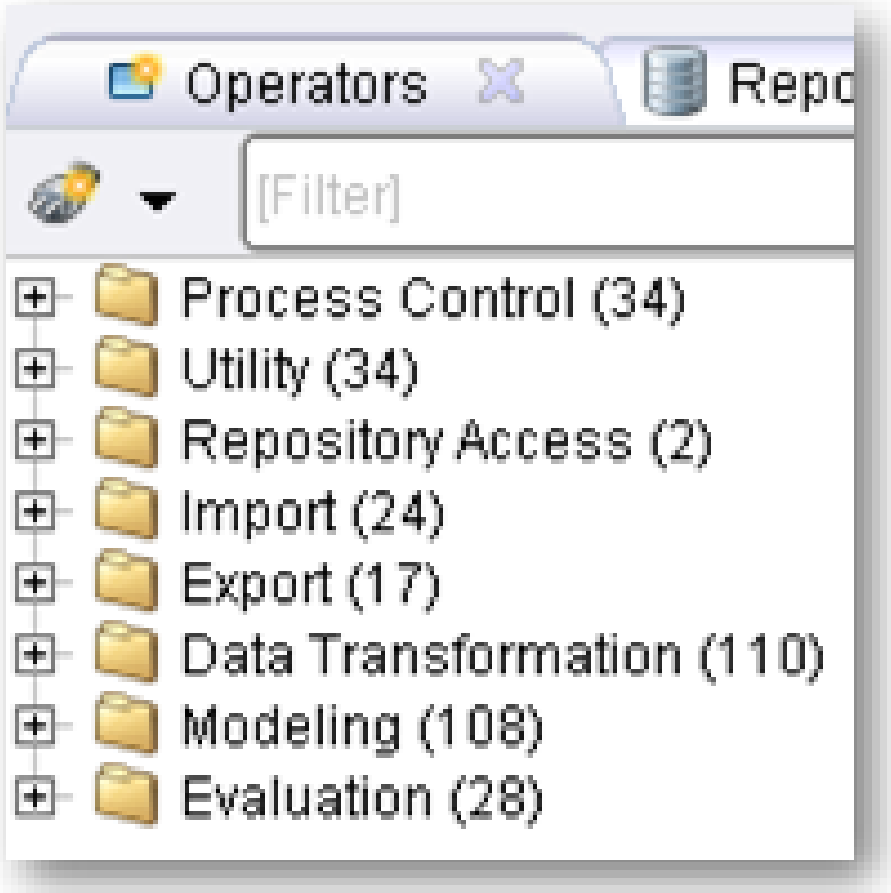


Perspektif Desain

- Perspektif dimana semua proses dibuat dan dikelola
- Pindah ke Perspektif Desain dengan Klik:

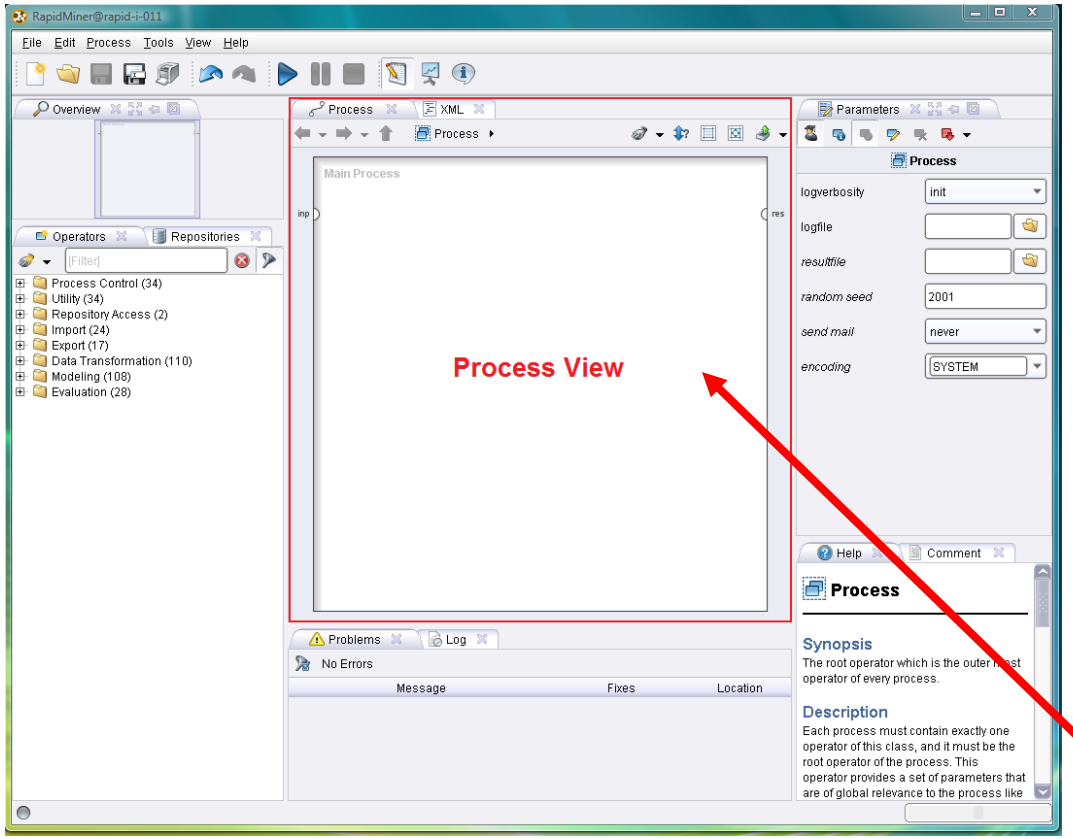


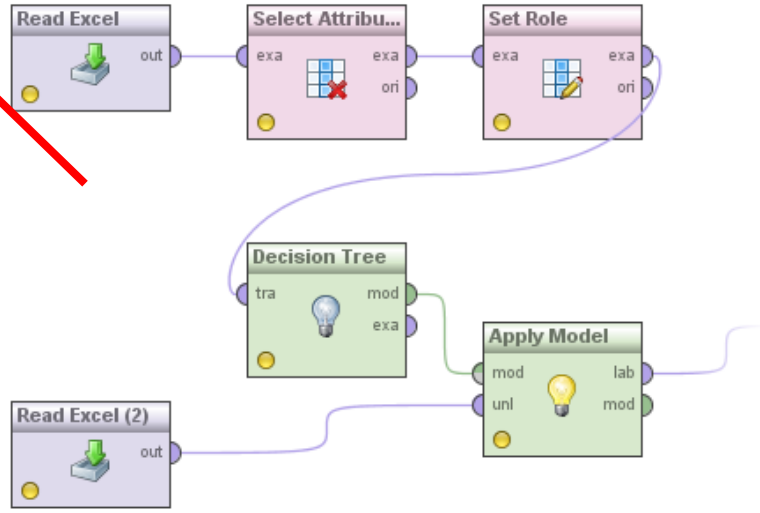
View Operator



- Process Control
  Untuk mengontrol aliran proses, seperti *loop* atau *conditional branch*
- Utility
  Untuk mengelompokkan *subprocess*, juga *macro* dan *logger*
- Repository Access
  Untuk membaca dan menulis repositori
- Import
  Untuk membaca data dari berbagai format eksternal
- Export
  Untuk menulis data ke berbagai format eksternal
- Data Transformation
  Untuk transformasi data dan metadata
- Modelling
  Untuk proses data mining seperti klasifikasi, regresi, clustering, asosiasi, dll
- Evaluation
  Untuk mengukur kualitas dan perfomansi dari model

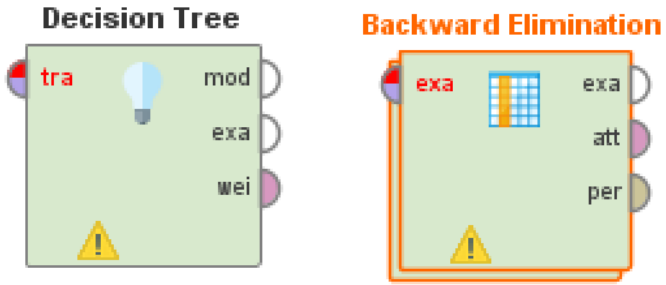
View Proses


RapidMiner@rapid-i-011
File Edit Process Tools View Help
Overview
Operators
Repositories
[Filter]
Process Control (34)
Utility (34)
Repository Access (2)
Import (24)
Export (17)
Data Transformation (110)
Modeling (108)
Evaluation (28)
Process
XML
Process
Main Process
inp
res
Process View
Problems
Log
No Errors
Message
Fixes
Location
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Help
Comment
Process
Synopsis
The root operator which is the outer most operator of every process.
Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like


Read Excel
out
Select Attribu...
exa
exa
ori
Set Role
exa
exa
ori
Decision Tree
tra
mod
exa
Apply Model
mod
lab
unl
mod
Read Excel (2)
out

Operator dan Proses

- Proses data mining pada dasarnya adalah proses analisis yang berisi alur kerja dari komponen data mining
- Komponen dari proses ini disebut operator, yang memiliki:
  1. Input
  2. Output
  3. Aksi yang dilakukan
  4. Parameter yang diperlukan

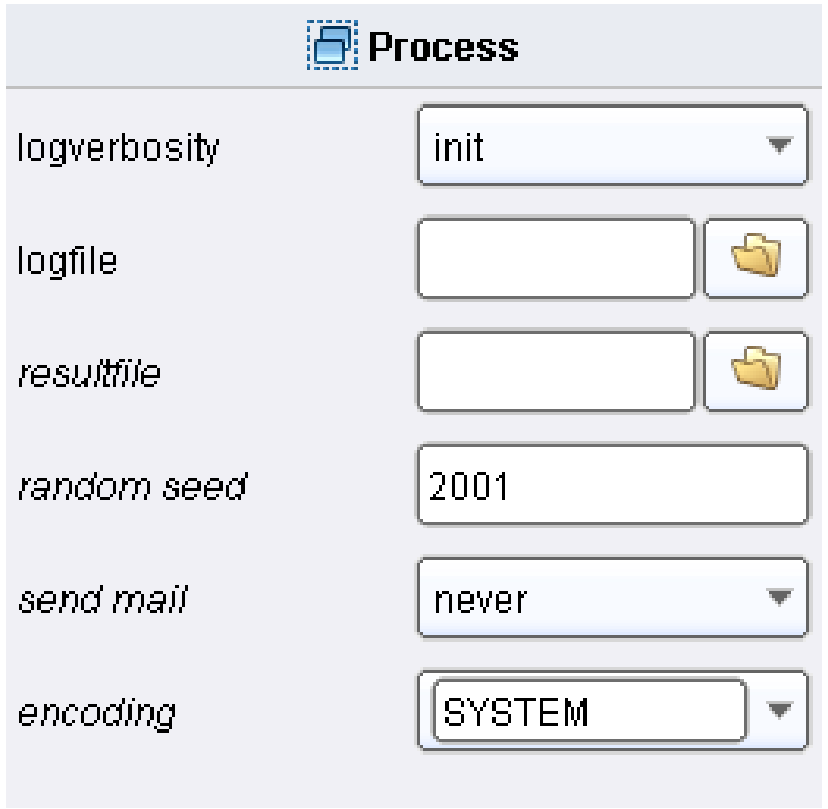


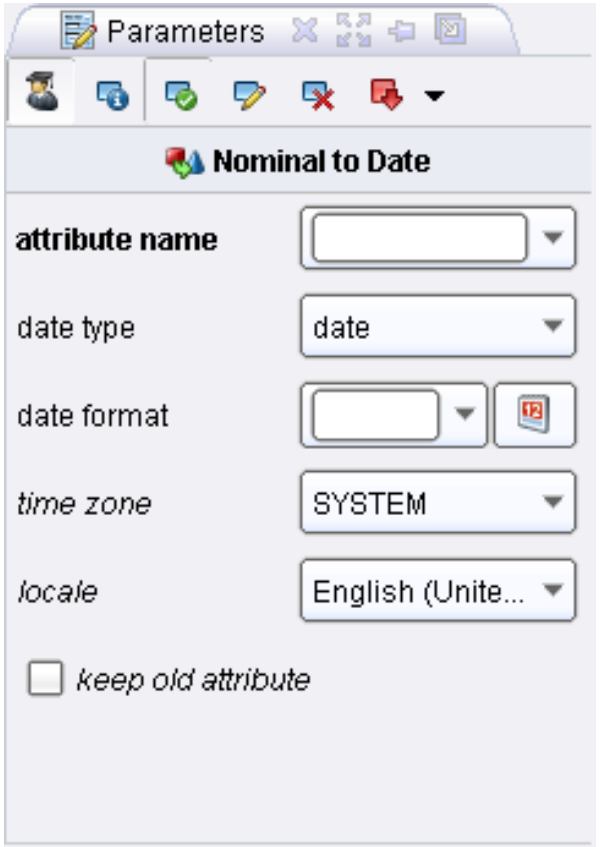
- Sebuah operator bisa disambungkan melalui port masukan (kiri) dan port keluaran (kanan)

