DATA MINING

Romi Satria Wahono
romi@romisatriawahono.net
http://romisatriawahono.net
08118228331

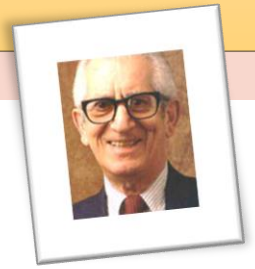
Romi Satria Wahono

- SMA Taruna Nusantara Magelang (1993)
- B.Eng, M.Eng and Ph.D in Software Engineering
  Saitama University Japan (1994-2004)
  Universiti Teknikal Malaysia Melaka (2014)
- Core Competency in Enterprise Architecture, Software Engineering and Machine Learning
- LIPI Researcher (2004-2007)
- Founder and CEO:
  - PT Brainmatics Cipta Informatika (2005)
  - PT IlmuKomputerCom Braindevs Sistema (2014)
- Professional Member of IEEE, ACM and PMI
- IT and Research Award Winners from WSIS (United Nations), Kemdikbud, Ristekdikti, LIPI, etc
- SCOPUS/ISI Indexed Q1 Journal Reviewer: Information and Software Technology, Journal of Systems and Software, Software: Practice and Experience, Empirical Software Engineering, etc
- Industrial IT Certifications: TOGAF, ITIL, CCAI, CCNA, etc
- Enterprise Architecture Consultant: KPK, RistekDikti, INSW, BPPT, Kemsos Kemenkeu (Itjend, DJBC, DJPK), Telkom, FIF, PLN, PJB, Pertamina EP, etc

Learning Design

Educational Objectives
(*Benjamin Bloom*)

- Cognitive
- Affective
- Psychomotor



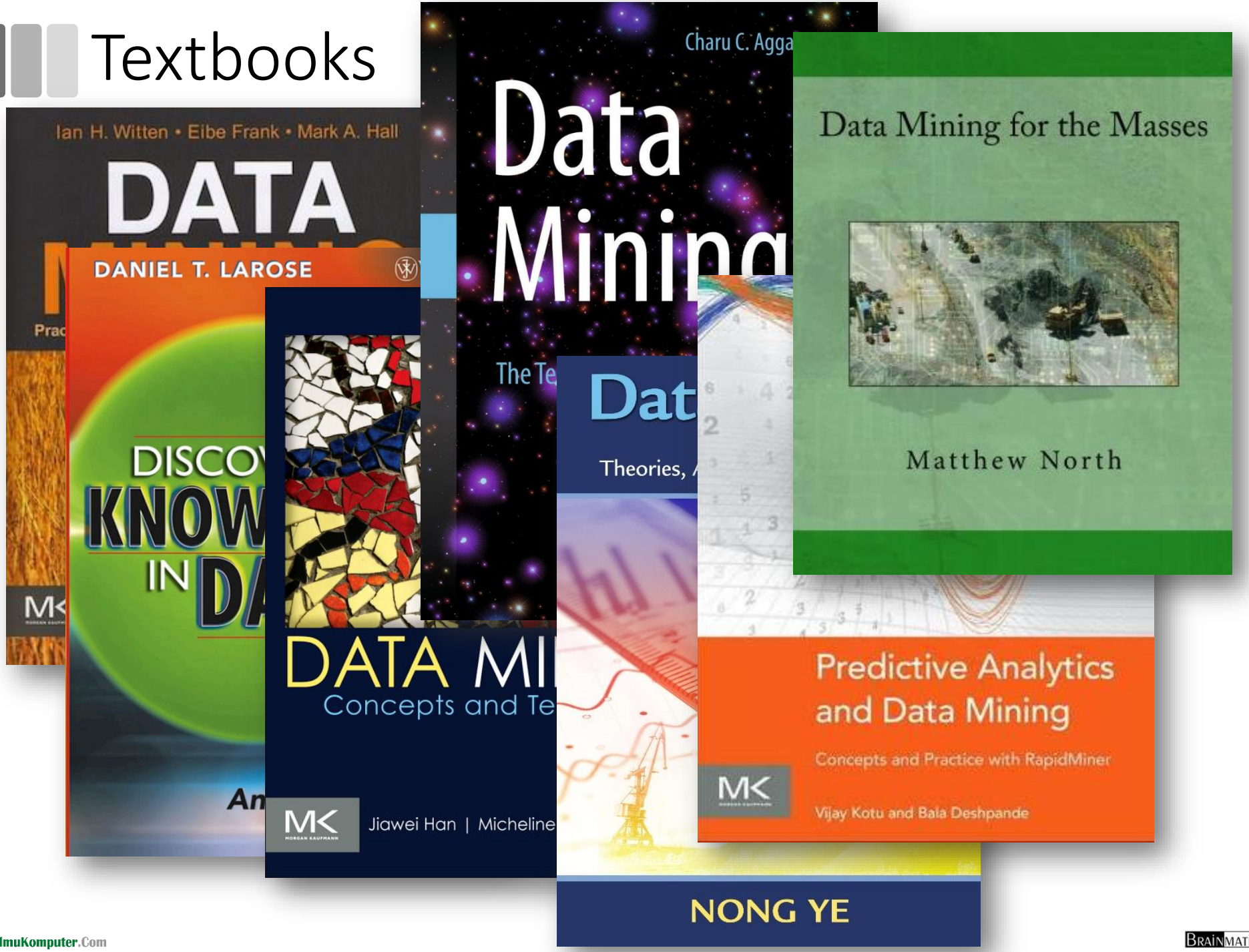
Criterion Referenced Instruction
(*Robert Mager*)

- Competencies
- Performance
- Evaluation

Minimalism
(*John Carroll*)

- Start Immediately
- Minimize the Reading
- Error Recognition
- Self-Contained

Textbooks



Pre-Test

1. Jelaskan perbedaan antara data, informasi dan pengetahuan!
2. Jelaskan apa yang anda ketahui tentang data mining!
3. Sebutkan peran utama data mining!
4. Sebutkan pemanfaatan dari data mining di berbagai bidang!
5. Pengetahuan apa yang bisa kita dapatkan dari data di bawah?

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	SMAN 7	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

Course Outline

1. Pengantar

1.1 Apa dan Mengapa Data Mining?
1.2 Peran Utama dan Metode Data Mining
1.3 Sejarah dan Penerapan Data Mining

2. Proses

2.1 Proses dan Tools Data Mining
2.2 Penerapan Proses Data Mining
2.3 Evaluasi Model Data Mining
2.4 Proses Data Mining berbasis CRISP-DM

3. Persiapan Data

3.1 Data Cleaning
3.2 Data Reduction
3.3 Data Transformation
3.4 Data Integration

4. Algoritma

4.1 Algoritma Klasifikasi
4.2 Algoritma Klastering
4.3 Algoritma Asosiasi
4.4 Algoritma Estimasi dan Forecasting

5. Text Mining

5.1 Text Mining Concepts
5.2 Text Clustering
5.3 Text Classification
5.4 Data Mining Laws

1. Pengantar Data Mining

1.1 Apa dan Mengapa Data Mining?

1.2 Peran Utama dan Metode Data Mining

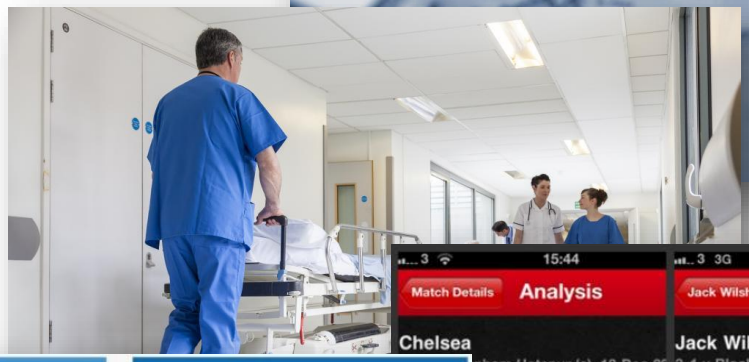
1.3 Sejarah dan Penerapan Data Mining

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
- …



Pertumbuhan Data

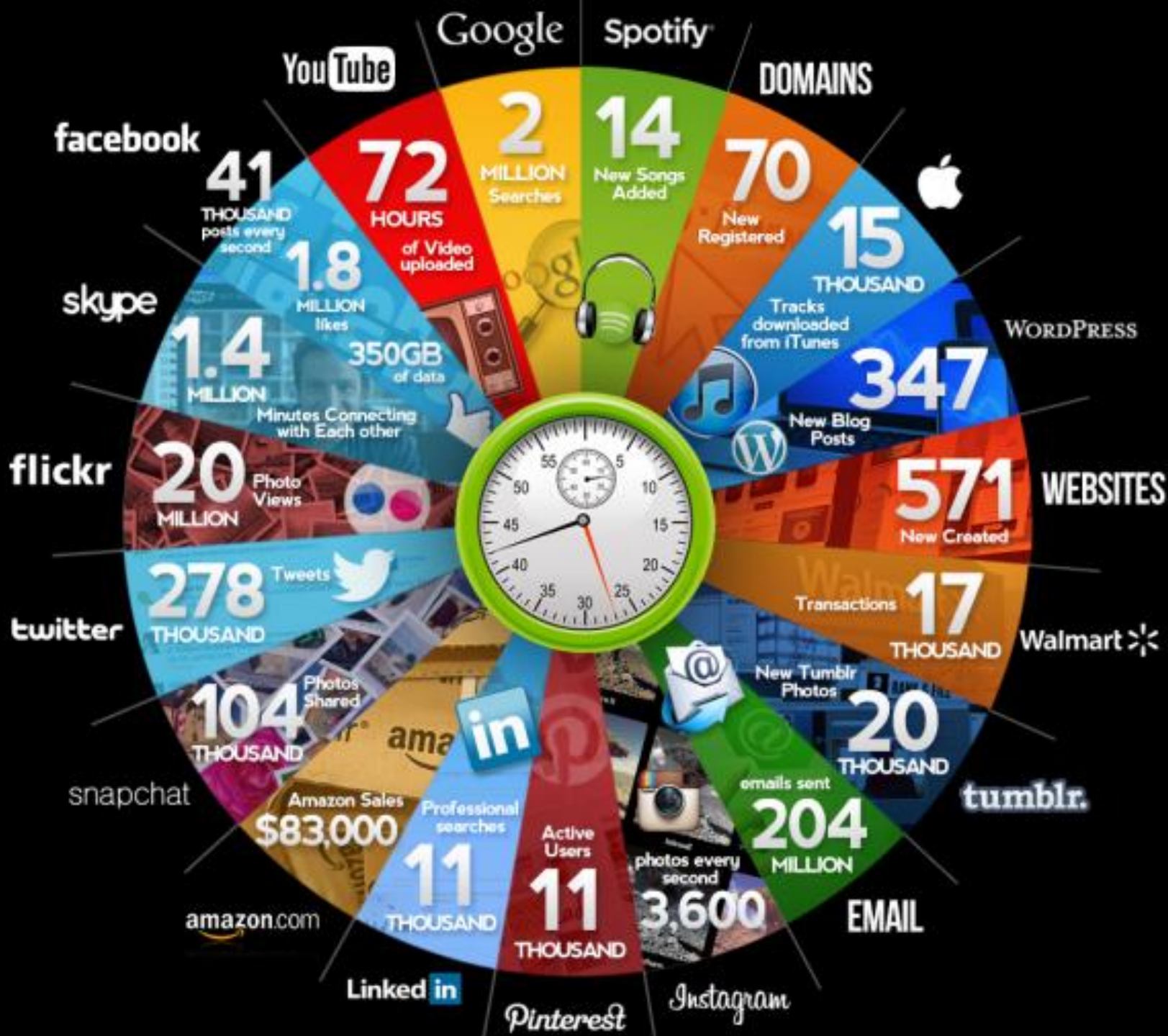
kilobyte (**kB**)	10^3
megabyte (**MB**)	10^6
gigabyte (**GB**)	10^9
terabyte (**TB**)	10^{12}
petabyte (**PB**)	10^{15}
exabyte (**EB**)	10^{18}
zettabyte (**ZB**)	10^{21}
yottabyte (**YB**)	10^{24}

Astronomi

- Sloan Digital Sky Survey
  - New Mexico, 2000
  - 140TB over 10 years
- Large Synoptic Survey Telescope
  - Chile, 2016
  - Will acquire 140TB every five days

Biologi dan Kedokteran

- European Bioinformatics Institute (EBI)
  - 20PB of data (genomic data doubles in size each year)
  - A single sequenced human genome can be around 140GB in size


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

Datangnya Tsunami Data

- Mobile Electronics market
  - 7B smartphone subscriptions in 2015

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}

- 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

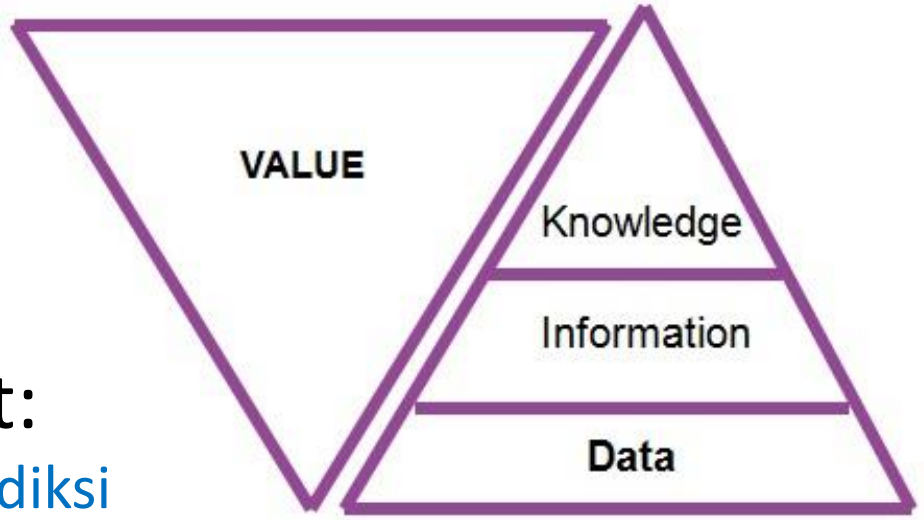
Kebanjiran Data tapi Miskin Pengetahuan

We are drowning in data,
but starving for knowledge!

(John Naisbitt, Megatrends, 1988)

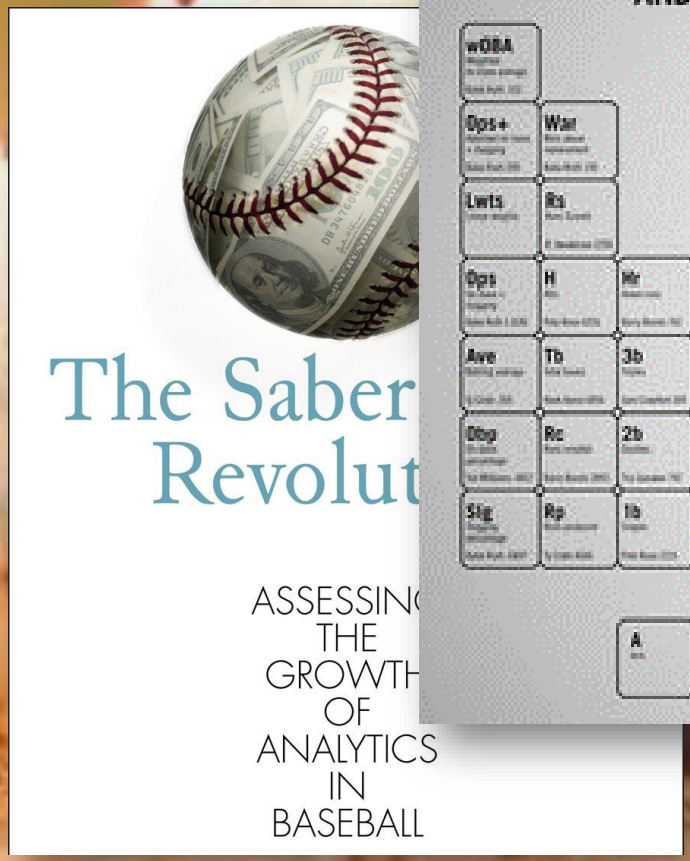
Mengubah Data Menjadi Pengetahuan

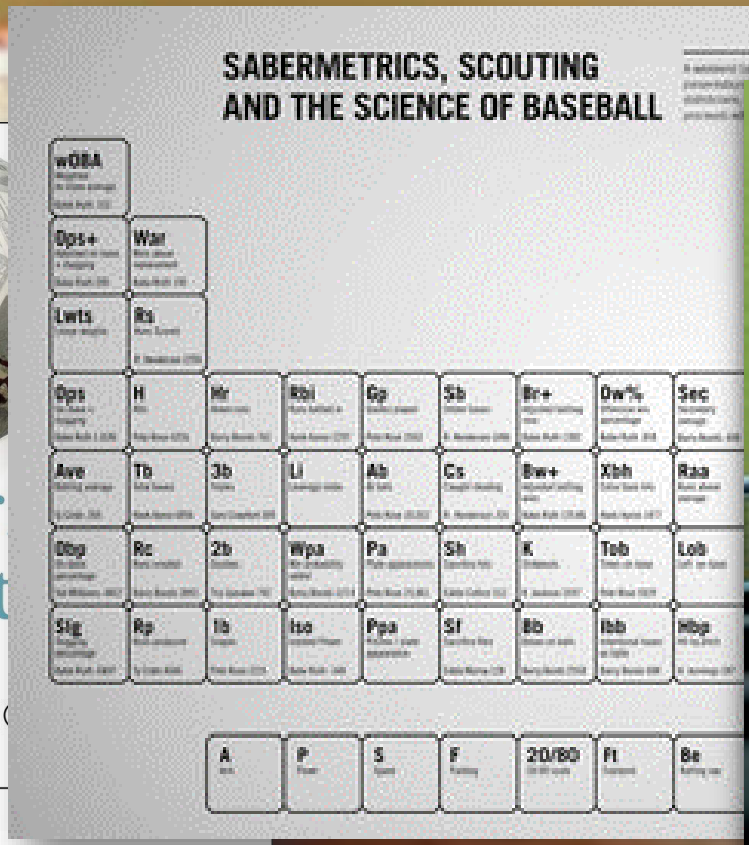
- Data harus kita olah menjadi pengetahuan supaya bisa bermanfaat bagi manusia

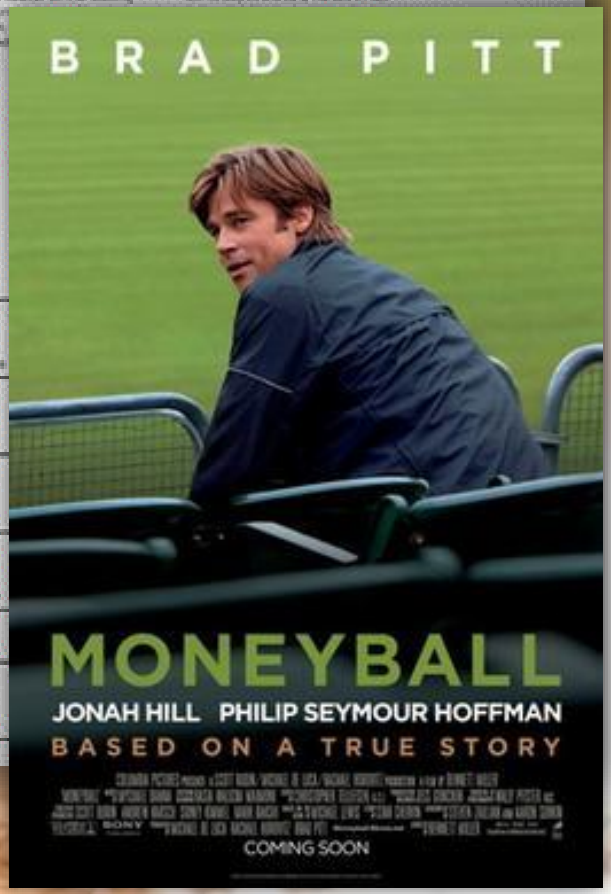


- Dengan pengetahuan tersebut, manusia dapat:
  - Melakukan estimasi dan prediksi apa yang terjadi di depan
  - Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
  - Membantu pengambilan keputusan dan pembuatan kebijakan

Winning An Unfair Game







How Do The Oakland A's Win With a Poor Team?

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

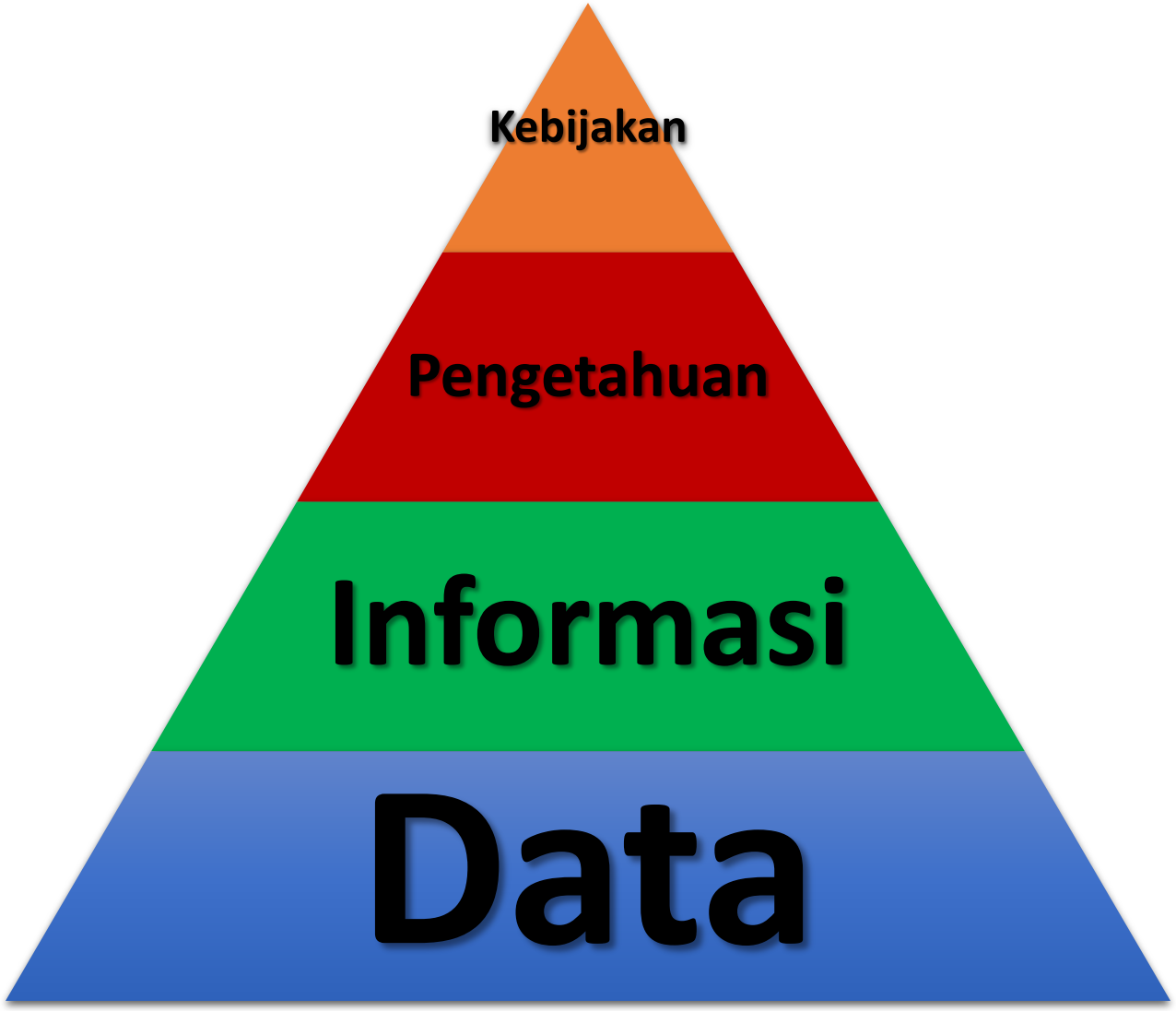
Pola Kebiasaan Kehadiran Mingguan Pegawai

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



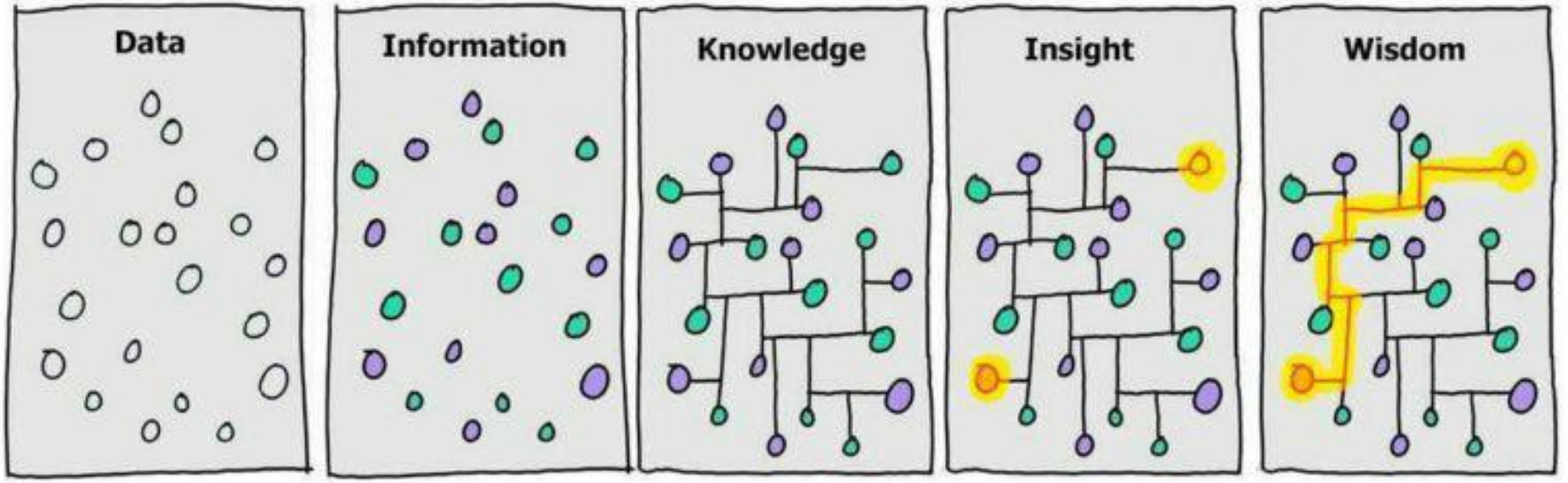
Kebijakan Penataan Jam Kerja Pegawai

Pola Kebiasaan Datang-Pulang Pegawai

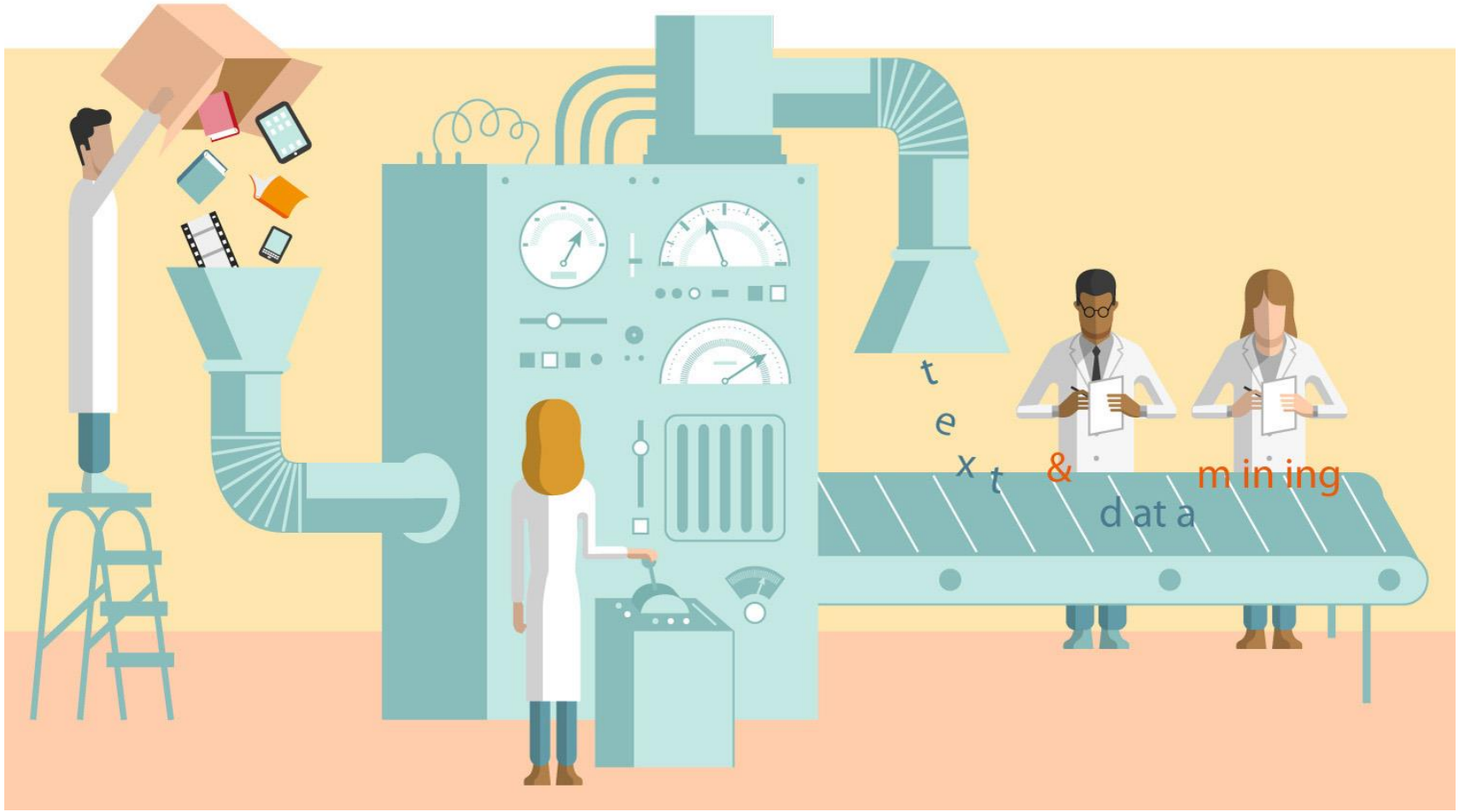
Informasi Rekap Kehadiran Pegawai

Data Absensi Pegawai

Data - Informasi – Pengetahuan - Kebijakan



Apa itu Data Mining?



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

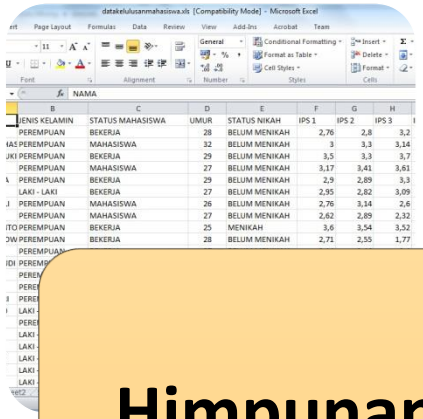
Apa itu Data Mining?

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

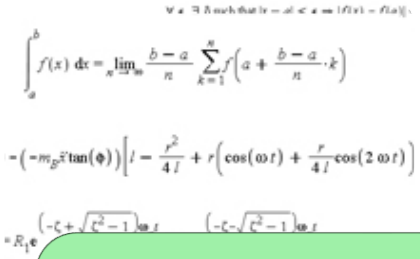
- Ekstraksi dari data ke pengetahuan:
  1. Data: fakta yang terekam dan tidak membawa arti
  2. Informasi: Rekap, rangkuman, penjelasan dan statistik dari data
  3. Pengetahuan: pola, rumus, aturan atau model yang muncul dari data

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

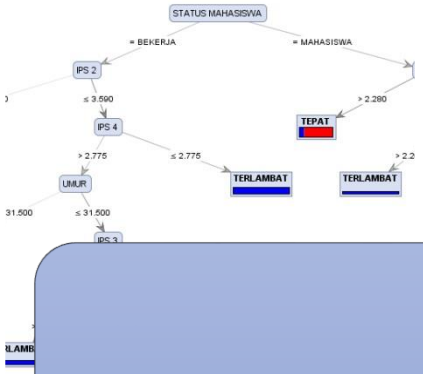
Konsep Proses Data Mining



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)*

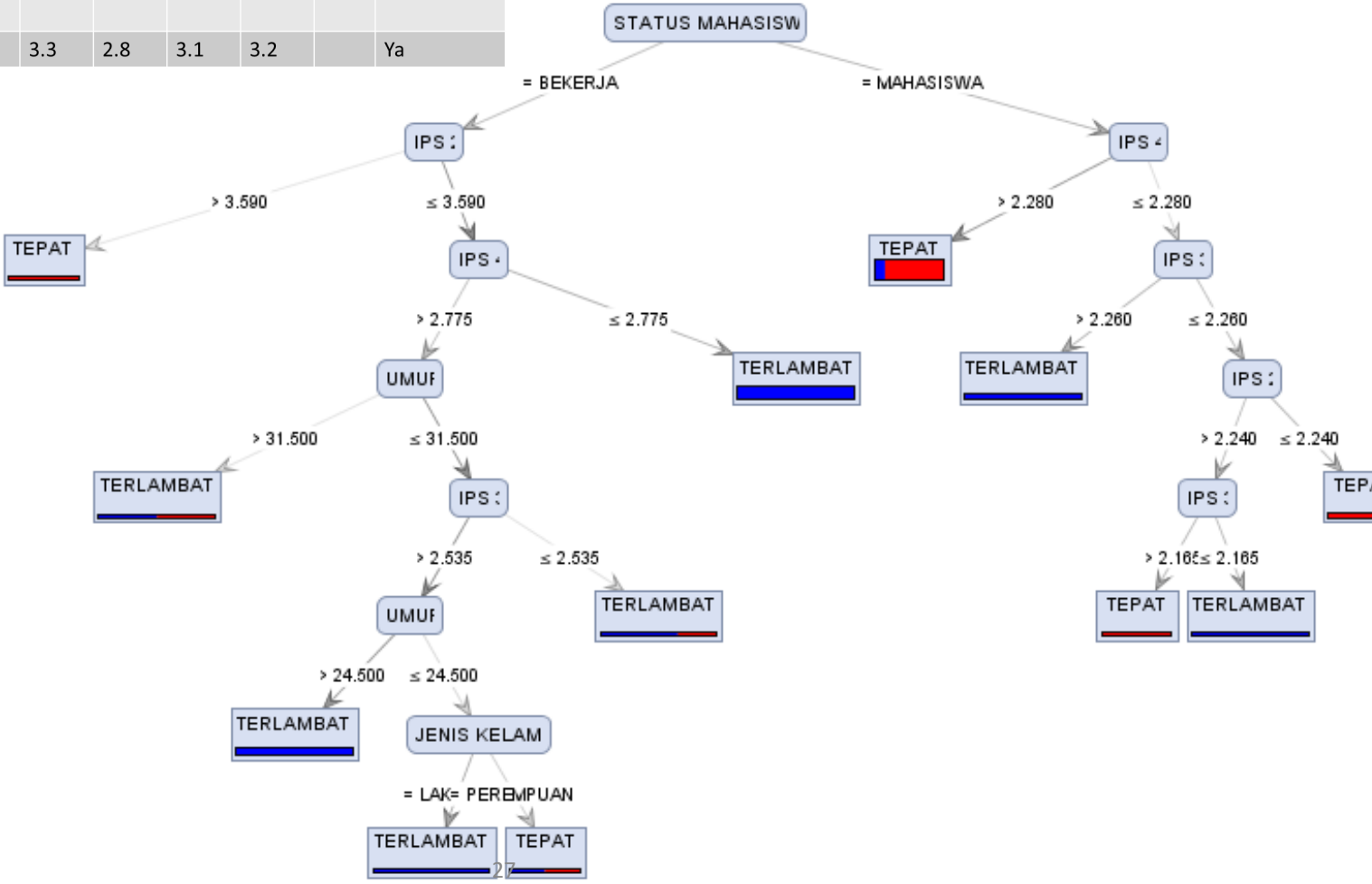
Contoh Data di Kampus

- 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



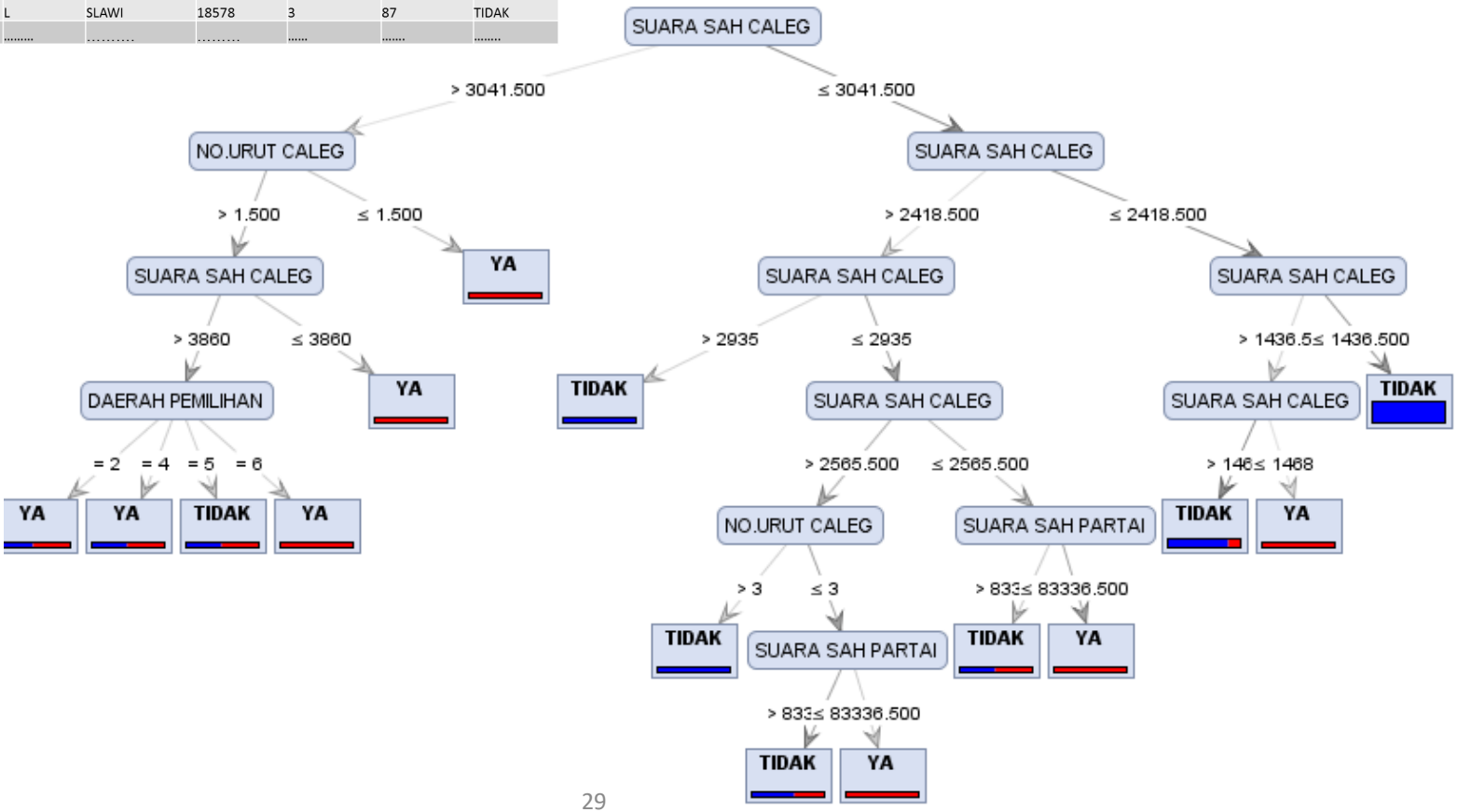
Contoh Data di Komisi Pemilihan Umum

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

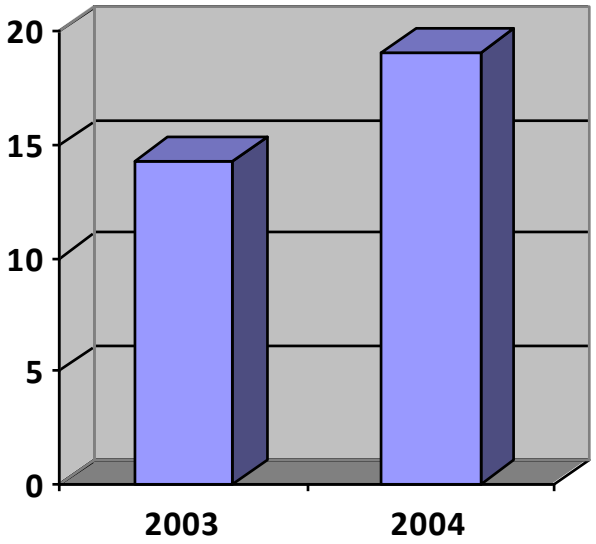
NAMA PARTAI POLITIK	NAMA CALON LEGESLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH CALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO,BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO,SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATI,SH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	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
……	……	……	……	……	……	……	……