- Indikator status dari operator:
  - **Lampu** status: **merah** (tak tersambung), **kuning** (lengkap tetapi belum dijalankan), **hijau** (sudah behasil dijalankan)
  - **Segitiga** warning: bila ada pesan status
  - **Breakpoint**: bila ada breakpoint sebelum/sesudahnya
  - **Comment**: bila ada komentar

View Parameter

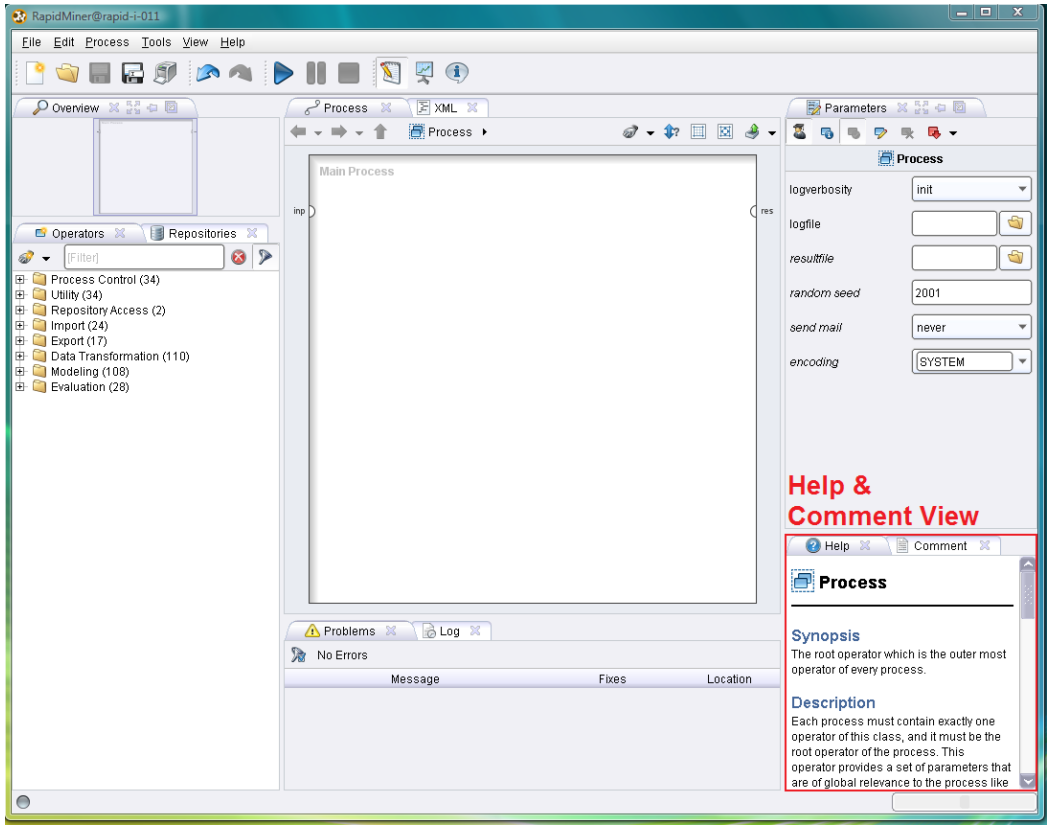
- Operator kadang memerlukan parameter untuk bisa berfungsi
- Setelah operator dipilih di view Proses, parameternya ditampilkan di view ini

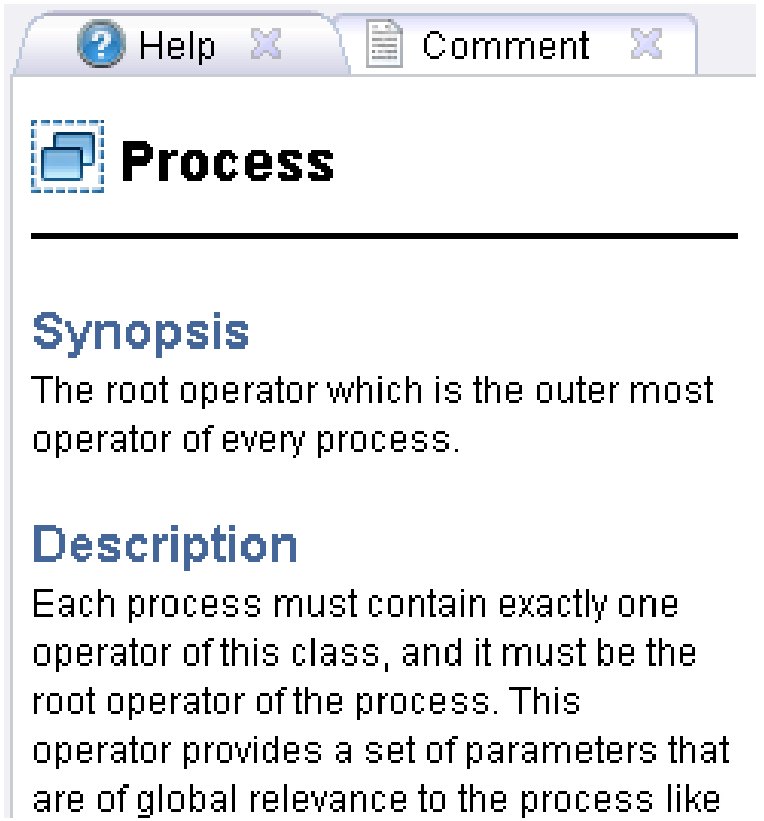




View Help dan View Comment

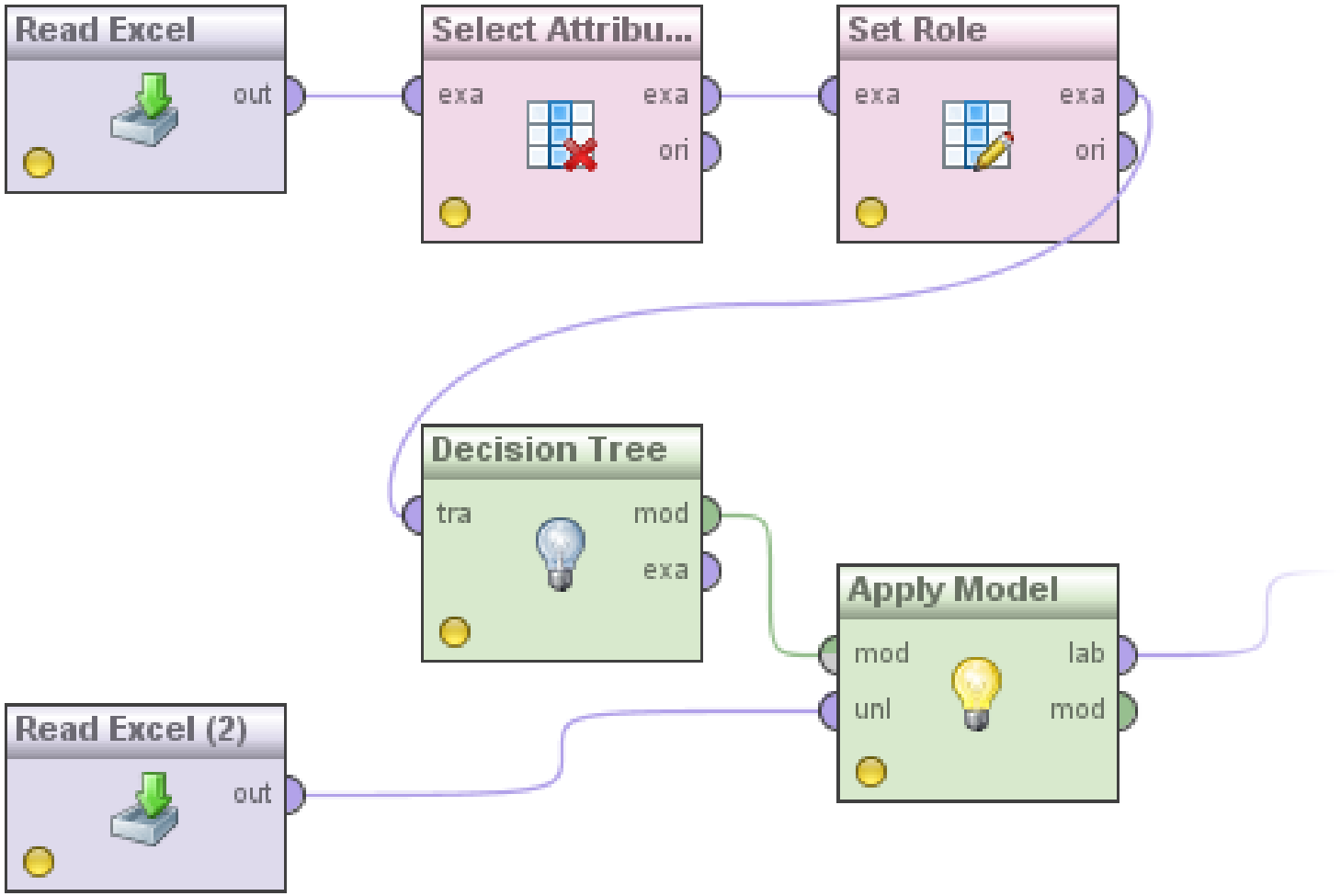
- View Help menampilkan deskripsi dari operator
- View Comment menampilkan komentar yang dapat diedit terhadap operator





Mendesain Proses

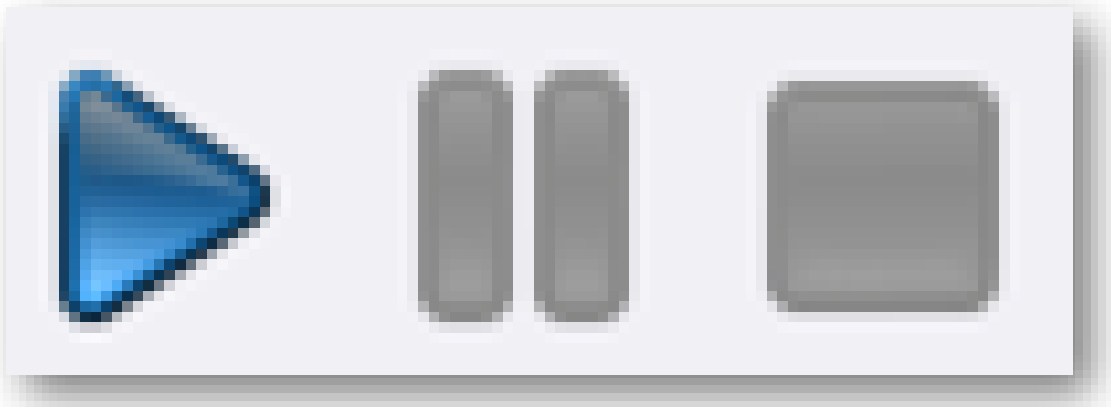
Kumpulan dan rangkaian fungsi-fungsi (operator)
yang bisa disusun secara visual (visual programming)



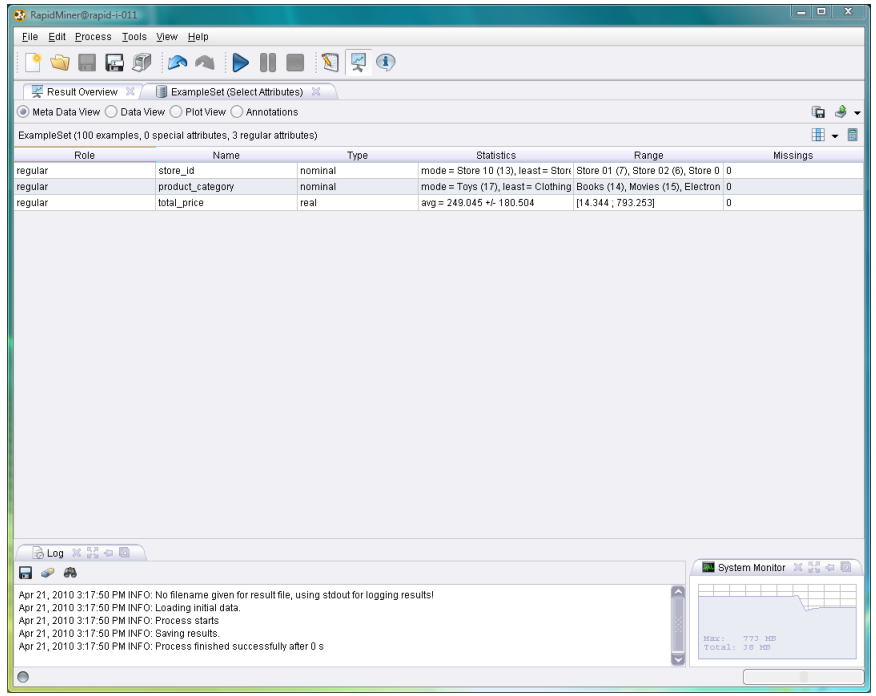
Menjalankan Proses

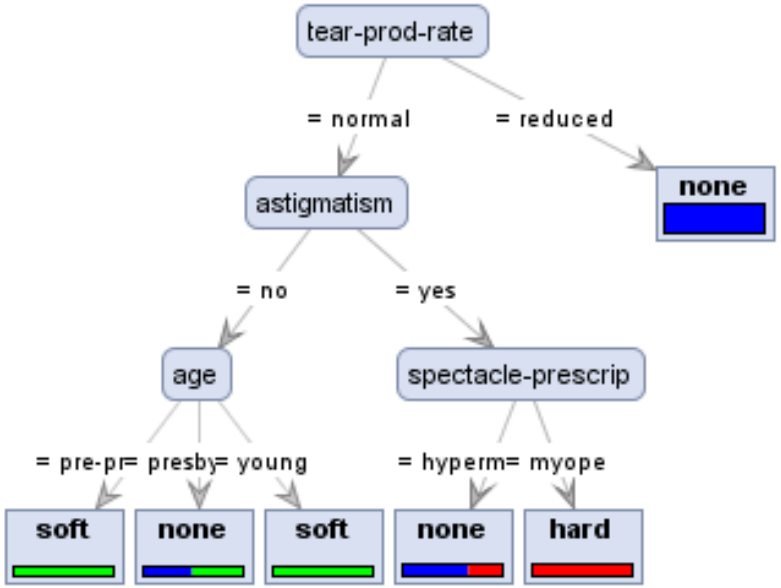
Proses dapat dijalankan dengan:

- Menekan tombol Play
- Memilih menu Process → Run
- Menekan kunci F11

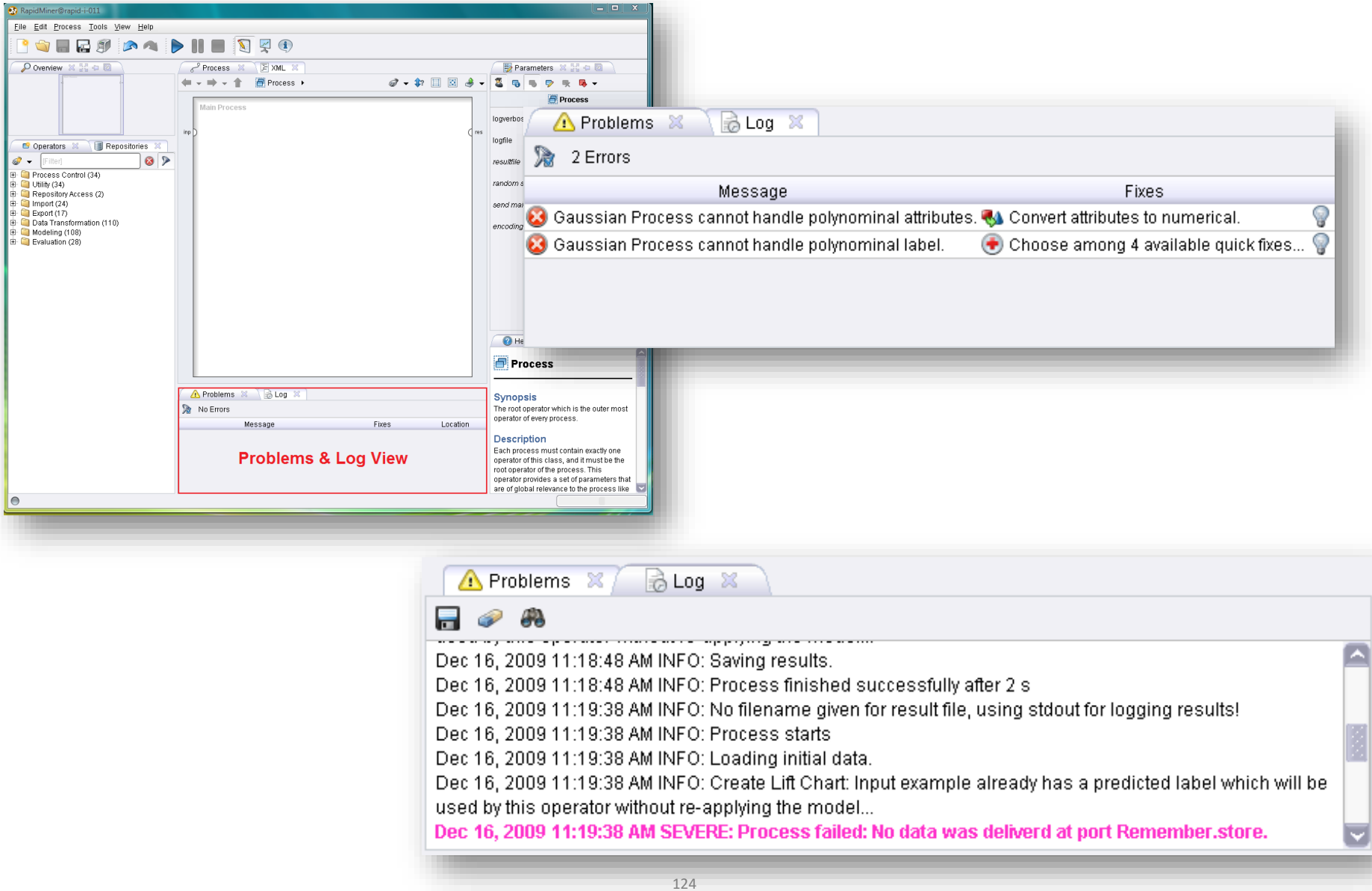


Melihat Hasil



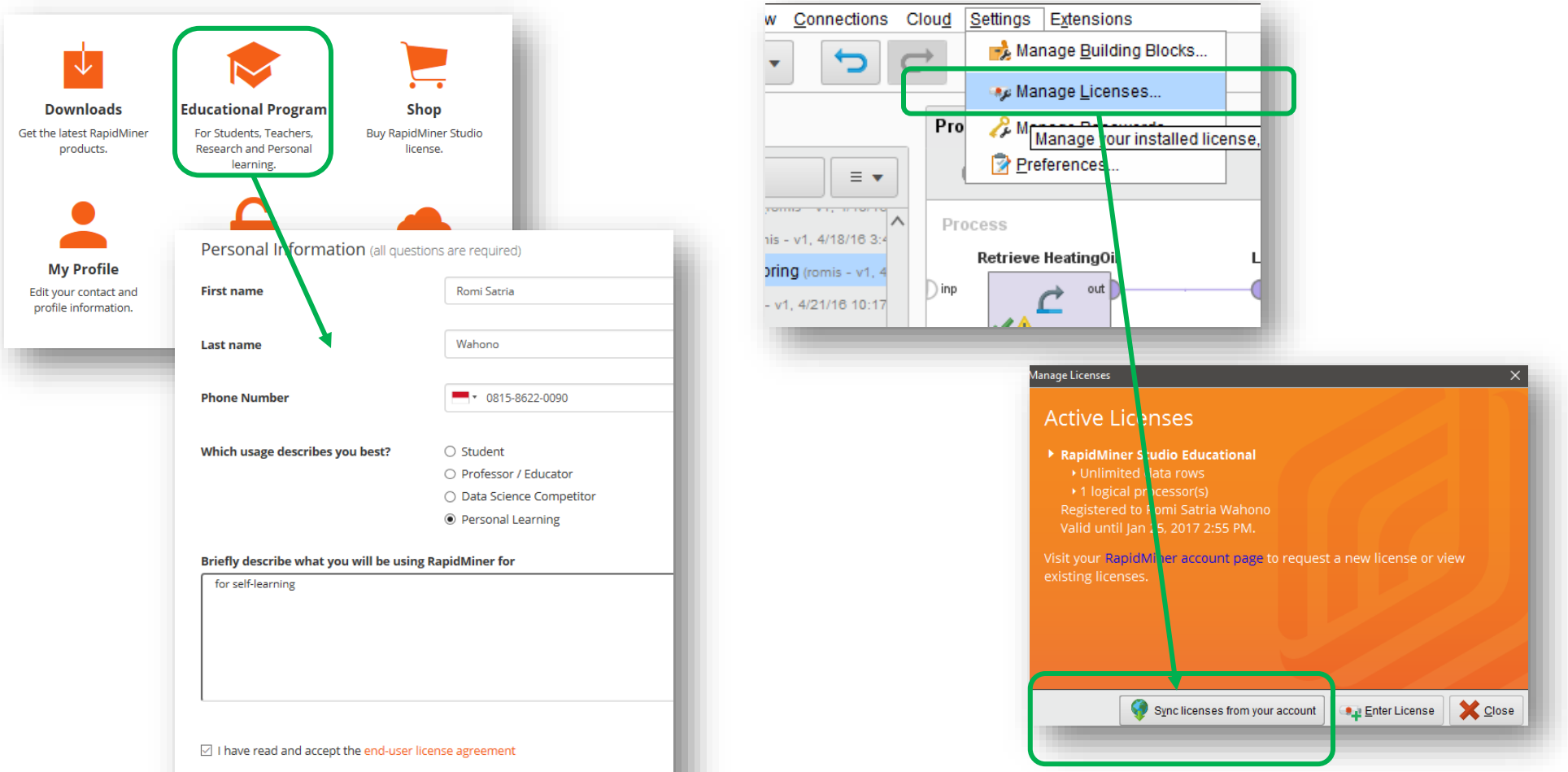


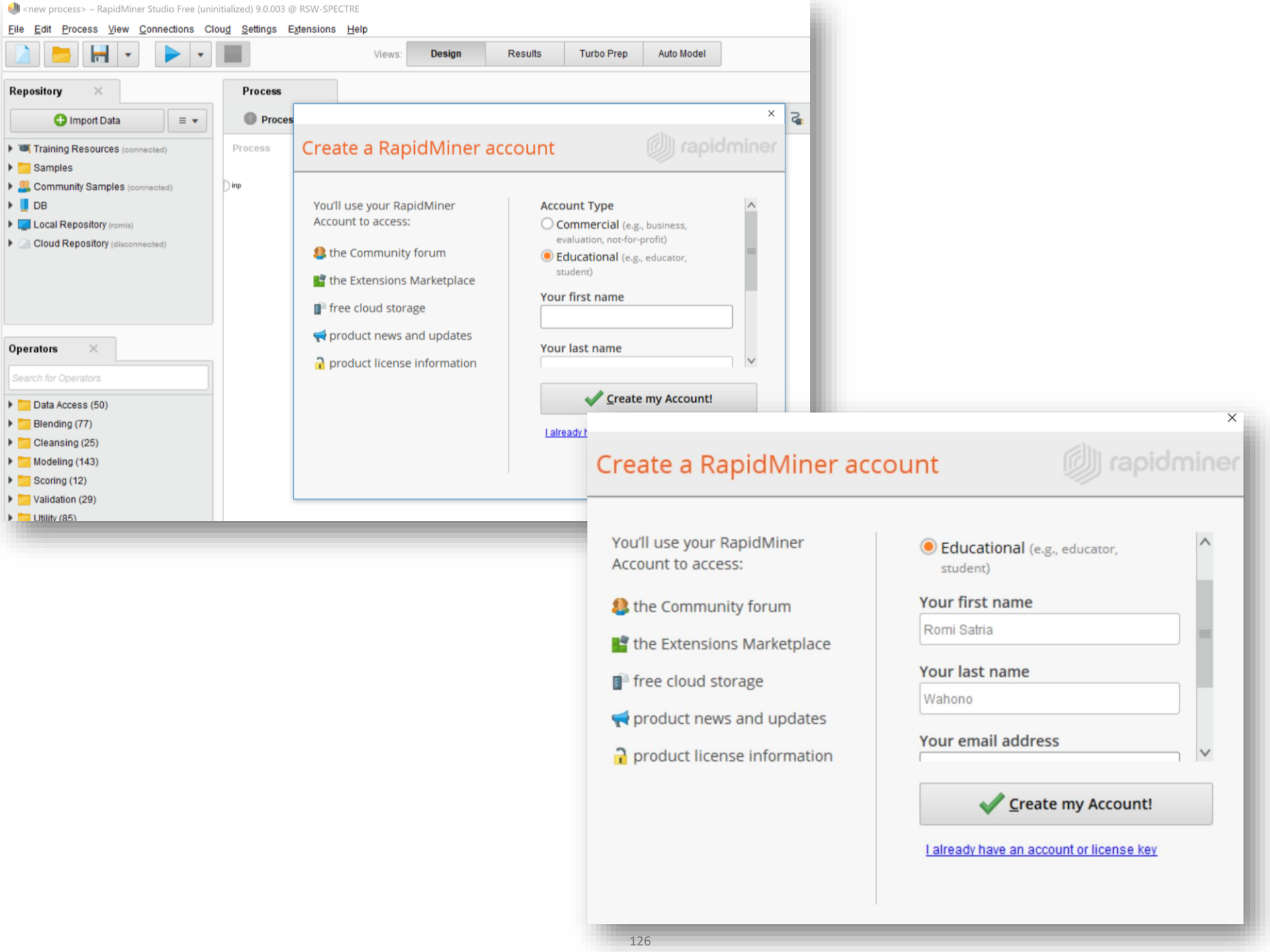
View Problems dan View Log



Instalasi dan Registrasi Lisensi Rapidminer

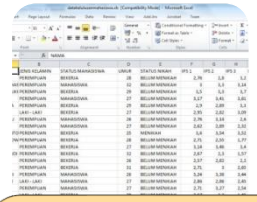
- Instal Rapidminer versi 9
- Registrasi account di rapidminer.com dan dapatkan lisensi Educational Program untuk mengolah data tanpa batasan record