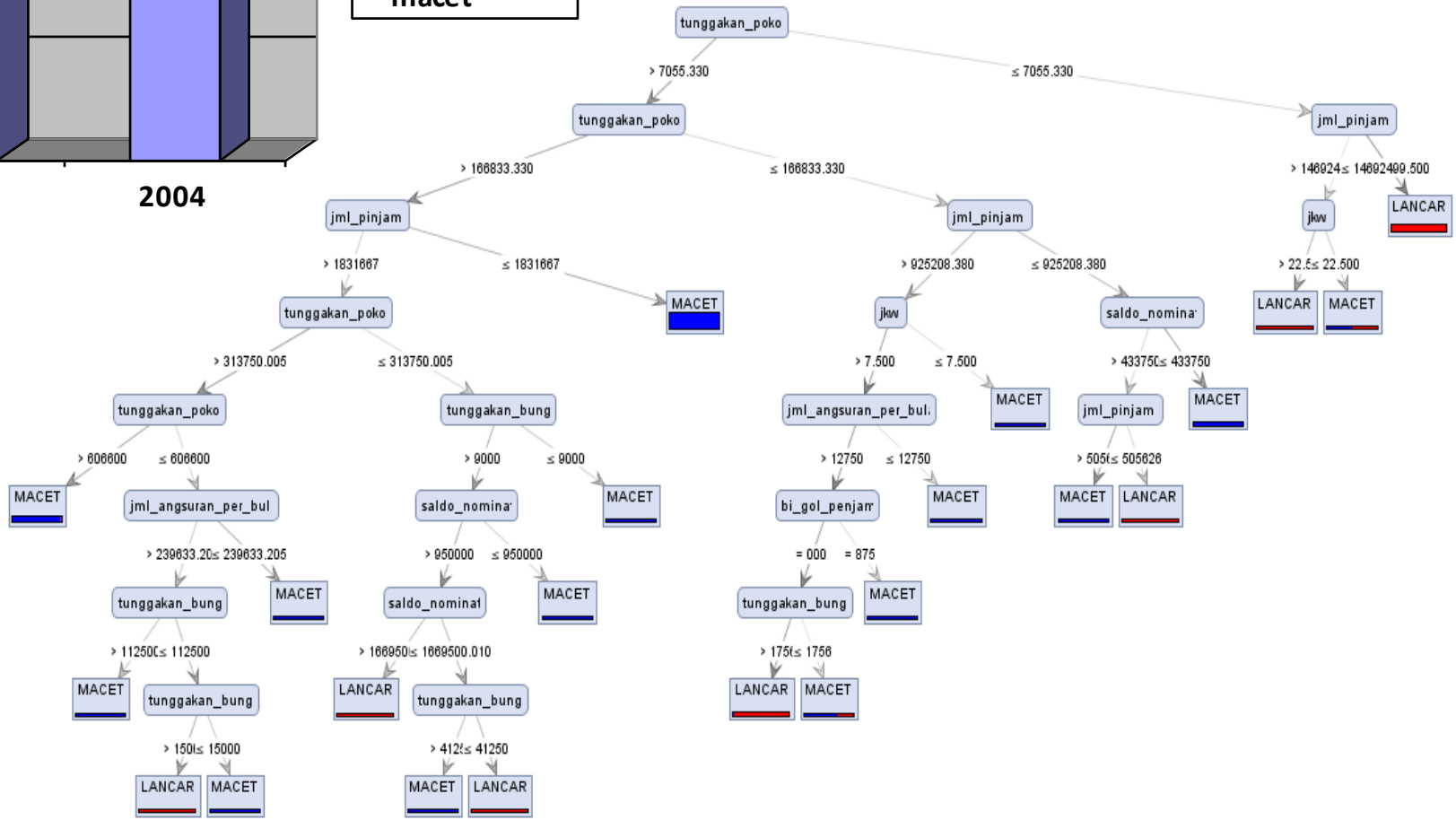
Prediksi Calon Legislatif DKI Jakarta

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
……	……	……	……	……	……	……	……

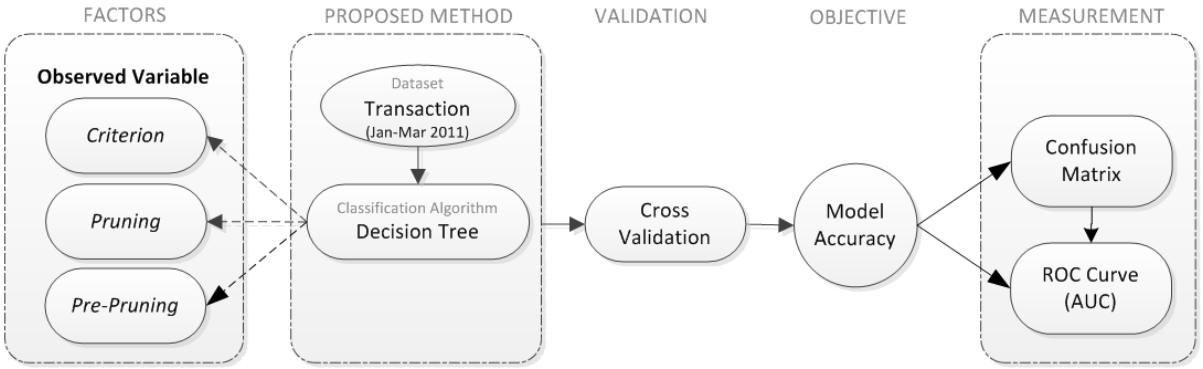


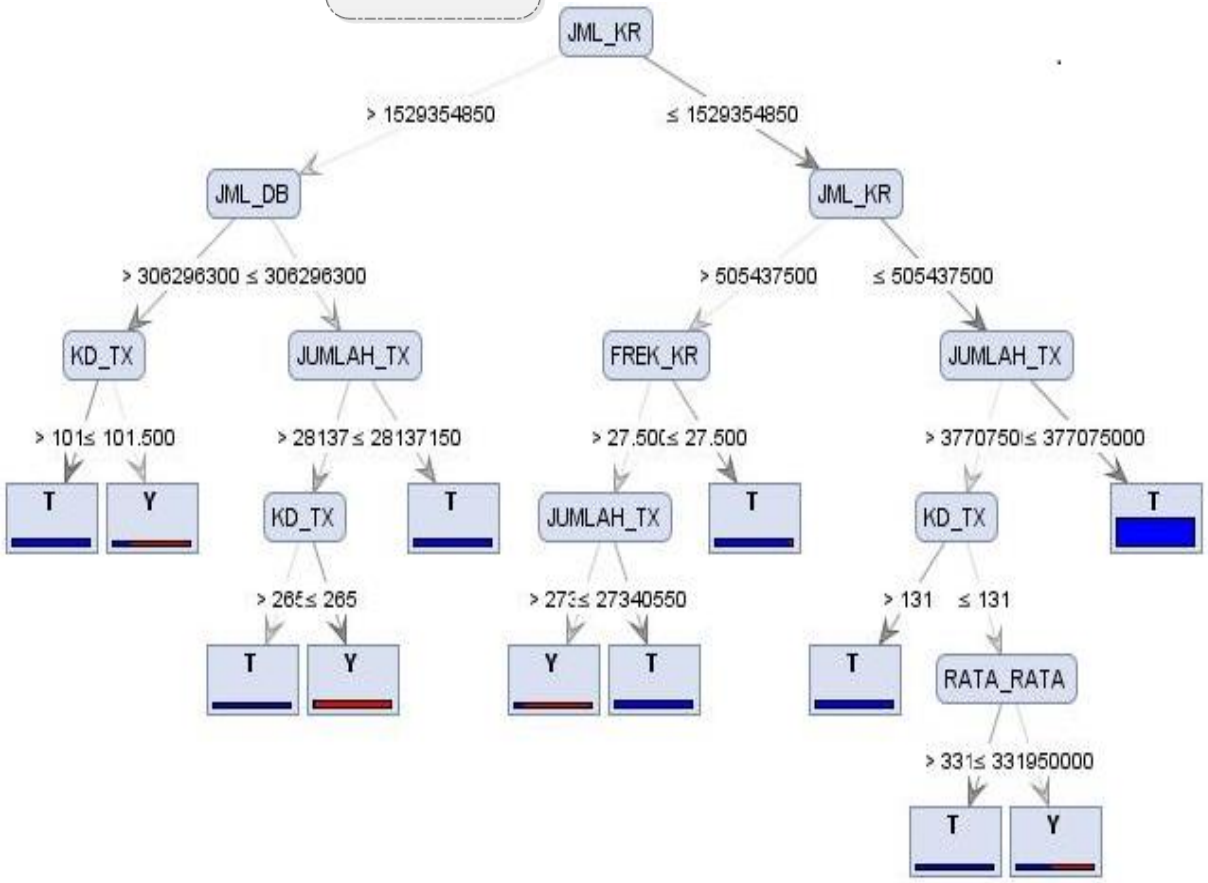
Penentuan Kelayakan Kredit





Deteksi Pencucian Uang

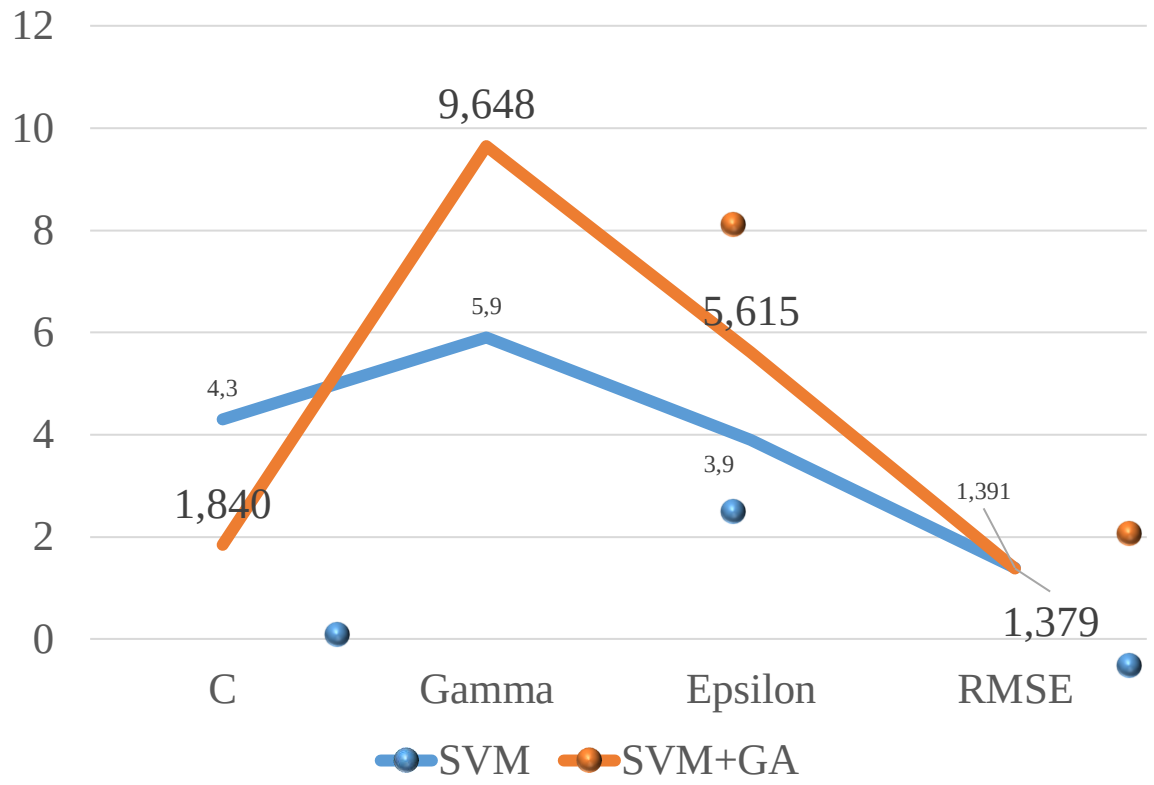




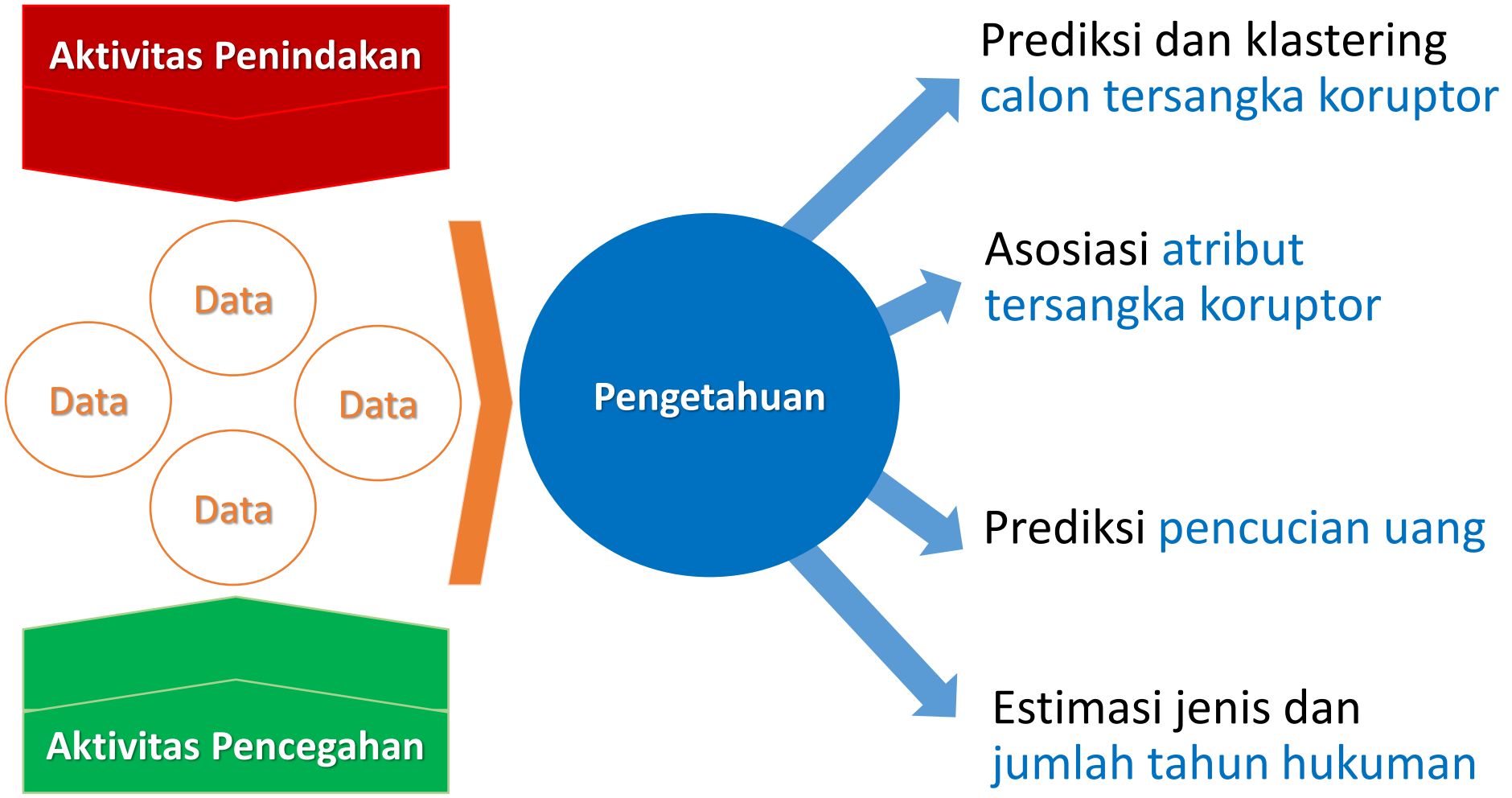
Prediksi Kebakaran Hutan

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

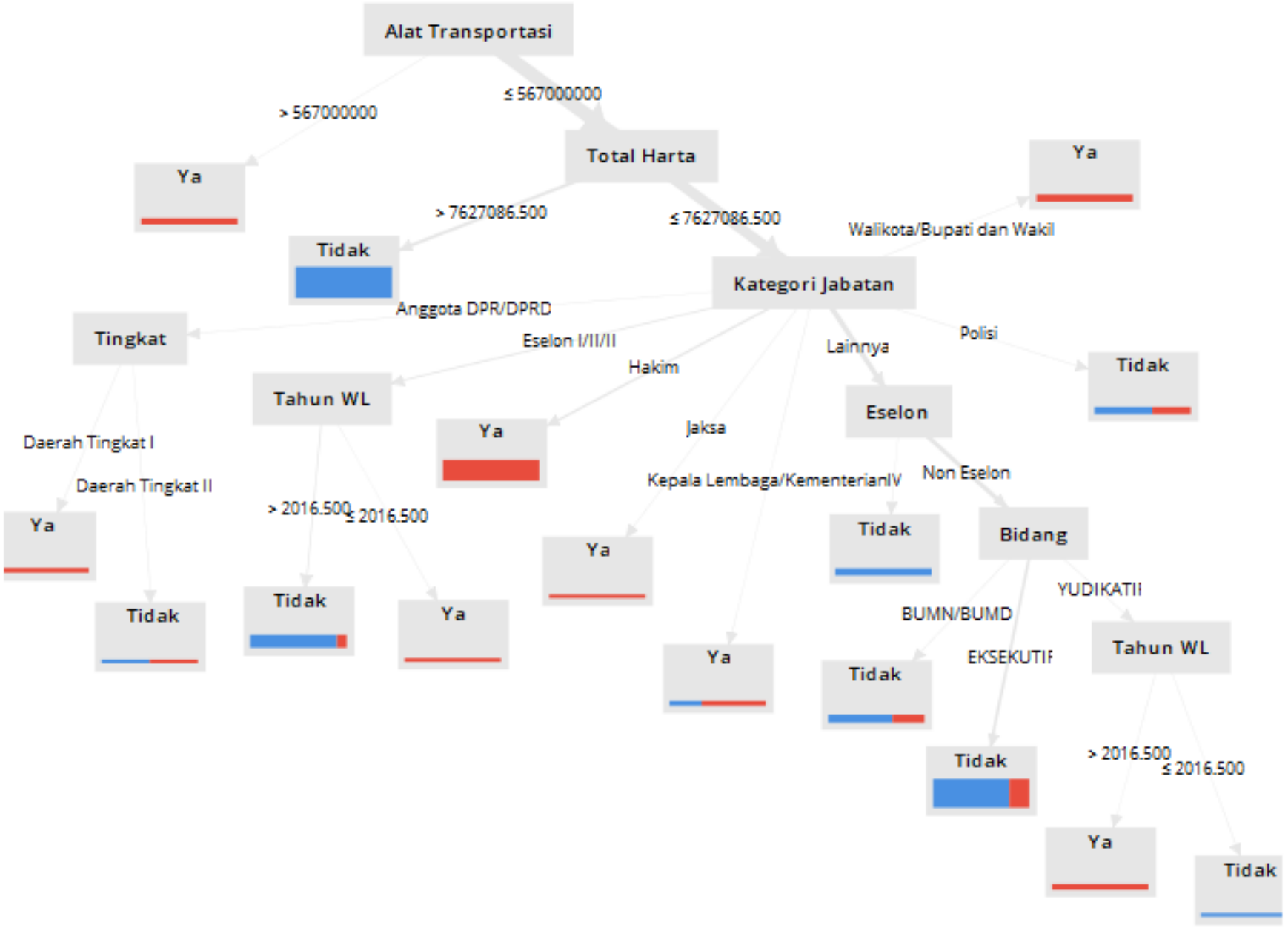
	SVM	SVM+GA
C	4.3	1,840
Gamma (γ)	5.9	9,648
Epsilon (ε)	3.9	5,615
RMSE	1.391	1.379



Profiling dan Prediksi Koruptor

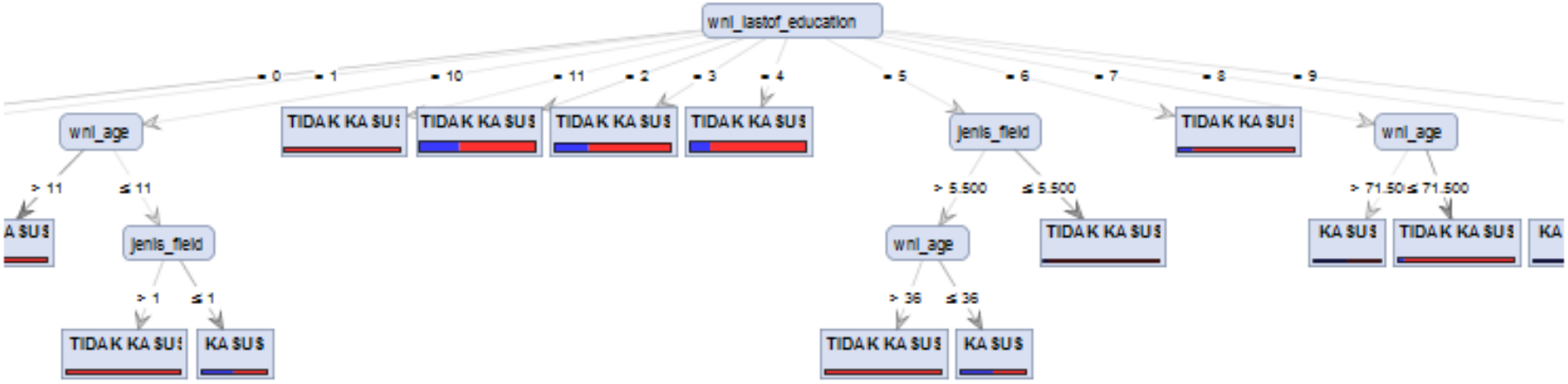


Pola Profil Tersangka Koruptor

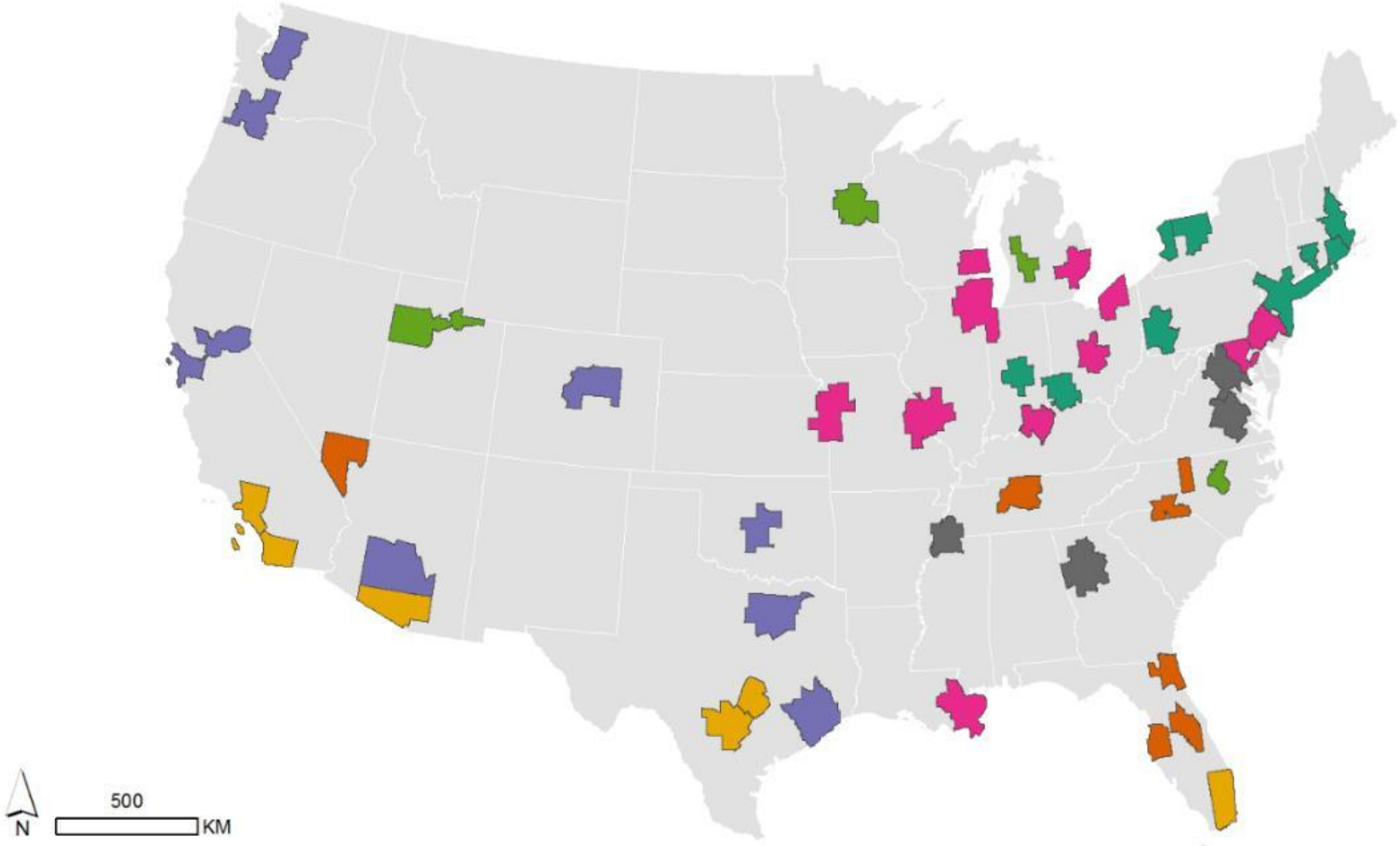


Profiling dan Deteksi Kasus TKI

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

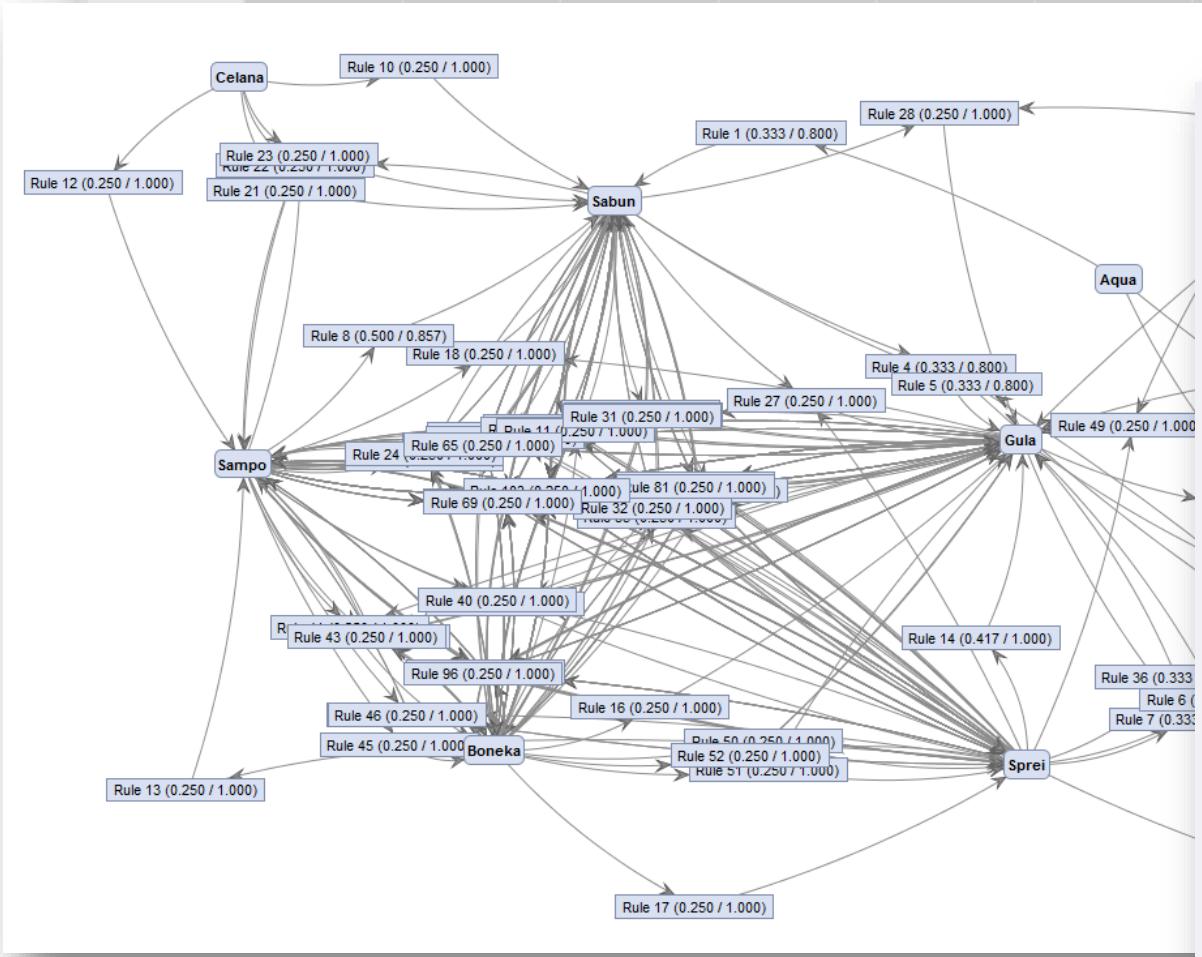


Group 1: Stability		Group 2: New South		Group 3: Hispanic Destinations	Group 4: Emerging Multiethnic		Group 5: Persistent Black Poverty		Group 6: Immigrant\ Educated	Group 7:New Old South
Boston	New York	Charlotte	Tampa	Austin	Dallas	Portland	Baltimore	Louisville	Grand Rapids	Atlanta
Buffalo	Pittsburgh	Greensboro	Las Vegas	Miami	Denver	Sacramento	Chicago	Milwaukee	Minneapolis	Memphis
Cincinnati	Providence	Jacksonville	Orlando	San Antonio	Houston	Seattle	Cleveland	New Orleans	Raleigh	Richmond
Hartford	Rochester	Nashville		Tuscon	Oklahoma City		Columbus	Philadelphia	Salt Lake City	Washington
Indianapolis				San Diego	Phoenix		Detroit	St. Louis		
				Los Angeles	San Francisco		Kansas City			

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)

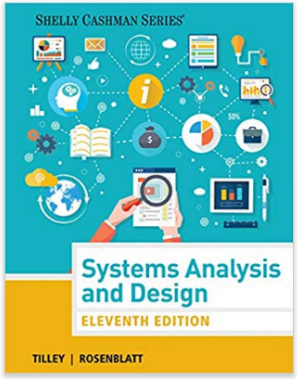
Pola Aturan Asosiasi di Amazon.com

Systems Analysis and Design (Shelly Cashman Series) 11th Edition

by Scott Tilley (Author), Harry J. Rosenblatt (Author)

28 customer reviews

Look inside



ISBN-13: 978-1305494602
ISBN-10: 9781305494602
Why is ISBN important?

Sell yours for a Gift Card
We'll buy it for up to **$7.49**
Learn More
Trade in now

Have one to sell? Sell on Amazon

Add to List

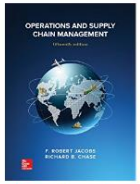
Share

Hardcover $19.04 - $110.00 | Other Sellers from $19.04

- Rent $19.04
- Buy used $55.60

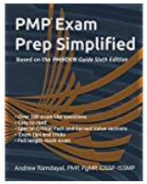
Customers who bought this item also bought

Principles of Marketing (17th Edition)
› Philip T. Kotler
14
Hardcover
$199.99



Operations and Supply Chain Management
› F. Robert Jacobs
201
Hardcover
$154.60

Sponsored products related to this item

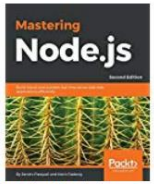


PMP Exam Prep Simplified: Based on PMBOK® Guide Sixth Edition
Andrew Ramdayal
28
Paperback
$36.09 prime

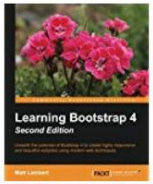
React Design Patterns and Best Practices: Build easy to scale modular applications ...
Michele Bertoli
5
Paperback
$41.27 prime

Node.js Design Patterns - Second Edition: Master best practices to build modular an...
Mario Casciaro
Get the best out of Node.js by mastering its most powerful components and patterns to create modular and scalable applications with ease
32
Paperback

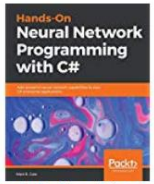
Vue.js 2 Design Patterns and Best Practices: Build enterprise-ready, modular Vue.js...
Paul Halliday
4
Paperback
$44.99 prime



Mastering Node.js - Second Edition: Build robust and scalable real-time server-side...
Sandro Pasquali
Expert techniques for building fast servers and scalable, real-time network applications with minimal effort, rewritten for Node.js 8 and Node.js 9.
7



Learning Bootstrap 4 - Second Edition
Matt Lambert
4
Paperback
$37.71 prime



Hands-On Neural Network Programming with C#: Add powerful neural network capabiliti...
Matt R. Cole
Just released
Paperback
$39.99 prime

From Stupid Apps to Smart Apps

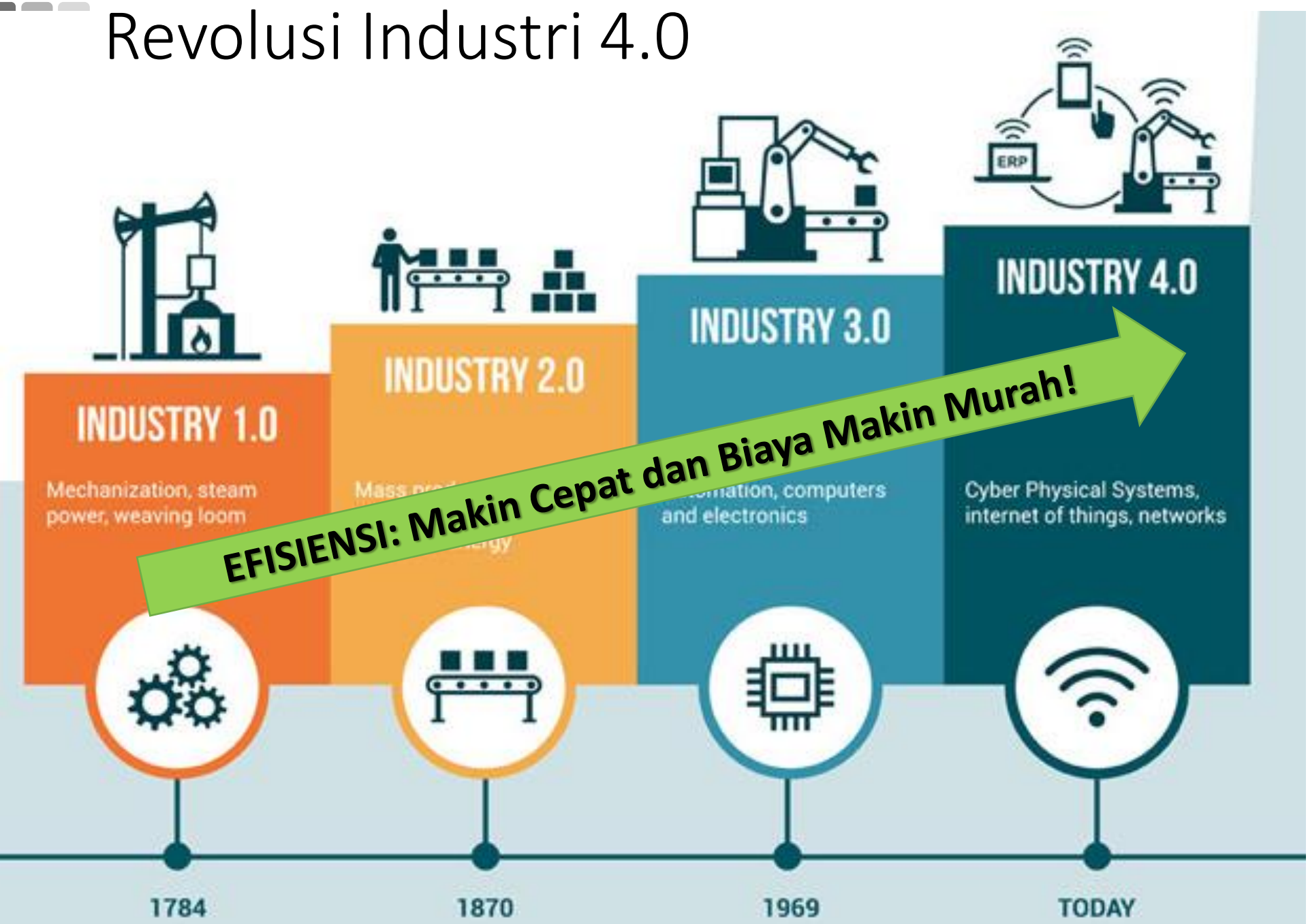
Stupid Applications

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

Smart Applications

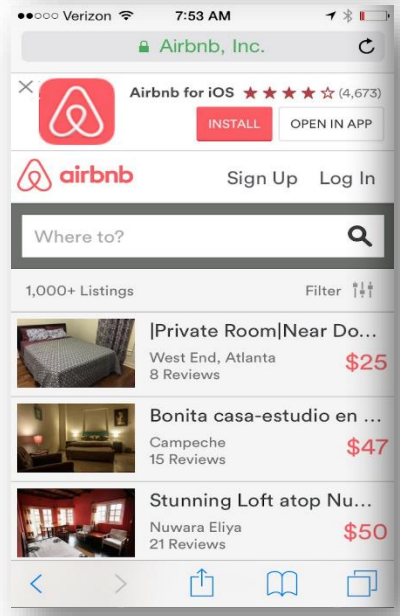
- Sistem Prediksi Kelulusan Mahasiswa
- Sistem Prediksi Hasil Pemilu
- Sistem Prediksi Koruptor
- Sistem Penentu Kelayakan Kredit

Revolusi Industri 4.0

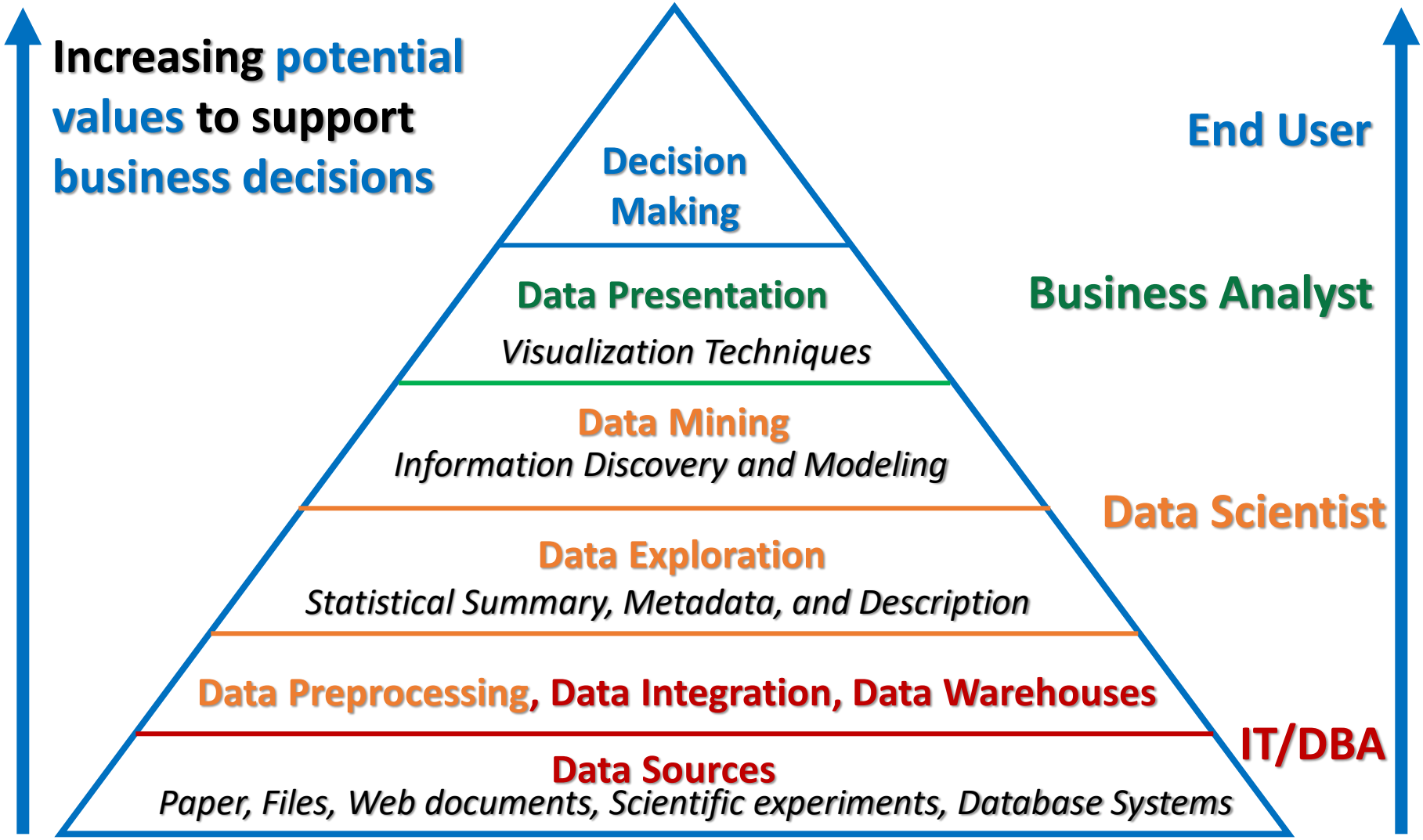


Perusahaan Pengolah Pengetahuan

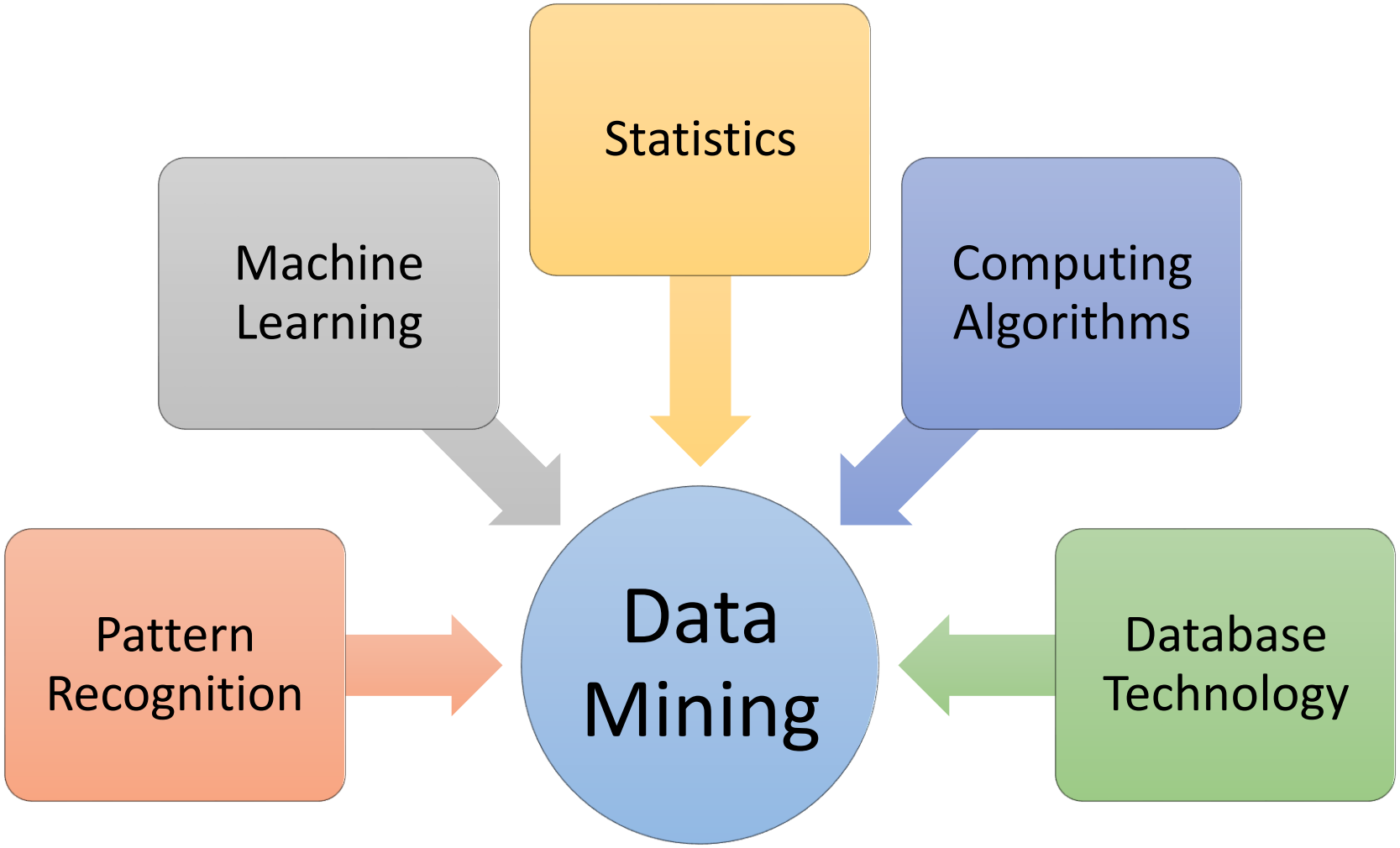
- Uber - the world's largest taxi company, owns no vehicles
- Google - world's largest media/advertising company, creates no content
- Alibaba - the most valuable retailer, has no inventory
- Airbnb - the world's largest accommodation provider, owns no real estate
- Gojek - perusahaan angkutan umum,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



Data Mining Tasks and Roles



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

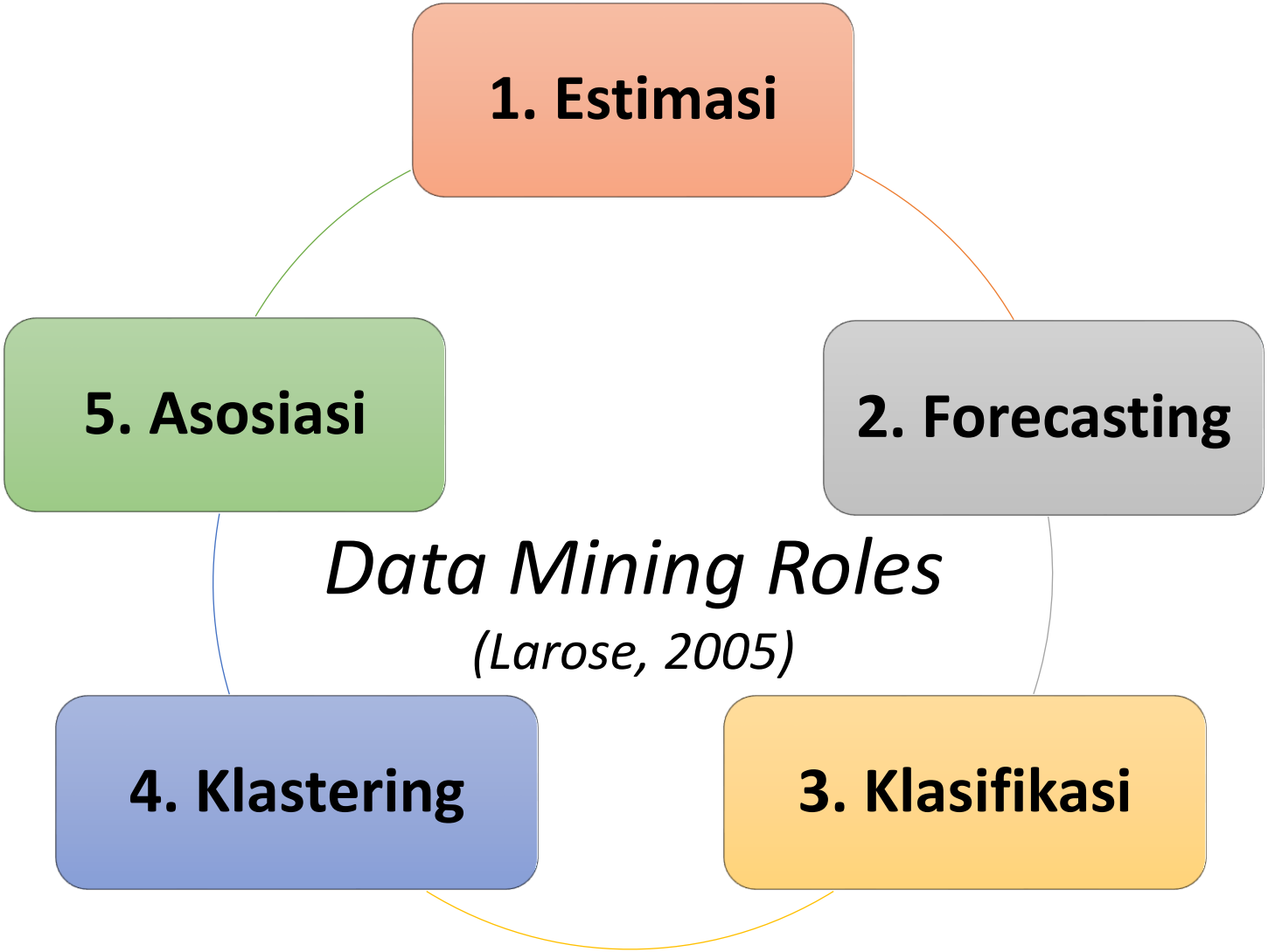
Latihan

1. Jelaskan dengan kalimat sendiri apa yang dimaksud dengan data mining?

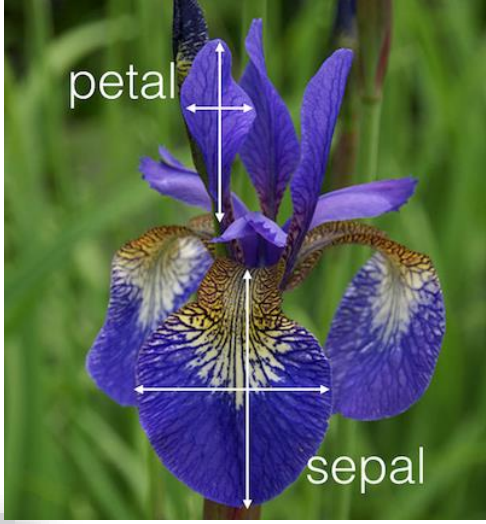
2. Sebutkan konsep alur proses data mining!