<new process> – RapidMiner Studio Free (uninitialized) 9.0.003 @ RSW-SPECTRE
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
DB
Local Repository (romis)
Cloud Repository (disconnected)
Operators
Search for Operators
Data Access (50)
Blending (77)
Cleansing (25)
Modeling (143)
Scoring (12)
Validation (29)
Utility (85)
Process
Create a RapidMiner account
rapidminer
You'll use your RapidMiner Account to access:
the Community forum
the Extensions Marketplace
free cloud storage
product news and updates
product license information
Account Type
Commercial (e.g., business, evaluation, not-for-profit)
Educational (e.g., educator, student)
Your first name
Your last name
Create my Account!
Create a RapidMiner account
rapidminer
You'll use your RapidMiner Account to access:
the Community forum
the Extensions Marketplace
free cloud storage
product news and updates
product license information
Educational (e.g., educator, student)
Your first name
Romi Satria
Your last name
Wahono
Your email address
Create my Account!
I already have an account or license key

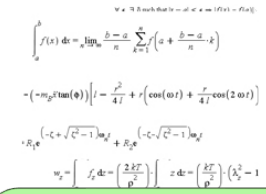
2.2 Penerapan Proses Data Mining

Proses Data Mining



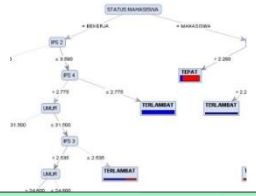
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



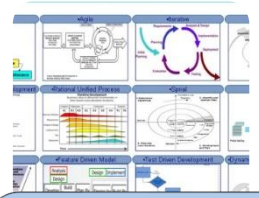
2. Metode Data Mining

(**Pilih Metode** Sesuai Karakter Data)



3. Pengetahuan

(**Pahami Model** dan Pengetahuan yg Sesuai)



4. Evaluation

(**Analisis Model** dan **Kinerja** Metode)