1.2 Peran Utama dan Metode Data Mining

Peran Utama Data Mining



Dataset (Himpunan Data)



Attribute/Feature/Dimension

Class/Label/Target

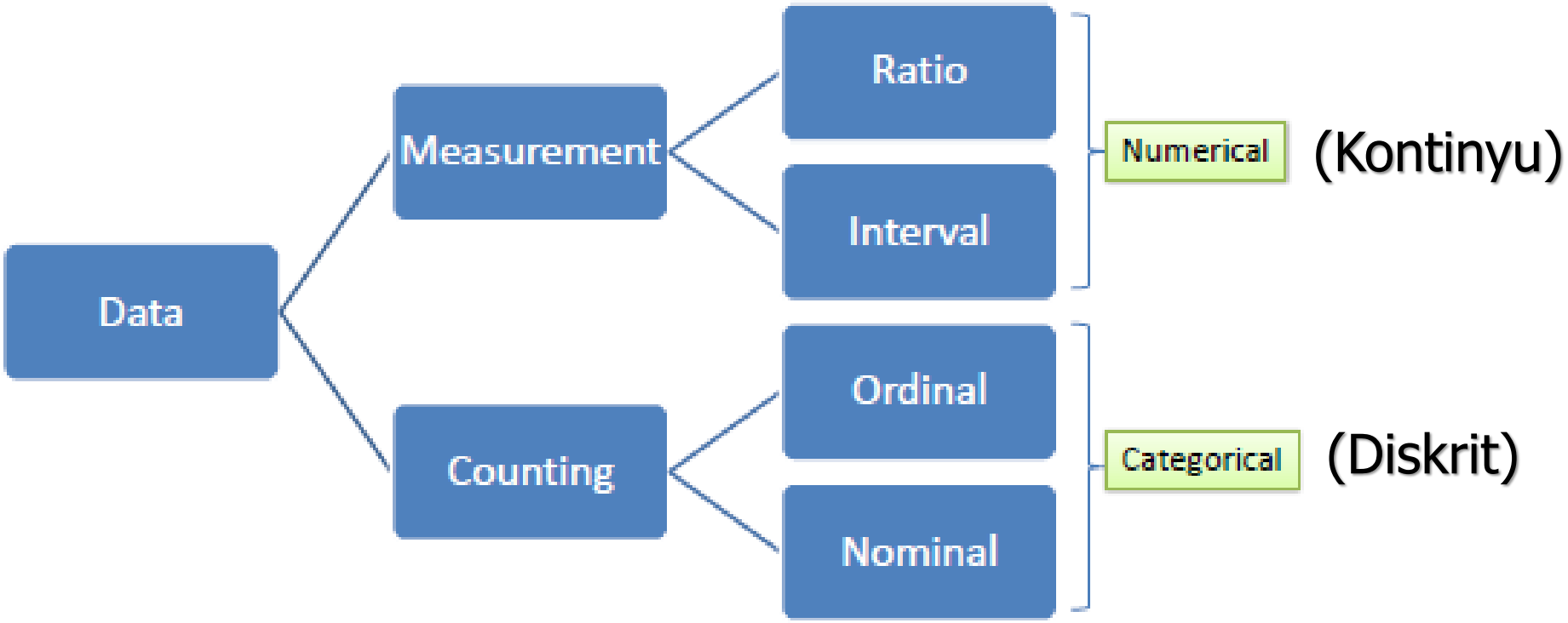
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris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

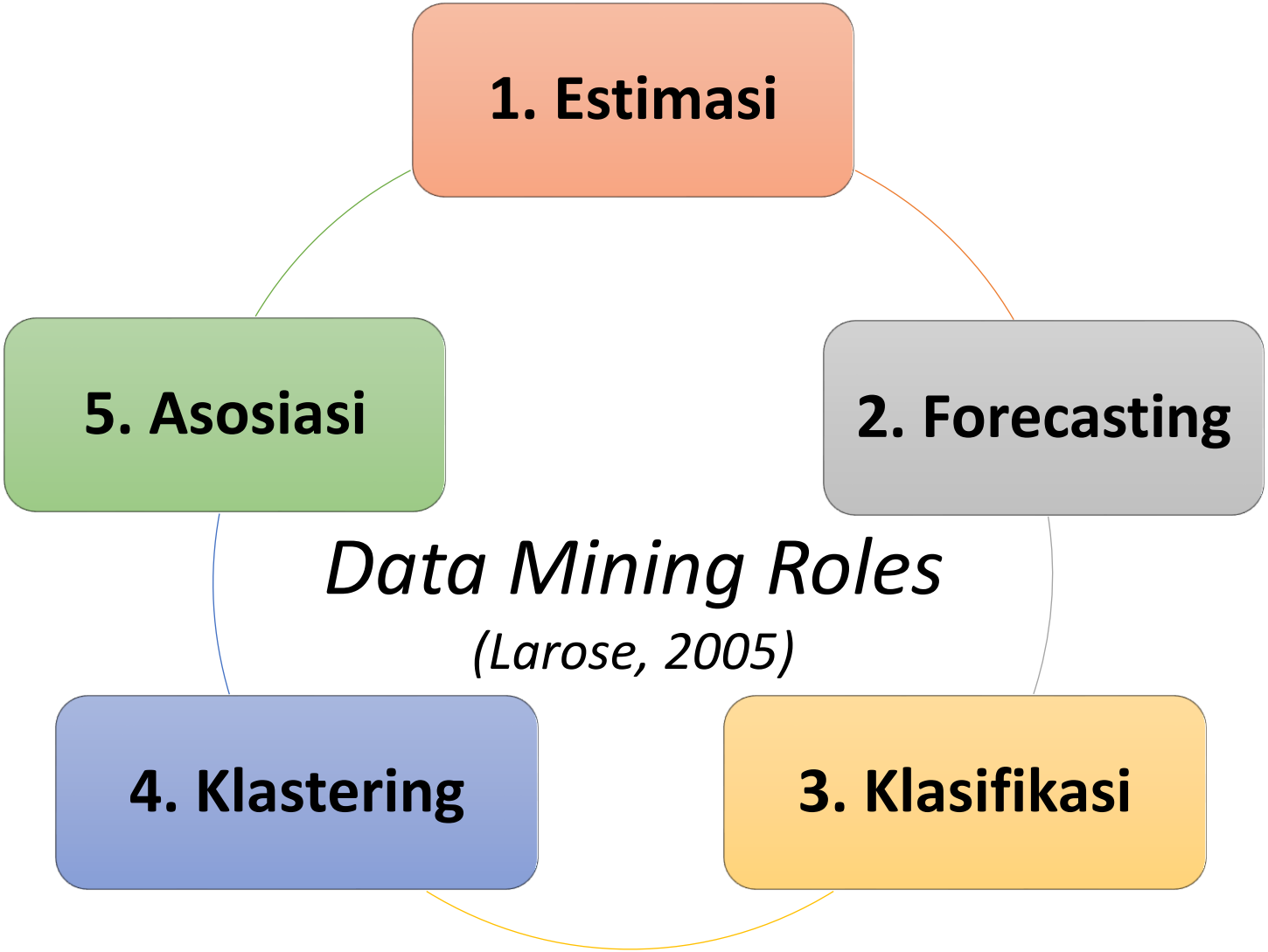
Numerik

Tipe Data


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

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

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*)

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

Pengetahuan

Contoh: Estimasi Performansi CPU

- **Example**: 209 different computer configurations

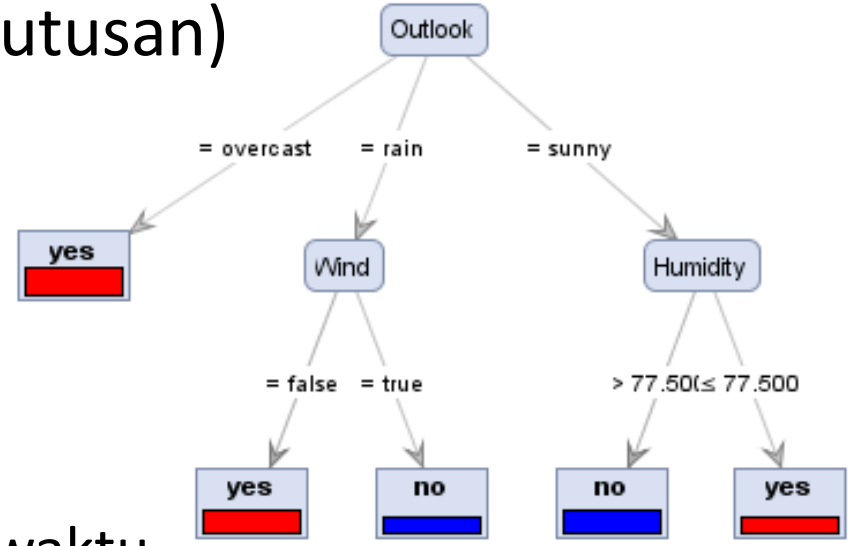
	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- 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

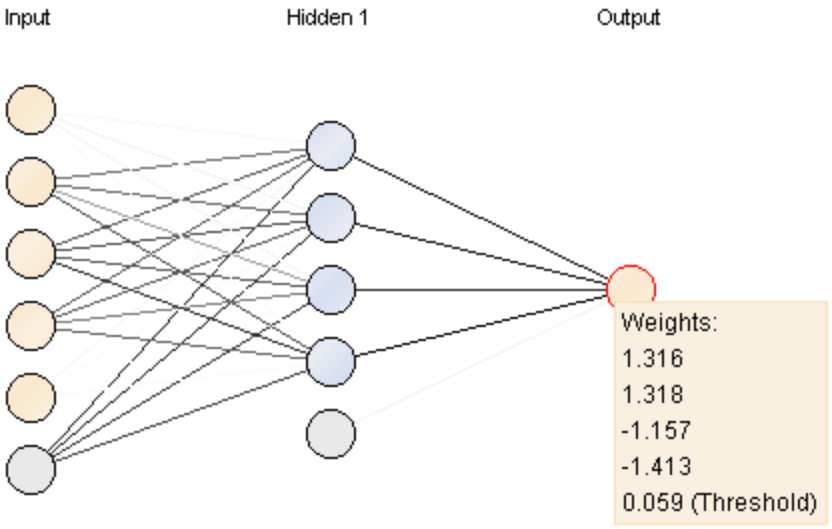
Label Time Series

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)

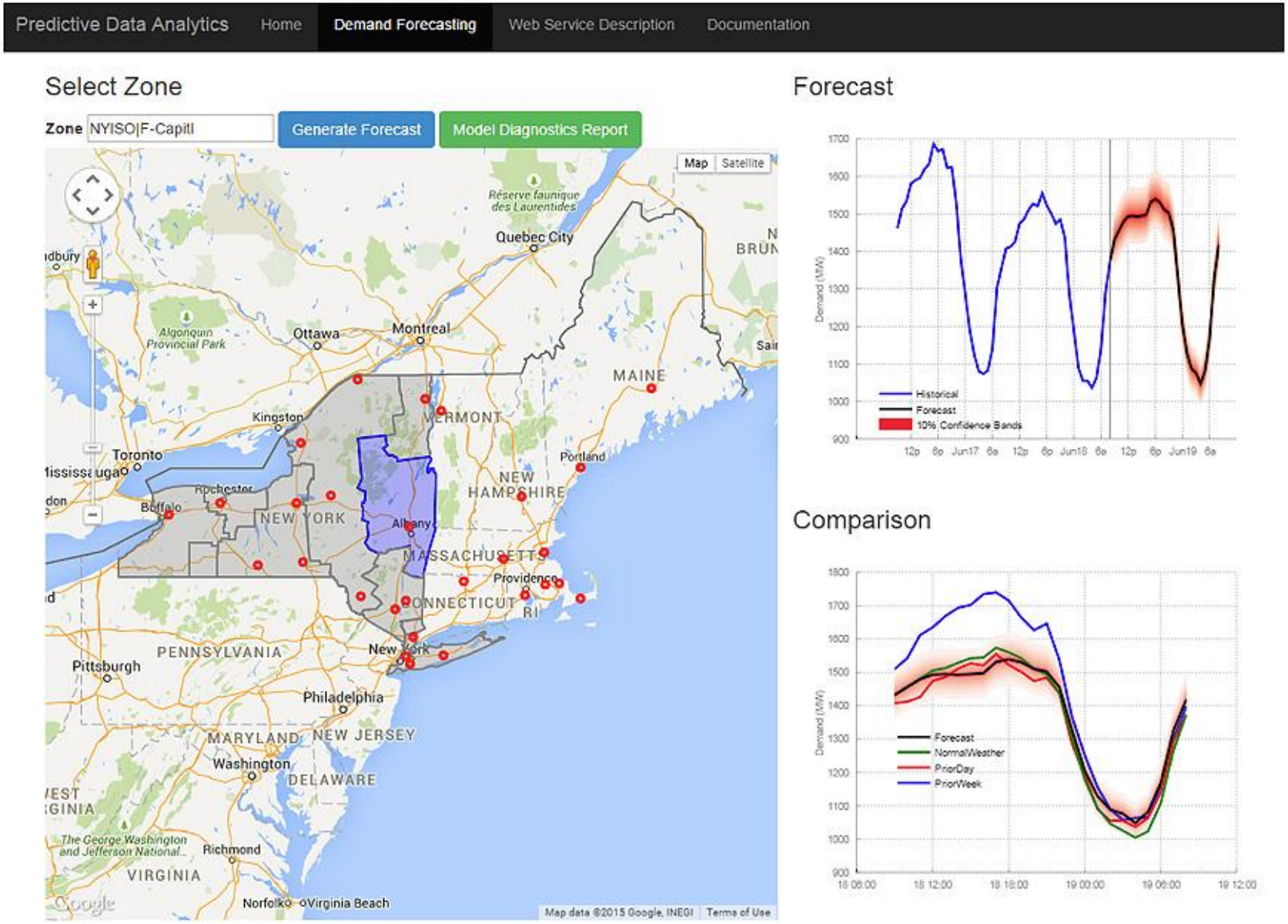
Pembelajaran dengan
Metode Forecasting (*Neural Network*)

Pengetahuan berupa Rumus Neural Network

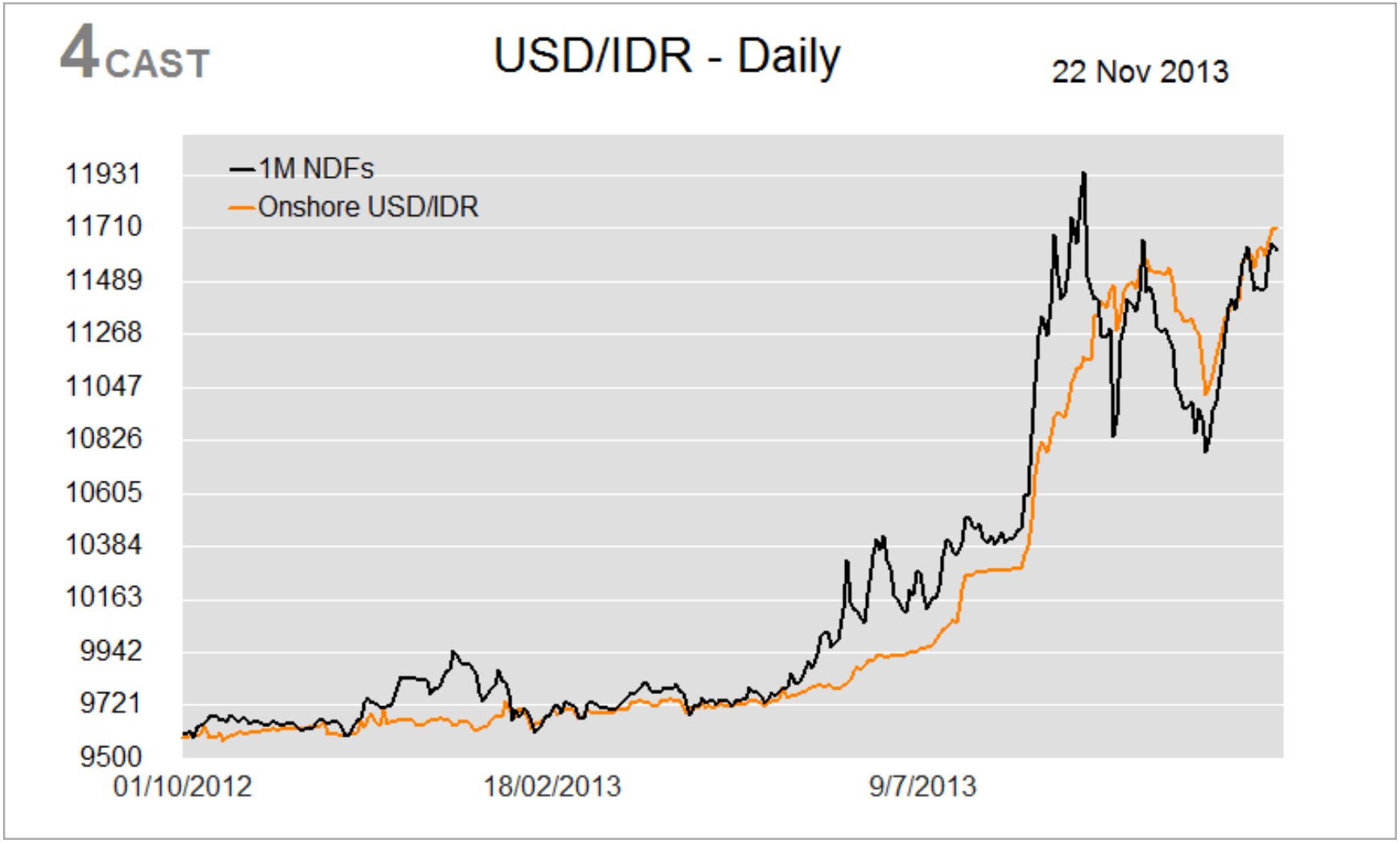


Prediction Plot

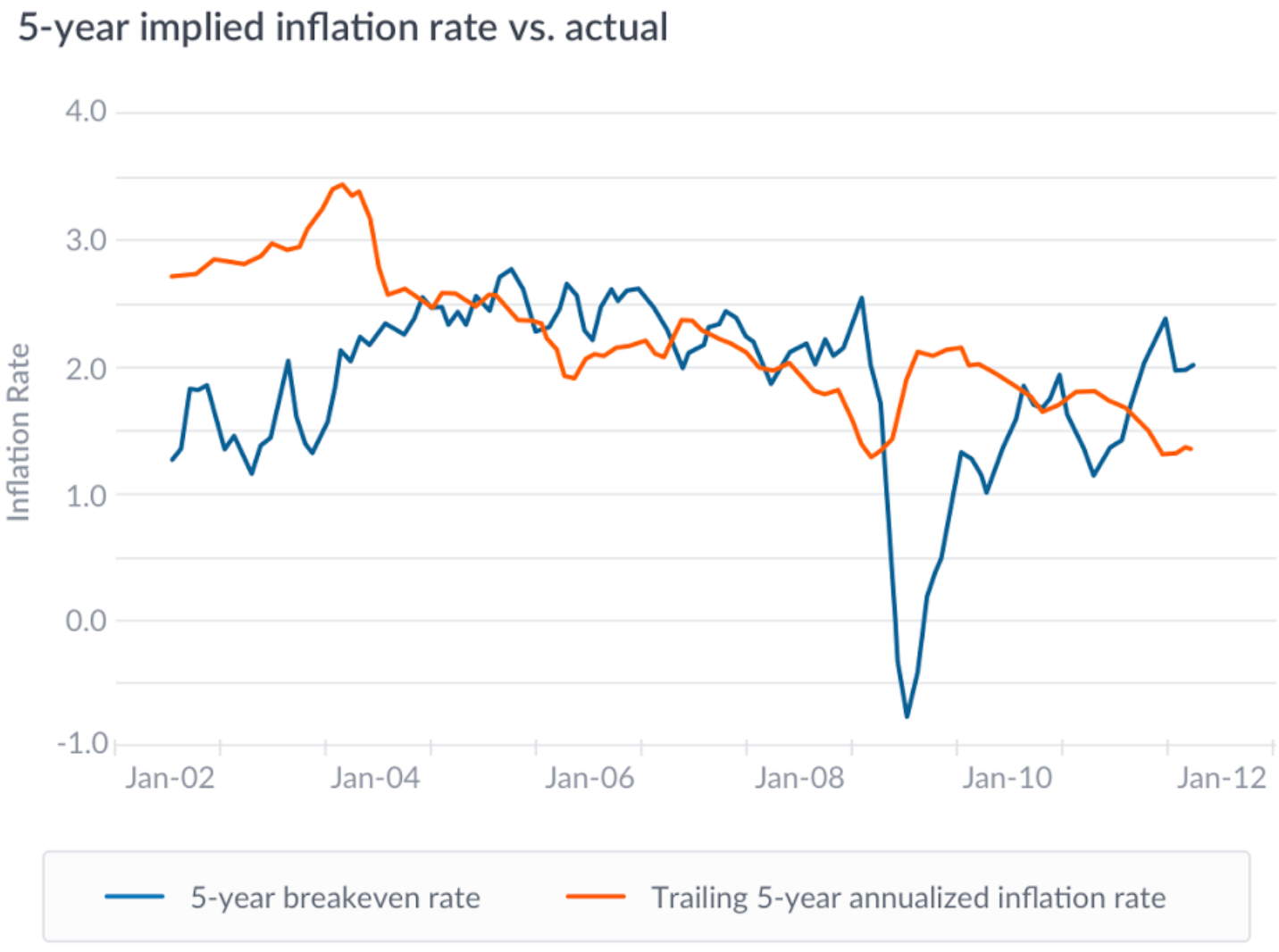
Forecasting Cuaca



Exchange Rate Forecasting



Inflation Rate Forecasting



*Source: Bloomberg

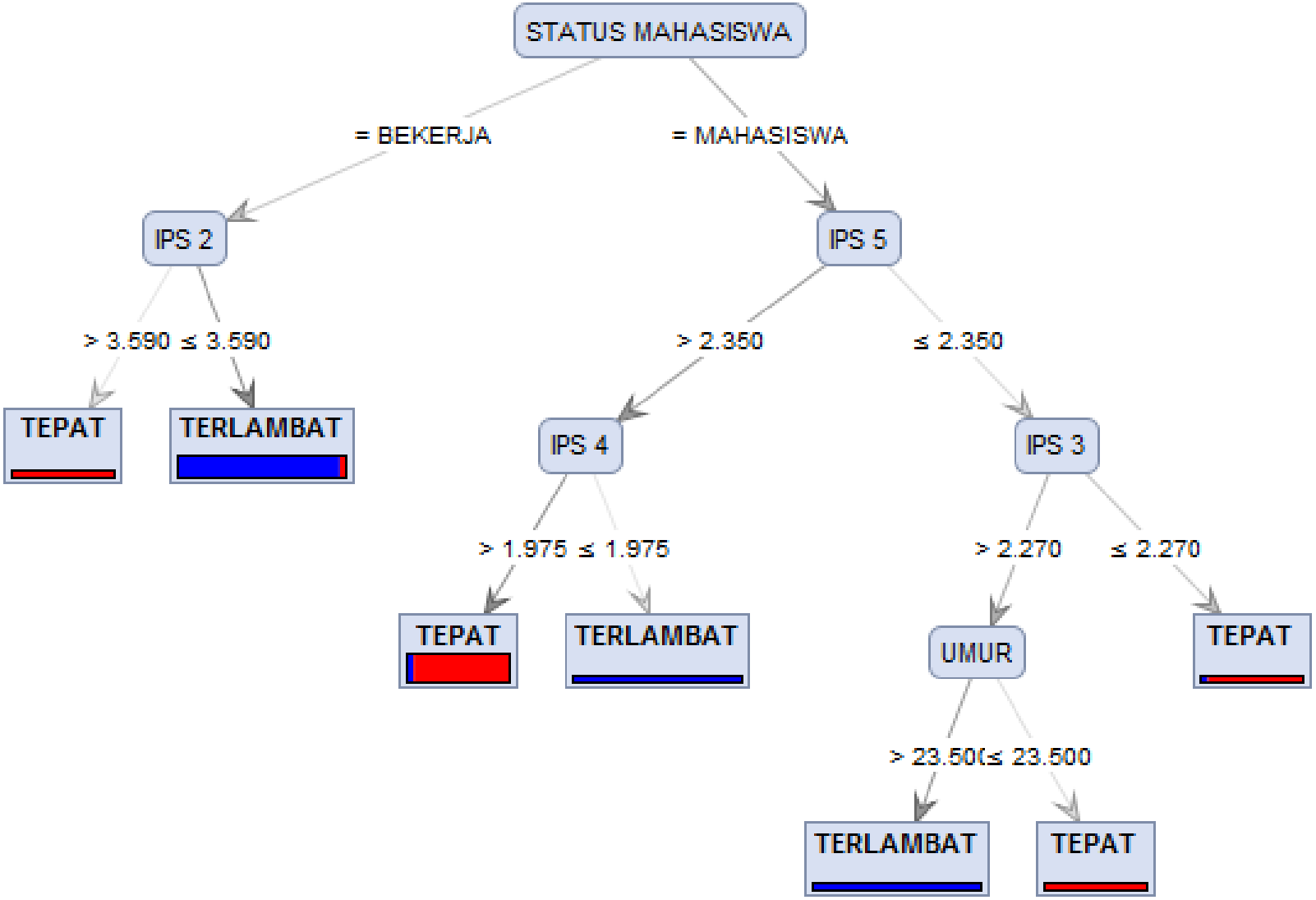
3. Klasifikasi Kelulusan Mahasiswa

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**:

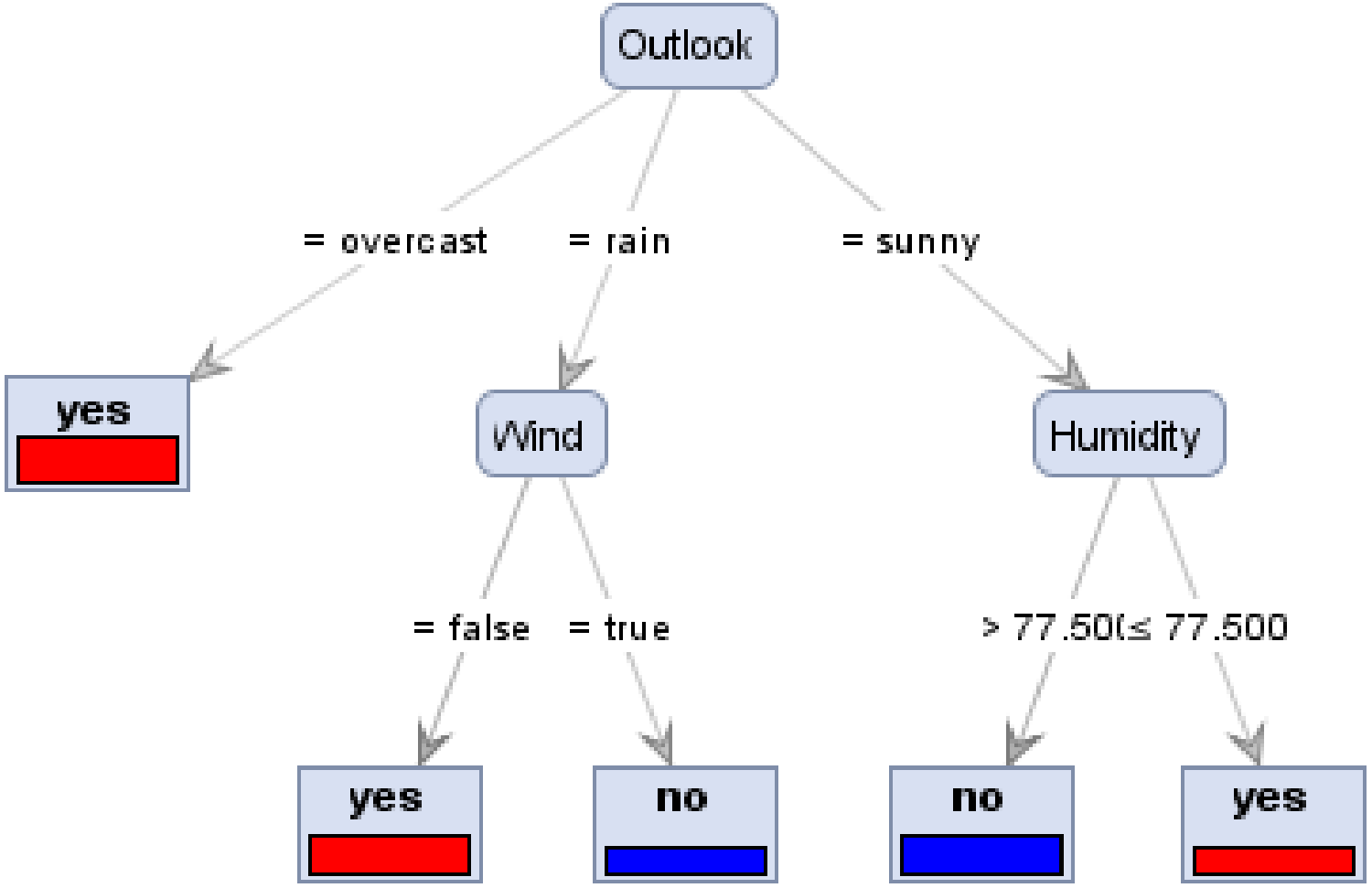
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false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



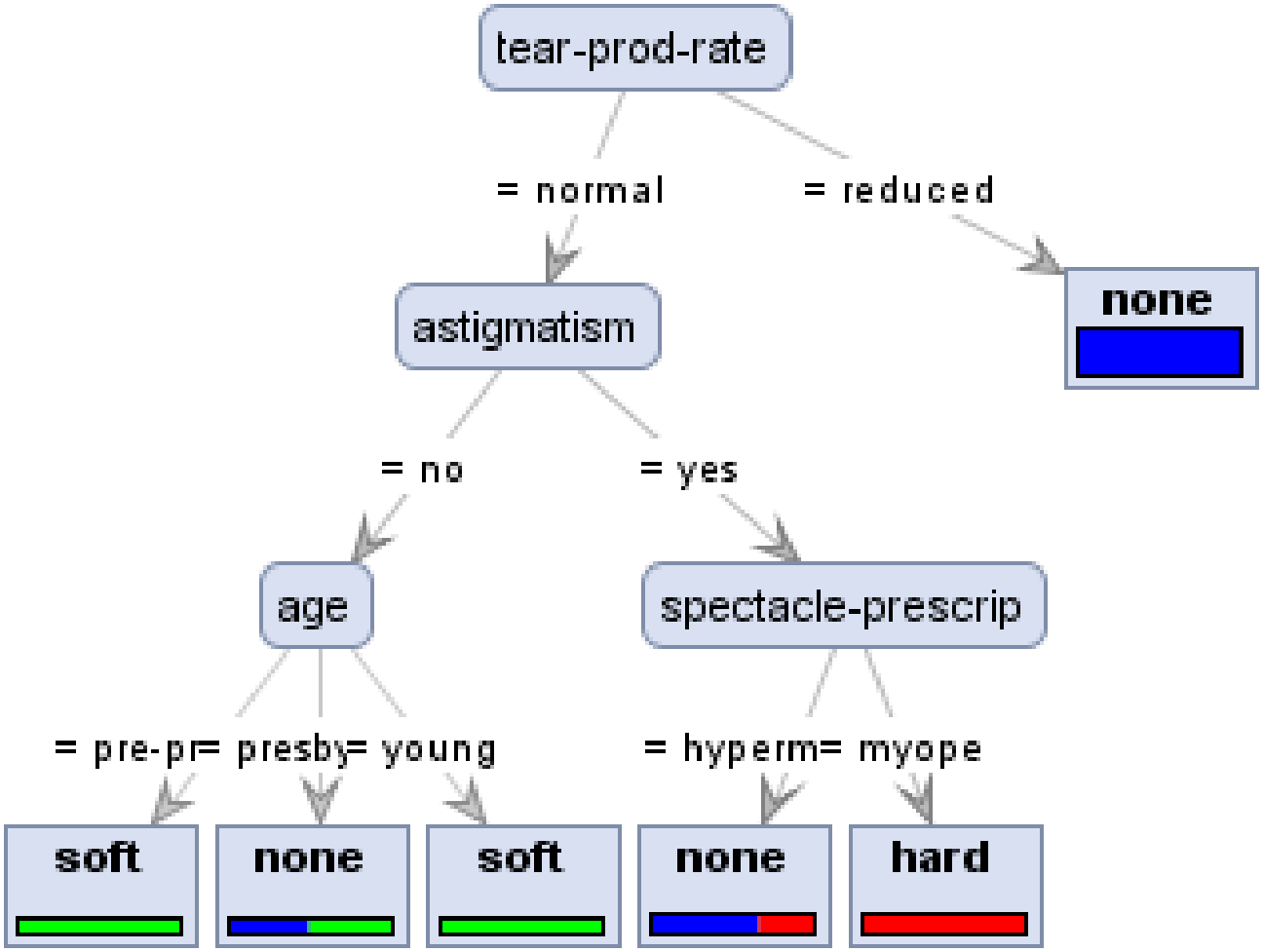
Contoh: Rekomendasi Contact Lens

- Input:

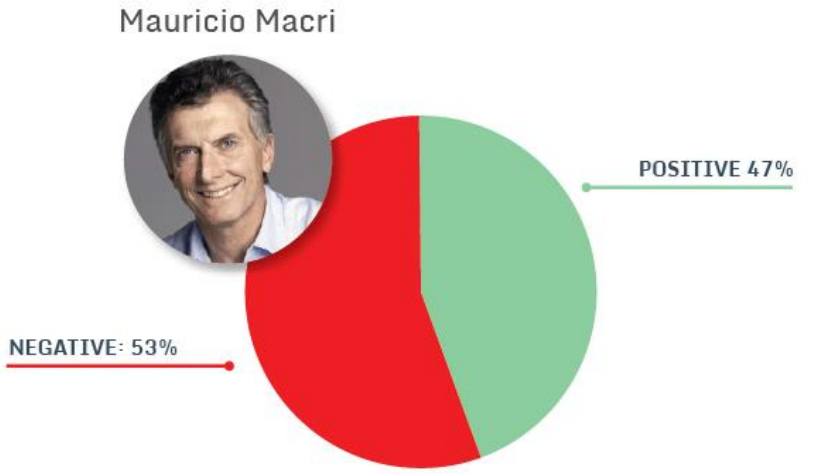
Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young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

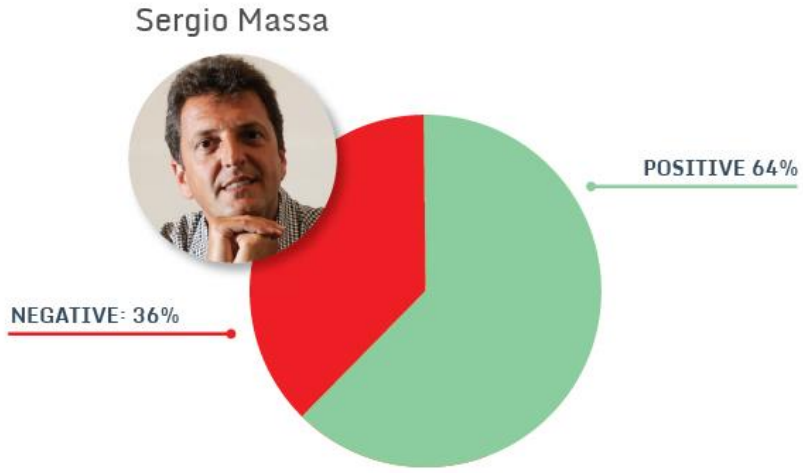
Contoh: Rekomendasi Contact Lens

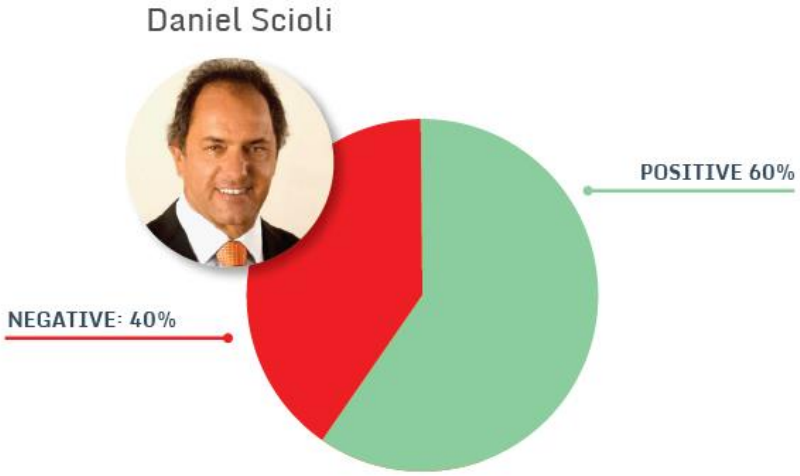
- Output/Model (Tree):



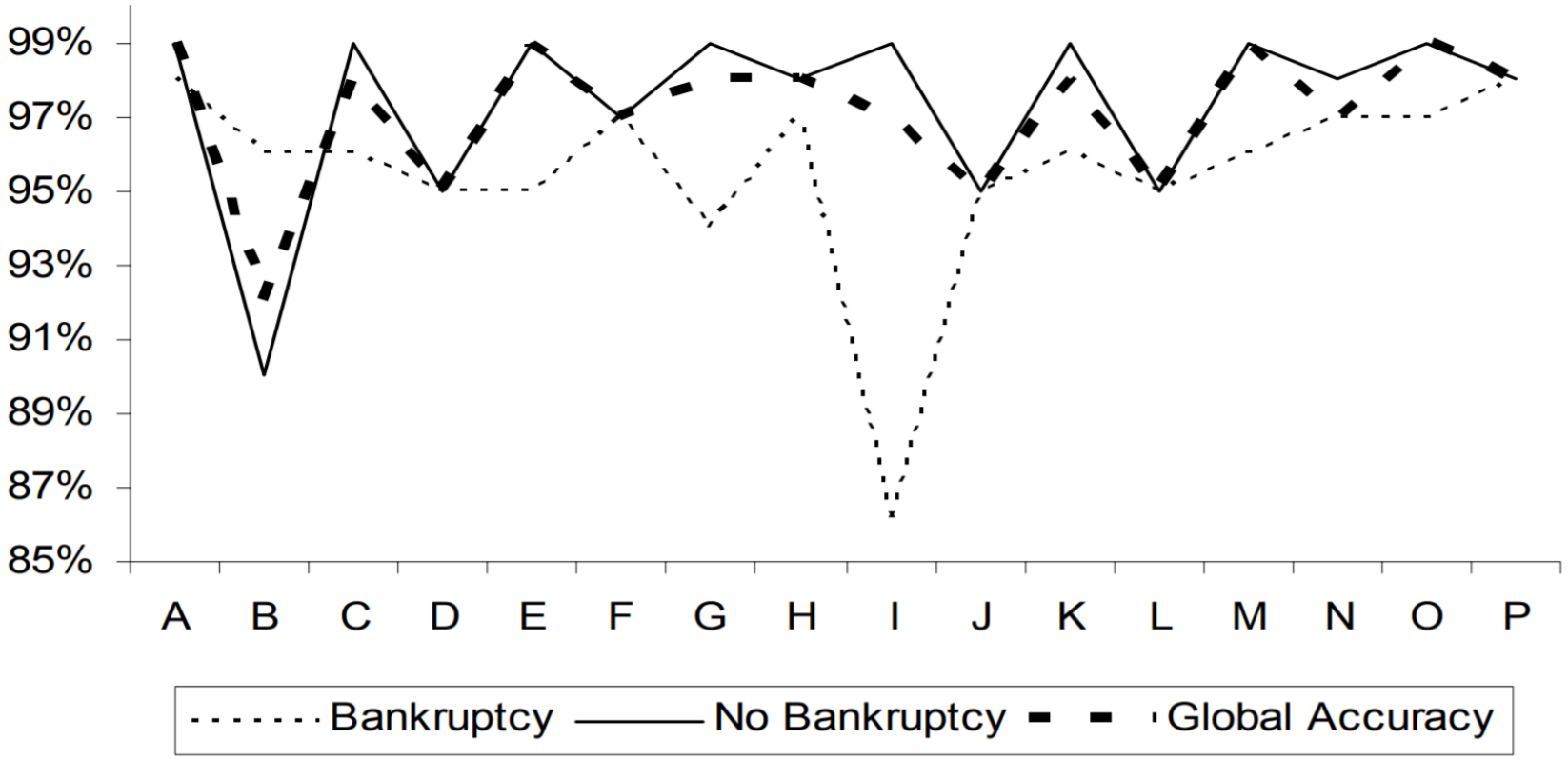
Klasifikasi Sentimen Analisis







Bankruptcy Prediction



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

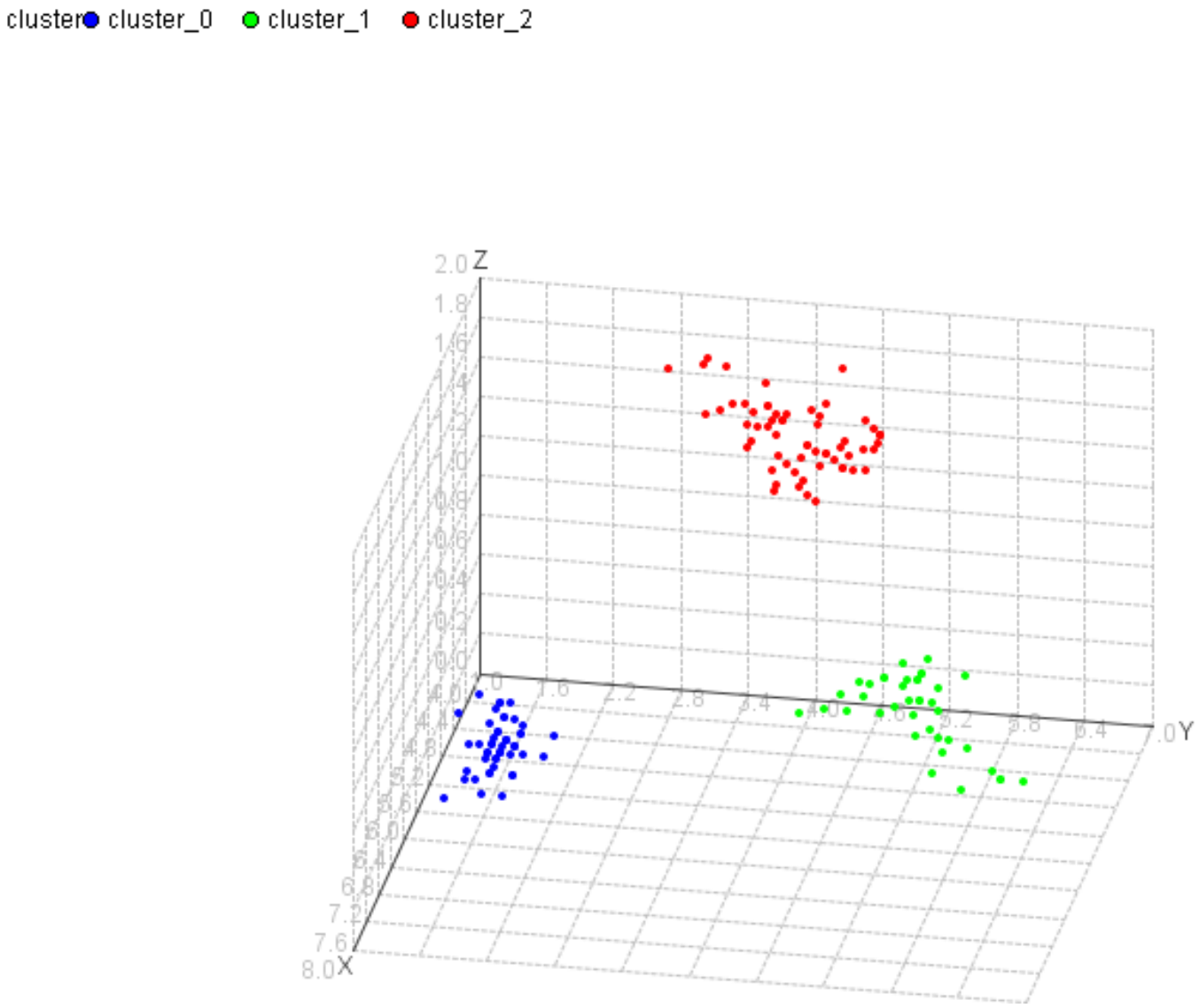
4. Klastering Bunga Iris

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*)

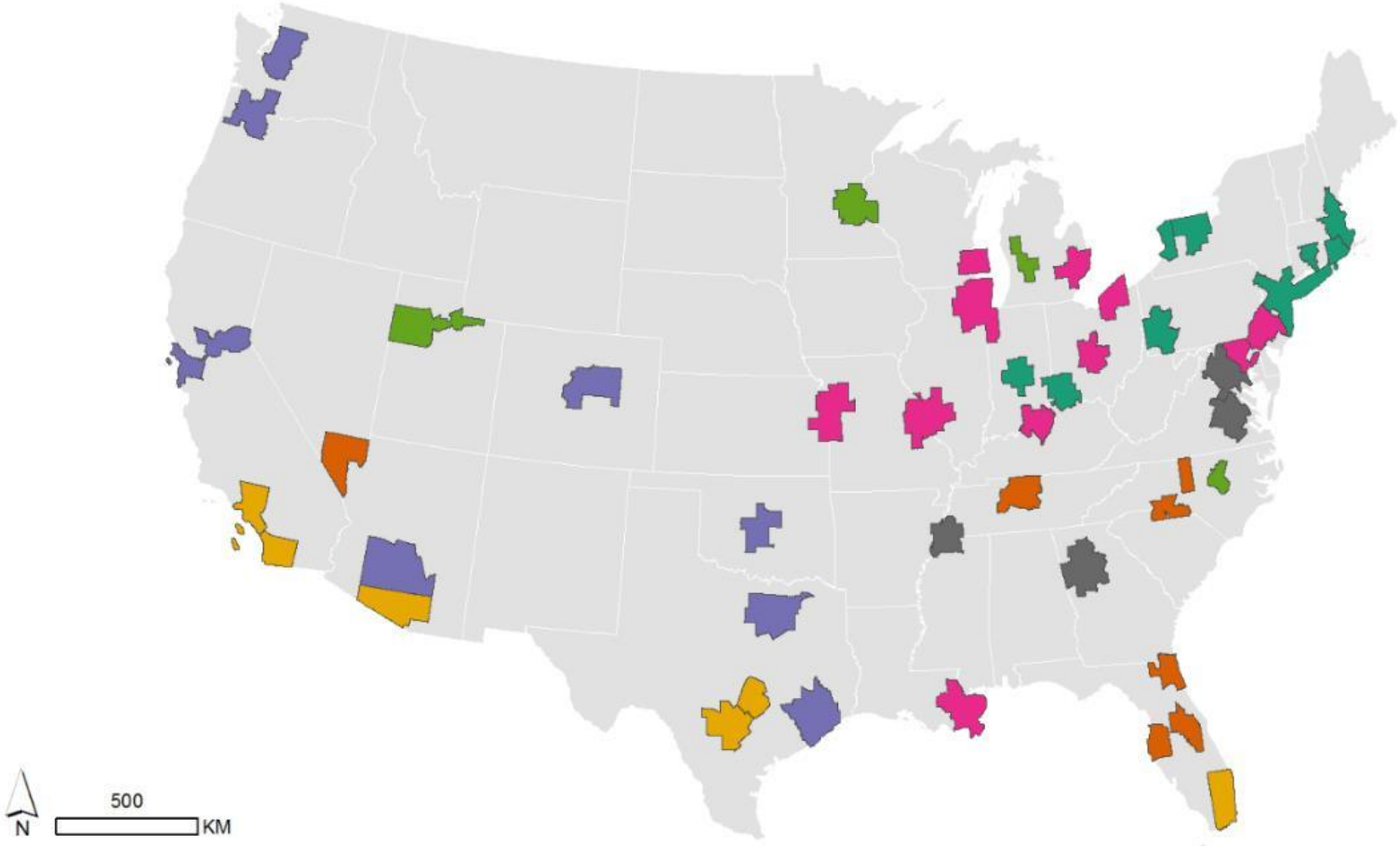
Pengetahuan (Model) Berupa Klaster



Klastering Jenis Pelanggan

Klastering Sentimen Warga

Poverty Rate Clustering



Group 1: Stability	Group 2: New South	Group 3: Hispanic Destinations	Group 4: Emerging Multiethnic	Group 5: Persistent Black Poverty	Group 6: Immigrant\ Educated	Group 7:New Old South
Boston	Charlotte	Austin	Dallas	Baltimore	Grand Rapids	Atlanta
Buffalo	Greensboro	Miami	Denver	Chicago	Minneapolis	Memphis
Cincinnati	Jacksonville	San Antonio	Houston	Cleveland	Raleigh	Richmond
Hartford	Nashville	Tuscon	Oklahoma City	Columbus	Salt Lake City	Washington
Indianapolis		San Diego	Phoenix	Detroit		
New York	Tampa	Los Angeles	San Francisco	Kansas City		
Pittsburgh	Las Vegas		Portland	Louisville		
Providence	Orlando		Sacramento	Milwaukee		
Rochester			Seattle	New Orleans		
				Philadelphia		
				St. Louis		