DATA PREPROCESSING

Data Cleaning
Data Integration
Data Reduction
Data Transformation

MODELING

Estimation
Prediction
Classification
Clustering
Association

MODEL

Formula
Tree
Cluster
Rule
Correlation

KINERJA

Akurasi
Tingkat Error
Jumlah Cluster

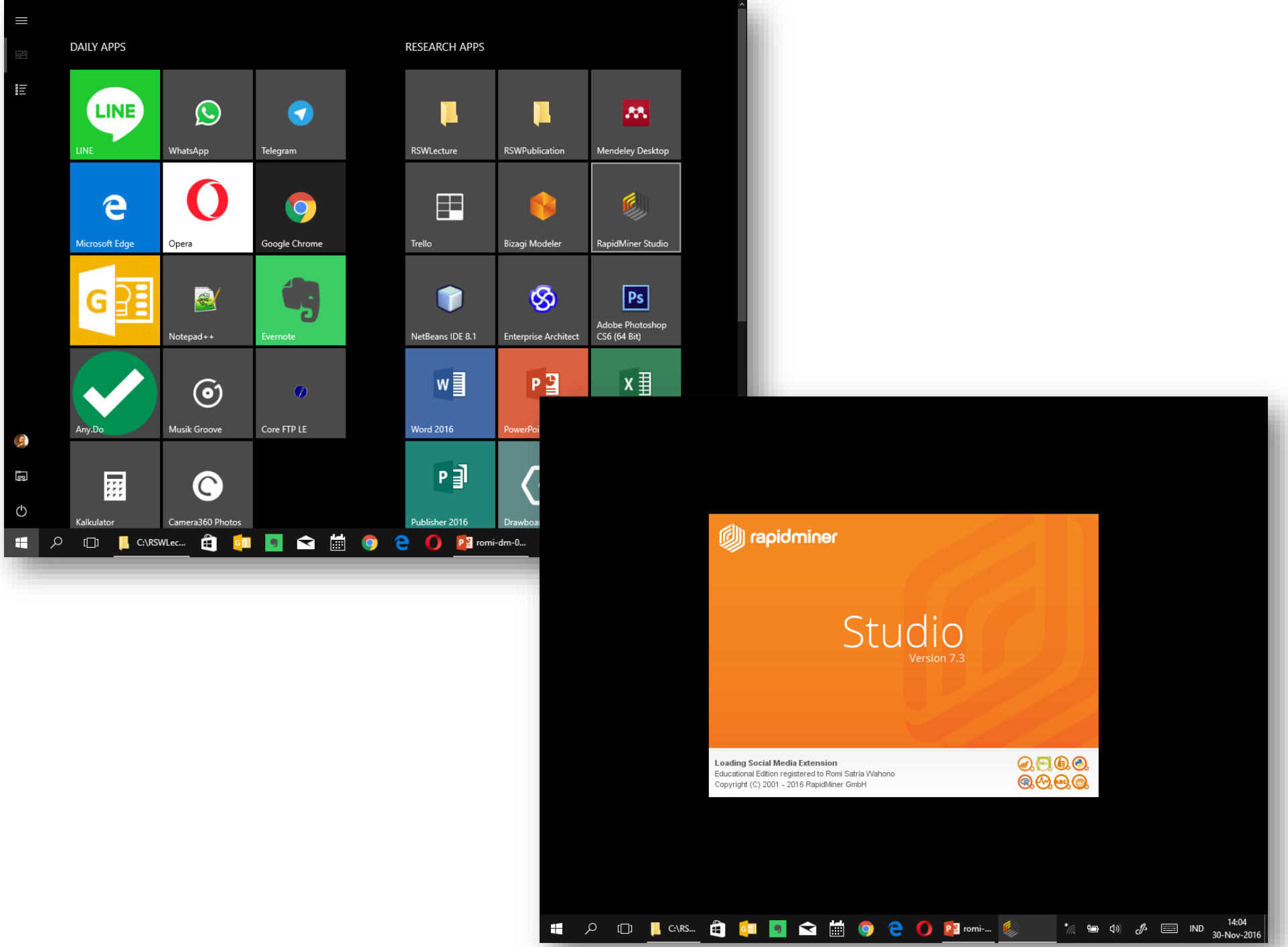
MODEL

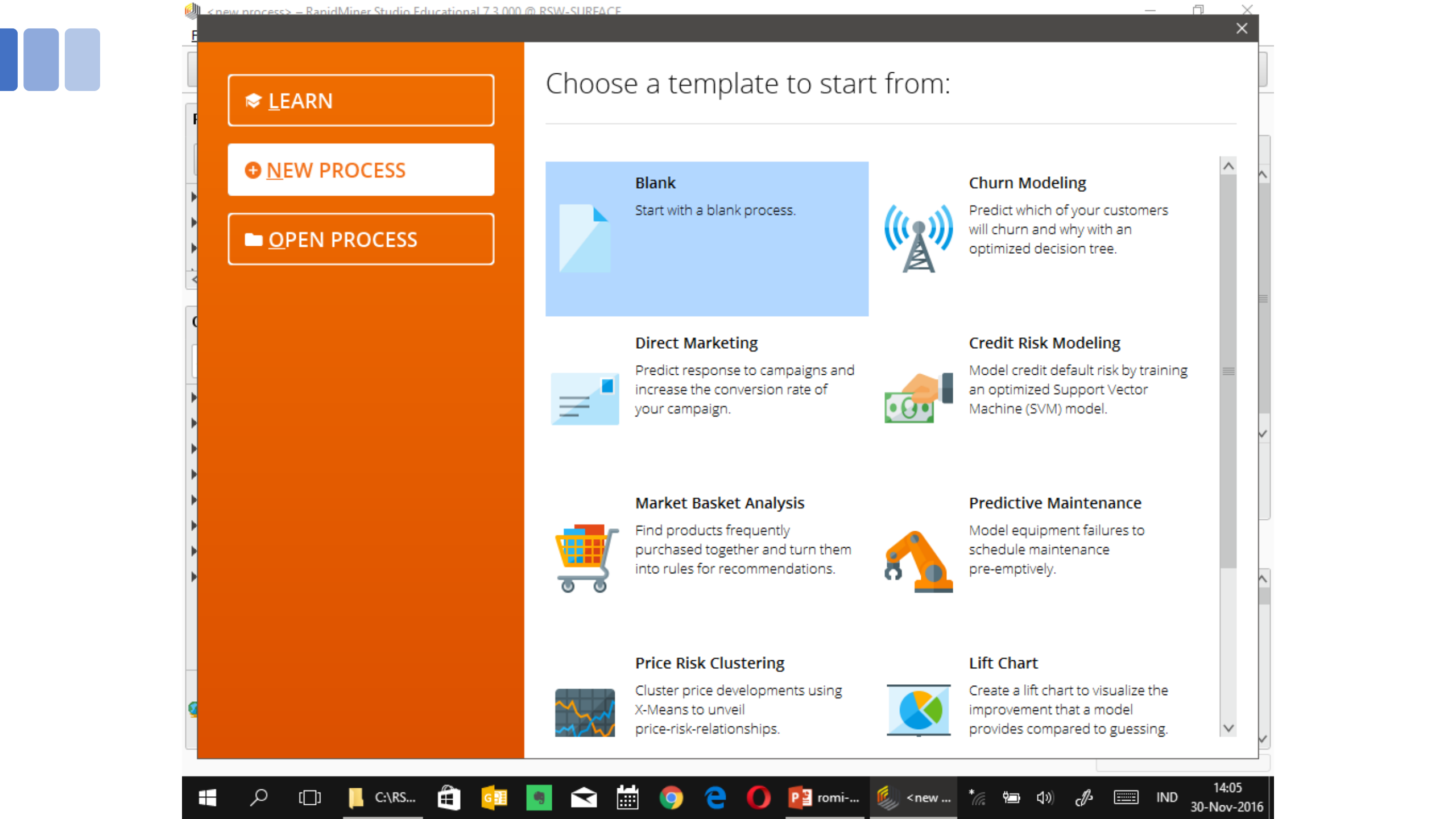
Atribute/Faktor
Korelasi
Bobot

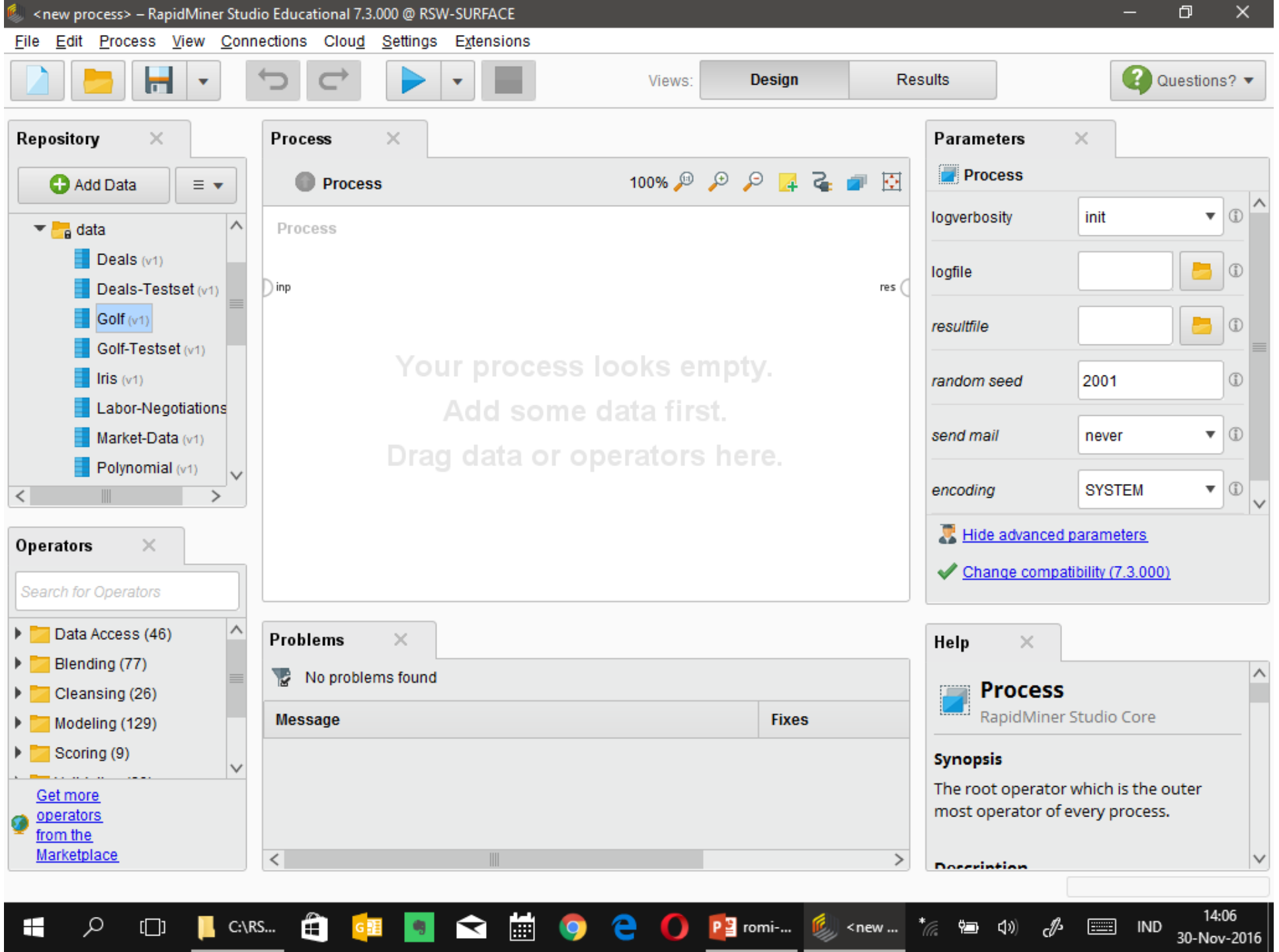
Latihan: Rekomendasi Main Golf

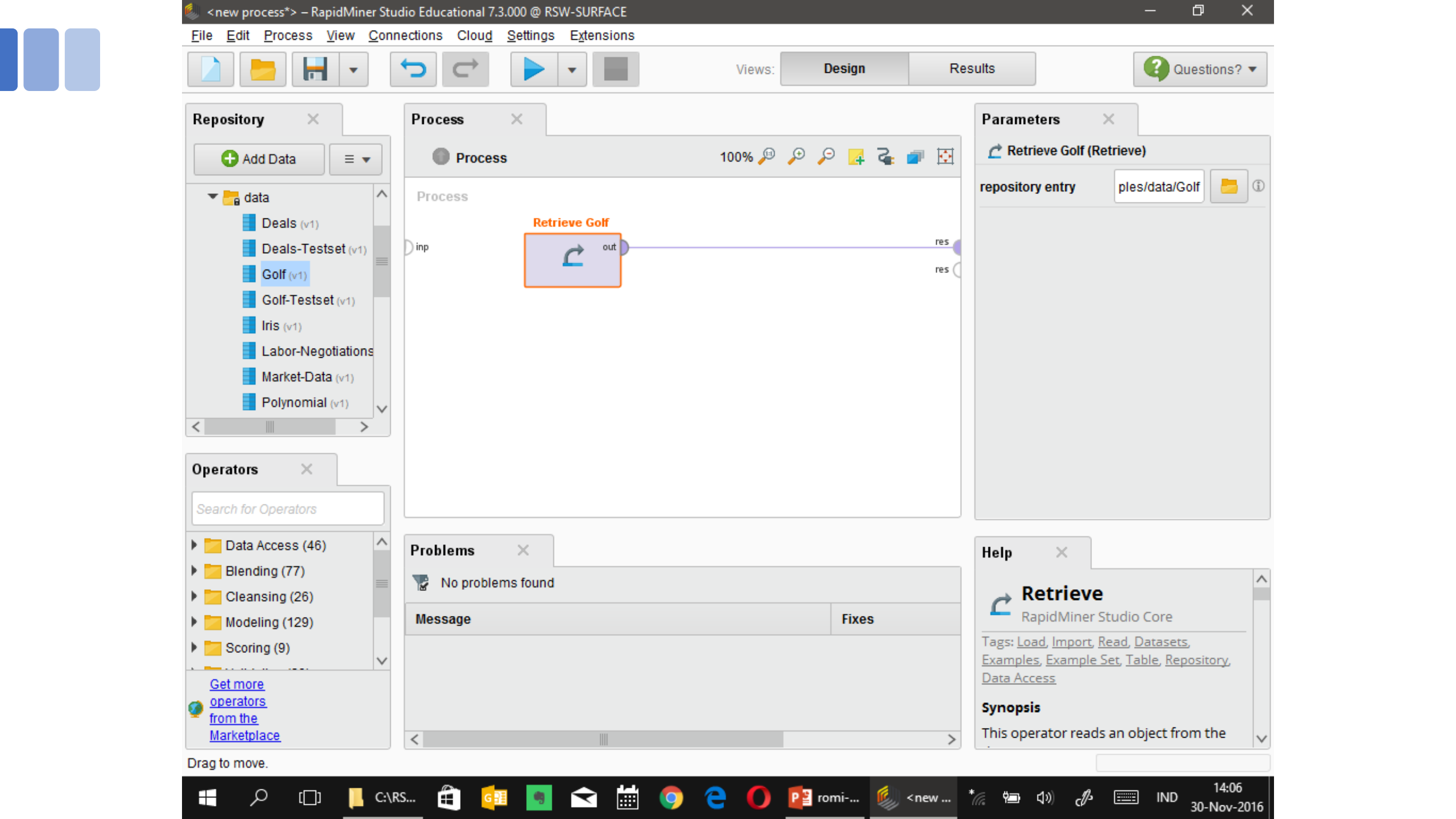
1. Lakukan training pada data golf (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree

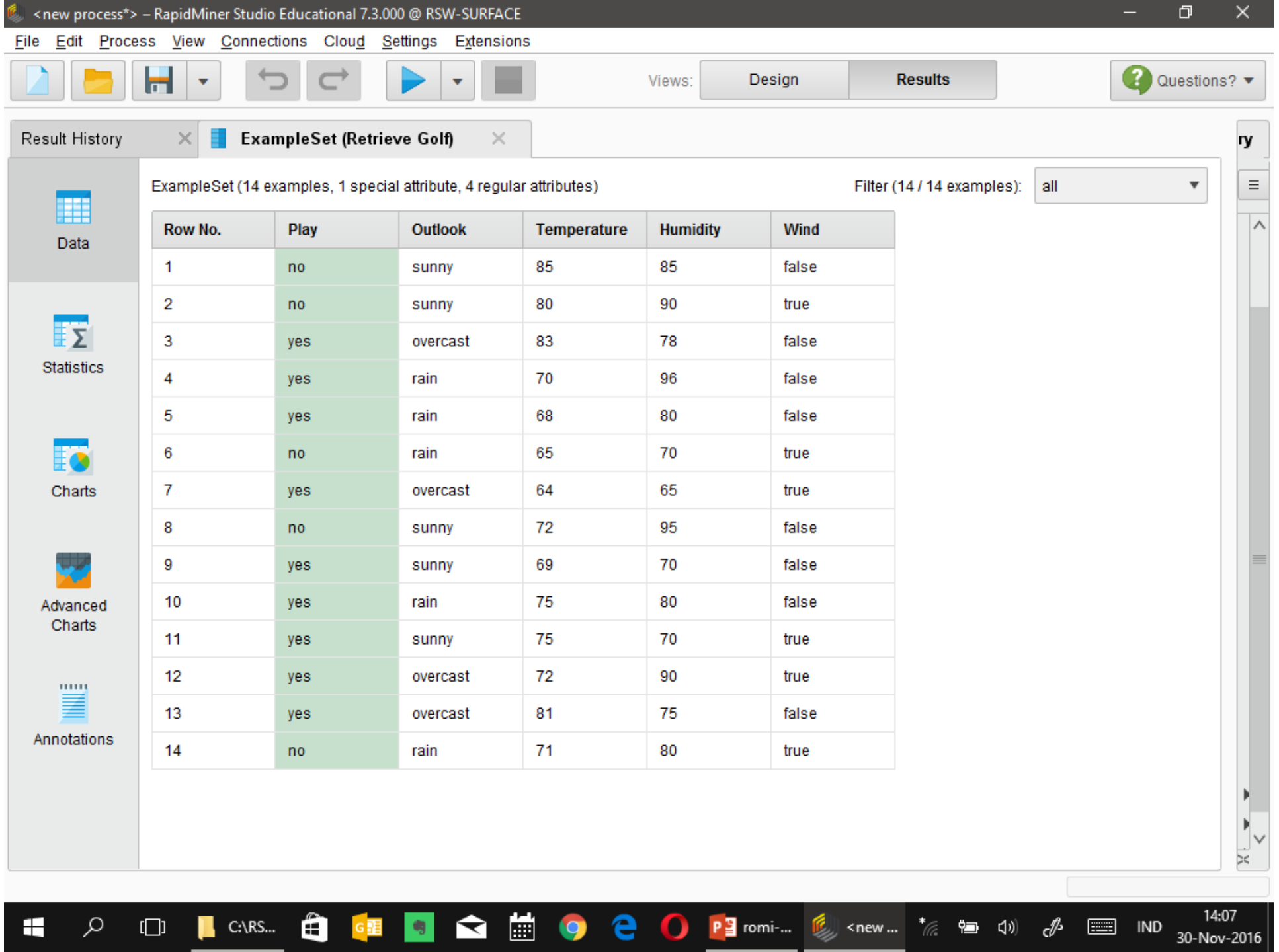
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk

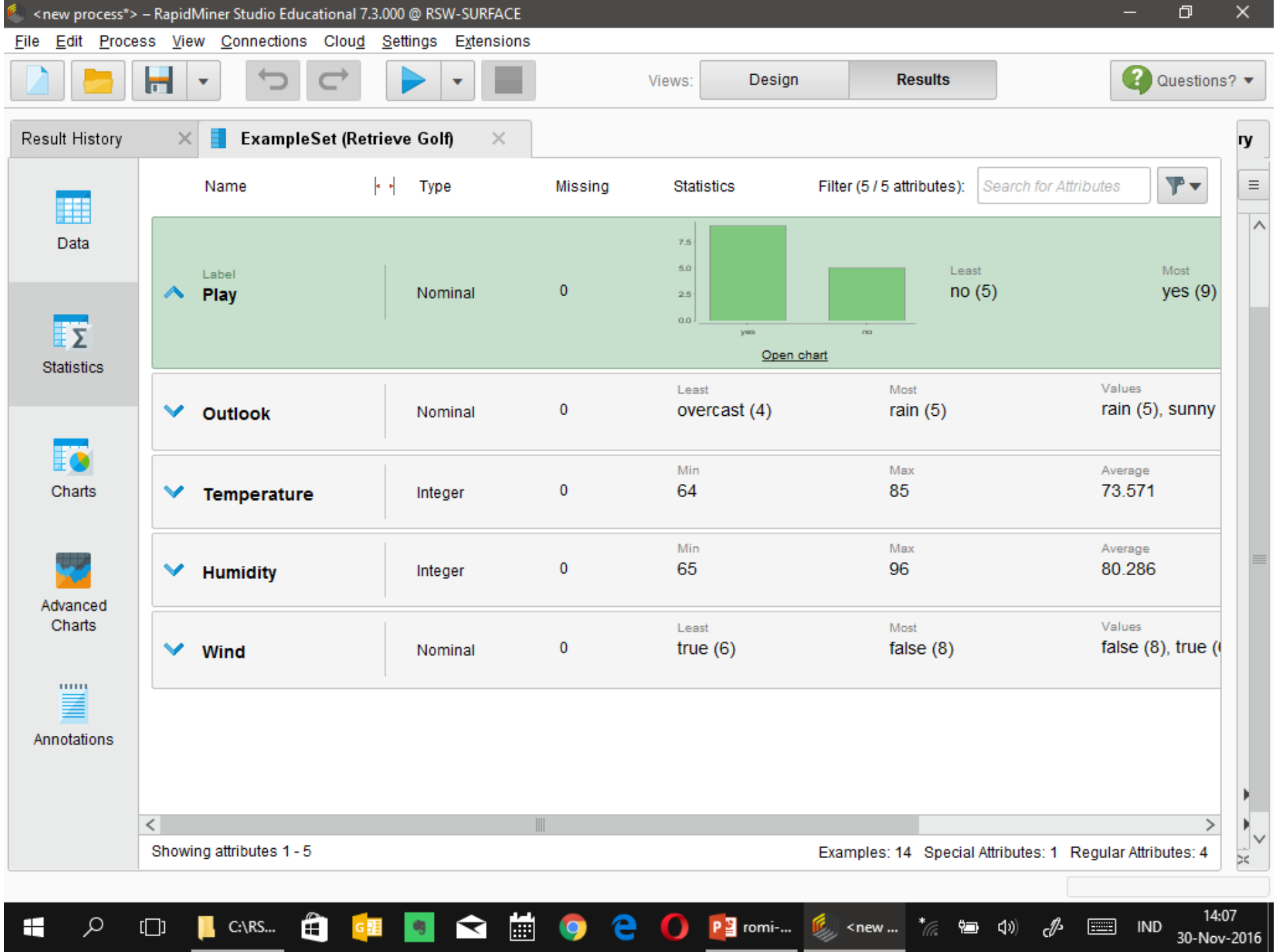

DAILY APPS
RESEARCH APPS
LINE
WhatsApp
Telegram
Microsoft Edge
Opera
Google Chrome
Notepad++
Evernote
Any.Do
Musik Groove
Core FTP LE
Kalkulator
Camera360 Photos
RSWLecture
RSWPublication
Mendeley Desktop
Trello
Bizagi Modeler
RapidMiner Studio
NetBeans IDE 8.1
Enterprise Architect
Adobe Photoshop CS6 (64 Bit)
Word 2016
Publisher 2016
rapidminer
Studio
Version 7.3
Loading Social Media Extension
Educational Edition registered to Romi Satria Wahono
Copyright (C) 2001 - 2016 RapidMiner GmbH
IND
14:04
30-Nov-2016