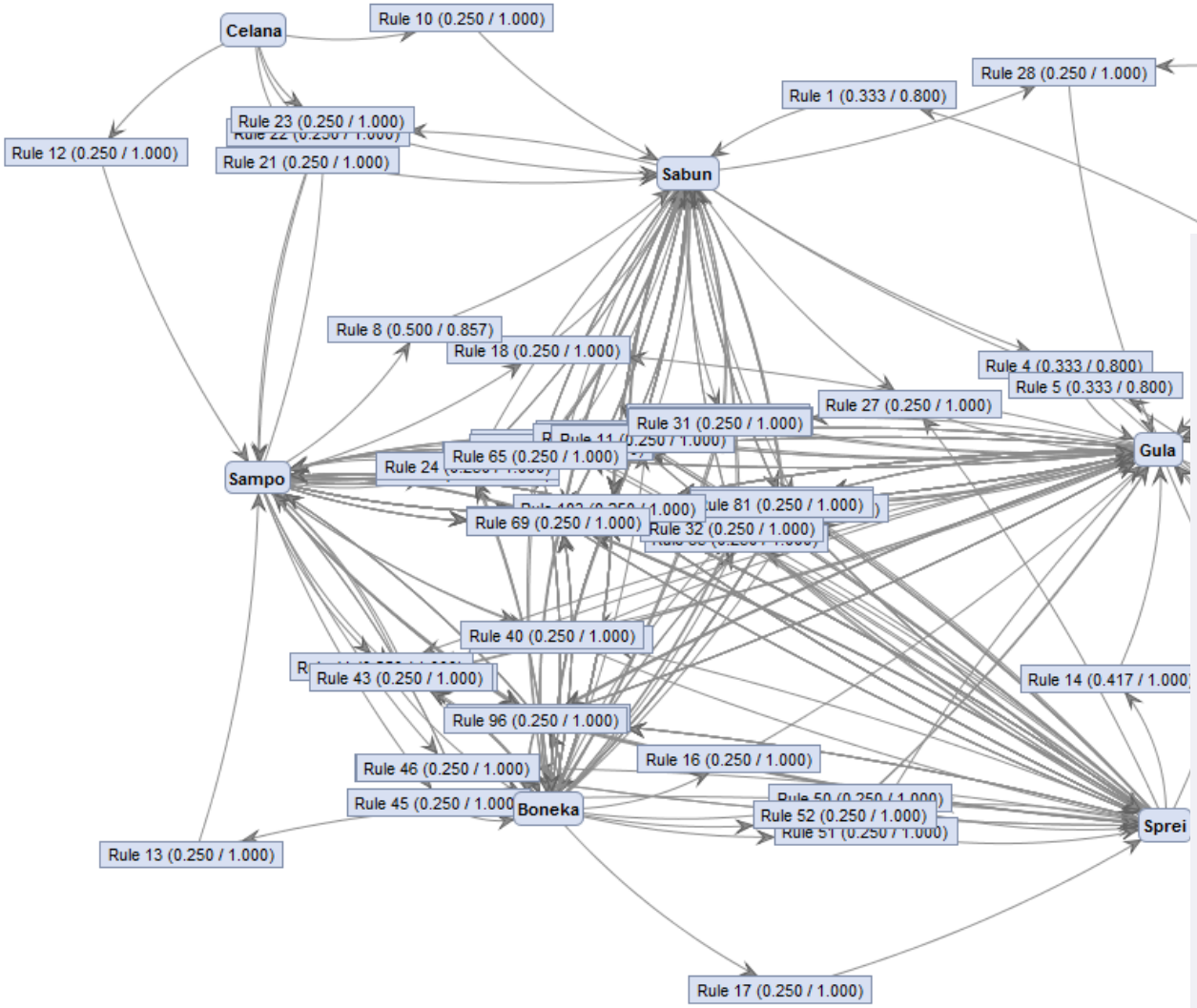
5. Aturan Asosiasi Pembelian Barang

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

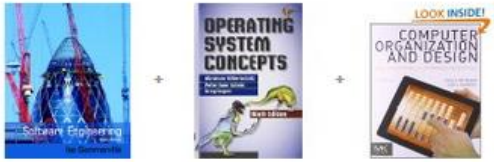
```
Association Rules
[Aqua] --> [Sabun] (confidence: 0.800)
[Sprei] --> [Kopi] (confidence: 0.800)
[Aqua] --> [Kopi] (confidence: 0.800)
[Sabun, Kopi] --> [Gula] (confidence: 0.800)
[Sabun, Gula] --> [Kopi] (confidence: 0.800)
[Sprei] --> [Kopi, Gula] (confidence: 0.800)
[Gula, Sprei] --> [Kopi] (confidence: 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

Frequently Bought Together



Price for all three: $387.88

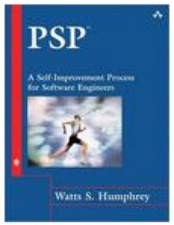
Add all three to Cart

Add all three to Wish List

Some of these items ship sooner than the others. Show details

- ☑ **This item:** Software Engineering (10th Edition) by Ian Sommerville Hardcover $169.67
- ☑ Operating System Concepts by Abraham Silberschatz Hardcover $144.03
- ☑ Computer Organization and Design, Fifth Edition: The Hardware/Software Interface (The Morgan Kaufmann ... by David A. Patterson Paperback $74.18

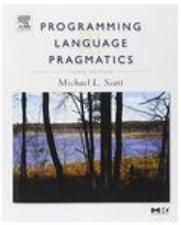
Customers Who Bought This Item Also Bought



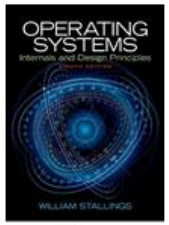
PSP(sm): A Self-Improvement Process for Software Engineers
› Watts S. Humphrey
★★★★☆ 12
Hardcover
$46.41 Prime

Computer Networking: A Top-Down Approach (6th Edition)
› James F. Kurose
★★★★☆ 131
Hardcover
$127.42 Prime

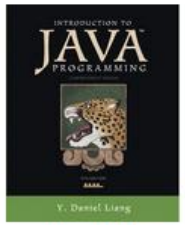
Computer Organization and Design, Fifth Edition: The Hardware/Software
David A. Patterson
★★★½☆ 42
Paperback
$74.18 Prime



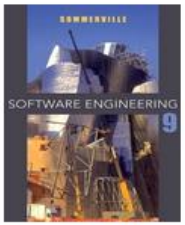
Programming Language Pragmatics, Third Edition
Michael L. Scott
★★★★☆ 24
Paperback
$60.54 Prime



Operating Systems: Internals and Design Principles (8th Edition)
› William Stallings
★★★★½ 10
Hardcover
$141.29 Prime



Introduction to Java Programming, Comprehensive Version (9th Edition)
Y. Daniel Liang
★★★★☆ 82
Paperback



Software Engineering (9th Edition)
Ian Sommerville
★★★☆☆ 29
Hardcover
$140.10 Prime

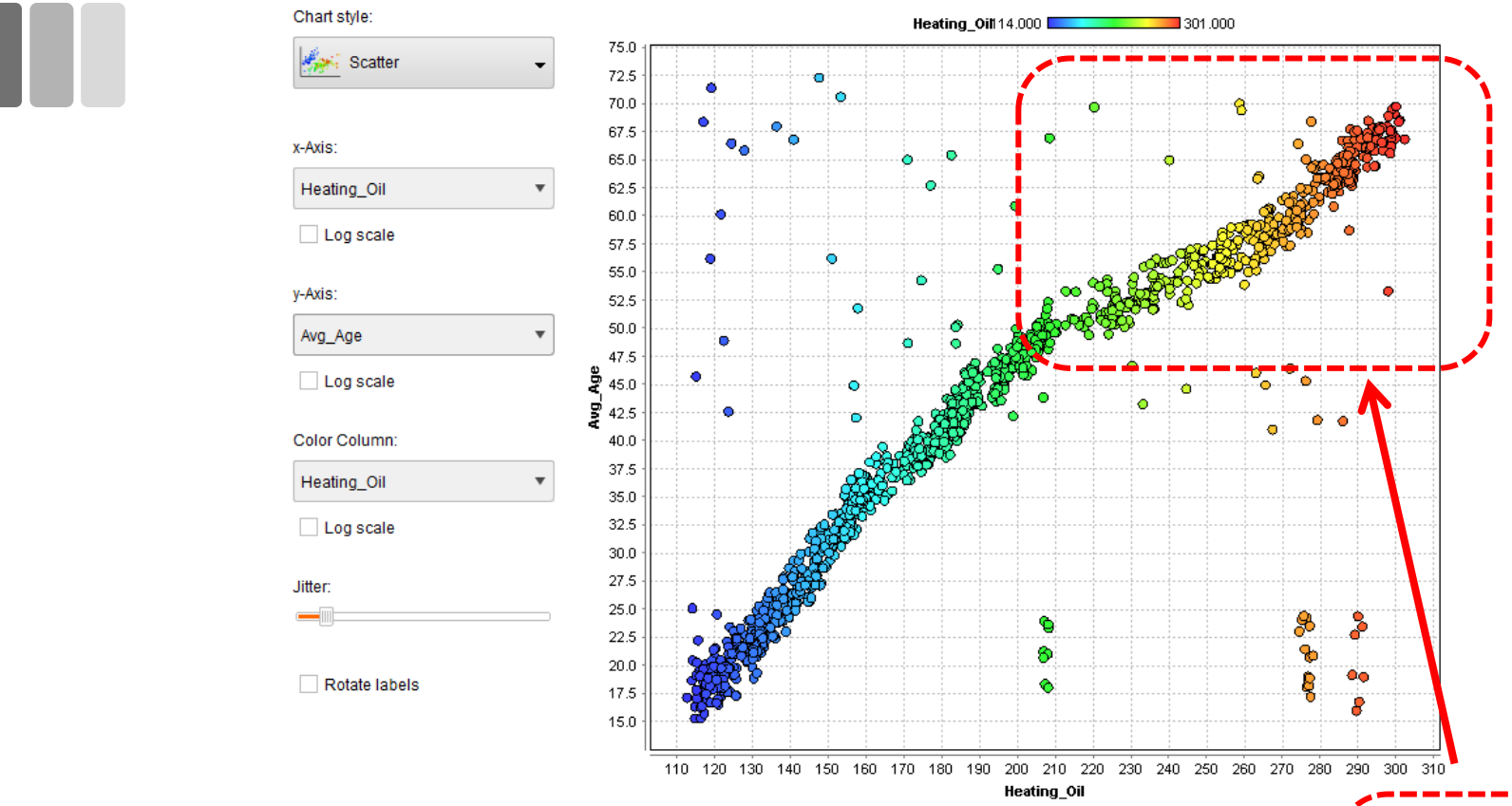
Show more

Heating Oil Consumption

Korelasi antara jumlah konsumsi minyak pemanas dengan faktor-faktor di bawah:

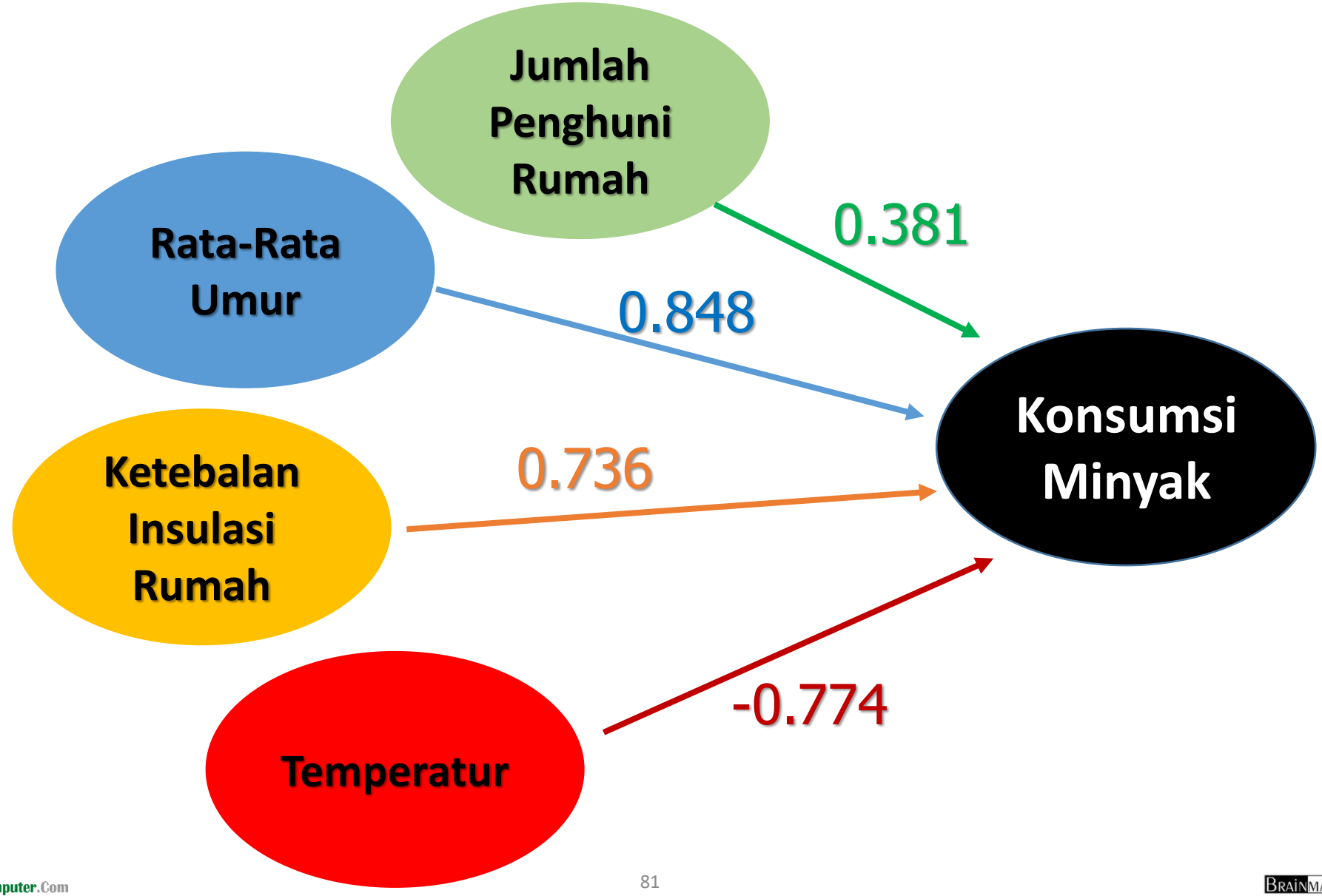
1. Insulation: Ketebalan insulasi rumah
2. Temperatur: Suhu udara sekitar rumah
3. Heating Oil: Jumlah konsumsi minyak pertahun perrumah
4. Number of Occupant: Jumlah penghuni rumah
5. Average Age: Rata-rata umur penghuni rumah
6. Home Size: Ukuran rumah

Row No.	Insulation	Temperature	Heating_Oil	Num_Occup...	Avg_Age	Home_Size
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28	6
4	9	50	196	4	45.100	3
5	2	80	131	5	20.800	2
6	5	76	129	3	21.500	3
7	5	72	131	4	23.500	3
8	6	88	161	2	38.200	6
9	5	77	184	3	42.500	3
10	10	42	225	3	51.100	1
11	6	90	178	2	42.100	2
12	3	83	121	1	19.800	2
13	10	43	186	5	45.100	6
14	8	59	206	2	50.100	8
15	4	86	179	5	41.400	6



Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

Korelasi 4 Variable terhadap Konsumsi Minyak

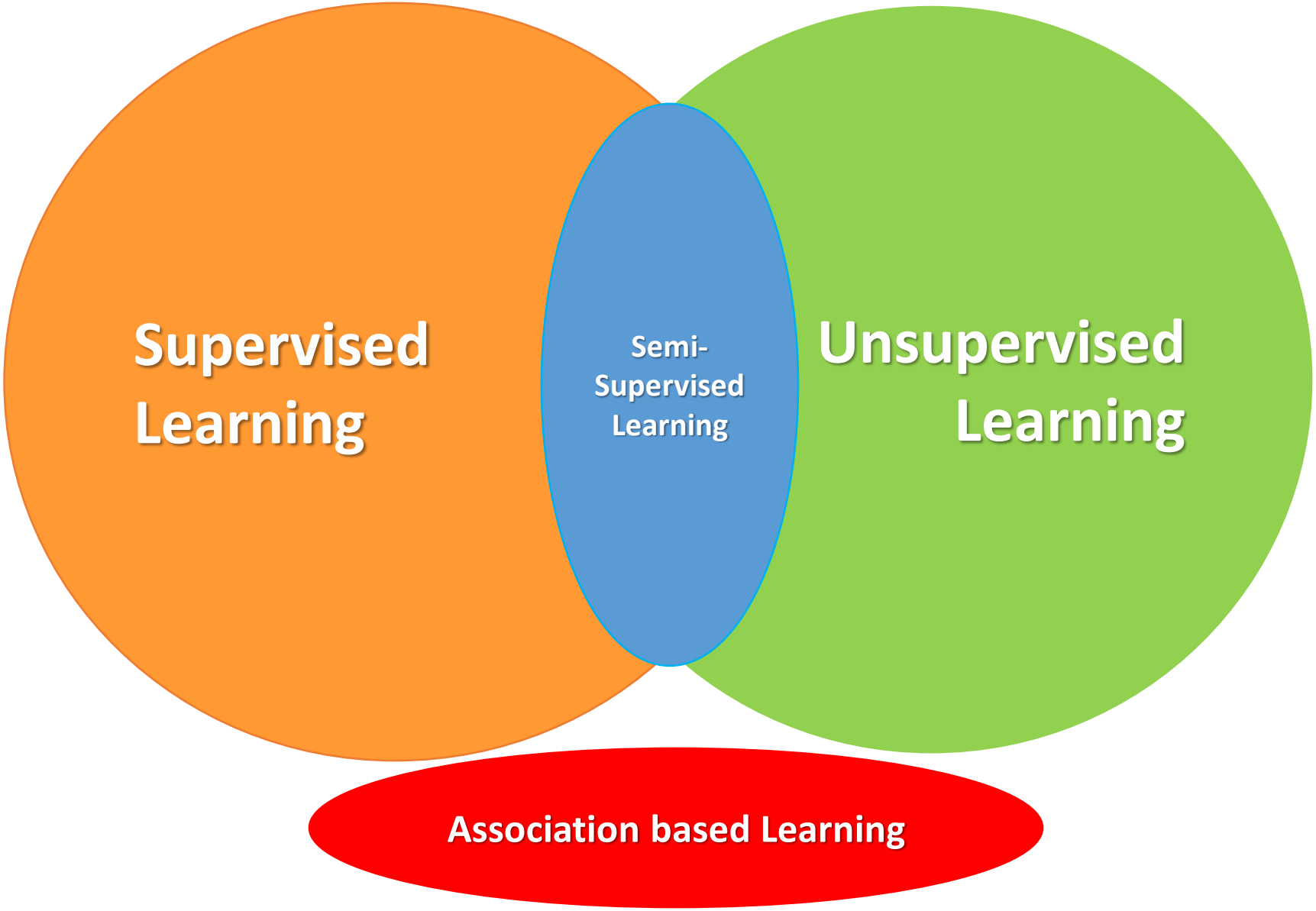


Insight Law (Data Mining Law 6)

Data mining amplifies perception in the business domain

- How does data mining produce insight? This law approaches the heart of data mining – why it must be a business process and not a technical one
  - Business problems are solved by people, not by algorithms
- The data miner and the business expert "see" the solution to a problem, that is the patterns in the domain that allow the business objective to be achieved
  - Thus data mining is, or assists as part of, a perceptual process
  - Data mining algorithms reveal patterns that are not normally visible to human perception
- Within the data mining process, the human problem solver interprets the results of data mining algorithms and integrates them into their business understanding

Metode Learning 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

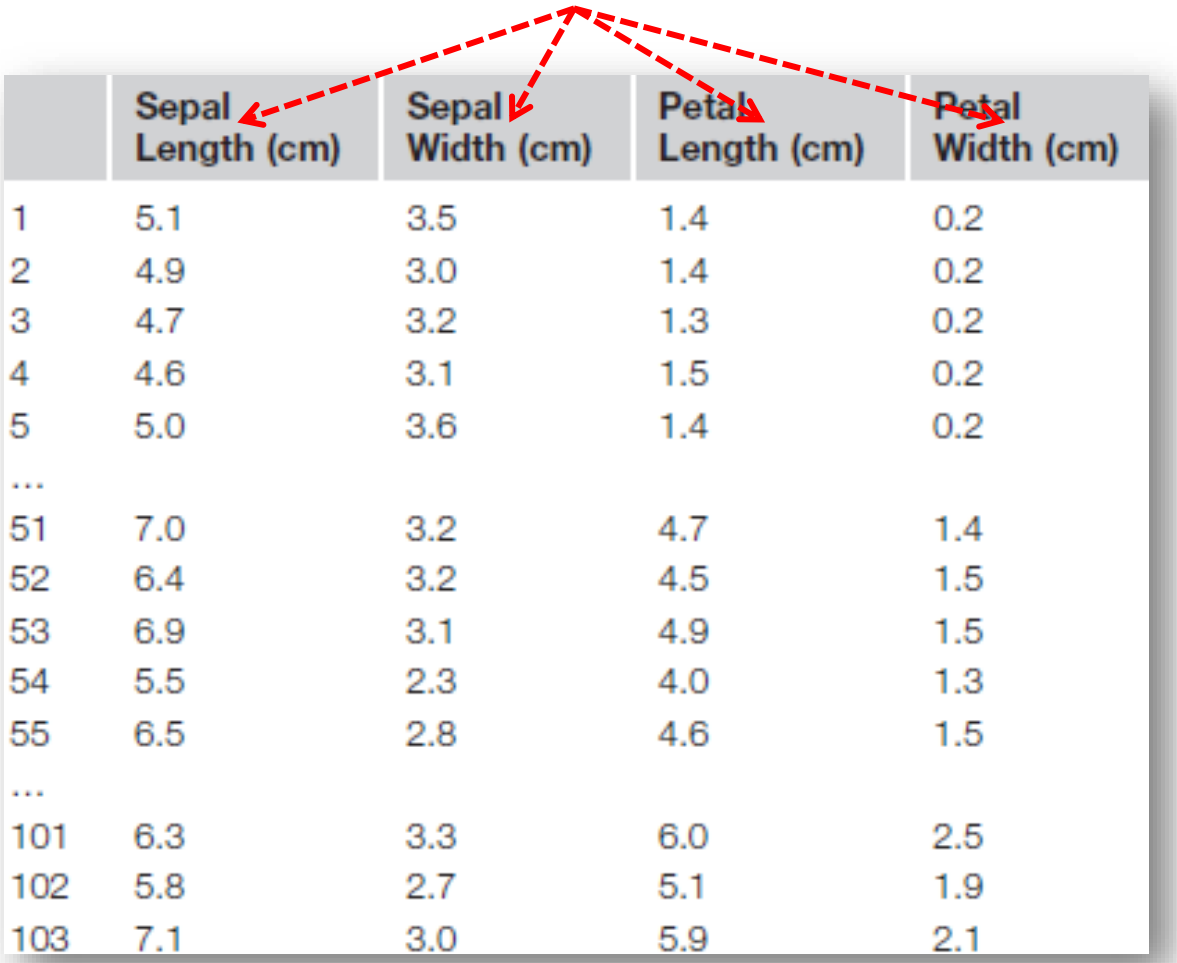
Numerik

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

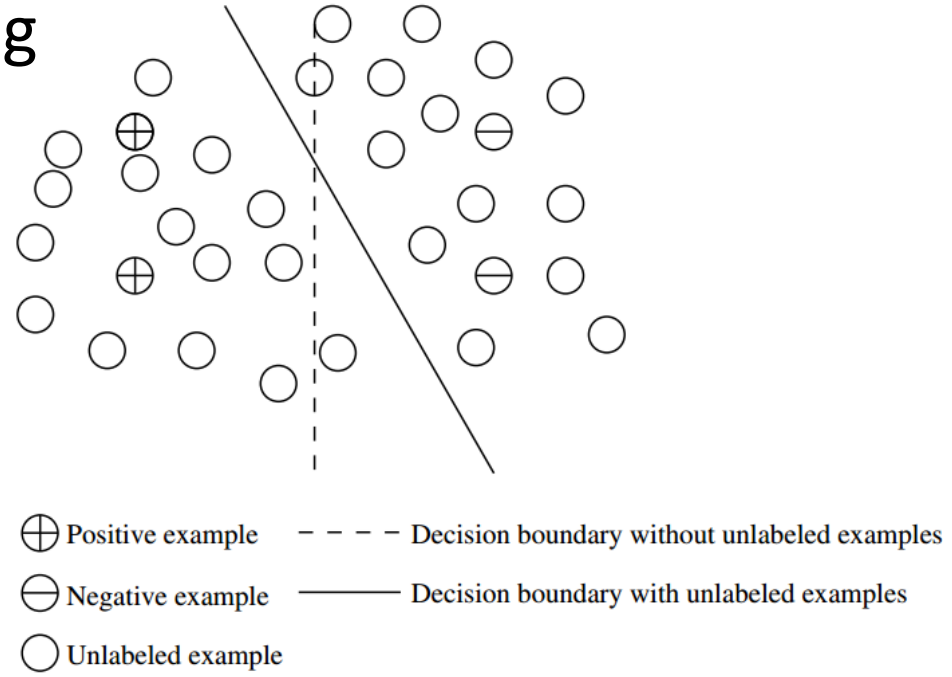
Attribute/Feature/Dimension



	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 mining yang menggunakan data dengan label dan tidak berlabel sekaligus dalam proses pembelajarannya



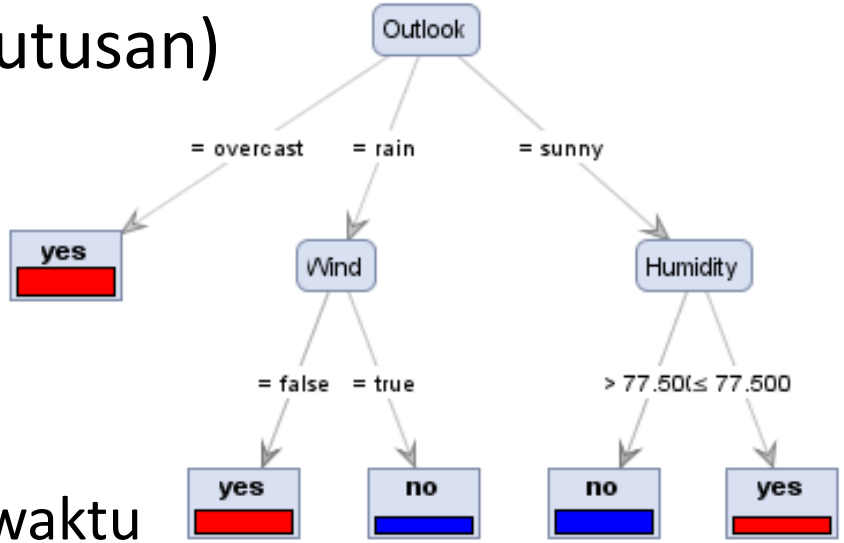
- Data yang memiliki kelas digunakan untuk membentuk model (pengetahuan), data tanpa label digunakan untuk membuat batasan antara kelas

Metode Data Mining

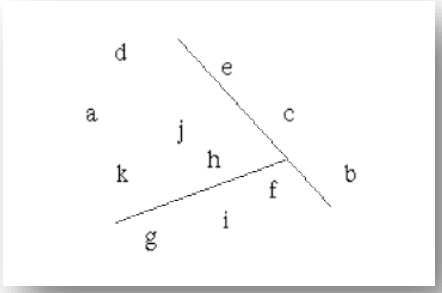
1. Estimation (Estimasi):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc
2. Forecasting (Prediksi/Peramalan):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized 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

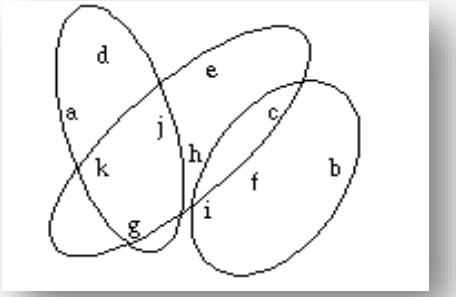
Output/Pola/Model/Knowledge

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. Tingkat Korelasi
4. Rule (Aturan)
   - IF ips3=2.8 THEN lulustepatwaktu
5. Cluster (Klaster)





Latihan

1. Sebutkan 5 peran utama data mining!
2. Jelaskan perbedaan estimasi dan forecasting!
3. Jelaskan perbedaan forecasting dan klasifikasi!
4. Jelaskan perbedaan klasifikasi dan klastering!
5. Jelaskan perbedaan klastering dan association!
6. Jelaskan perbedaan estimasi dan klasifikasi!
7. Jelaskan perbedaan estimasi dan klastering!
8. Jelaskan perbedaan supervised dan unsupervised learning!
9. Sebutkan tahapan utama proses data mining!

1.3 Sejarah dan Penerapan Data Mining

Evolution of Sciences

- Sebelum 1600: Empirical science
  - Disebut sains kalau bentuknya kasat mata
- 1600-1950: Theoretical science
  - Disebut sains kalau bisa dibuktikan secara matematis atau eksperimen
- 1950s-1990: Computational science
  - Seluruh disiplin ilmu bergerak ke komputasi
  - Lahirnya banyak model komputasi
- 1990-sekarang: Data science
  - Kultur manusia menghasilkan data besar
  - Kemampuan komputer untuk mengolah data besar
  - Datangnya data mining sebagai arus utama sains

(Jim Gray and Alex Szalay, The World Wide Telescope: An Archetype for Online Science, Communication of ACM, 45(11): 50-54, Nov. 2002)

Forbes / Tech

JAN 20, 2016 @ 02:39 PM 15,446 VIEWS The Little Black

Report: Why "Data Scientist" Is The Best Job To Pu 2016

Gregory Ferenstein, CONTRIBUTOR

FULL BIO

Opinions expressed

(Ferenstein Wire)
jobs in America, a
company review s
voluntary reviews
company's massiv
a composite score
openings, and car

According to the r
Scientist is an imp

CNBC

JOBS

ECONOMY | WORLD ECONOMY | US ECONOMY | THE FED | CENTRAL BANKS | JOBS |

Data science jobs top Glassdoor survey for best work-life balance

Uptin Saiidi | @uptin
Tuesday, 4 Oct 2016 | 3:40 AM ET
CNBC

glassdoor Jobs Companies Salaries Interviews Sign In
Search Jobs or Companies... Employers | Try Free Job Postings

- Employees' Choice Awards
- Other Lists
  - Oddball Interview Questions
  - Best Jobs
  - Best Cities for Jobs
- Trends

Additional Resources
Award FAQ
Trends FAQ
Free Employer Account
Press Center

25 Best Jobs in America 2.5k Shares

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

United States 2016

1

Data Scientist

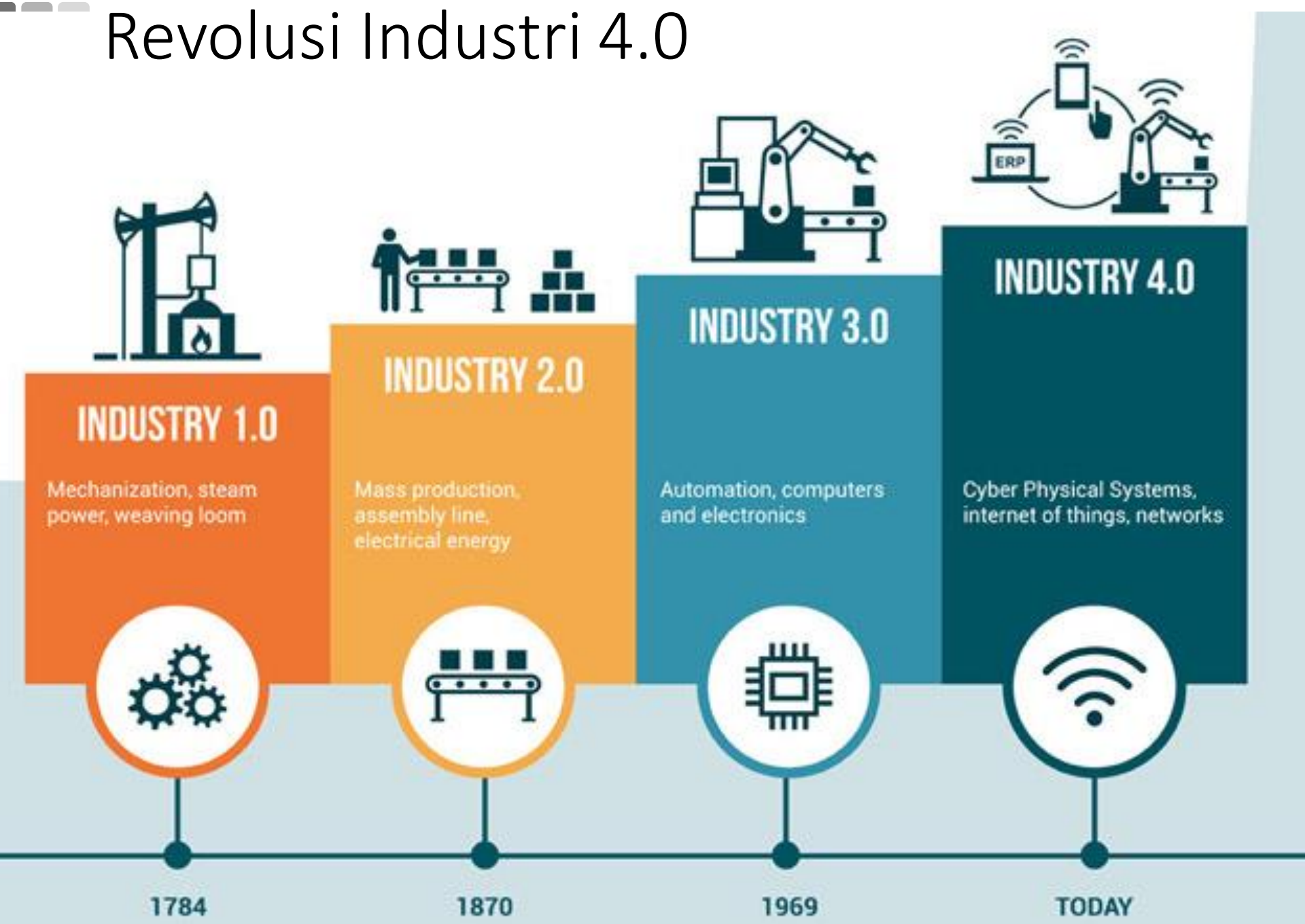
Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

2

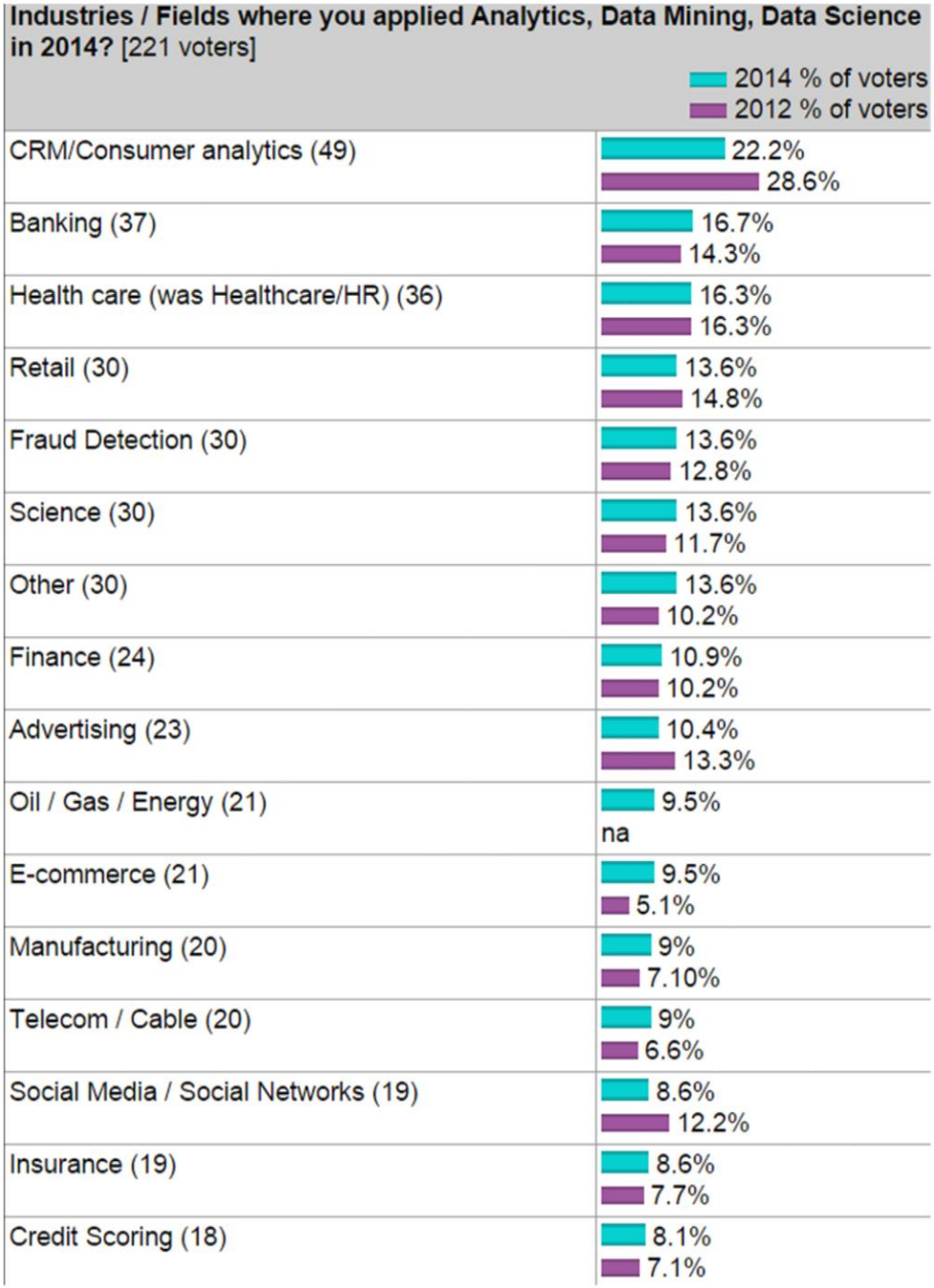
Tax Manager

Job Openings	1,574
Median Base Salary	$108,000
Career Opportunity	3.9
Job Score	4.7

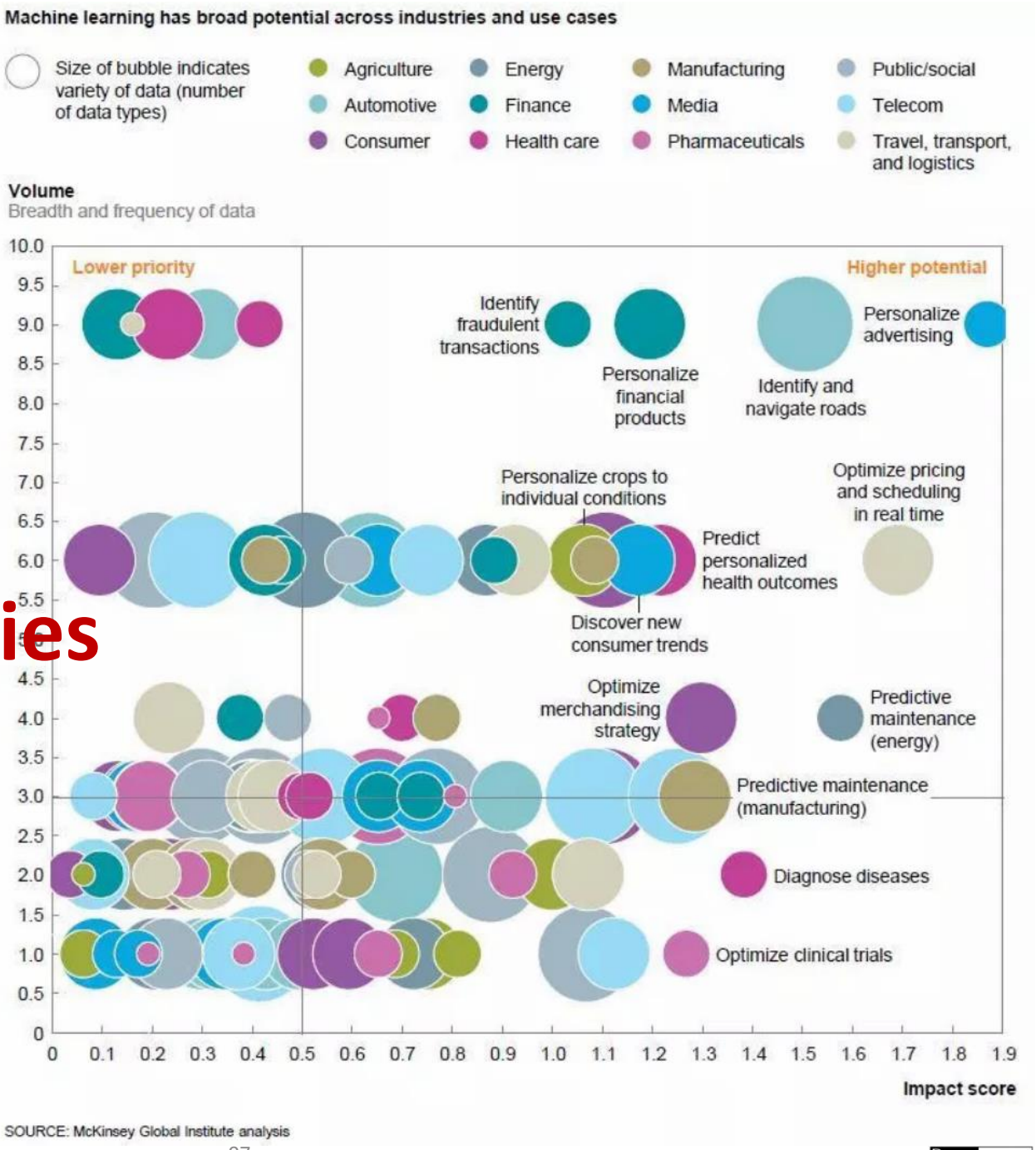
Revolusi Industri 4.0



Industries / Fields where you applied Analytics, Data Mining, Data Science in 2014? [221 voters]



Data Mining Use Cases across Industries

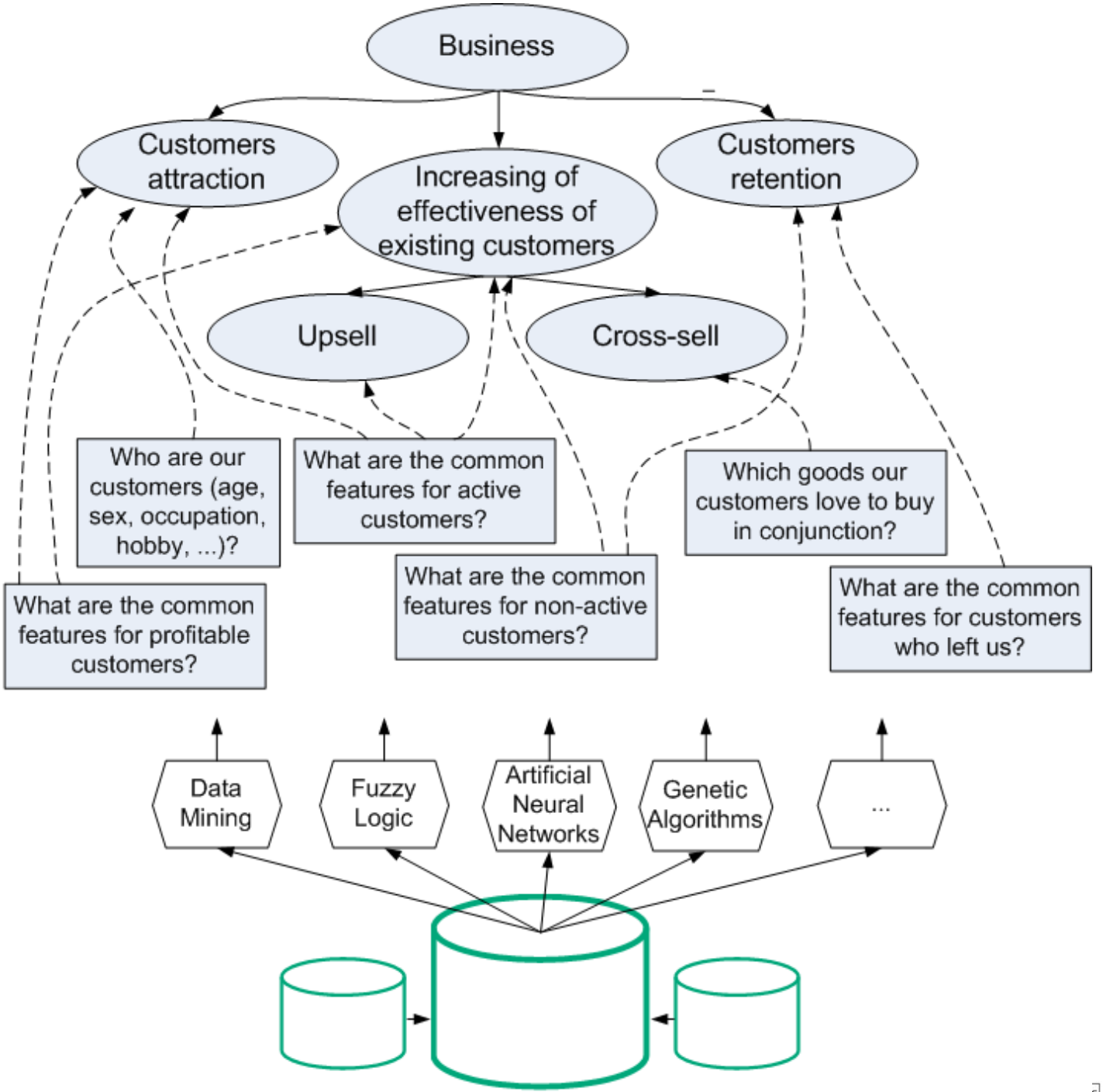


Business

Knowledge

Methods

Technology



Business Goals Law (Data Mining Law 1)

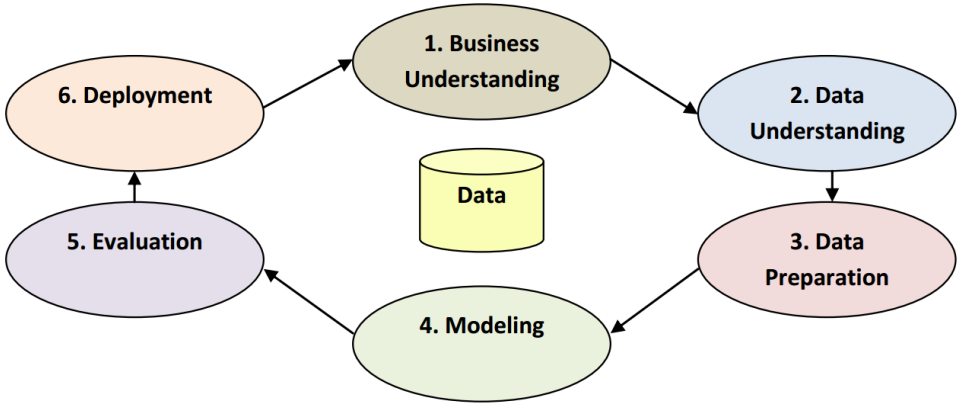
Business objectives are the origin of every data mining solution

- This defines the field of data mining: data mining is concerned with solving business problems and achieving business goals
- Data mining is not primarily a technology; it is a process, which has one or more business objectives at its heart
- Without a business objective, there is no data mining
- The maxim: "Data Mining is a Business Process"

Business Knowledge Law (Data Mining Law 2)

Business knowledge is central to every step of the data mining process

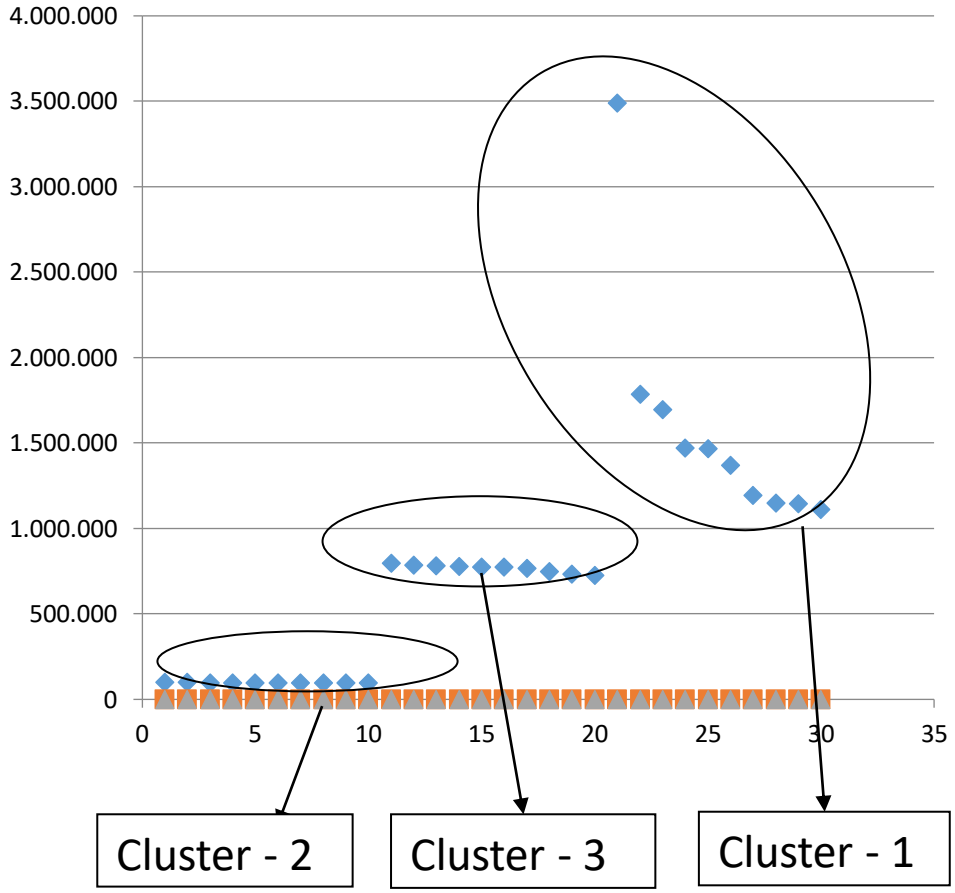
- A naive reading of CRISP-DM would see business knowledge used at the start of the process in defining goals, and at the end of the process in guiding deployment of results
- This would be to miss a key property of the data mining process, that business knowledge has a central role in every step



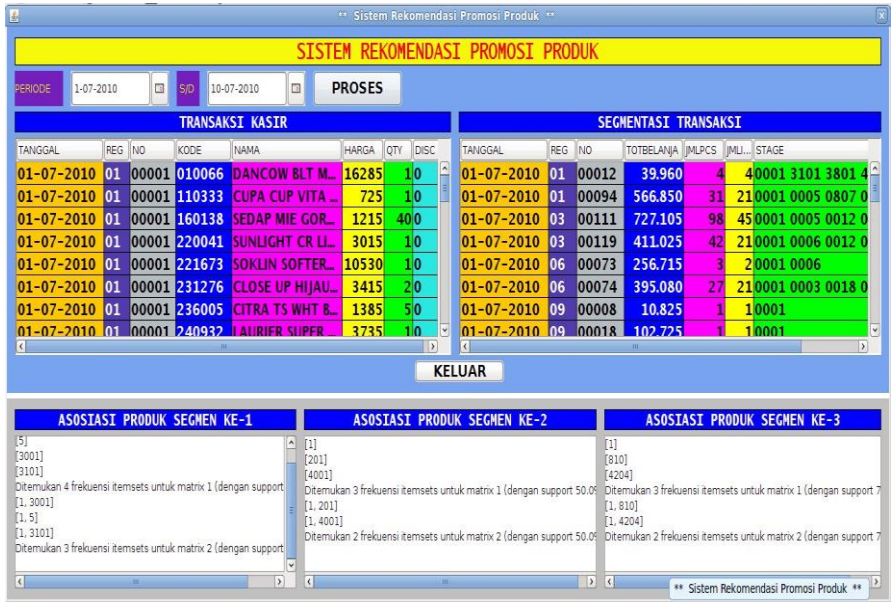
Private and Commercial Sector

- Marketing: product recommendation, market basket analysis, product targeting, customer retention
- Finance: investment support, portfolio management, price forecasting
- Banking and Insurance: credit and policy approval, money laundry detection
- Security: fraud detection, access control, intrusion detection, virus detection
- Manufacturing: process modeling, quality control, resource allocation
- Web and Internet: smart search engines, web marketing
- Software Engineering: effort estimation, fault prediction
- Telecommunication: network monitoring, customer churn prediction, user behavior analysis

Use Case: Product Recommendation

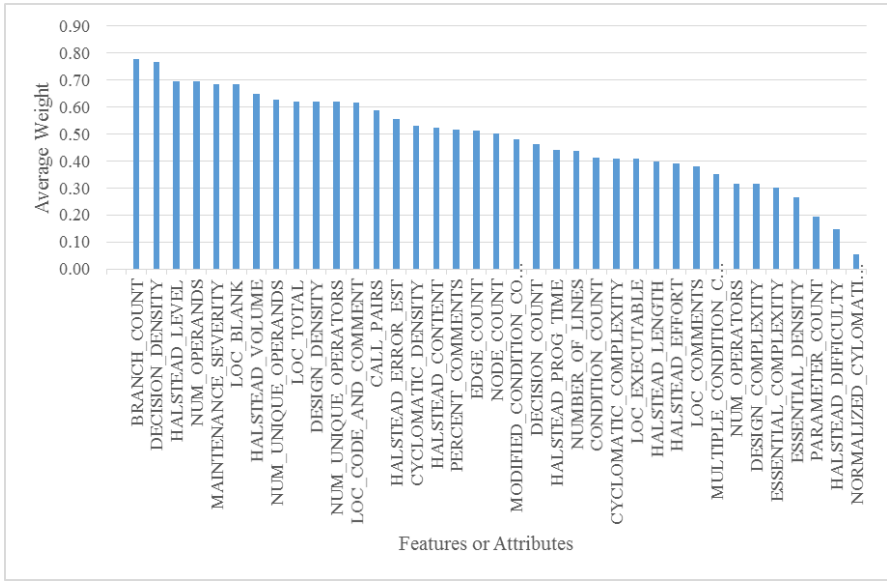


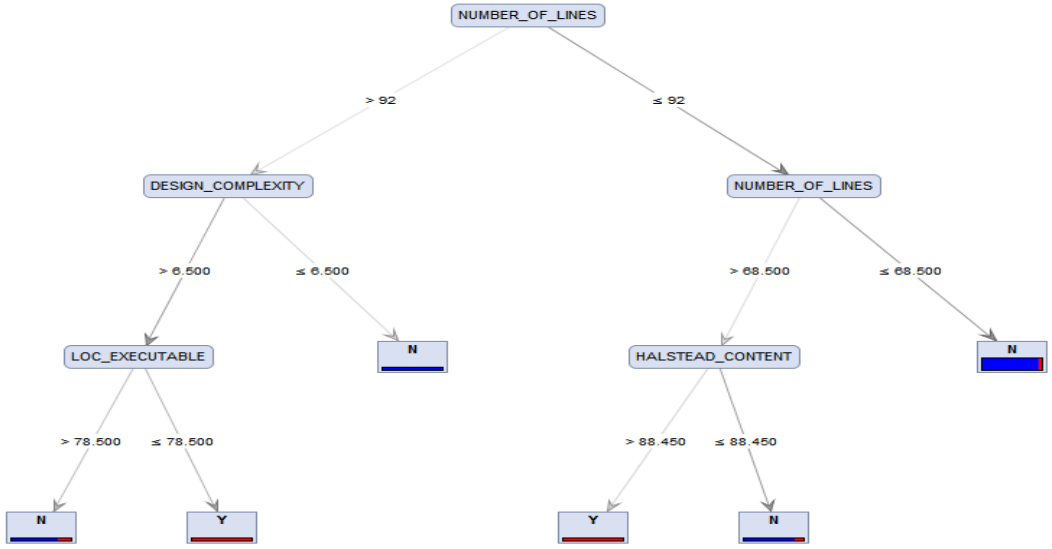
- Tot.Belanja
- Jml.Pcs
- Jml.Item



Use Case: Software Fault Prediction

- The cost of capturing and correcting defects is expensive
  - $14,102 per defect in post-release phase *(Boehm & Basili 2008)*
  - $60 billion per year (NIST 2002)
- Industrial methods of manual software reviews activities can find only 60% of defects (Shull et al. 2002)
- The probability of detec software fault predictio models is higher (71%) t software reviews (60%)





Public and Government Sector

- Finance: exchange rate forecasting, sentiment analysis
- Taxation: adaptive monitoring, fraud detection
- Medicine and Healt Care: hypothesis discovery, disease prediction and classification, medical diagnosis
- Education: student allocation, resource forecasting
- Insurance: worker's compensation analysis
- Security: bomb, iceberg detection
- Transportation: simulation and analysis, load estimation
- Law: legal patent analysis, law and rule analysis
- Politic: election prediction

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

Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases *(G. Piatetsky-Shapiro and W. Frawley, 1991)*
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining *(U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)*
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Conferences dan Journals Data Mining

Conferences

- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- (IEEE) Int. Conf. on Data Mining (ICDM)
- European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Int. Conf. on Web Search and Data Mining (WSDM)

Journals

- ACM Transactions on Knowledge Discovery from Data (TKDD)
- ACM Transactions on Information Systems (TOIS)
- IEEE Transactions on Knowledge and Data Engineering
- Springer Data Mining and Knowledge Discovery
- International Journal of Business Intelligence and Data Mining (IJBIDM)

2. Proses Data Mining

2.1 Proses dan Tools Data Mining

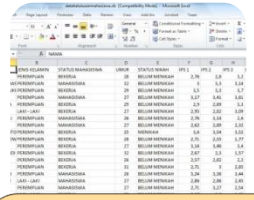
2.2 Penerapan Proses Data Mining

2.3 Evaluasi Model Data Mining

2.4 Proses Data Mining berbasis CRISP-DM

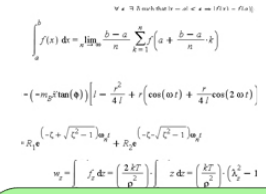
2.1 Proses dan Tools Data Mining

Proses Data Mining



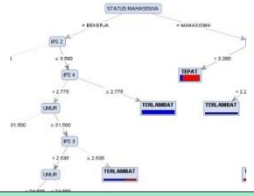
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



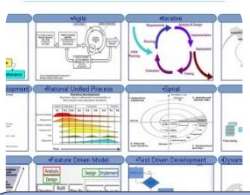
2. Metode Data Mining

(**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

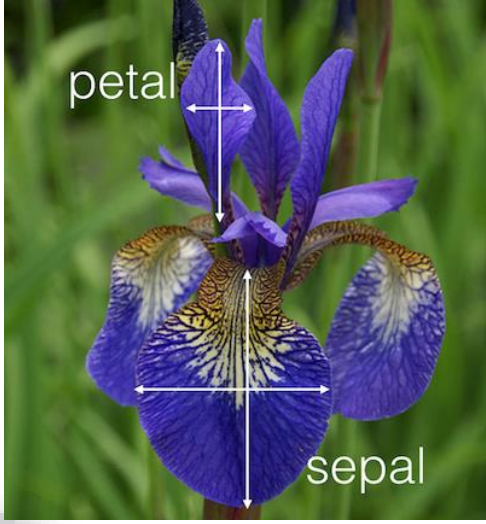
- Life Sciences (82)
- Physical Sciences (43)
- CS / Engineering (111)
- Social Sciences (23)
- Business (21)
- Game (10)
- Other (67)

 # 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

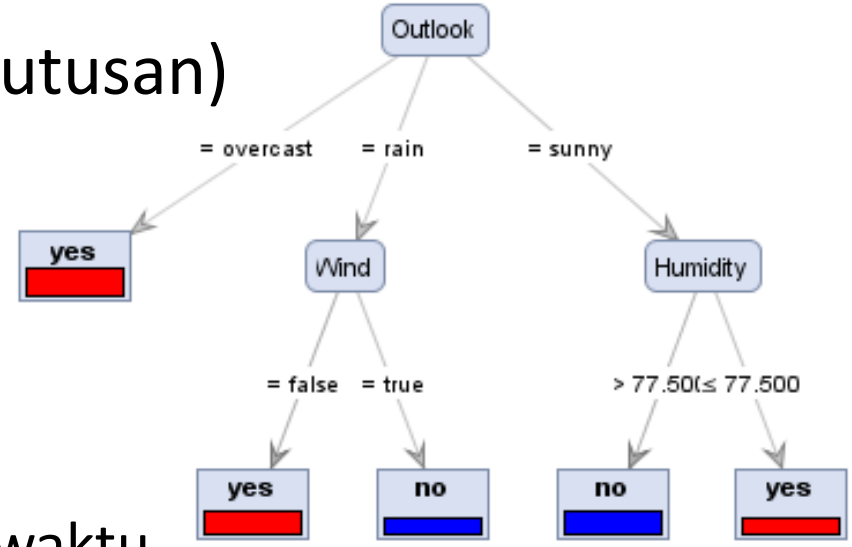
Numerik

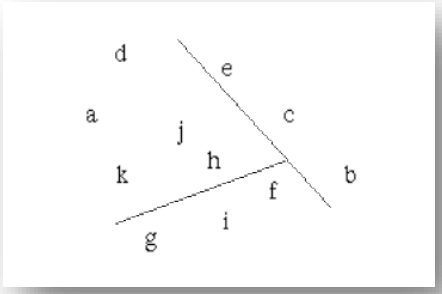
2. Metode Data Mining

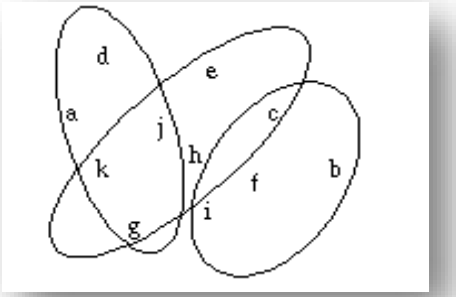
1. Estimation (Estimasi):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc
2. Forecasting (Prediksi/Peramalan):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized 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

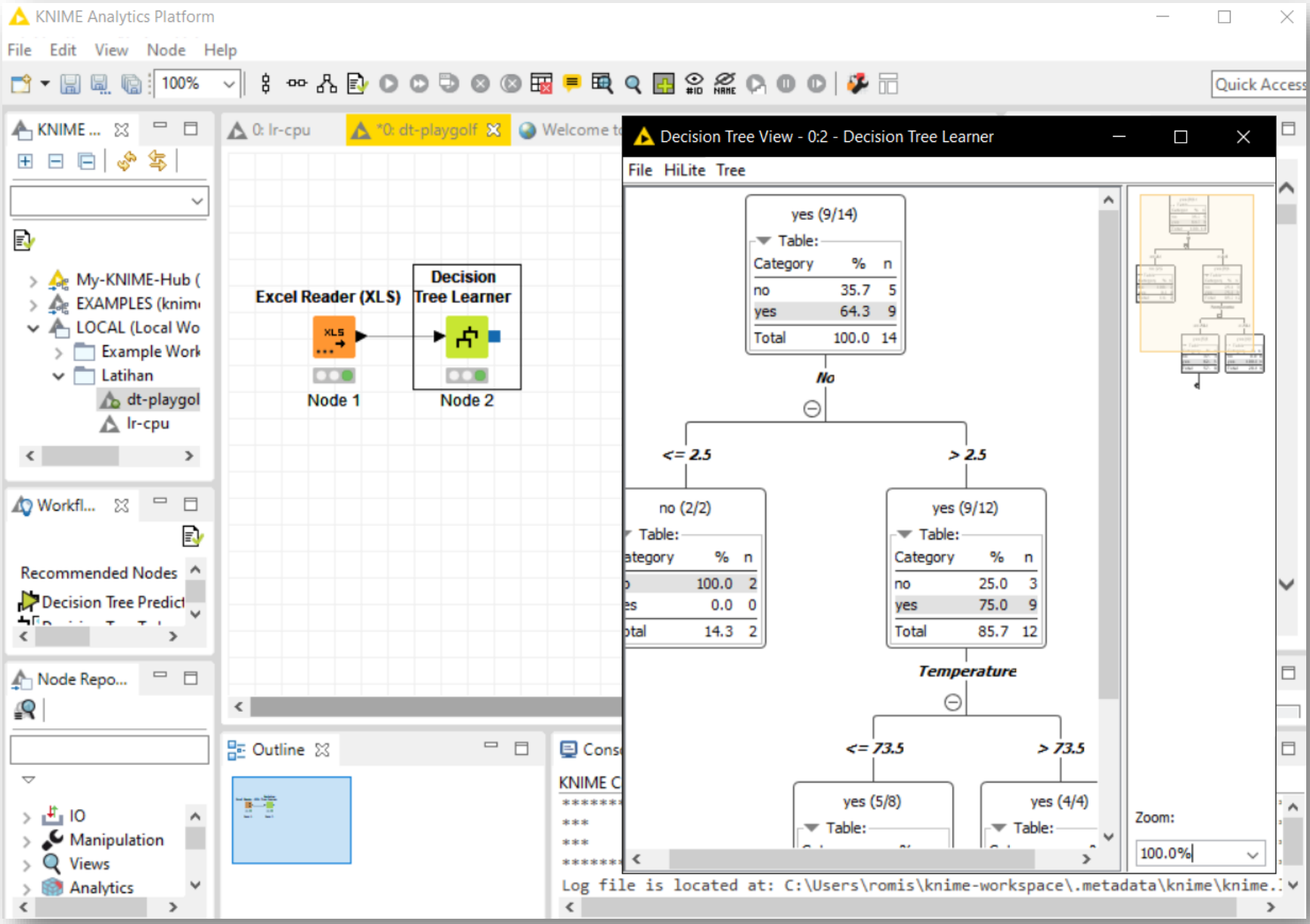
Magic Quadrant for Data Science Platform

(Gartner, 2017)