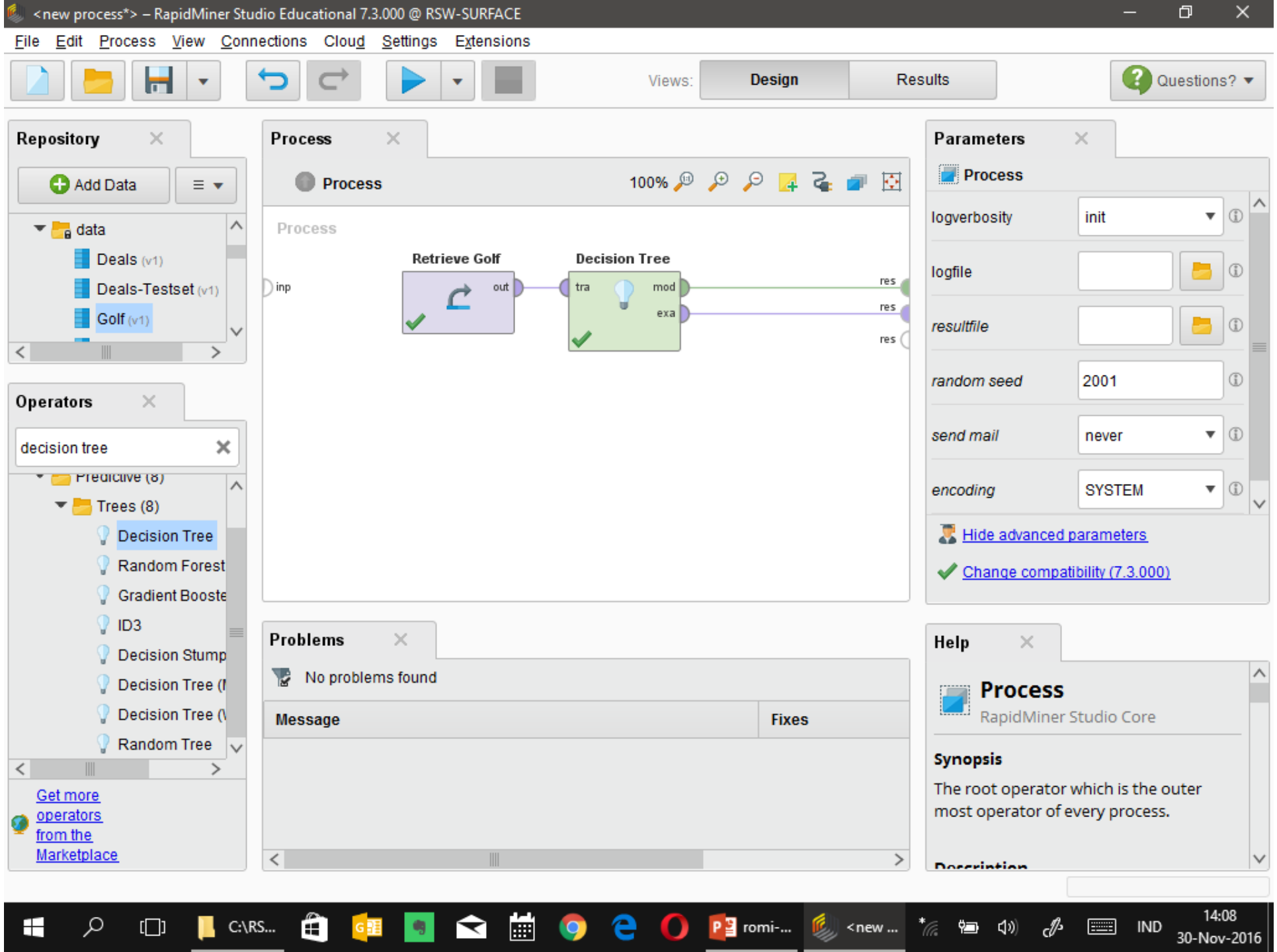

LEARN
NEW PROCESS
OPEN PROCESS
Choose a template to start from:
Blank
Start with a blank process.
Churn Modeling
Predict which of your customers will churn and why with an optimized decision tree.
Direct Marketing
Predict response to campaigns and increase the conversion rate of your campaign.
Credit Risk Modeling
Model credit default risk by training an optimized Support Vector Machine (SVM) model.
Market Basket Analysis
Find products frequently purchased together and turn them into rules for recommendations.
Predictive Maintenance
Model equipment failures to schedule maintenance pre-emptively.
Price Risk Clustering
Cluster price developments using X-Means to unveil price-risk-relationships.
Lift Chart
Create a lift chart to visualize the improvement that a model provides compared to guessing.

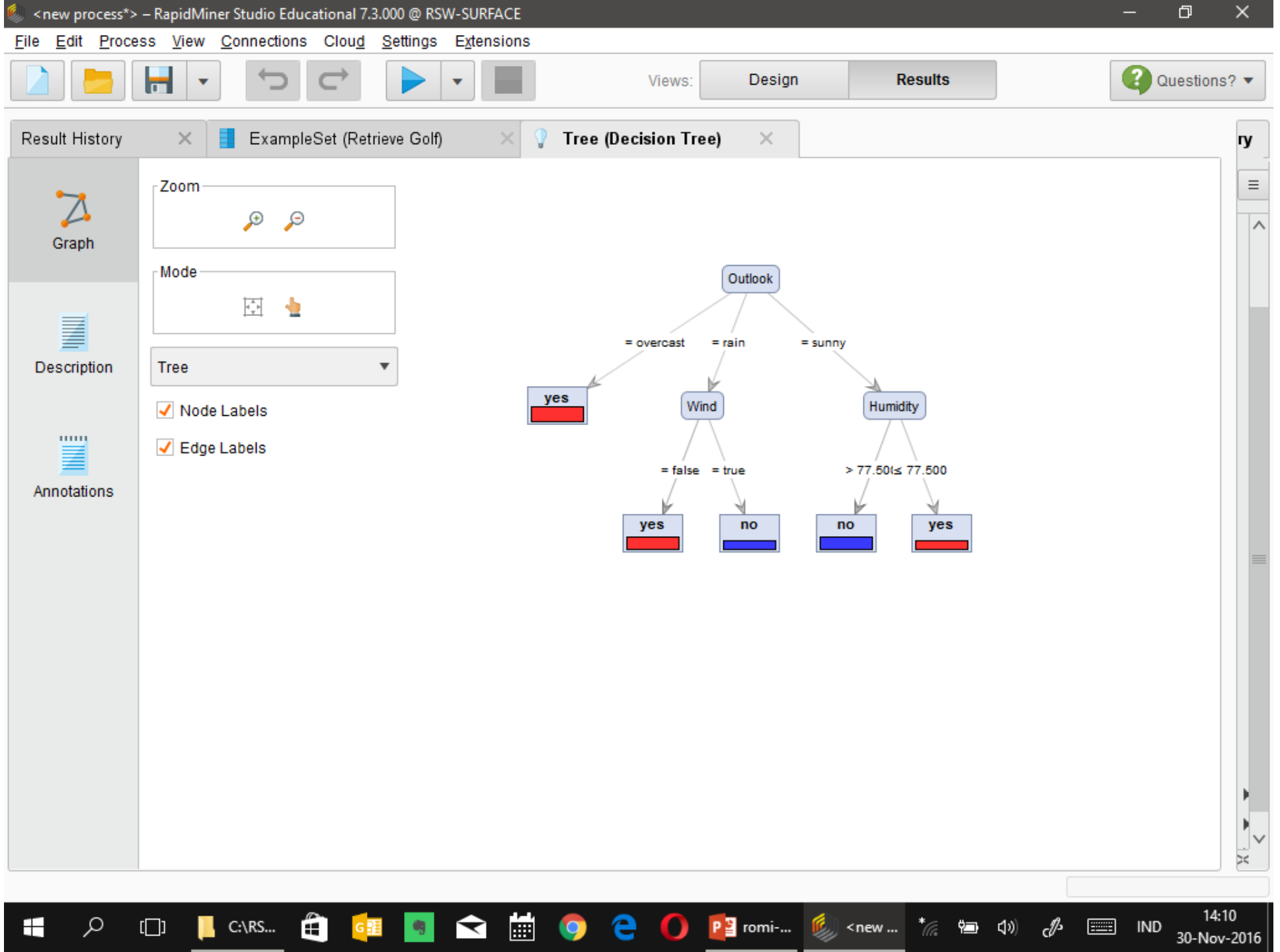

<new process> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
data
Deals (v1)
Deals-Testset (v1)
Golf (v1)
Golf-Testset (v1)
Iris (v1)
Labor-Negotiations
Market-Data (v1)
Polynomial (v1)
Process
Process
100%
Process
inp
res
Your process looks empty.
Add some data first.
Drag data or operators here.
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Hide advanced parameters
Change compatibility (7.3.000)
Operators
Search for Operators
Data Access (46)
Blending (77)
Cleansing (26)
Modeling (129)
Scoring (9)
Get more operators from the Marketplace
Problems
No problems found
Message
Fixes
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
IND
14:06
30-Nov-2016