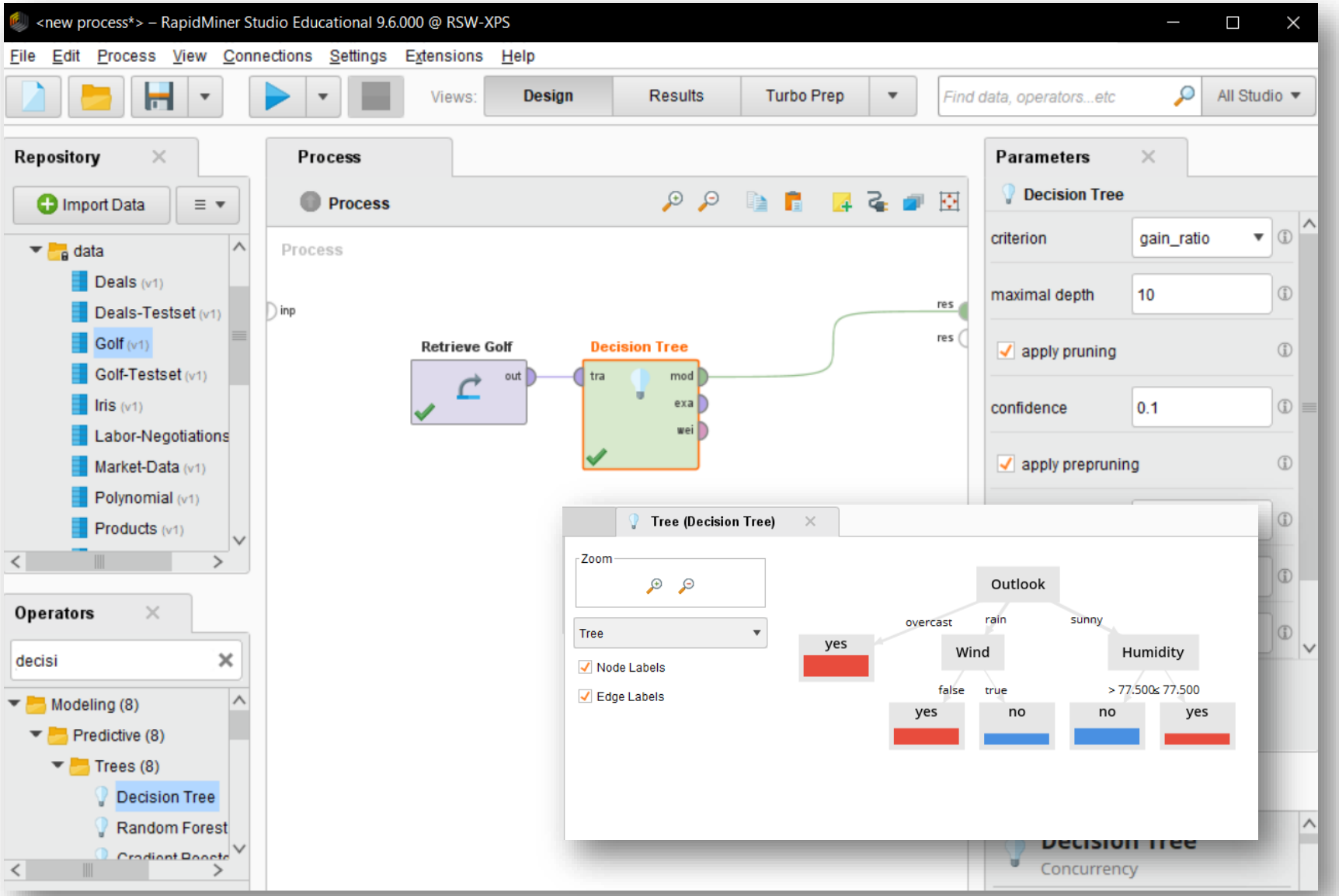
Magic Quadrant for Data Science Platform

(Gartner, 2018)

KNIME


KNIME Analytics Platform
File Edit View Node Help
Quick Access
KNIME ...
My-KNIME-Hub (
EXAMPLES (knim
LOCAL (Local Wo
Example Work
Latihan
dt-playgol
lr-cpu
Workfl...
Recommended Nodes
Decision Tree Predict
Node Repo...
IO
Manipulation
Views
Analytics
0: lr-cpu
*0: dt-playgolf
Welcome to
Excel Reader (XLS)
Decision Tree Learner
Node 1
Node 2
Outline
Decision Tree View - 0:2 - Decision Tree Learner
File HiLite Tree
yes (9/14)
Table:
Category % n
no 35.7 5
yes 64.3 9
Total 100.0 14
No
<= 2.5
> 2.5
no (2/2)
Table:
100.0 2
0.0 0
14.3 2
yes (9/12)
Table:
Category % n
no 25.0 3
yes 75.0 9
Total 85.7 12
Temperature
<= 73.5
> 73.5
yes (5/8)
Table:
yes (4/4)
Table:
Zoom:
100.0%
Log file is located at: C:\Users\romis\knime-workspace\.metadata\knime\knime..

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
Concurrency

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

Fitur Rapidminer

- Mendukung proses dan metode data mining: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi
- Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI
- Mengintegrasikan berbagai tools data mining lain termasuk Weka, R, Hadoop, Python, dsb
- Dukungan Format data: Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL, Ingres, Excel, Access, SPSS, CSV files dan berbagai format lain

Atribut Pada Rapidminer

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

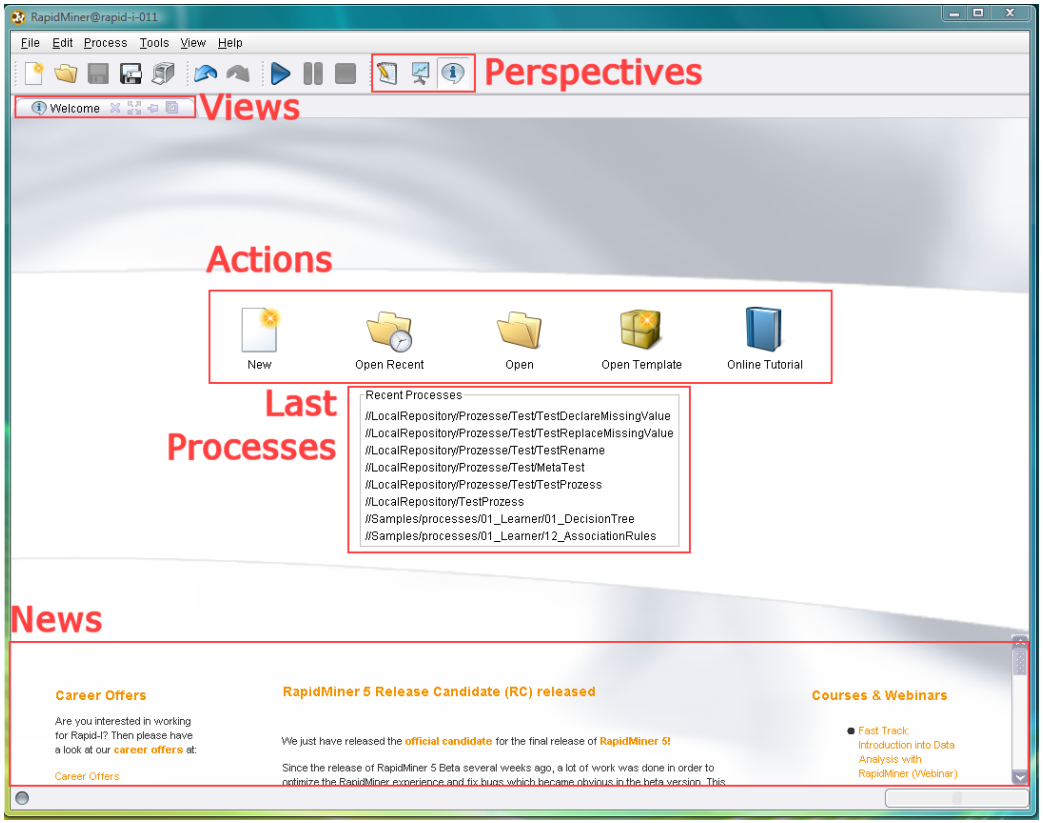
2. Atribut target: atribut yang menjadi tujuan untuk diisi oleh proses data mining
   - Label, cluster, weight

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

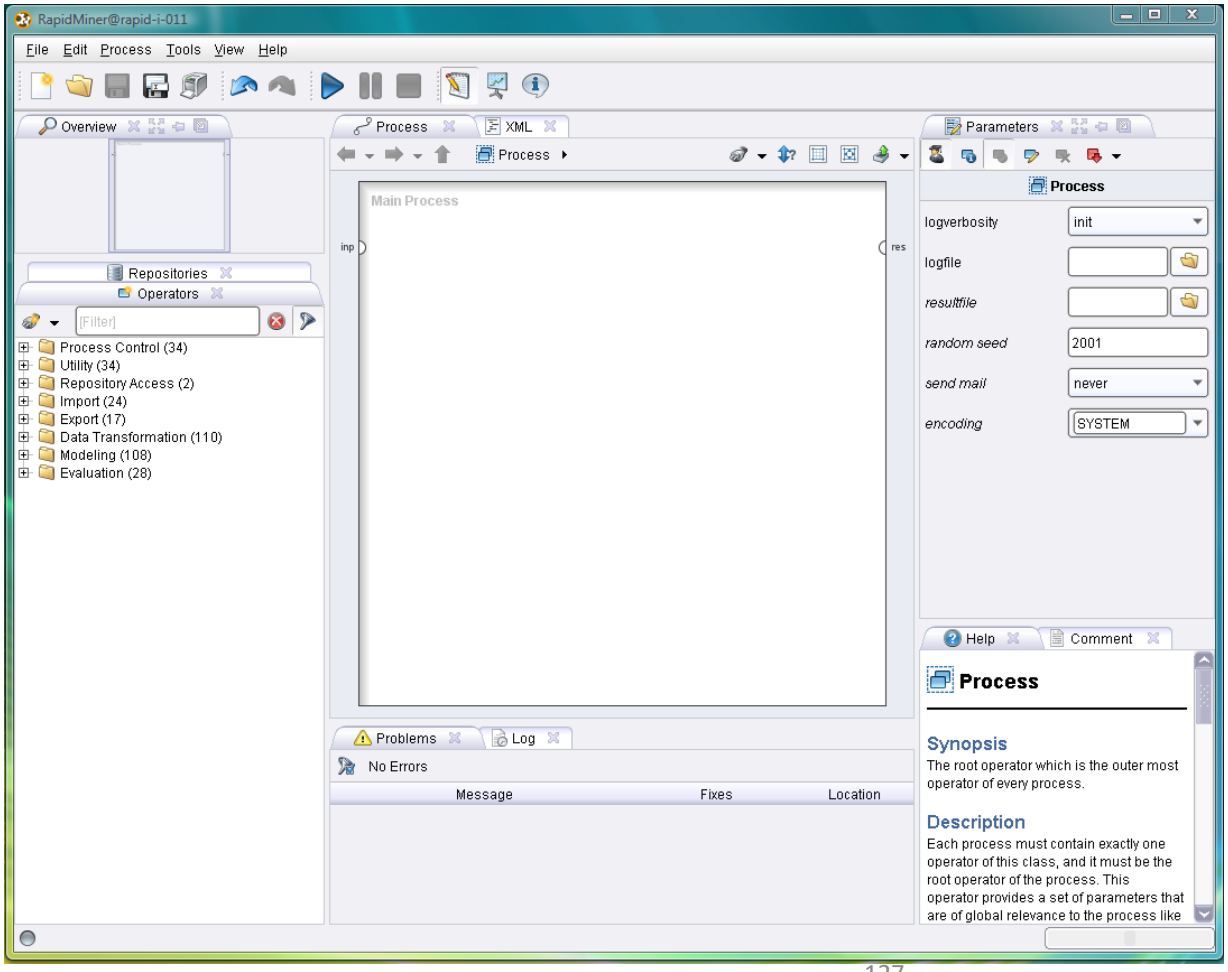
Perspektif dan View

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



Perspektif Desain

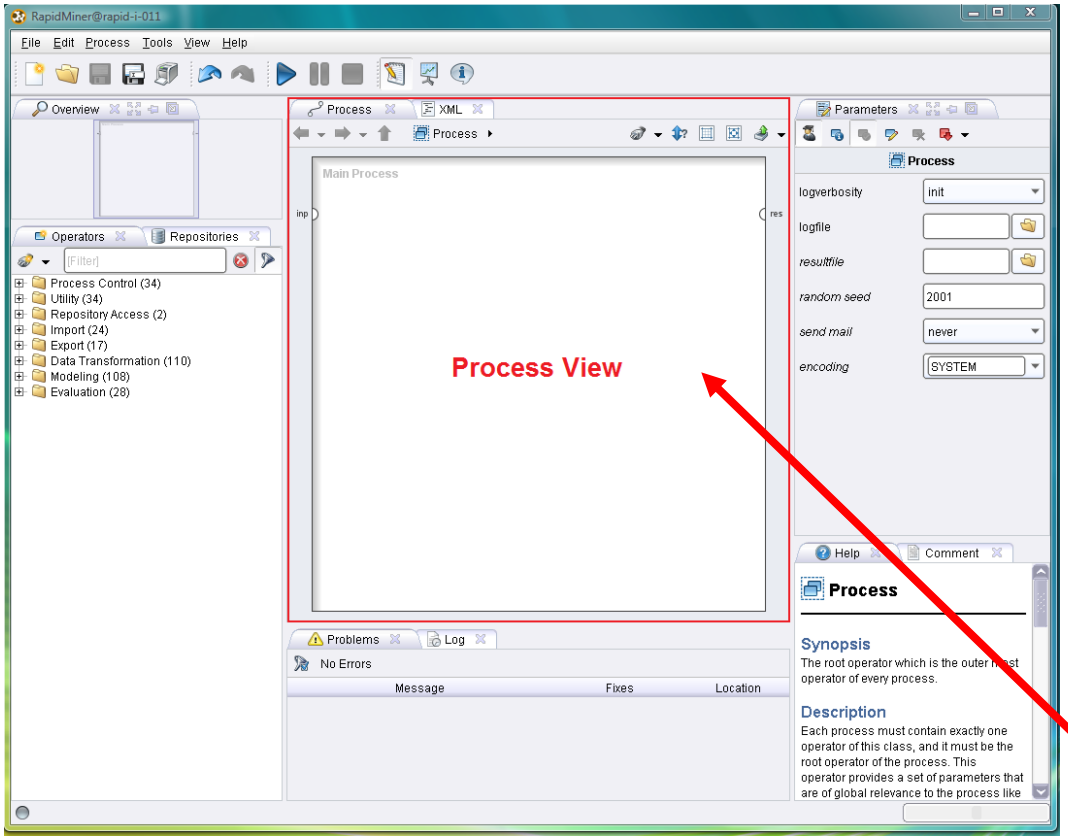
- Perspektif dimana semua proses dibuat dan dikelola
- Pindah ke Perspektif Desain dengan Klik:

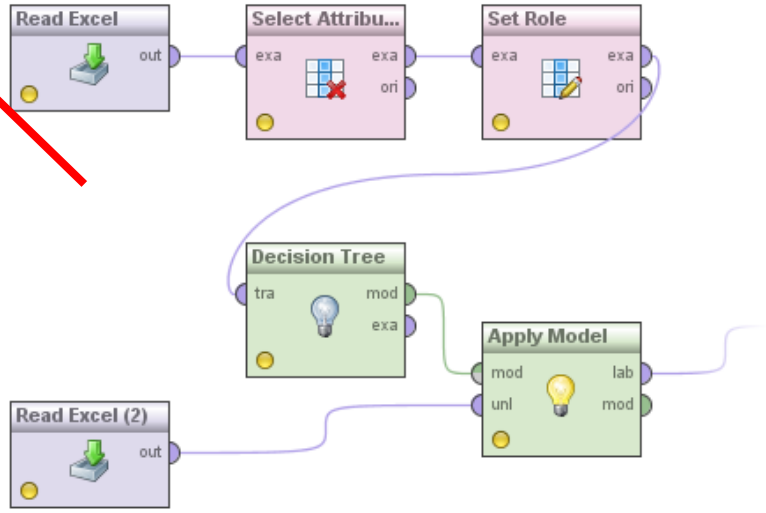


View Operator

- 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

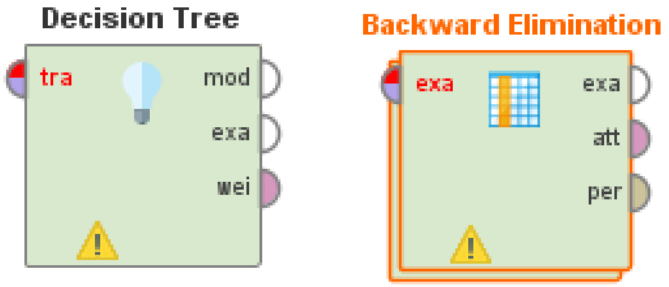
View Proses


RapidMiner@rapid-i-011
File Edit Process Tools View Help
Overview
Operators
Repositories
Process Control (34)
Utility (34)
Repository Access (2)
Import (24)
Export (17)
Data Transformation (110)
Modeling (108)
Evaluation (28)
Process
XML
Main Process
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

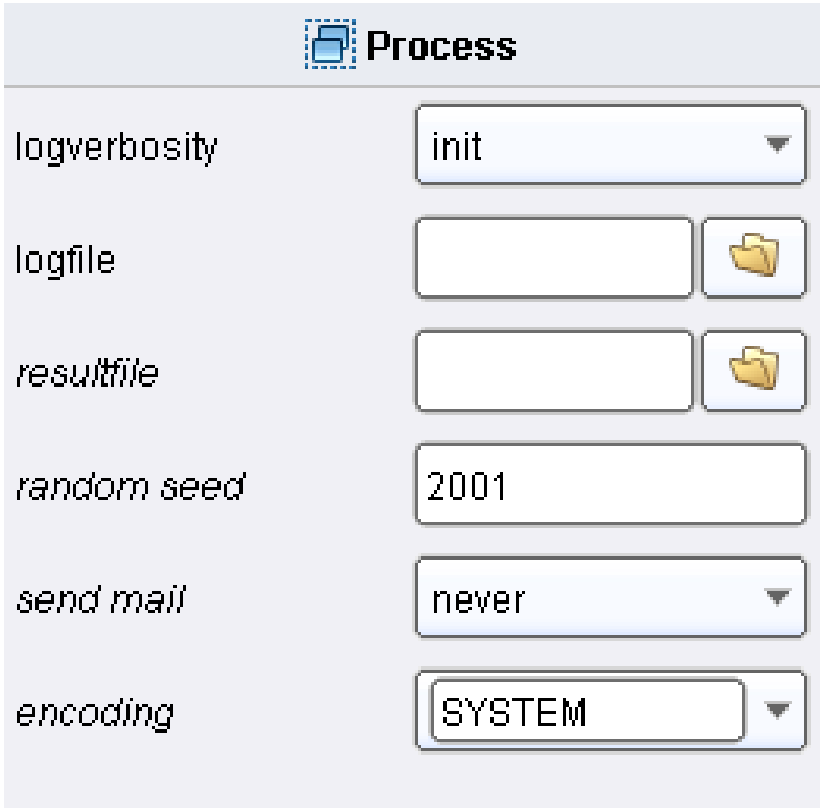


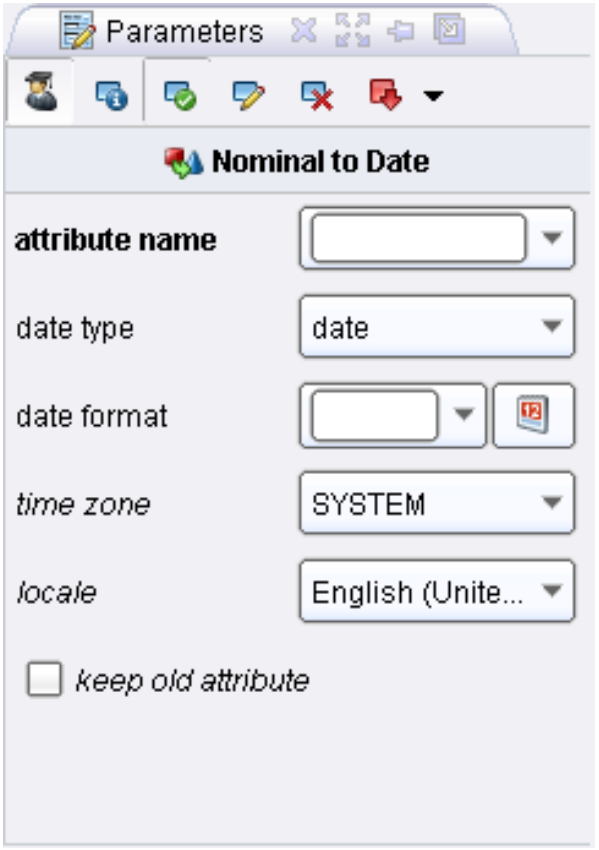
- Sebuah operator bisa disambungkan melalui port masukan (kiri) dan port keluaran (kanan)

- 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

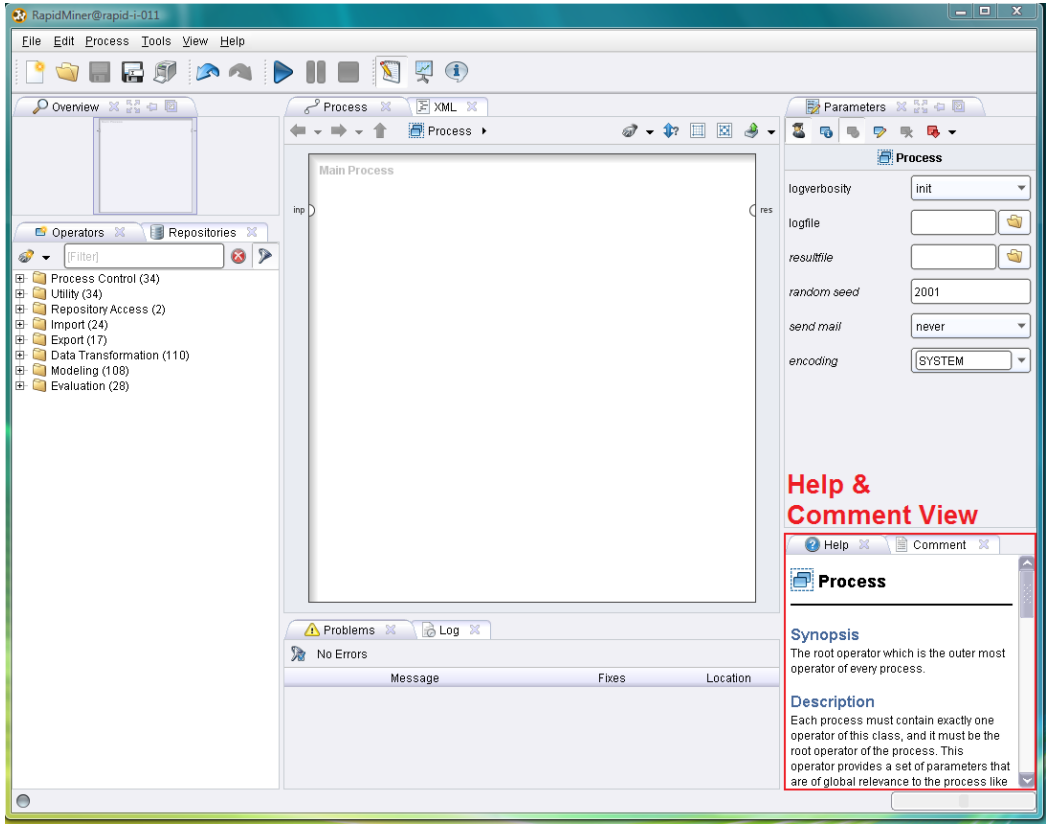
- Operator kadang memerlukan parameter untuk bisa berfungsi
- Setelah operator dipilih di view Proses, parameternya ditampilkan di view ini

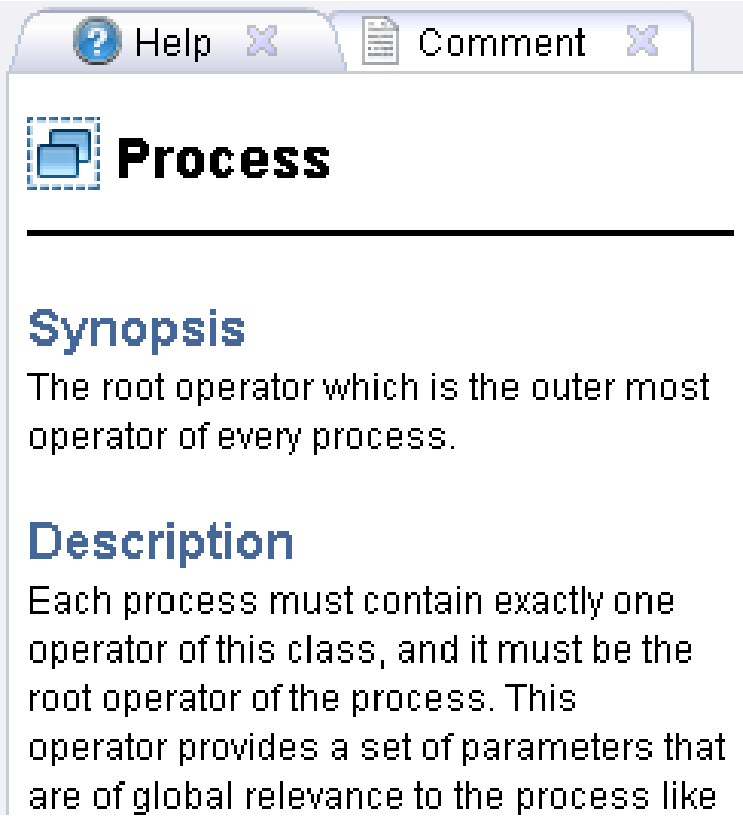




View Help dan View Comment

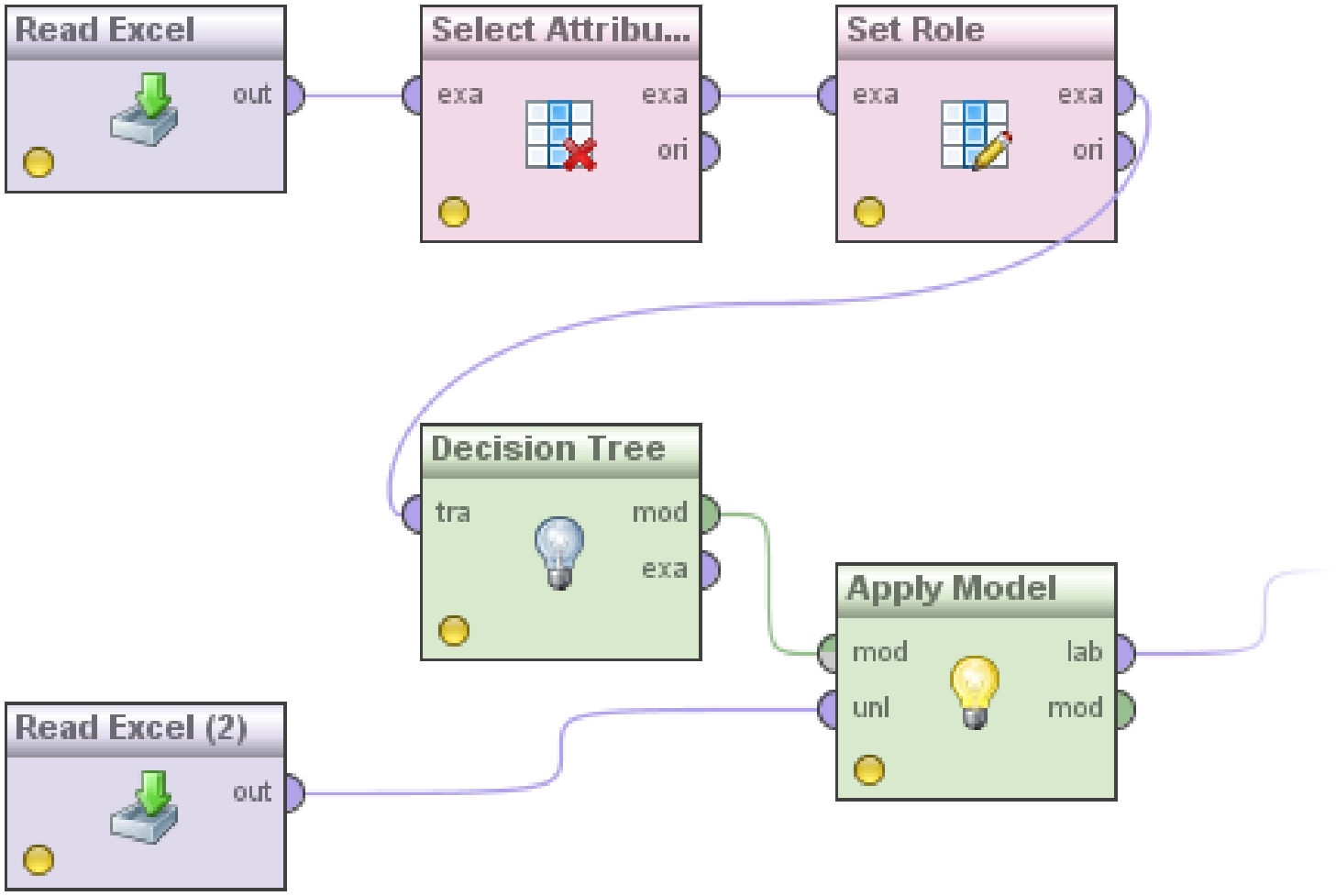
- View Help menampilkan deskripsi dari operator
- View Comment menampilkan komentar yang dapat diedit terhadap operator





Mendesain Proses

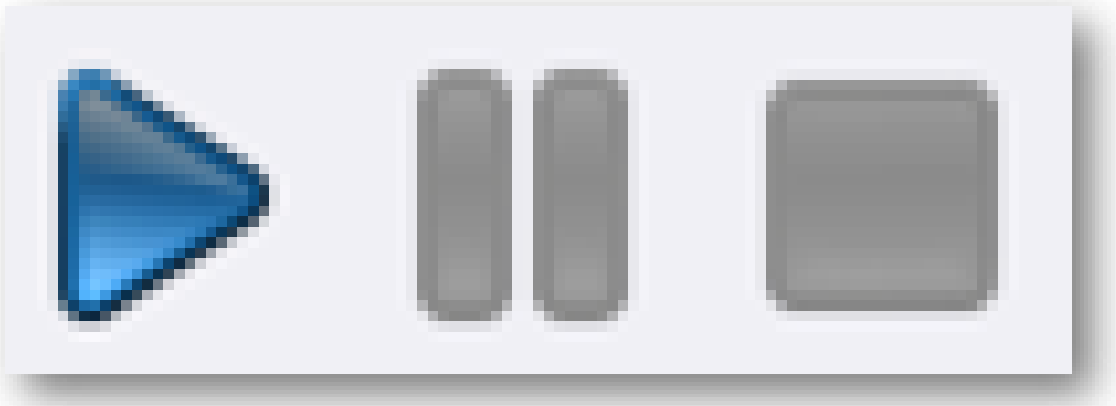
Kumpulan dan rangkaian fungsi-fungsi (operator) yang bisa disusun secara visual (visual programming)



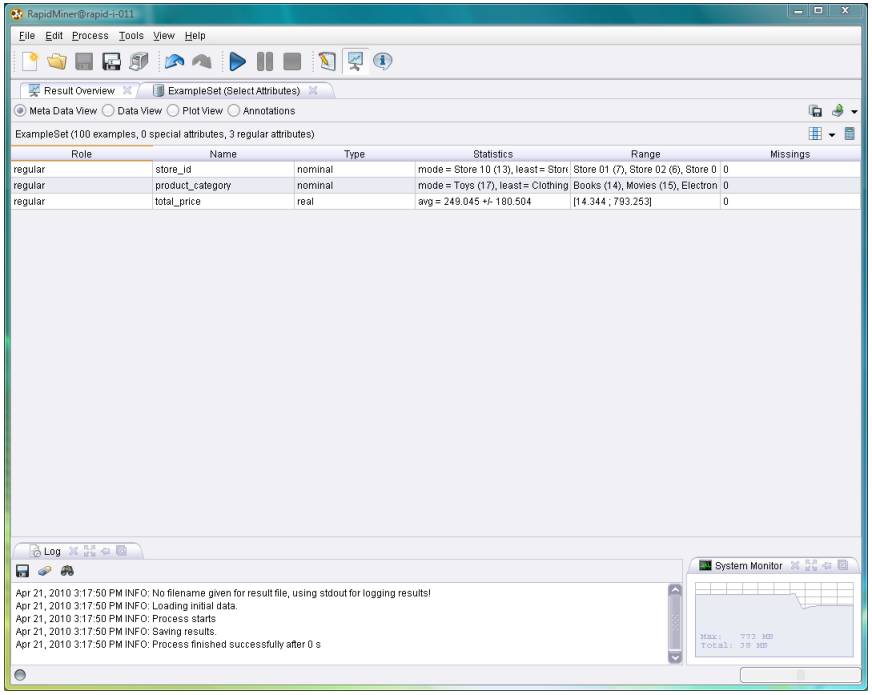
Menjalankan Proses

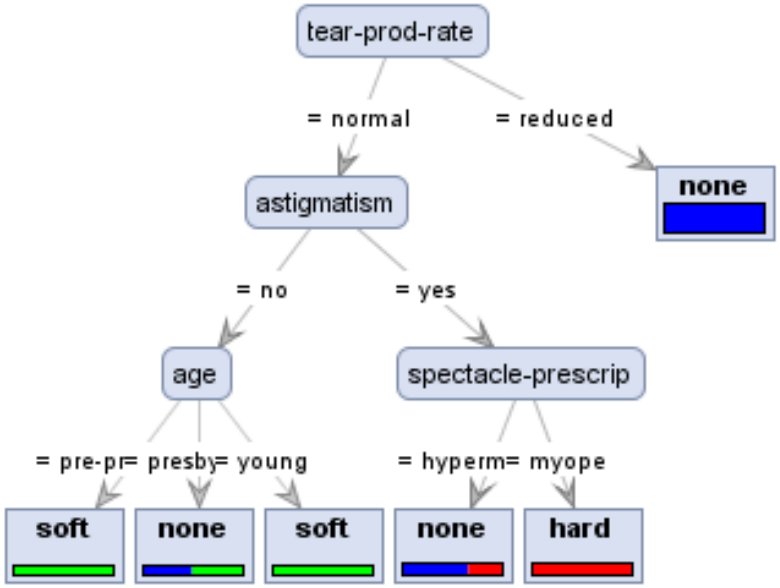
Proses dapat dijalankan dengan:

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

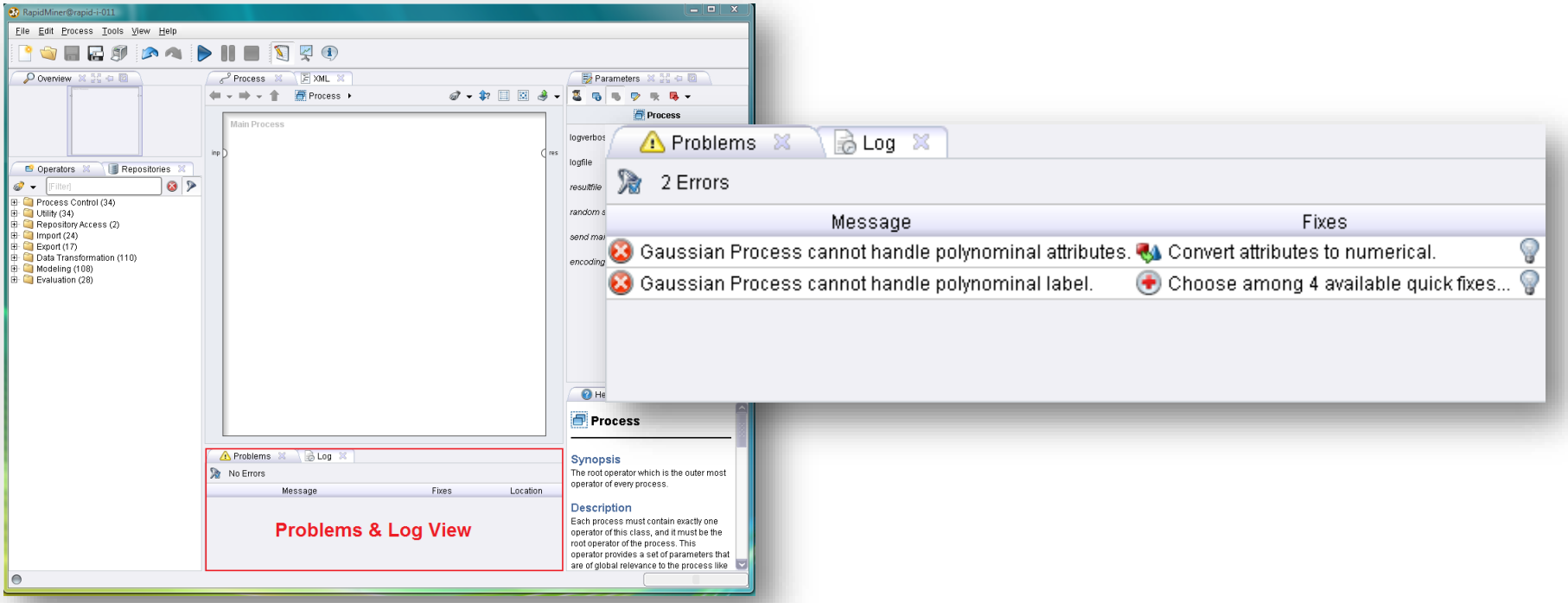


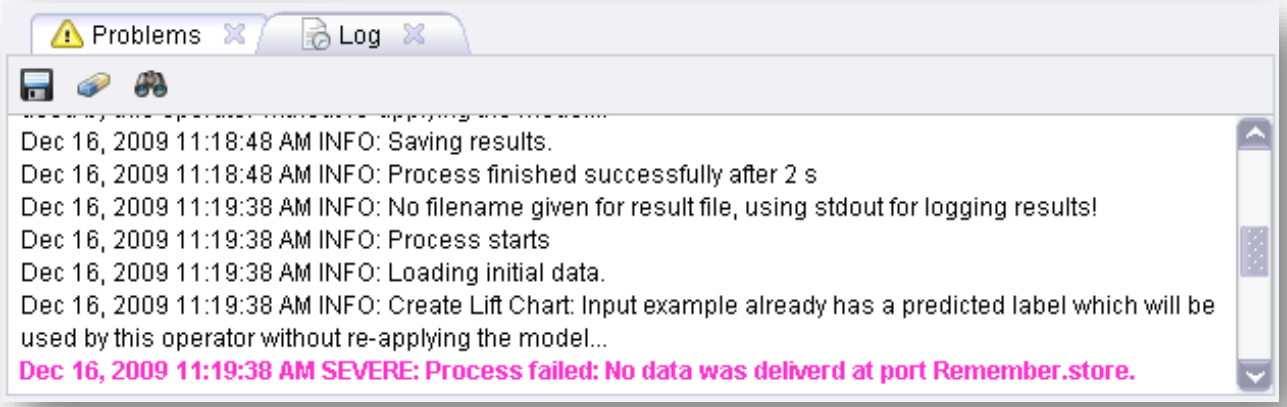
Melihat Hasil





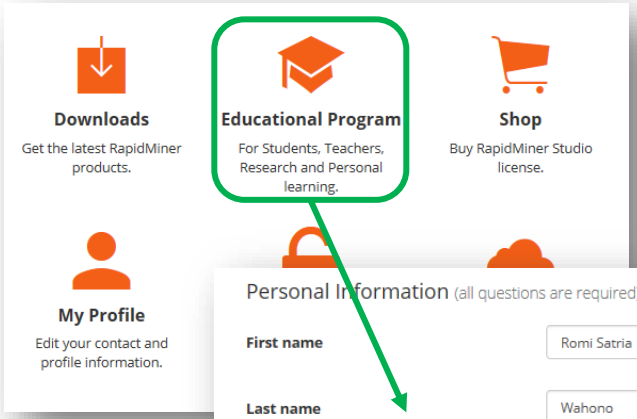
View Problems dan View Log

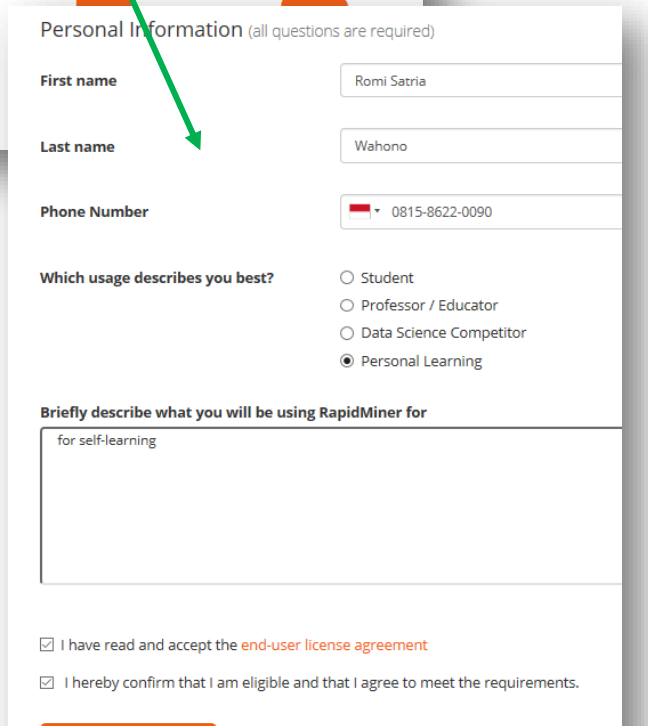


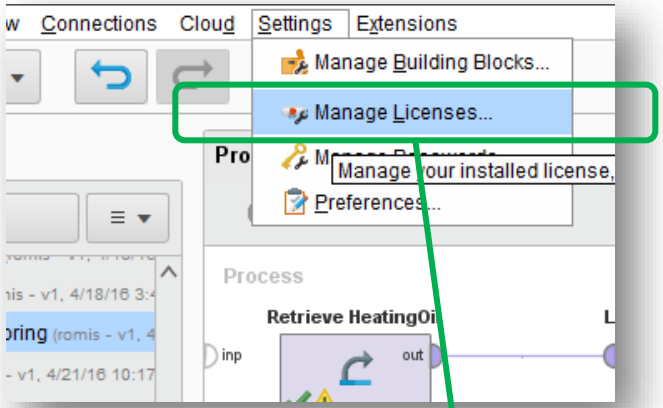


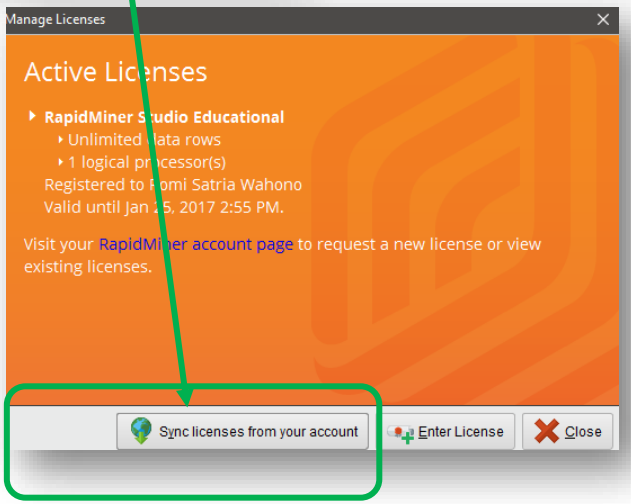
Instalasi dan Registrasi Lisensi Rapidminer

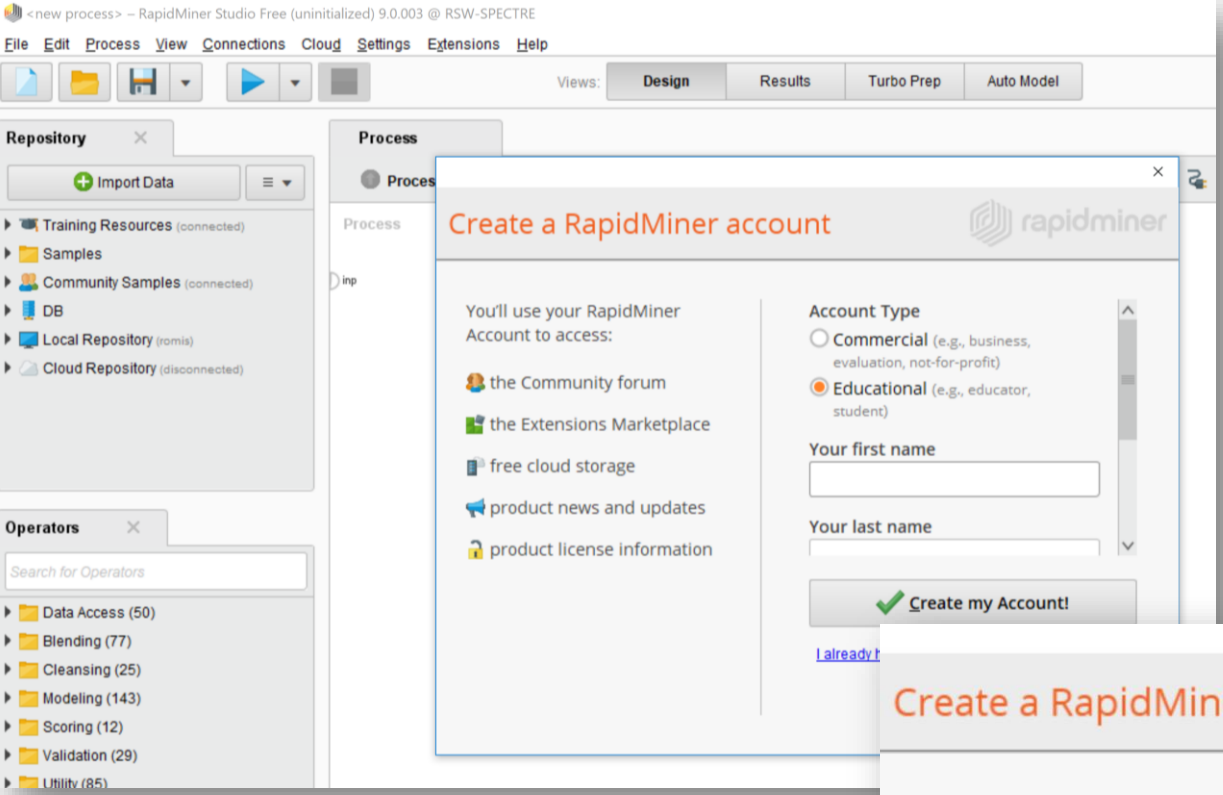
- Instal Rapidminer versi 9
- Registrasi account di rapidminer.com dan dapatkan lisensi Educational Program untuk mengolah data tanpa batasan record

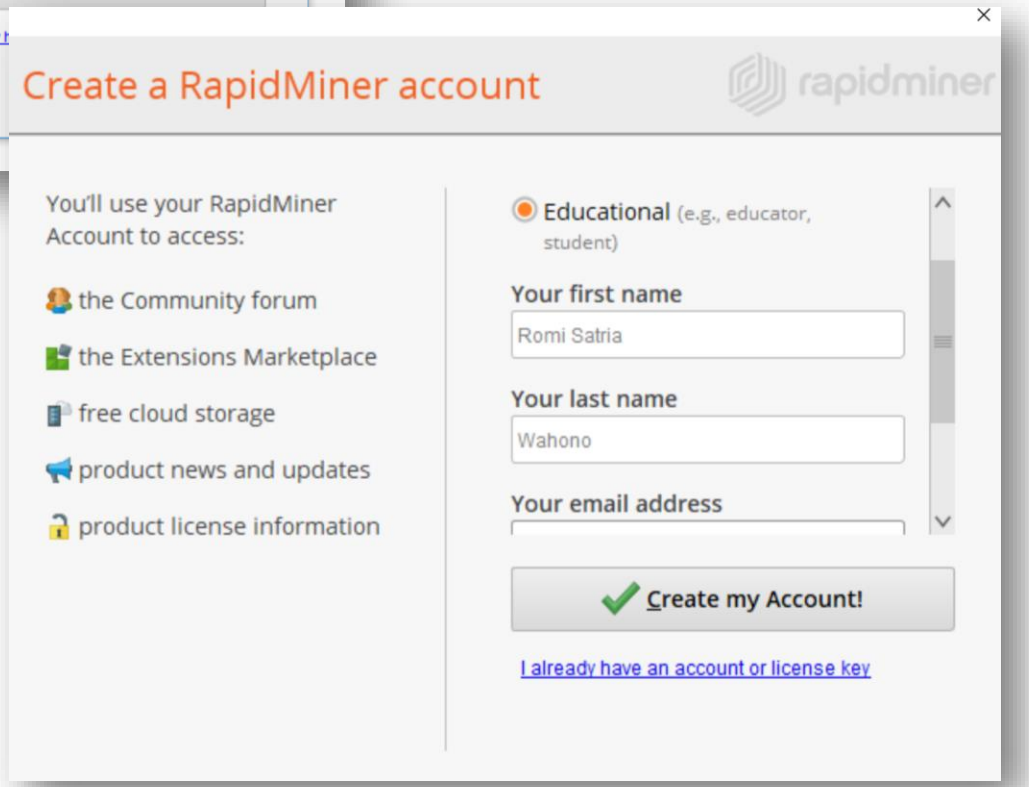






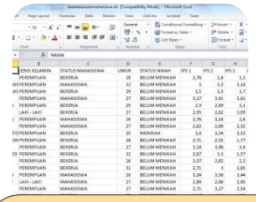



<new process> – RapidMiner Studio Free (uninitialized) 9.0.003 @ RSW-SPECTRE
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
DB
Local Repository (romis)
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

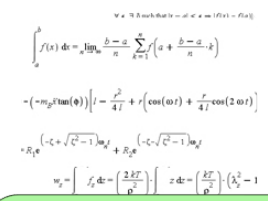
2.2 Penerapan Proses Data Mining

Proses Data Mining



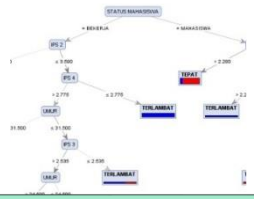
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



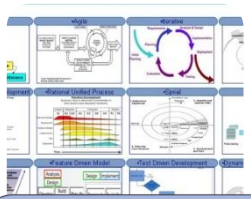
2. Metode Data Mining

(**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

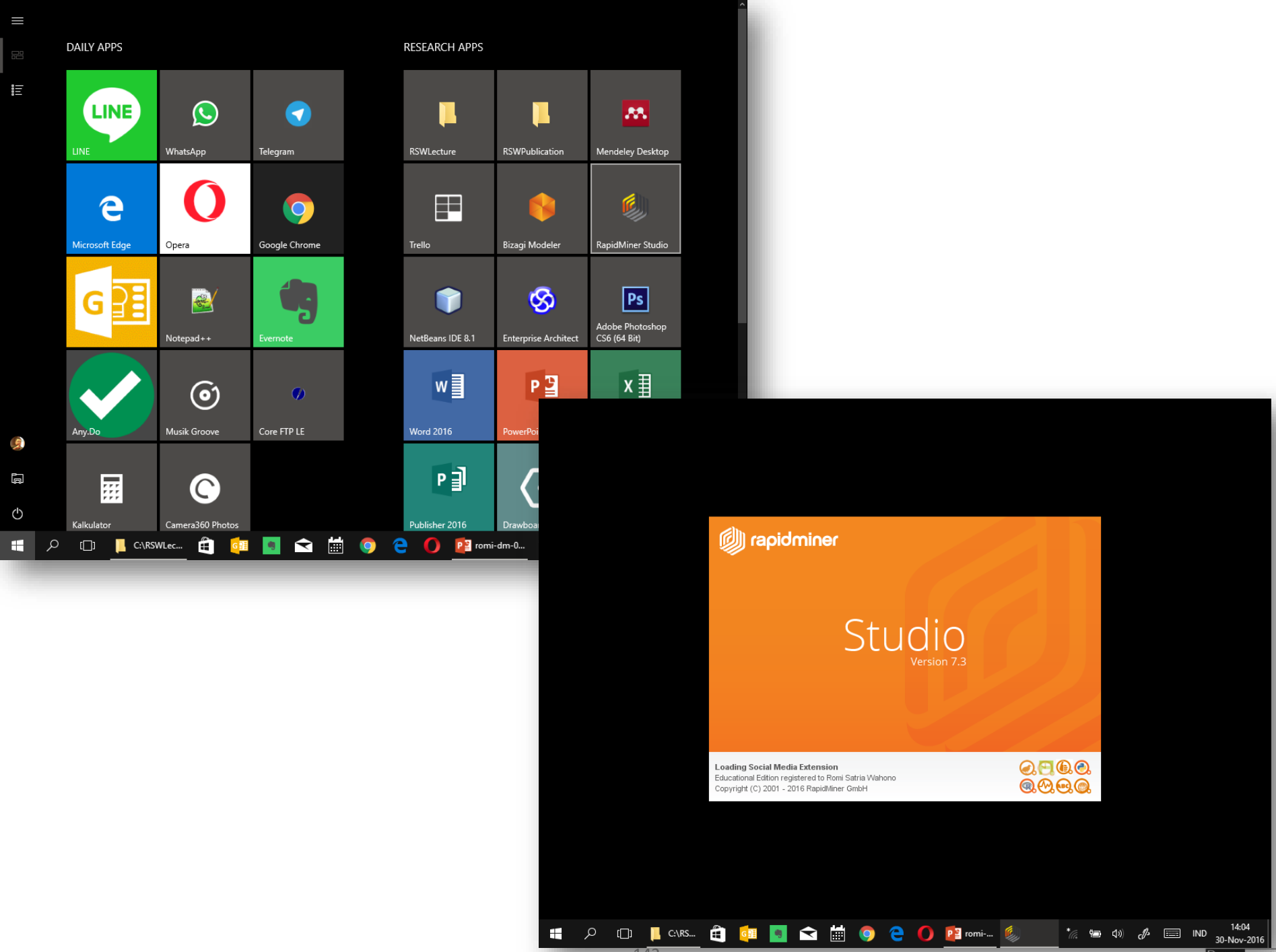
MODEL

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



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

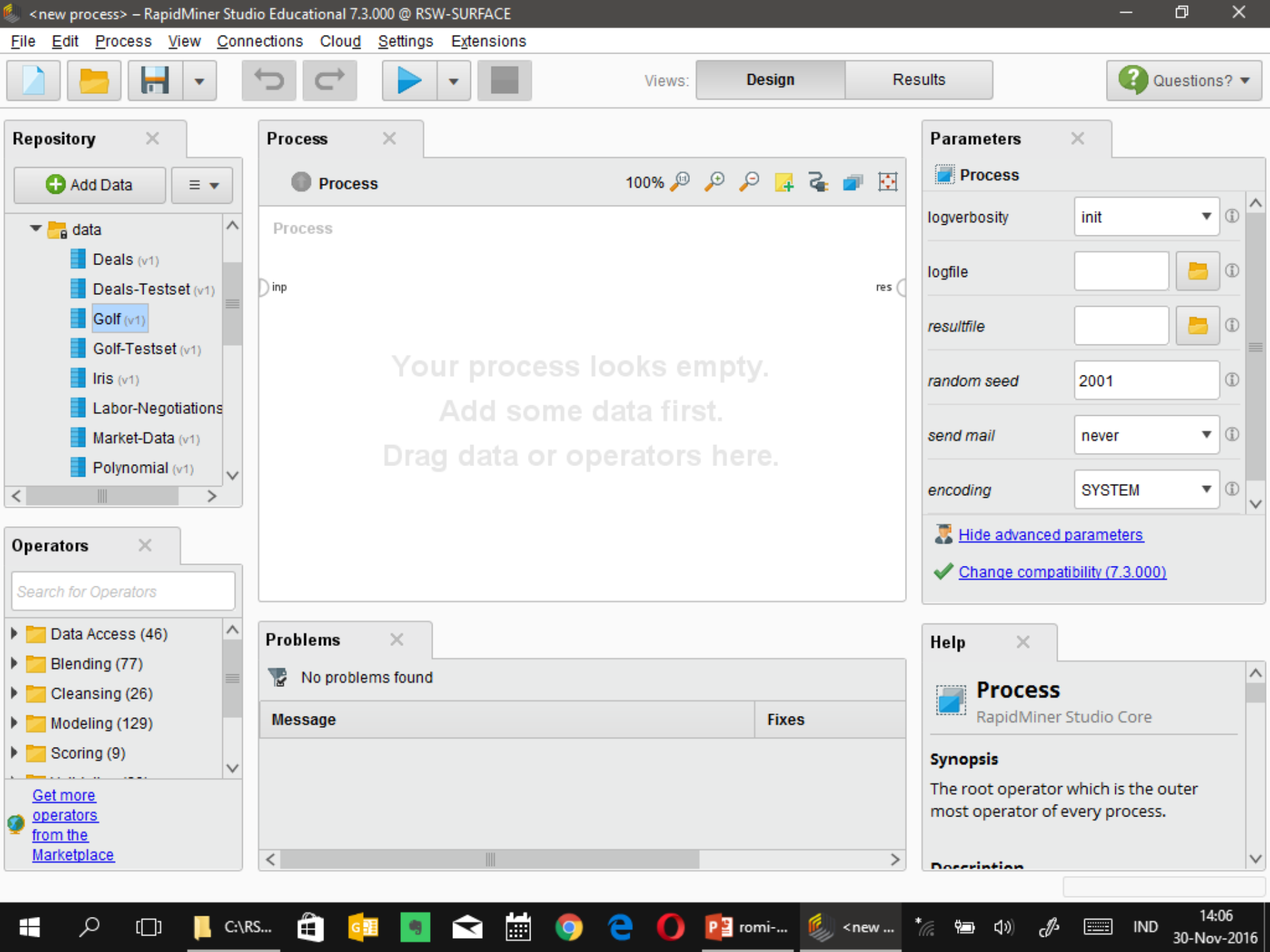
Model equipment failures to schedule maintenance pre-emptively.

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)
Operators
Search for Operators
Data Access (46)
Blending (77)
Cleansing (26)
Modeling (129)
Scoring (9)
Get more operators from the Marketplace
Process
Process
100%
Process
inp
res
Your process looks empty.
Add some data first.
Drag data or operators here.
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
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

Result History | ExampleSet (Retrieve Golf)

Data | Statistics | Charts | Advanced Charts | Annotations

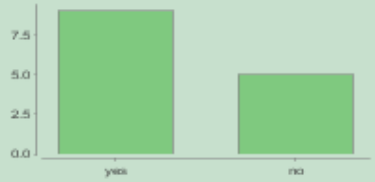
ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

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

Result History | ExampleSet (Retrieve Golf)

Data
Statistics
Charts
Advanced Charts
Annotations

Name	Type	Missing	Statistics		
Label **Play**	Nominal	0	 Open chart	Least no (5)	Most yes (9)
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 (

Filter (5 / 5 attributes): Search for Attributes

Showing attributes 1 - 5

Examples: 14 Special Attributes: 1 Regular Attributes: 4

<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
> 77.50
≤ 77.500
yes
no
no
yes
IND
14:10
30-Nov-2016

Result History | ExampleSet (Retrieve Golf) | **Tree (Decision Tree)**

Graph

Description

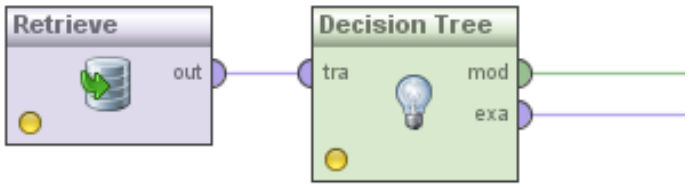
Annotations

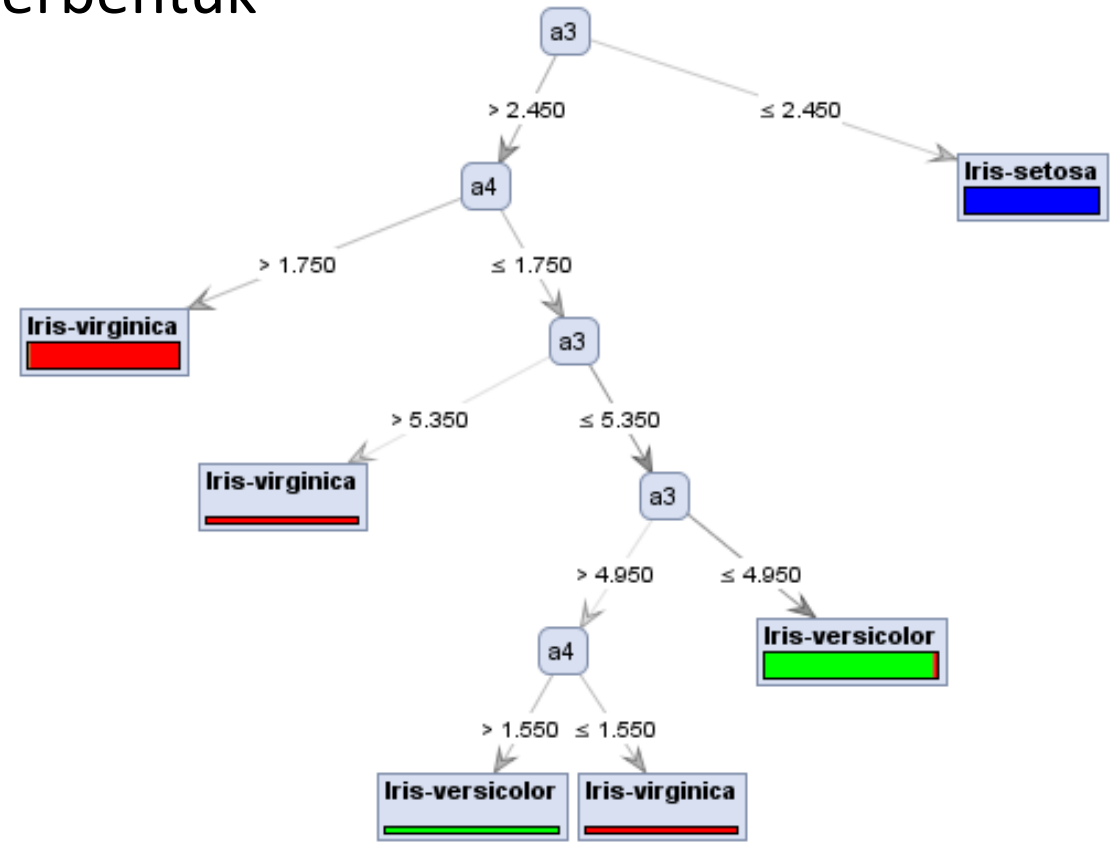
Tree