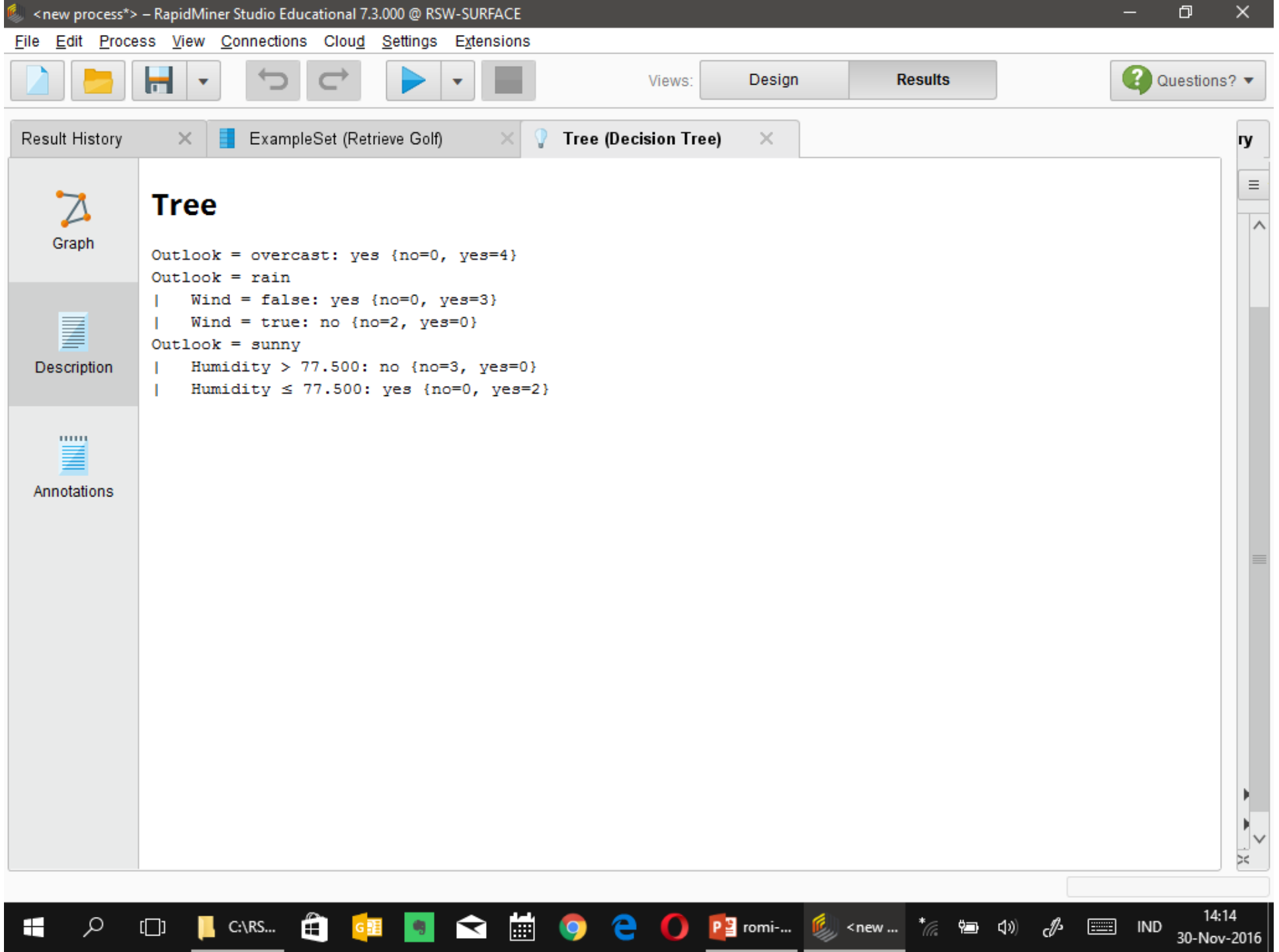

<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
data
Deals (v1)
Deals-Testset (v1)
Golf (v1)
Golf-Testset (v1)
Iris (v1)
Labor-Negotiations
Market-Data (v1)
Polynomial (v1)
Process
Process
100%
Process
Retrieve Golf
inp
out
res
res
Parameters
Retrieve Golf (Retrieve)
repository entry
ples/data/Golf
Operators
Search for Operators
Data Access (46)
Blending (77)
Cleansing (26)
Modeling (129)
Scoring (9)
Get more operators from the Marketplace
Problems
No problems found
Message
Fixes
Help
Retrieve
RapidMiner Studio Core
Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access
Synopsis
This operator reads an object from the
Drag to move.
C:\RS...
romi-...
<new ...
IND
14:06
30-Nov-2016


<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Result History
ExampleSet (Retrieve Golf)
Data
Statistics
Charts
Advanced Charts
Annotations
ExampleSet (14 examples, 1 special attribute, 4 regular attributes)
Filter (14 / 14 examples): all
Row No. | Play | Outlook | Temperature | Humidity | Wind
1 | no | sunny | 85 | 85 | false
2 | no | sunny | 80 | 90 | true
3 | yes | overcast | 83 | 78 | false
4 | yes | rain | 70 | 96 | false
5 | yes | rain | 68 | 80 | false
6 | no | rain | 65 | 70 | true
7 | yes | overcast | 64 | 65 | true
8 | no | sunny | 72 | 95 | false
9 | yes | sunny | 69 | 70 | false
10 | yes | rain | 75 | 80 | false
11 | yes | sunny | 75 | 70 | true
12 | yes | overcast | 72 | 90 | true
13 | yes | overcast | 81 | 75 | false
14 | no | rain | 71 | 80 | true
C:\RS...
romi-...
<new ...
IND
14:07
30-Nov-2016


<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Result History
ExampleSet (Retrieve Golf)
Data
Statistics
Charts
Advanced Charts
Annotations
Name Type Missing Statistics Filter (5 / 5 attributes): Search for Attributes
Label
Play Nominal 0 Least no (5) Most yes (9)
Open chart
Outlook Nominal 0 Least overcast (4) Most rain (5) Values rain (5), sunny
Temperature Integer 0 Min 64 Max 85 Average 73.571
Humidity Integer 0 Min 65 Max 96 Average 80.286
Wind Nominal 0 Least true (6) Most false (8) Values false (8), true
Showing attributes 1 - 5
Examples: 14 Special Attributes: 1 Regular Attributes: 4
IND
14:07
30-Nov-2016

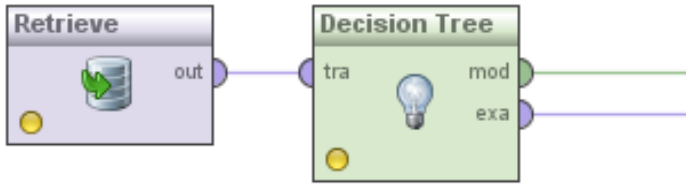

<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
data
Deals (v1)
Deals-Testset (v1)
Golf (v1)
Operators
decision tree
Predictive (8)
Trees (8)
Decision Tree
Random Forest
Gradient Booste
ID3
Decision Stump
Decision Tree (
Decision Tree (
Random Tree
Get more operators from the Marketplace
Process
Process
100%
Process
Retrieve Golf
Decision Tree
inp
out
tra
mod
exa
res
res
res
Problems
No problems found
Message
Fixes
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Hide advanced parameters
Change compatibility (7.3.000)
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description
C:\RS...
romi-...
<new ...
IND
14:08
30-Nov-2016

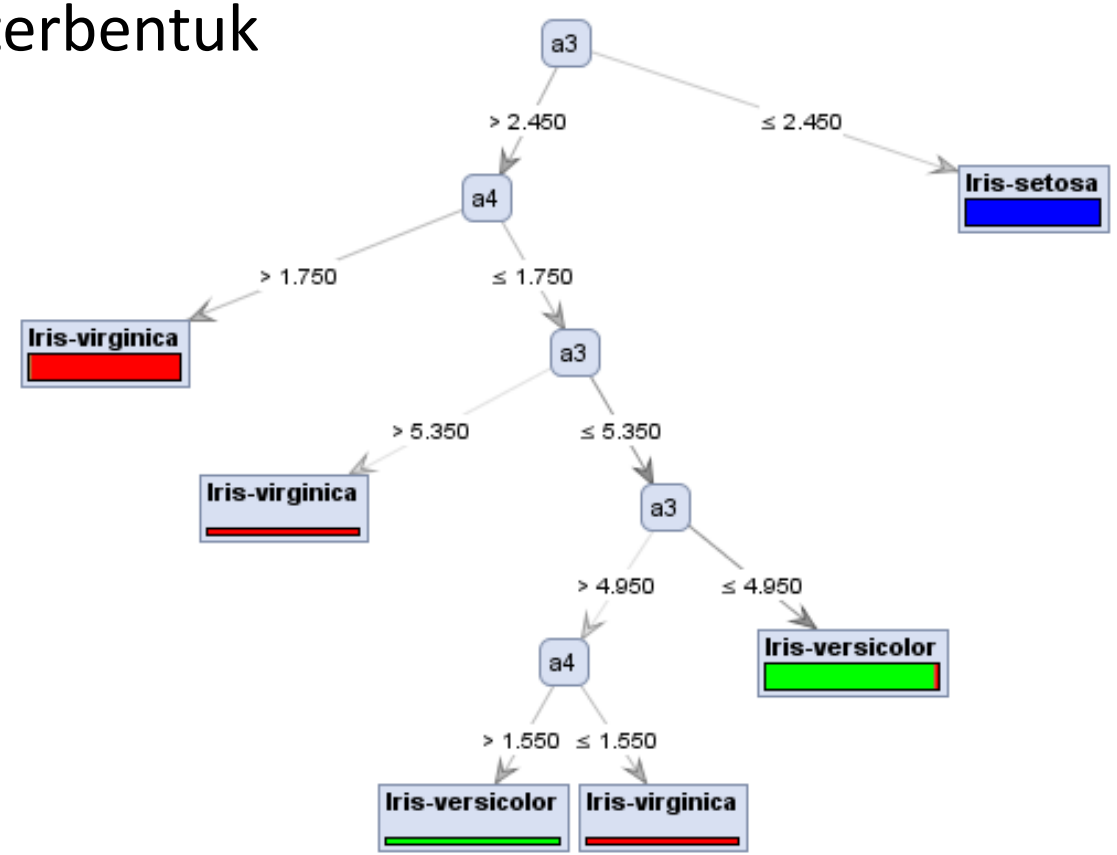

<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Result History
ExampleSet (Retrieve Golf)
Tree (Decision Tree)
Graph
Description
Annotations
Zoom
Mode
Tree
Node Labels
Edge Labels
Outlook
= overcast
= rain
= sunny
yes
Wind
Humidity
= false
= true
> 77.500
≤ 77.500
yes
no
no
yes
IND
14:10
30-Nov-2016


<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Result History
ExampleSet (Retrieve Golf)
Tree (Decision Tree)
Graph
Description
Annotations
Tree
Outlook = overcast: yes {no=0, yes=4}
Outlook = rain
| Wind = false: yes {no=0, yes=3}
| Wind = true: no {no=2, yes=0}
Outlook = sunny
| Humidity > 77.500: no {no=3, yes=0}
| Humidity ≤ 77.500: yes {no=0, yes=2}
IND
14:14
30-Nov-2016

Latihan: Penentuan Jenis Bunga Iris

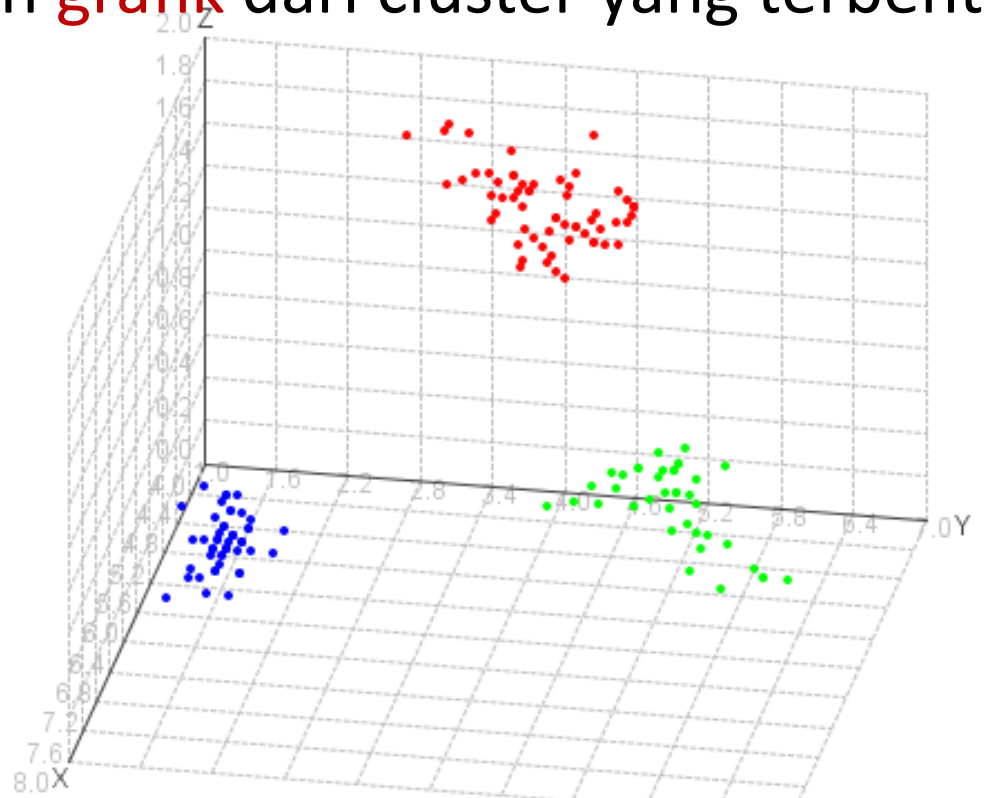
1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk





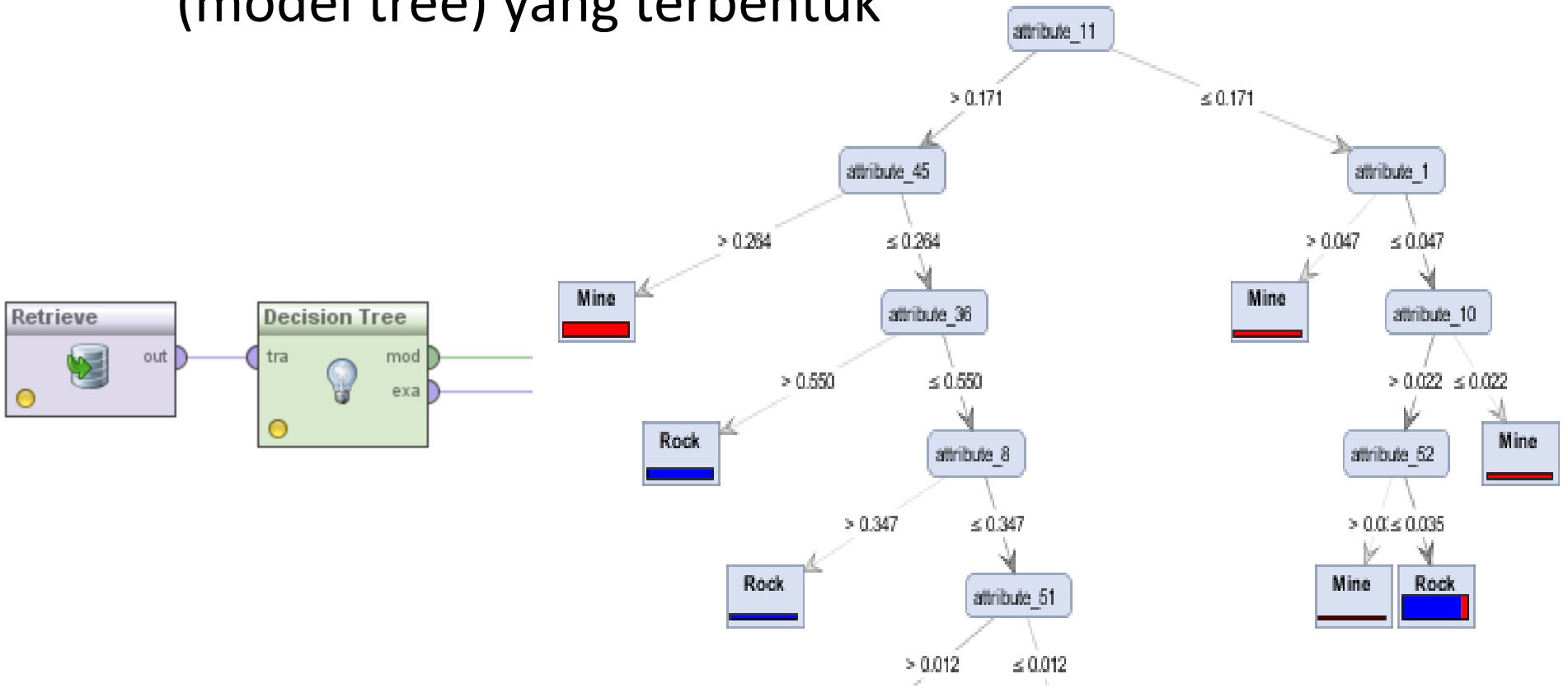
Latihan: Klastering Jenis Bunga Iris

1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan algoritma k-Means
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk
3. Tampilkan grafik dari cluster yang terbentuk



Latihan: Penentuan Mine/Rock

1. Lakukan training pada data Sonar (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree (C4.5)
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk



Data Profile dan Kinerja Marketing

Mana Atribut yang Layak jadi Class dan Tidak?

NIP	Gender	Universitas	Program Studi	IPK	Usia	Hasil Penjualan	Status Keluarga	Jumlah Anak	Kota Tinggal
1001	L	UI	Komunikasi	3.1	21	100jt	Single	0	Jakarta
1002	P	UNDIP	Informatika	2.9	26	50jt	menikah	1	Bekasi
…	…	…	…	…	…	…	…	…	…

NIP	Gender	Hasil Penjualan Produk A	Hasil Penjualan Produk B	Hasil Penjualan Layanan C	Hasil Penjualan Layanan D	Total Hasil Penjualan
1001	L	10	20	50	30	100jt
1002	P	10	10	5	25	50jt
…	…	…	…	…	…	…

Data Profile dan Kinerja Marketing

Mana Atribut yang Layak jadi Class dan Tidak?

Tahun	Total Hasil Penjualan	Total Pengeluaran Marketing	Total Keuntungan
1990	100jt	98jt	2jt
1991	120jt	100	20jt
…	…		

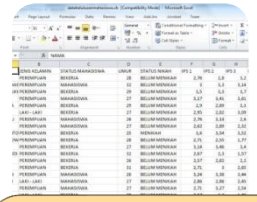
Data Profil dan Kinerja Dosen

NIP	Gender	Universitas	Program Studi	Absensi	Usia	Jumlah Penelitian	Status Keluarga	Disiplin	Kota Tinggal
1001	L	UI	Komunikasi	98%	21	3	Single	Baik	Jakarta
1002	P	UNDIP	Informatika	50%	26	4	menikah	Buruk	Bekasi
…	…	…	…	…	…	…	…	…	…

NIP	Gender	Universitas	Program Studi	Jumlah Publikasi Jurnal	Jumlah Publikasi Konferensi	Total Publikasi Penelitian
1001	L	UI	Komunikasi	5	3	8
1002	P	UNDIP	Informatika	2	1	3
…	…	…	…	…	…	…

2.3 Evaluasi Model Data Mining

Proses Data Mining

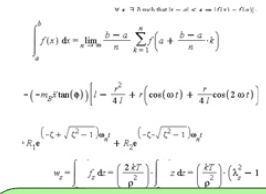


1. Himpunan Data

(**Pahami** dan Persiapkan Data)

DATA PREPROCESSING

Data Cleaning
Data Integration
Data Reduction
Data Transformation

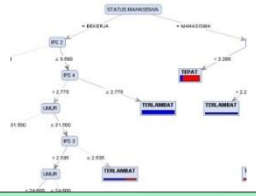


2. Metode Data Mining

(**Pilih Metode** Sesuai Karakter Data)

MODELING

Estimation
Prediction
Classification
Clustering
Association

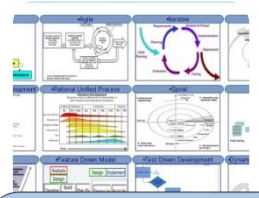


3. Pengetahuan

(**Pahami Model** dan Pengetahuan yg Sesuai)

MODEL

Formula
Tree
Cluster
Rule
Correlation



4. Evaluation

(**Analisis Model** dan **Kinerja** Metode)

KINERJA

Akurasi
Tingkat Error
Jumlah Cluster

MODEL

Atribute/Faktor
Korelasi
Bobot

Evaluasi Kinerja Model Data Mining

1. Estimation:
   - Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   - Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
   - Confusion Matrix: Accuracy
   - ROC Curve: Area Under Curve (AUC)
4. Clustering:
   - Internal Evaluation: Davies–Bouldin index, Dunn index,
   - External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:
   - Lift Charts: Lift Ratio
   - Precision and Recall (F-measure)

Evaluasi Model Data Mining

- Pembagian dataset, perbandingan 90:10 atau 80:20:
  - Data Training
  - Data Testing

- Data training untuk pembentukan model, dan data testing digunakan untuk pengujian model

- Pemisahan data training dan testing
  1. Data dipisahkan secara manual
  2. Data dipisahkan otomatis dengan operator Split Data
  3. Data dipisahkan otomatis dengan X Validation

Confusion Matrix → Accuracy

accuracy: 90.00%

	true MACET	true LANCAR	class precision
pred. MACET	53	4	92.98%
pred. LANCAR	6	37	86.05%
class recall	89.83%	90.24%	

- pred MACET- true MACET: Jumlah data yang diprediksi macet dan kenyataannya macet (**TP**)
- pred LANCAR-true LANCAR: Jumlah data yang diprediksi lancar dan kenyataannya lancar (**TN**)
- pred MACET-true LANCAR: Jumlah data yang diprediksi macet tapi kenyataannya lancer (**FP**)
- pred LANCAR-true MACET: Jumlah data yang diprediksi lancar tapi kenyataanya macet (**FN**)

$$\text{Accuracy} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}} = \frac{53 + 37}{53 + 37 + 4 + 6} = \frac{90}{100} = 90\%$$

Precision and Recall, and F-measures

- Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

- Recall: completeness – what % of positive tuples did the classifier label as positive?
  - Perfect score is 1.0
  - Inverse relationship between precision & recall

$$recall = \frac{TP}{TP + FN}$$

- F measure (F1 or F-score): harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- Fß: weighted measure of preci and recall
  - assigns ß times as much weight to recall as to precision

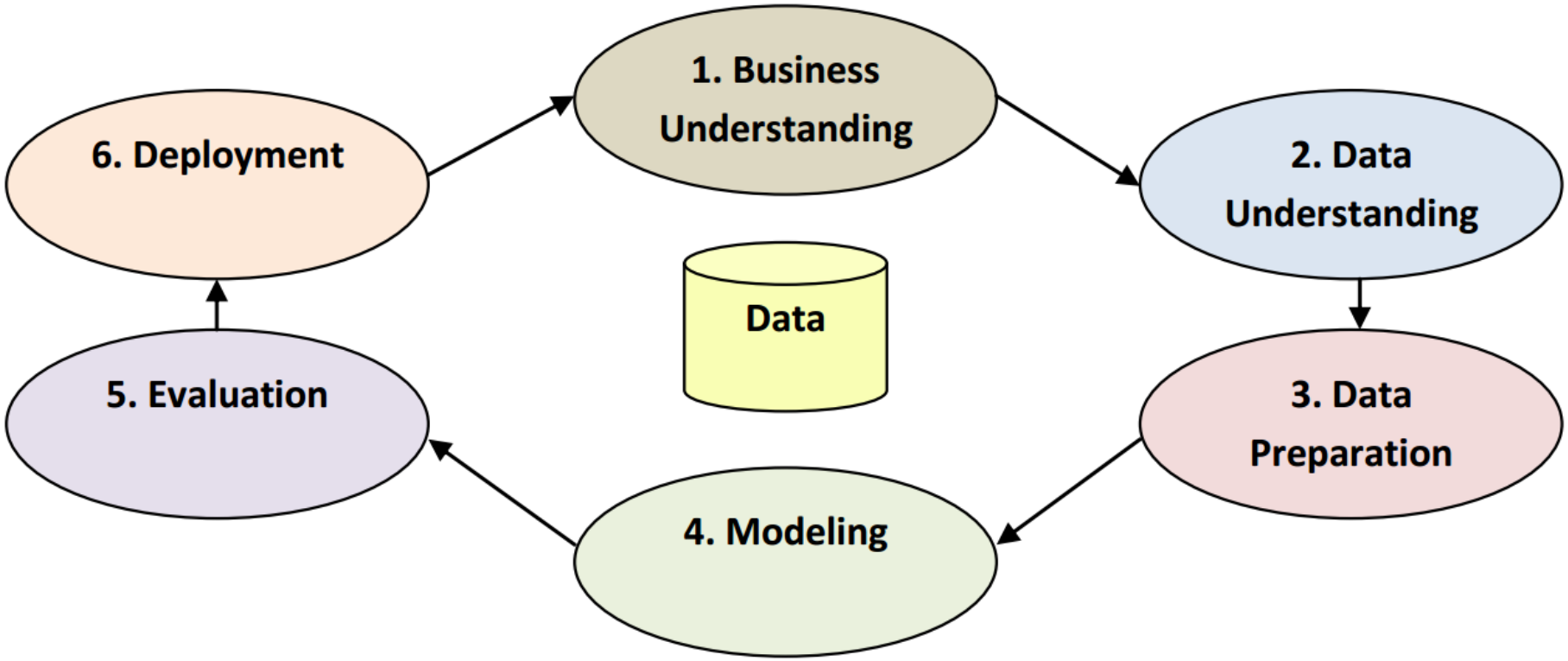
$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

2.4 Proses Data Mining berbasis Metodologi CRISP-DM

Data Mining Standard Process

- 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)*

CRISP-DM


1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
Data

1. Business Understanding

- Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole
- Translate these goals and restrictions into the formulation of a data mining problem definition
- Prepare a preliminary strategy for achieving these objectives
- Designing what you are going to build

2. Data Understanding

- Collect the data
- Use exploratory data analysis to familiarize yourself with the data and discover initial insights
- Evaluate the quality of the data
- If desired, select interesting subsets that may contain actionable patterns

3. Data Preparation

- Prepare from the initial raw data the final data set that is to be used for all subsequent phases
- Select the cases and variables you want to analyze and that are appropriate for your analysis
- Perform data cleaning, integration, reduction and transformation, so it is ready for the modeling tools

4. Modeling

- Select and apply appropriate modeling techniques
- Calibrate model settings to optimize results
- Remember that often, several different techniques may be used for the same data mining problem
- If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique

5. Evaluation

- Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field
- Determine whether the model in fact achieves the objectives set for it in the first phase
- Establish whether some important facet of the business or research problem has not been accounted for sufficiently
- Come to a decision regarding use of the data mining results

6. Deployment

- Make use of the models created:
  - model creation does not signify the completion of a project
- Example of a simple deployment:
  - Generate a report
- Example of a more complex deployment:
  - Implement a parallel data mining process in another department
- For businesses, the customer often carries out the deployment based on your model

CRISP-DM: Detail Flow