```
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}
```

Latihan: Penentuan Jenis Bunga Iris

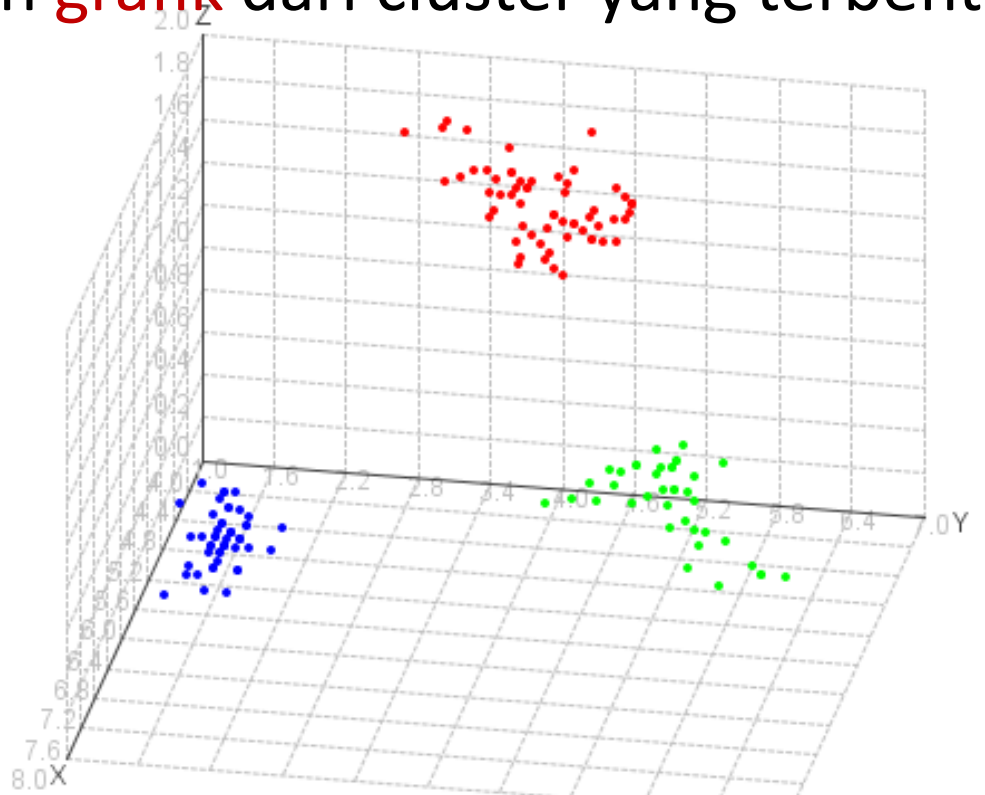
1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk





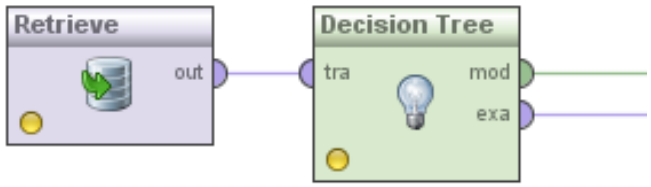
Latihan: Klastering Jenis Bunga Iris

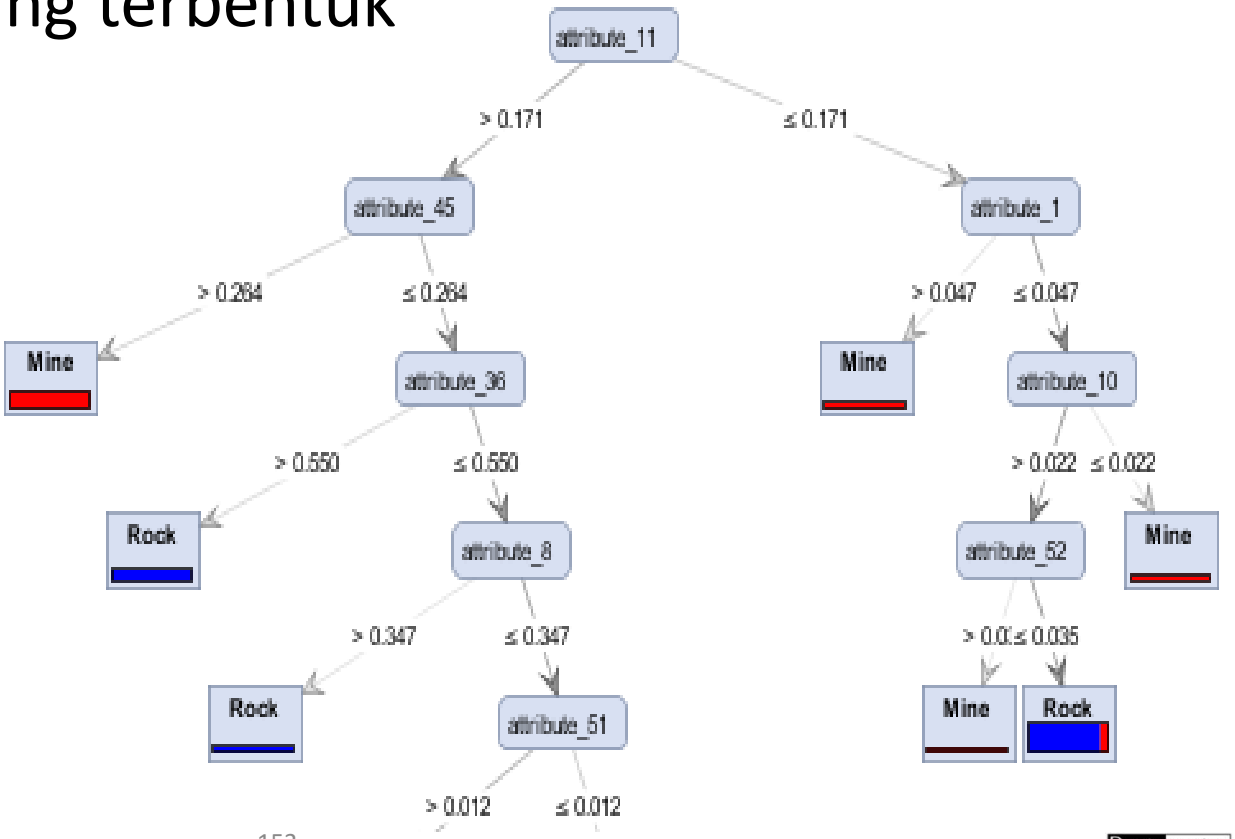
1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan algoritma k-Means
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk
3. Tampilkan grafik dari cluster yang terbentuk



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

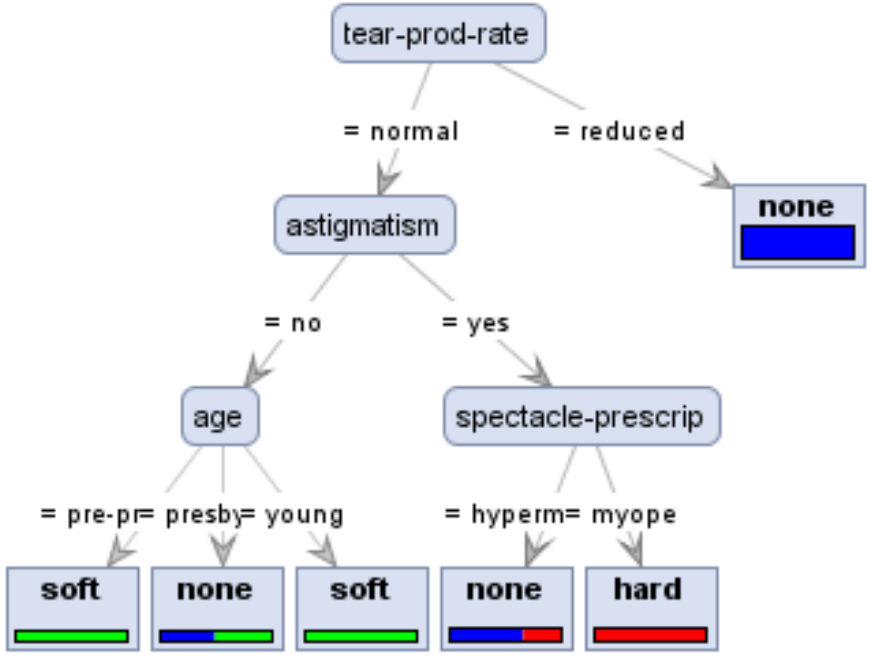




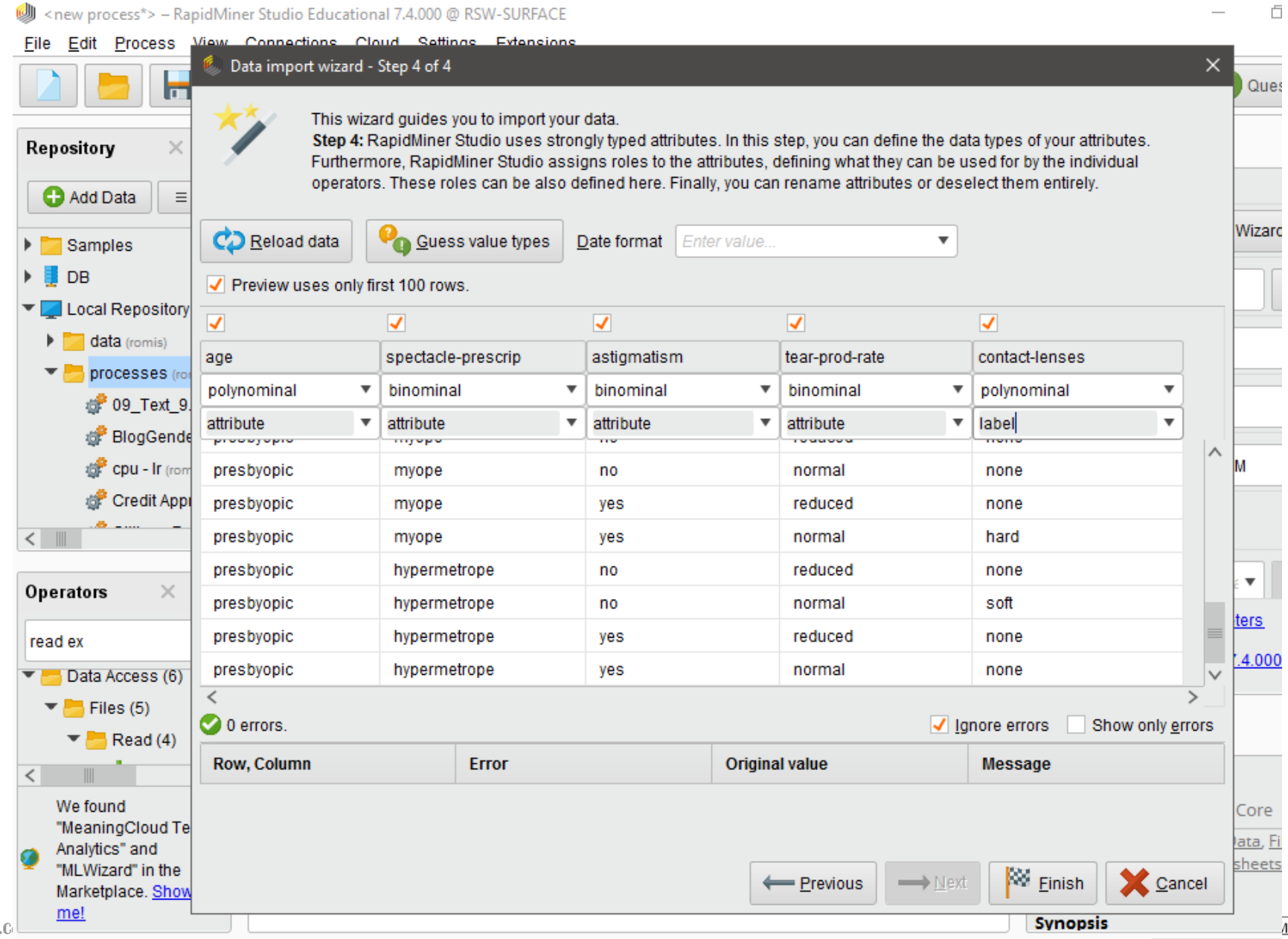
Latihan: Rekomendasi Contact Lenses

1. Lakukan training pada data Contact Lenses (contact-lenses.xls) dengan menggunakan algoritma decision tree
2. Gunakan operator Read Excel (*on the fly*) atau langsung menggunakan fitur Import Data (*persistent*)
3. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk

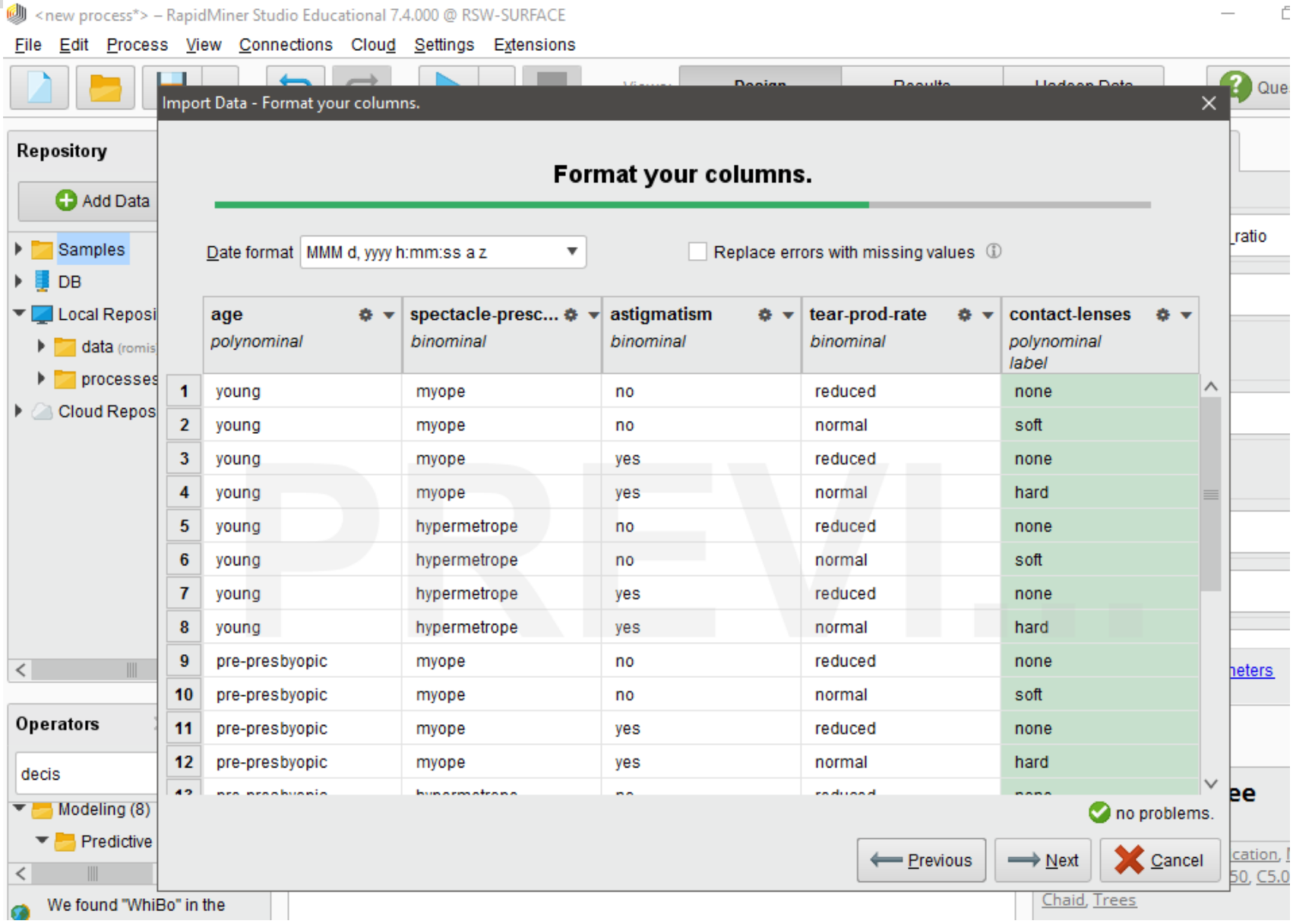
Row No.	contact-len...	age	spectacle-p...	astigmatism	tear-prod-rate
1	none	young	myope	no	reduced
2	soft	young	myope	no	normal
3	none	young	myope	yes	reduced
4	hard	young	myope	yes	normal
5	none	young	hypermetrop	no	reduced
6	soft	young	hypermetrop	no	normal
7	none	young	hypermetrop	yes	reduced
8	hard	young	hypermetrop	yes	normal
9	none	pre-presbyo	myope	no	reduced
10	soft	pre-presbyo	myope	no	normal
11	none	pre-presbyo	myope	yes	reduced
12	hard	pre-presbyo	myope	yes	normal
13	none	pre-presbyo	hypermetrop	no	reduced
14	soft	pre-presbyo	hypermetrop	no	normal



Read Excel Operator



Import Data Function

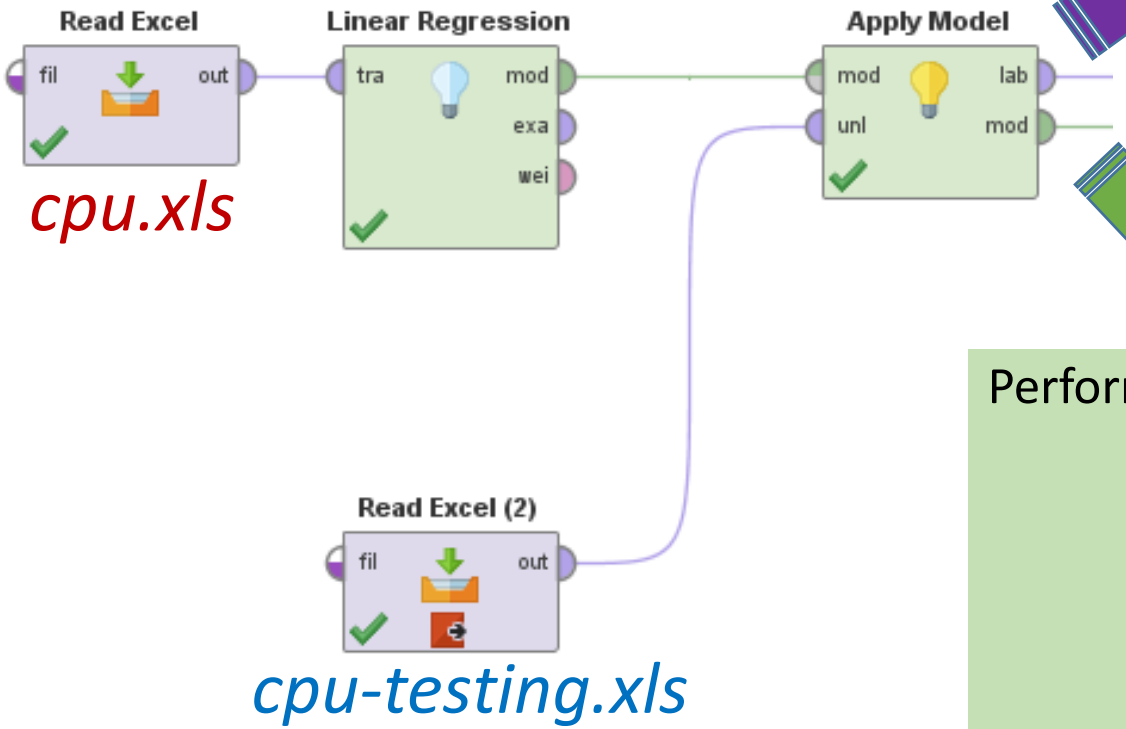


Latihan: Estimasi Performance CPU

1. Lakukan training pada data CPU (cpu.xls) dengan menggunakan algoritma linear regression
2. Lakukan pengujian terhadap data baru (cpu-testing.xls), untuk model yang dihasilkan dari tahapan 1. Data baru berisi 10 setting konfigurasi, yang belum diketahui berapa performancenya
3. Amati hasil estimasi performance dari 10 setting konfigurasi di atas

	A	B	C	D	E	F
1	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX
2	480.0	1000.0	4000.0	.0	.0	.0
3	30.0	8000.0	64000.0	128.0	12.0	176.0
4	180.0	262.0	4000.0	.0	1.0	3.0
5	180.0	512.0	4000.0	.0	1.0	3.0
6	180.0	262.0	4000.0	.0	1.0	3.0
7	180.0	512.0	4000.0	.0	1.0	3.0
8	124.0	1000.0	8000.0	.0	1.0	8.0
9	98.0	1000.0	8000.0	32.0	2.0	8.0
10	125.0	2000.0	8000.0	.0	2.0	14.0
11	480.0	512.0	8000.0	32.0	.0	.0

Estimasi Performace cpu-testing.xls



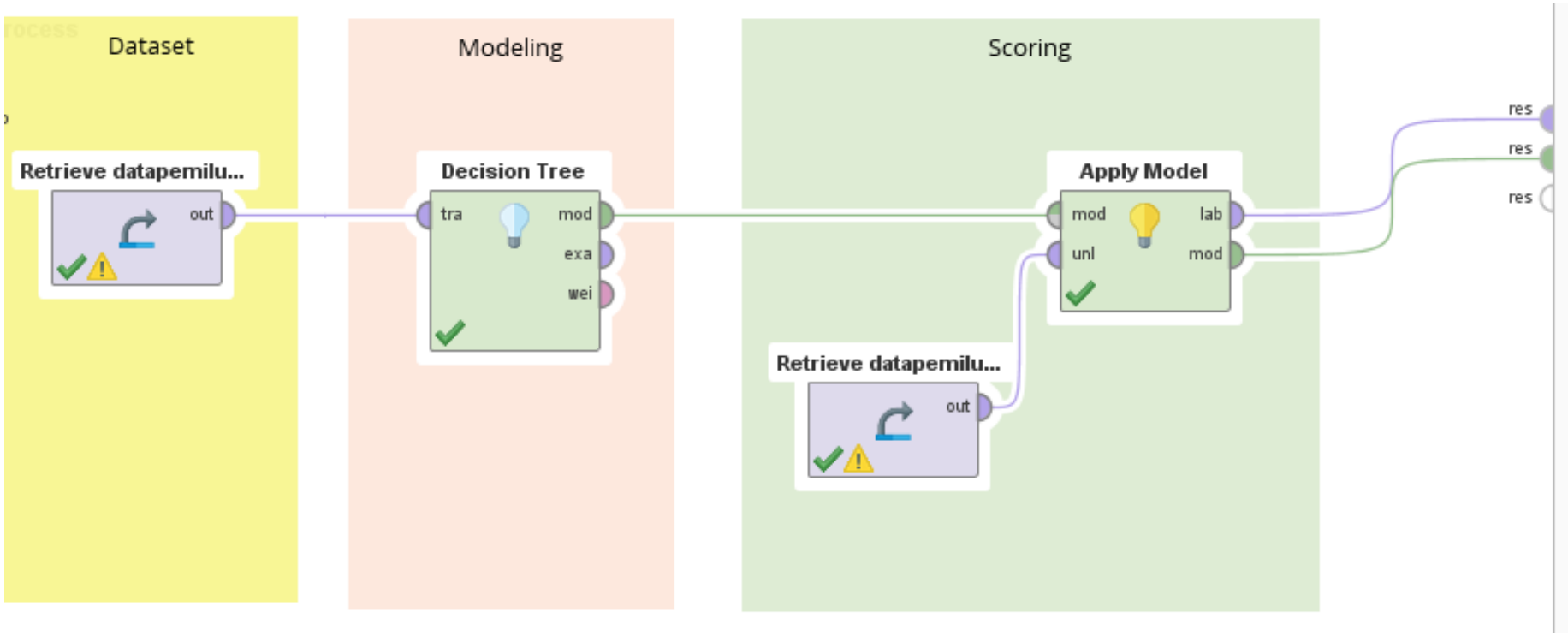
Row No.	prediction(cl...	MYCT	MMIN
1	9.025	480	1000
2	628.047	30	8000
3	-10.854	180	262
4	-6.669	180	512
5	-10.854	180	262
6	-6.669	180	512
7	21.681	124	1000
8	41.254	98	1000

Performance CPU = 0.038 * MYCT
+ 0.017 * MMIN
+ 0.004 * MMAX
+ 0.603 * CACH
+ 1.291 * CHMIN
+ 0.906 * CHMAX
- 43.975

Latihan: Prediksi Elektabilitas Caleg

1. Lakukan training pada data pemilu (datapemilukpu.xls) dengan algoritma yang tepat
2. Data bisa ditarik dari Import Data atau operator Read Excel
3. Tampilkan himpunan data (dataset) dan pengetahuan (pola/model) yang terbentuk
4. Gunakan model yang dihasilkan untuk memprediksi datapemilukpu-testing.xls

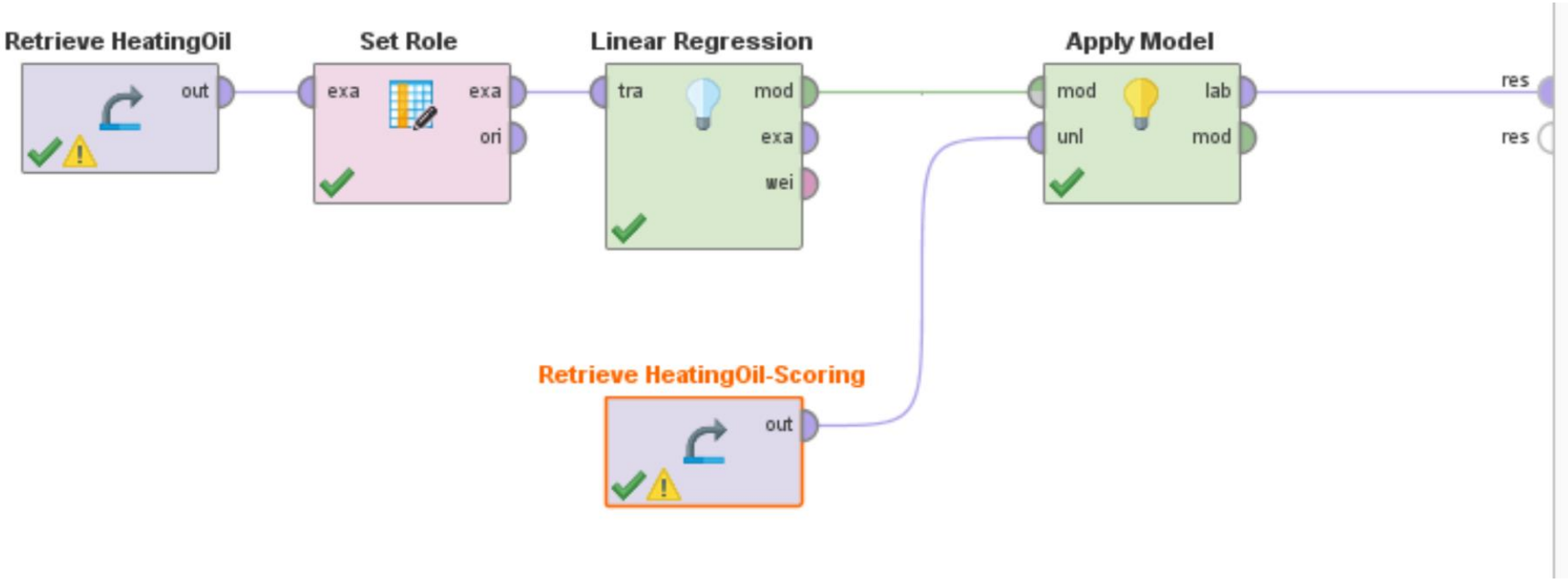
Proses Prediksi Elektabilitas Caleg



Latihan: Estimasi Konsumsi Minyak

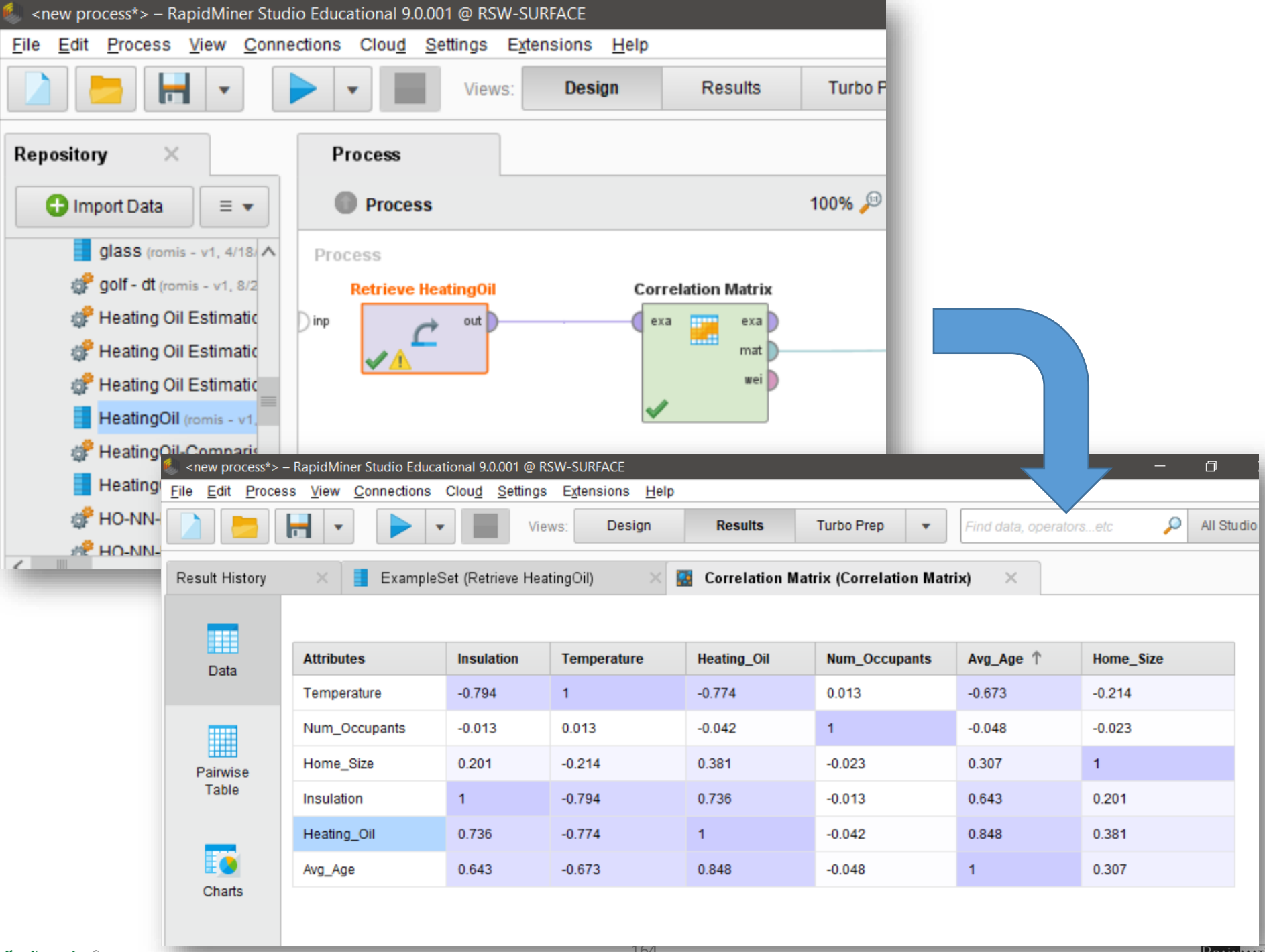
1. Lakukan training pada data konsumsi minyak (HeatingOil.csv)
   - Dataset jumlah konsumsi minyak untuk alat pemanas ruangan di rumah pertahun perrumah
   - Atribut:
     - Insulation: Ketebalan insulasi rumah
     - Temperatur: Suhu udara sekitar rumah
     - Heating Oil: Jumlah konsumsi minyak pertahun perrumah
     - Number of Occupant: Jumlah penghuni rumah
     - Average Age: Rata-rata umur penghuni rumah
     - Home Size: Ukuran rumah
2. Gunakan operator Set Role untuk memilih Label (Heating Oil), tidak langsung dipilih pada saat Import Data
3. Pilih metode yang tepat supaya menghasilkan model
4. Apply model yang dihasilkan ke data pelanggan baru di file HeatingOil-Scoring.csv, supaya kita bisa mengestimasi berapa kebutuhan konsumsi minyak mereka, untuk mengatur stok penjualan minyak

Proses Estimasi Konsumsi Minyak



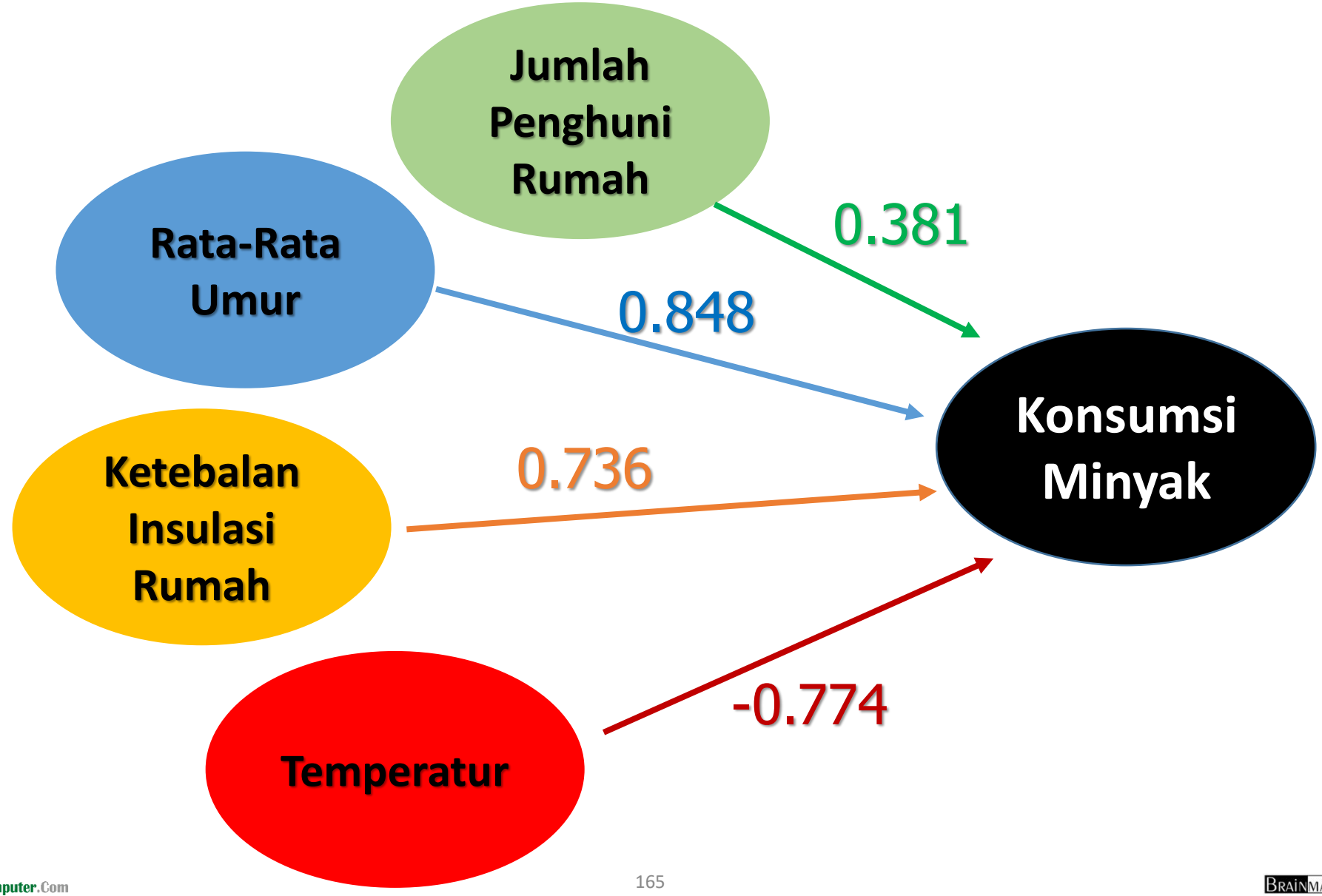
Latihan: Matrix Correlation Konsumsi Minyak

1. Lakukan training pada data konsumsi minyak (HeatingOil.csv)
   - Dataset jumlah konsumsi minyak untuk alat pemanas ruangan di rumah pertahun perrumah
   - Atribut:
     - Insulation: Ketebalan insulasi rumah
     - Temperatur: Suhu udara sekitar rumah
     - Heating Oil: Jumlah konsumsi minyak pertahun perrumah
     - Number of Occupant: Jumlah penghuni rumah
     - Average Age: Rata-rata umur penghuni rumah
     - Home Size: Ukuran rumah
2. Tujuannya ingin mendapatkan informasi tentang atribut apa saja yang paling berpengaruh pada konsumsi minyak



Attributes	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Num_Occupants	-0.013	0.013	-0.042	1	-0.048	-0.023
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307

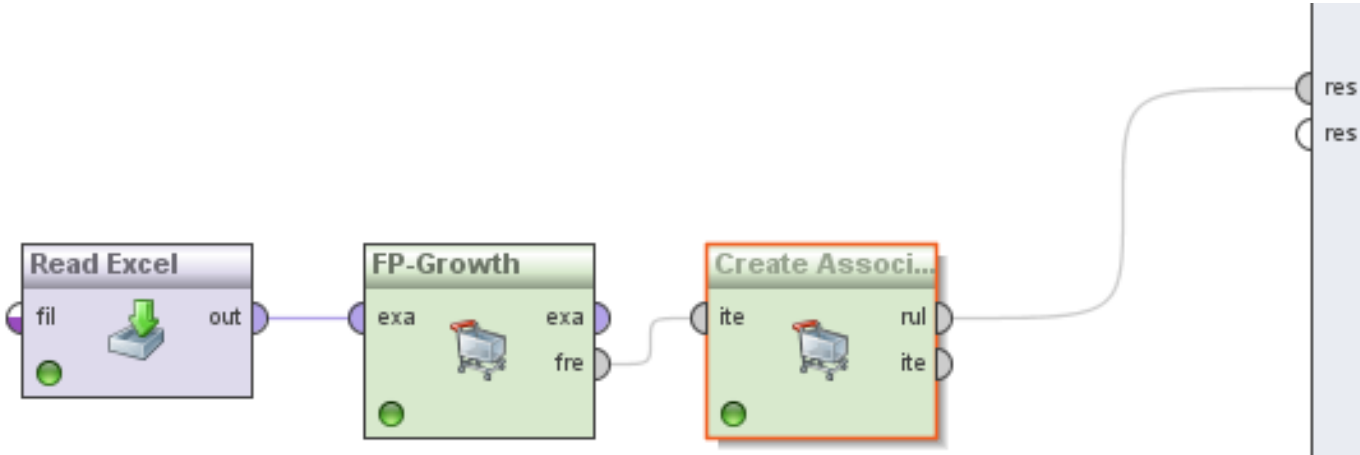
Tingkat Korelasi 4 Atribut terhadap Konsumsi Minyak



Latihan: Aturan Asosiasi Data Transaksi

1. Lakukan training pada data transaksi (transaksi.xlsx)

2. Pilih metode yang tepat supaya menghasilkan pola



Show rules matching

all of these conclusions:

Sabun
Kopi
Sampo
Gula
Sprei
Boneka

Min. Criterion:

confidence

Min. Criterion Value:

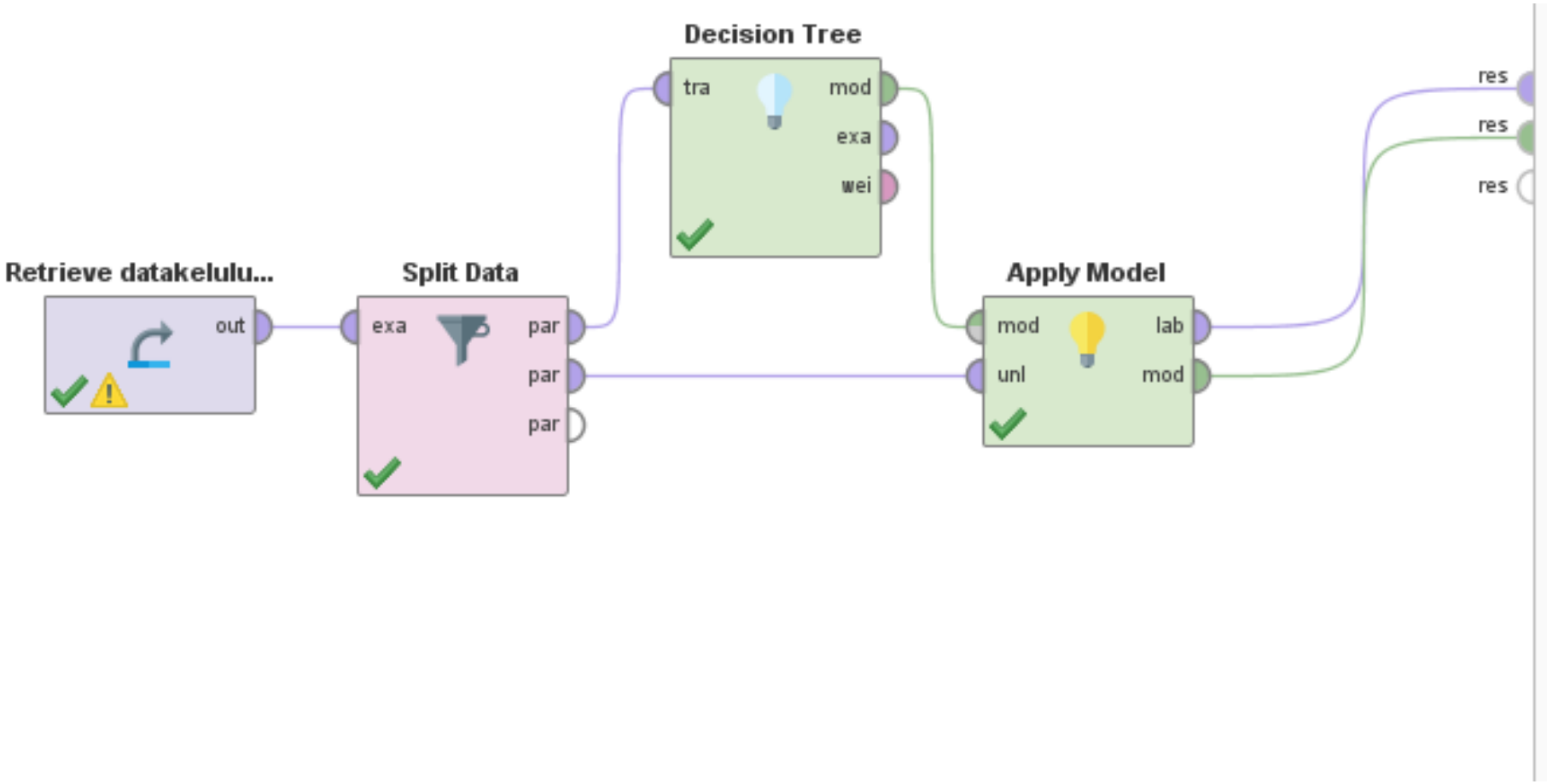
No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain
7	Sampo	Sabun	0.500	0.857	0.947	-0.667
8	Kopi	Gula	0.500	0.857	0.947	-0.667
9	Boneka	Sabun	0.250	1	1	-0.250
10	Celana	Sabun	0.250	1	1	-0.250
11	Gula	Kopi	0.500	1	1	-0.500
12	Boneka	Sampo	0.250	1	1	-0.250
13	Celana	Sampo	0.250	1	1	-0.250
14	Boneka	Sprei	0.250	1	1	-0.250
15	Kopi, Sampo	Sabun	0.250	1	1	-0.250
16	Sabun, Gula	Kopi	0.333	1	1	-0.333
17	Sabun, Sprei	Sampo	0.250	1	1	-0.250
18	Sampo, Sprei	Sabun	0.250	1	1	-0.250
19	Boneka	Sabun, Sampo	0.250	1	1	-0.250
20	Sabun, Boneka	Sampo	0.250	1	1	-0.250
21	Sampo, Boneka	Sabun	0.250	1	1	-0.250
22	Celana	Sabun, Sampo	0.250	1	1	-0.250

Latihan: Klasifikasi Data Kelulusan Mahasiswa

1. Lakukan training pada data kelulusan mahasiswa (datakelulusanmahasiswa.xls)

2. Gunakan operator Split Data untuk memecah data secara otomatis menjadi dua dengan perbandingan 0.9:0.1, di mana 0.9 untuk training dan 0.1 untuk testing

3. Pilih metode yang tepat supaya menghasilkan pola yang bisa menguji data testing 10%


Decision Tree
tra
mod
exa
wei
res
res
res
Retrieve datakelulu...
out
Split Data
exa
par
par
par
Apply Model
mod
unl
lab
mod

Latihan: Forecasting Harga Saham

1. Lakukan training pada data Harga Saham (hargasaham-training.xls) dengan menggunakan algoritma yang tepat
2. Tampilkan himpunan data (dataset) dan pengetahuan (model regresi) yang terbentuk
3. Lakukan pengujian terhadap data baru (hargasaham-testing.xls), untuk model yang dihasilkan dari tahapan 1
4. Lakukan visualisasi berupa grafik dari data yang terbentuk dengan menggunakan Line atau Spline

Result History

ExampleSet (Apply Model)

Data

Statistics

Visualizations

Annotations

Plot

Plot 1

Plot type

Spline

X-Axis column

Date

Value columns

prediction(Close), C..

Aggregate data

Color Group

-

Stacking

No stacking

Plot style

Add new plot

General

X-Axis

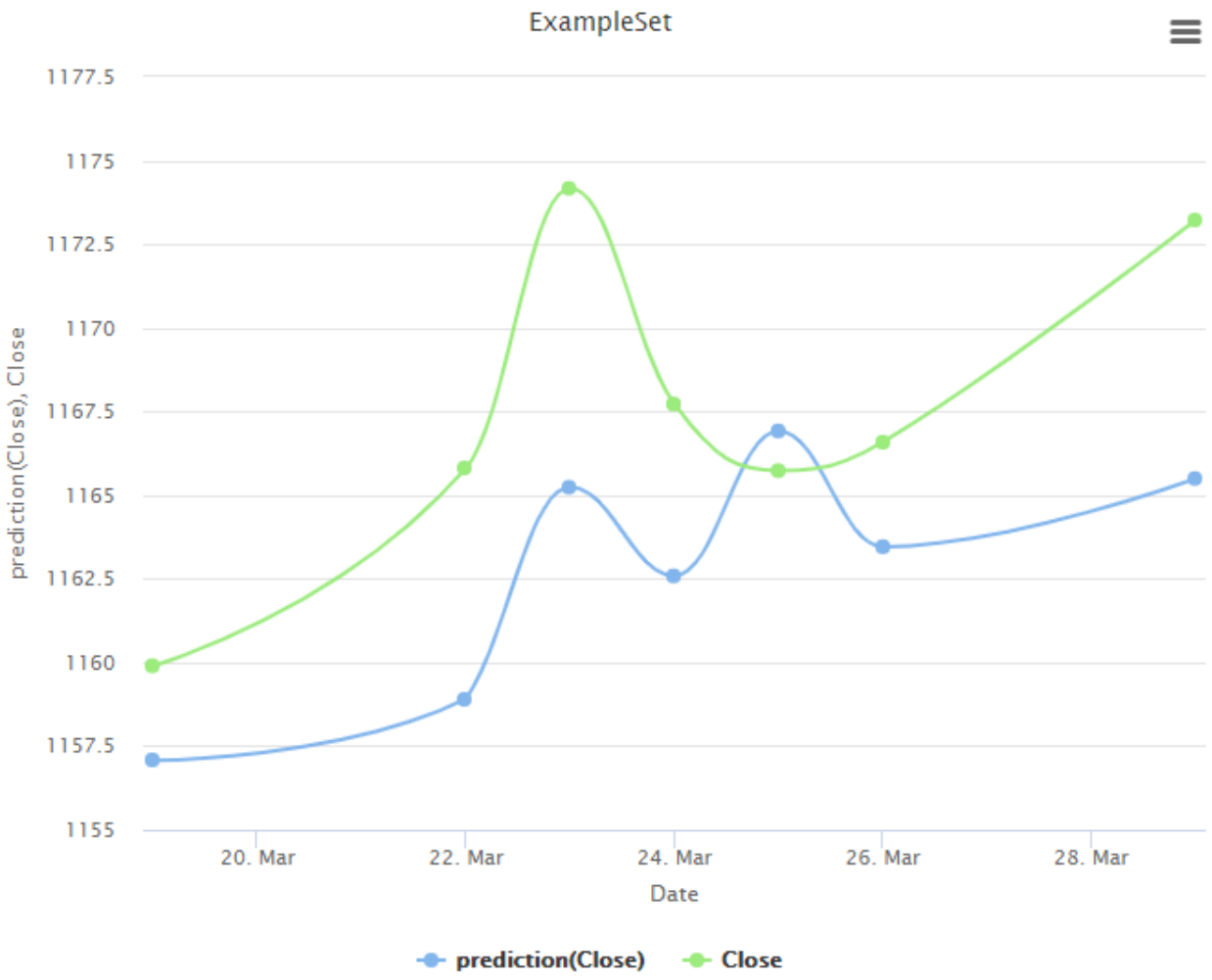


Chart style:

Scatter Multiple

x-Axis:

Date

Log scale

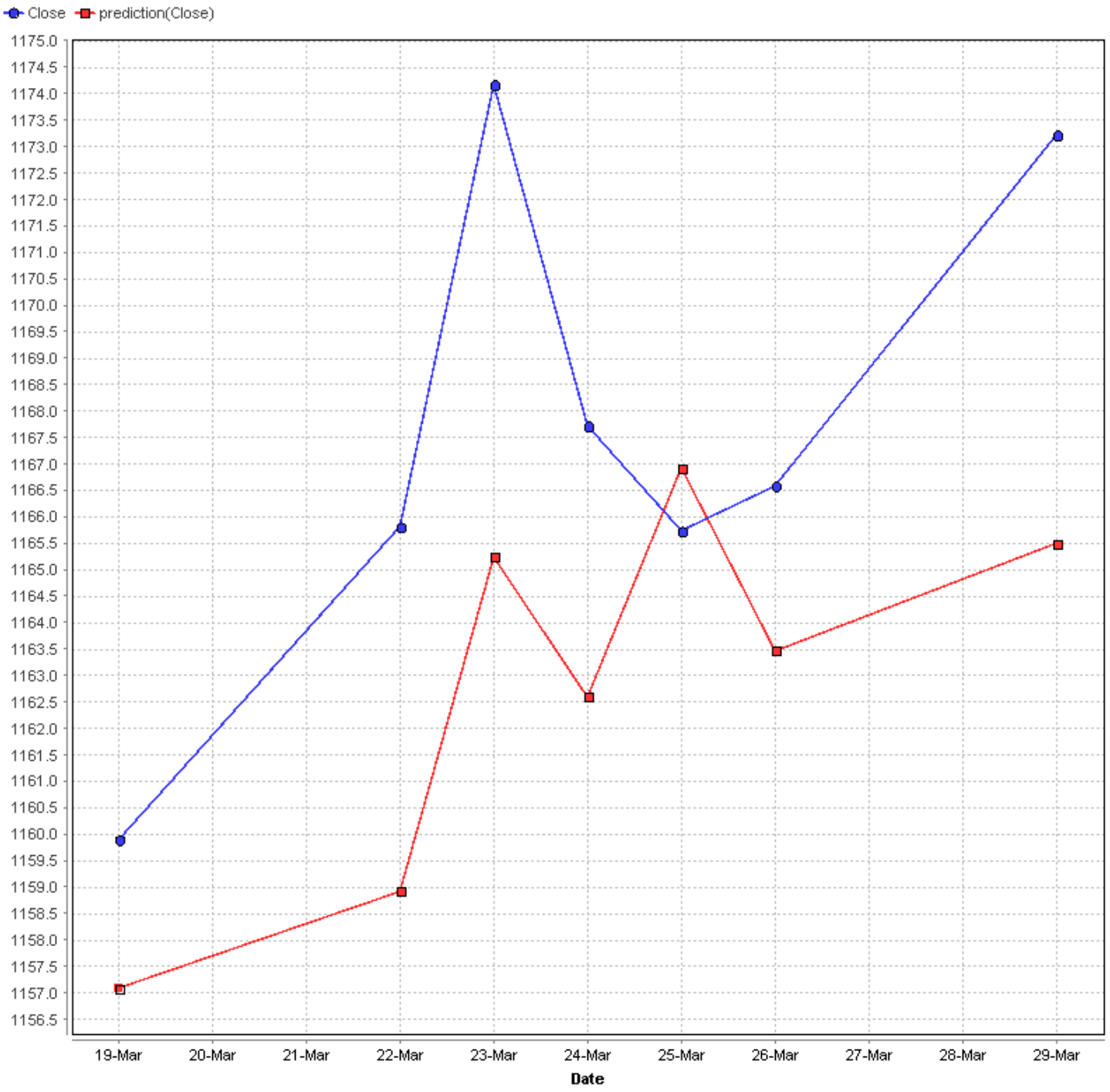
y-Axis:

Date
Open
High
Low
Volume
Close
prediction(Close)

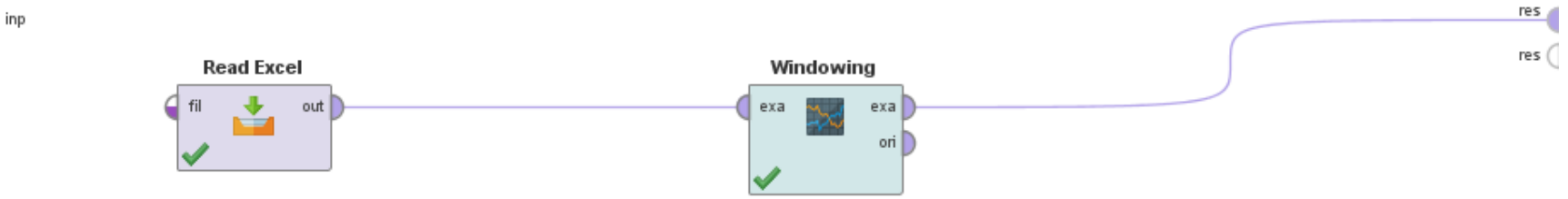
Jitter:

Rotate labels

Points and Lines...



Latihan: Forecasting Harga Saham (Univariat)



Date	inputYt
Jan 1, 2009	0.709
Feb 1, 2009	1.886
Mar 1, 2009	1.293
Apr 1, 2009	0.822
May 1, 2009	-0.173
Jun 1, 2009	0.552
Jul 1, 2009	1.169
Aug 1, 2009	1.604
Sep 1, 2009	0.949
Oct 1, 2009	0.080
Nov 1, 2009	-0.040
Dec 1, 2009	1.381
Jan 1, 2010	0.761

Date	label	inputYt-5	inputYt-4	inputYt-3	inputYt-2	inputYt-1	inputYt-0
Jun 1, 2009	1.169	0.709	1.886	1.293	0.822	-0.173	0.552
Jul 1, 2009	1.604	1.886	1.293	0.822	-0.173	0.552	1.169
Aug 1, 2009	0.949	1.293	0.822	-0.173	0.552	1.169	1.604
Sep 1, 2009	0.080	0.822	-0.173	0.552	1.169	1.604	0.949
Oct 1, 2009	-0.040	-0.173	0.552	1.169	1.604	0.949	0.080
Nov 1, 2009	1.381	0.552	1.169	1.604	0.949	0.080	-0.040
Dec 1, 2009	0.761	1.169	1.604	0.949	0.080	-0.040	1.381
Jan 1, 2010	2.312	1.604	0.949	0.080	-0.040	1.381	0.761
Feb 1, 2010	1.795	0.949	0.080	-0.040	1.381	0.761	2.312
Mar 1, 2010	0.586	0.080	-0.040	1.381	0.761	2.312	1.795
Apr 1, 2010	-0.077	-0.040	1.381	0.761	2.312	1.795	0.586
May 1, 2010	0.613	1.381	0.761	2.312	1.795	0.586	-0.077

Window size = 6
Step size = 1
Horizon = 1

Using data from 6 rows (Jan 2009 – Jun 2009) of the window, a learner can be trained to predict the label which is the value of the time series in the next time step (Jul 2009) and so on.

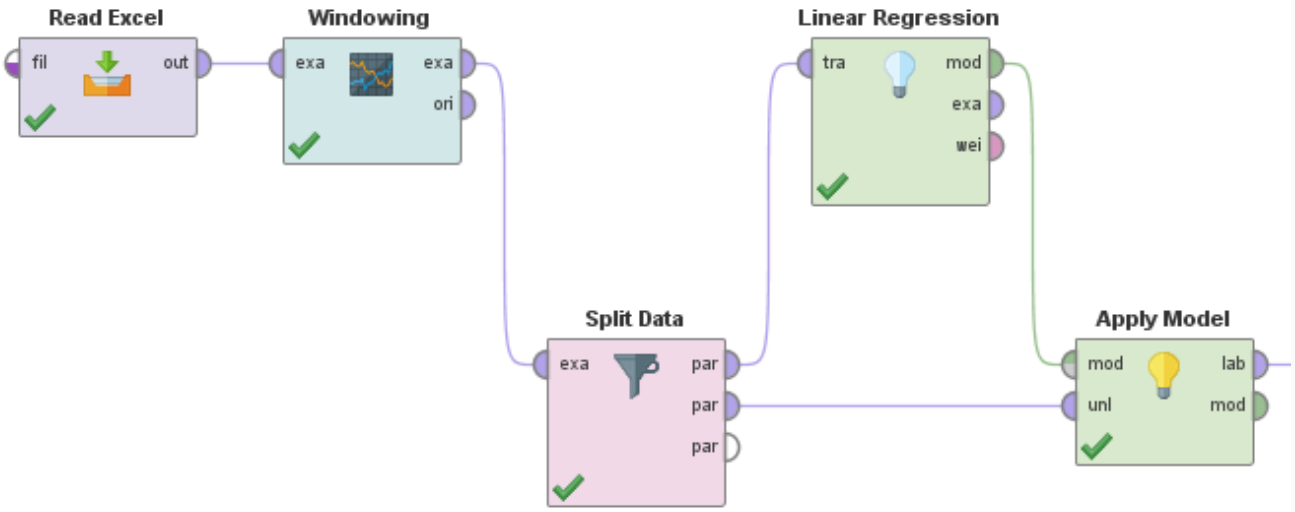
Parameter dari Windowing

- Window size: Determines how many "attributes" are created for the cross-sectional data
  - Each row of the original time series within the window width will become a new attribute
  - We choose w = 6
- Step size: Determines how to advance the window
  - Let us use s = 1
- Horizon: Determines how far out to make the forecast
  - If the window size is 6 and the horizon is 1, then the seventh row of the original time series becomes the first sample for the "label" variable
  - Let us use h = 1

Latihan

- Lakukan training dengan menggunakan linear regression pada dataset hargasaham-training-uni.xls
- Gunakan Split Data untuk memisahkan dataset di atas, 90% training dan 10% untuk testing
- Harus dilakukan proses Windowing pada dataset
- Plot grafik antara label dan hasil prediksi dengan menggunakan chart

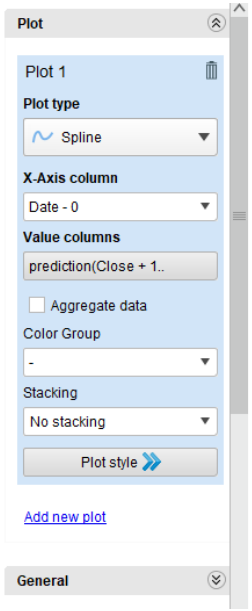
Forecasting Harga Saham (Data Lampau)



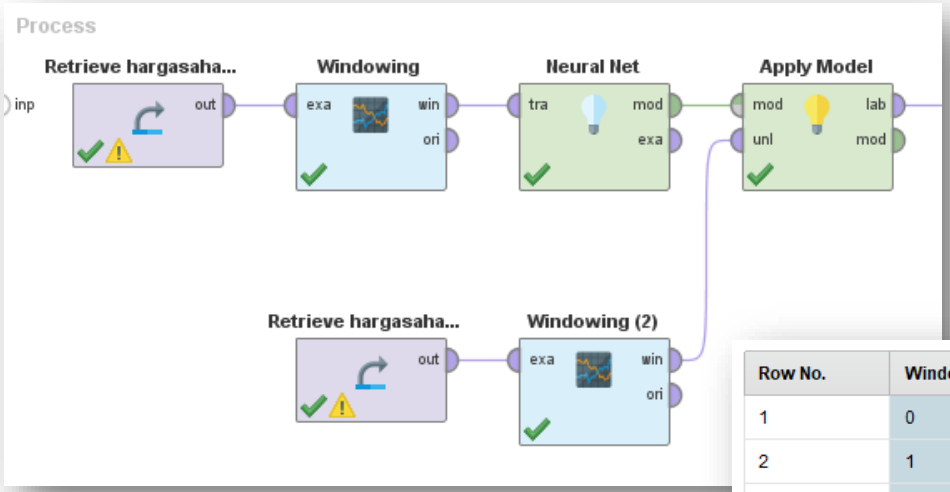
Parameters

Windowing

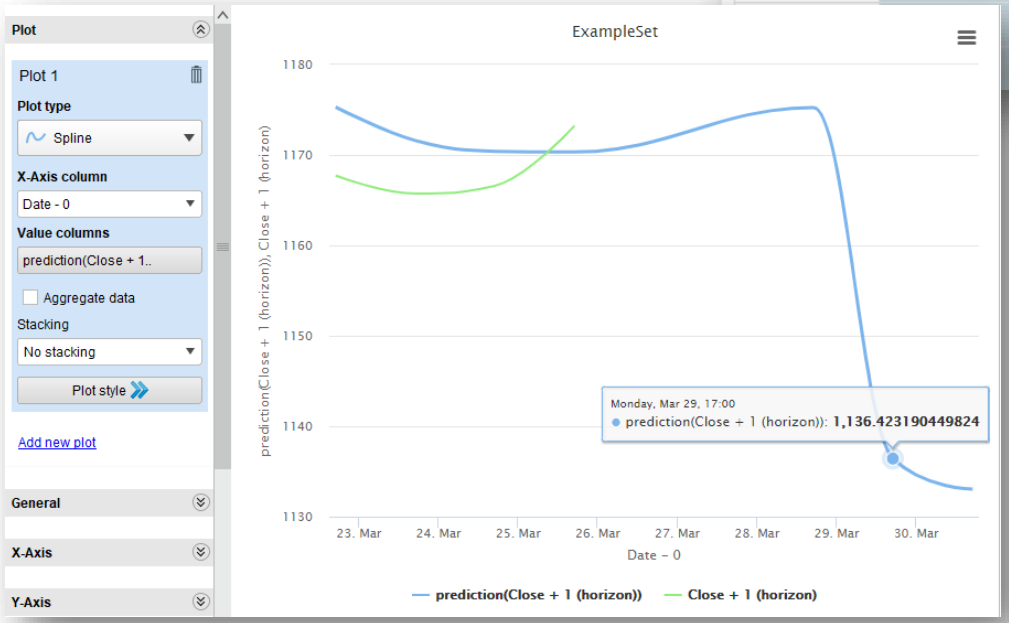
attribute filter type	all
invert selection	
include special attributes	
has indices	
window size	5
no overlapping windows	
step size	1
create horizon (labels)	
horizon attribute	Close
horizon size	1
horizon offset	0

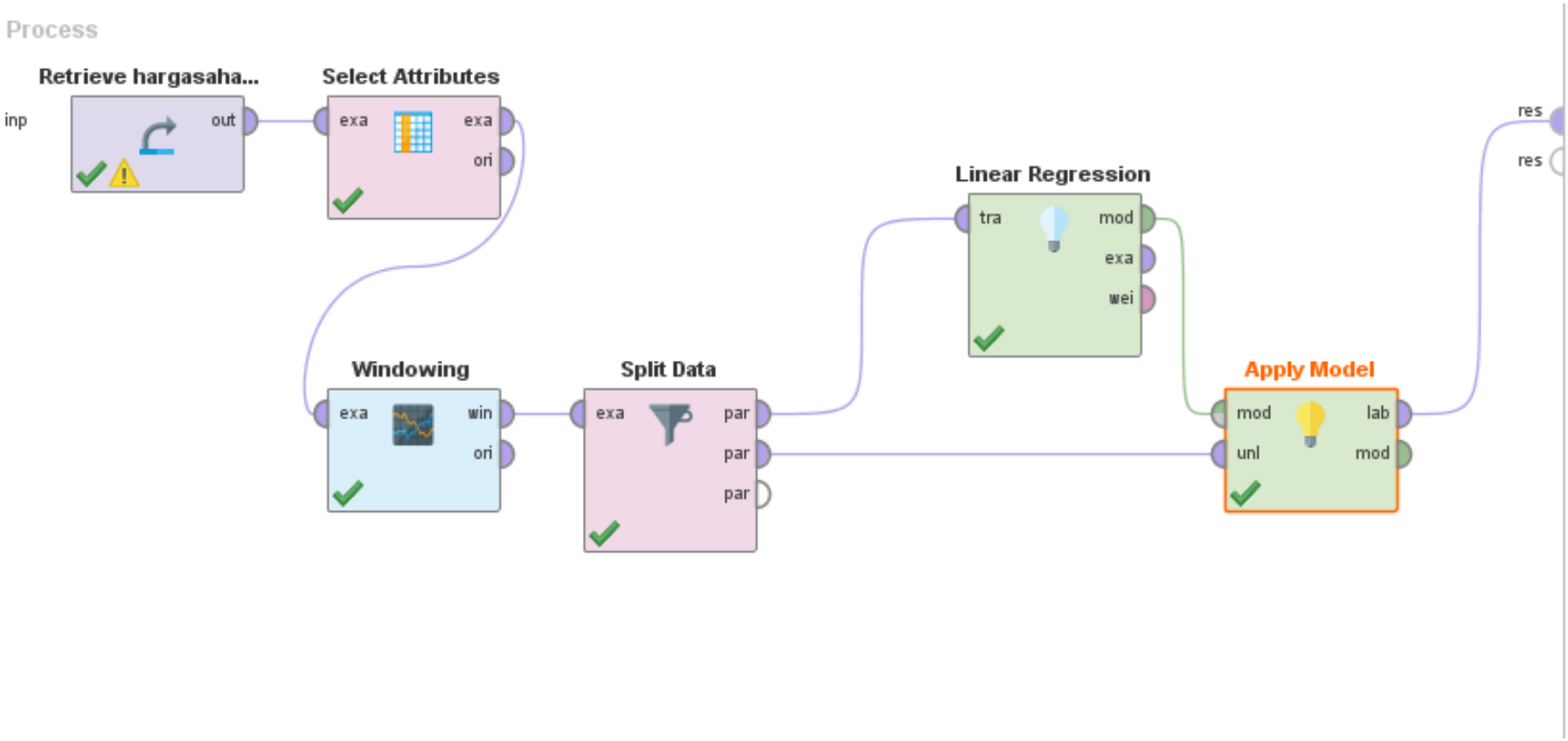


Forecasting Harga Saham (Data Masa Depan)



Row No.	Window id	Close + 1 (h...	prediction(C...	Date - 2	Date - 1	Date - 0	Close - 2
1	0	1167.720	1175.293	Mar 19, 2010	Mar 22, 2010	Mar 23, 2010	1159.900
2	1	1165.730	1171.606	Mar 22, 2010	Mar 23, 2010	Mar 24, 2010	1165.810
3	2	1166.590	1170.406	Mar 23, 2010	Mar 24, 2010	Mar 25, 2010	1174.170
4	3	1173.220	1170.324	Mar 24, 2010	Mar 25, 2010	Mar 26, 2010	1167.720
5	4	?	1175.235	Mar 25, 2010	Mar 26, 2010	Mar 29, 2010	1165.730
		?	1136.423	Mar 26, 2010	Mar 29, 2010	Mar 30, 2010	1166.590
		?	1133.052	Mar 29, 2010	Mar 30, 2010	Mar 31, 2010	1173.220




Process
Retrieve hargasaha...
inp
out
Select Attributes
exa
exa
ori
Linear Regression
tra
mod
exa
wei
Windowing
exa
win
ori
Split Data
exa
par
par
par
Apply Model
mod
unl
lab
mod
res
res

Latihan: Penentuan Kelayakan Kredit

1. Lakukan training dengan algoritma yang tepat pada dataset: creditapproval-training.xls

2. Ujicoba model yang dibentuk dari training di atas ke dataset di bawah: creditapproval-testing.xls

Latihan: Deteksi Kanker Payudara

1. Lakukan training pada data kanker payudara (breasttissue.xls)

2. Dataset adalah di sheet 2, sedangkan sheet 1 berisi penjelasan tentangd ata

3. Bagi dataset dengan menggunakan operator Split Data, 90% untuk training dan 10% untuk testing

4. Pilih metode yang tepat supaya menghasilkan pola, analisis pola yang dihasilkan

Latihan: Deteksi Serangan Jaringan

1. Lakukan training pada data serangan jaringan (intrusion-training.xls)

2. Pilih metode yang tepat supaya menghasilkan pola

Latihan: Klasifikasi Resiko Kredit

1. Lakukan training pada data resiko kredit (CreditRisk.csv)
   (http://romisatriawahono.net/lecture/dm/dataset/)

2. Pilih metode yang tepat supaya menghasilkan pola

Latihan: Klasifikasi Music Genre

1. Lakukan training pada data Music Genre (musicgenre-small.csv)

2. Pilih metode yang tepat supaya menghasilkan pola

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
…	…	…	…	…	…	…

Competency Check

1. Dataset – Methods – Knowledge
    1. Dataset Main Golf (Klasifikasi)
    2. Dataset Iris (Klasifikasi)
    3. Dataset Iris (Klastering)
    4. Dataset CPU (Estimasi)
    5. Dataset Pemilu (Klasifikasi)
    6. Dataset Heating Oil (Asosiasi, Estimasi)
    7. Dataset Transaksi (Association)
    8. Dataset Harga Saham (Forecasting) (Uni dan Multi)

Tugas: Mencari dan Mengolah Dataset

- Pahami berbagai dataset yang ada di folder dataset
- Gunakan rapidminer untuk mengolah dataset tersebut sehingga menjadi pengetahuan
- Pilih algoritma yang sesuai dengan jenis data pada dataset

Tugas: Menguasai Satu Metode DM

1. Pahami dan kuasai satu metode data mining dari berbagai literature:
   1. Naïve Bayes
   2. k Nearest Neighbor
   3. k-Means
   4. C4.5
   5. Neural Network
   6. Logistic Regression
   7. FP Growth
   8. Fuzzy C-Means
   9. Self-Organizing Map
   0. Support Vector Machine
2. Rangkumkan dengan detail dalam bentuk slide, dengan format:
   1. Definisi
   2. Tahapan Algoritma (lengkap dengan formulanya)
   3. Penerapan Tahapan Algoritma untuk Studi Kasus Dataset Main Golf, Iris, Transaksi, CPU, dsb
      (hitung manual (gunakan excel) dan tidak dengan menggunakan rapidminer, harus sinkron dengan tahapan algoritma)
3. Presentasikan di depan kelas pada mata kuliah berikutnya dengan bahasa manusia yang baik dan benar

Tugas: Kembangkan Code dari Algoritma DM

1. Kembangkan Java Code dari algoritma yang dipilih
2. Gunakan hanya 1 class (file) dan beri nama sesuai nama algoritma, boleh membuat banyak method dalam class tersebut
3. Buat account di Trello.Com dan register ke https://trello.com/b/ZOwroEYg/course-assignment
4. Buat card dengan nama sendiri dan upload semua file (pptx, xlsx, pdf, etc) laporan ke card tersebut
5. Deadline: sehari sebelum pertemuan berikutnya

Algoritma k-Means

Format Template Tugas

Definisi

- K-means adalah ..... (John, 2016)

Tahapan Algoritma k-Means

1. Siapkan dataset

2. Tentukan A dengan rumus A = x + y

3. Tentukan B dengan rumus B = d + e

4. Ulangi proses 1-2-3 sampai tidak ada perubahan

1. Siapkan dataset

2. Tentukan A

- blablabla

3. Tentukan B

- blablabla

4. Iterasi 1

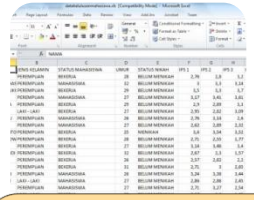
- blablabla

4. Iterasi 2 ... dst

- blablabla

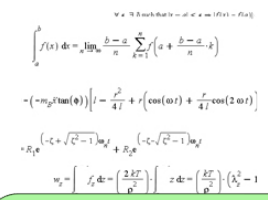
2.3 Evaluasi Model Data Mining

Proses Data Mining



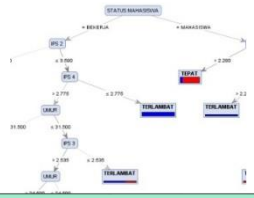
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



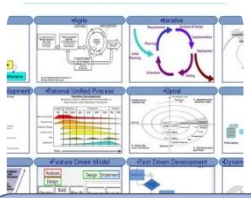
2. Metode Data Mining

(**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

Evaluasi 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

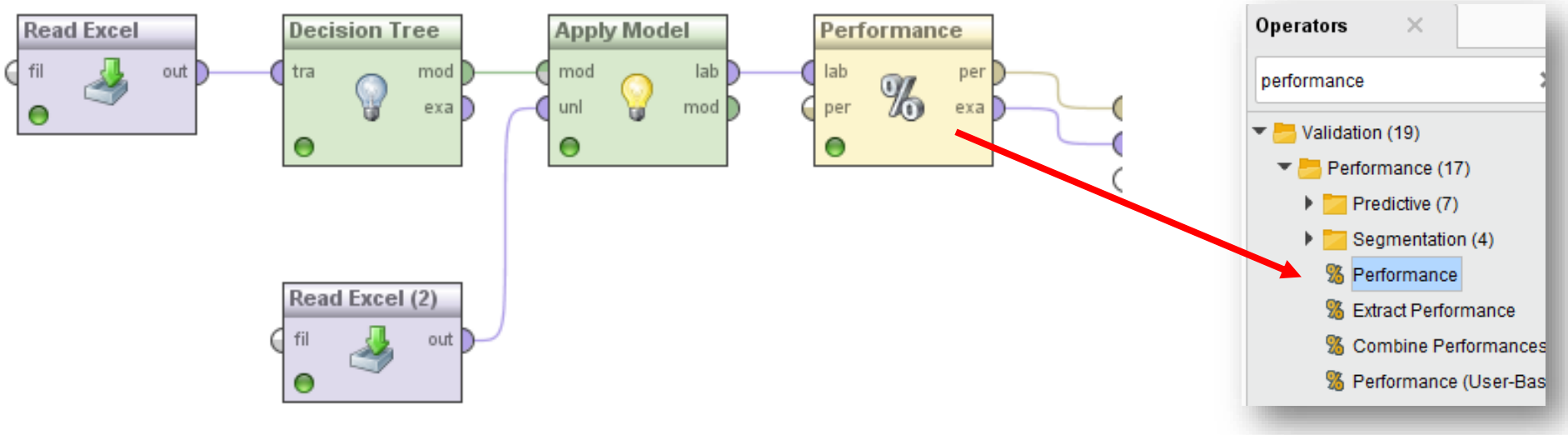
- Data training untuk pembentukan model, dan data testing digunakan untuk pengujian model

- 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

2.3.1 Pemisahan Data Manual

Latihan: Penentuan Kelayakan Kredit

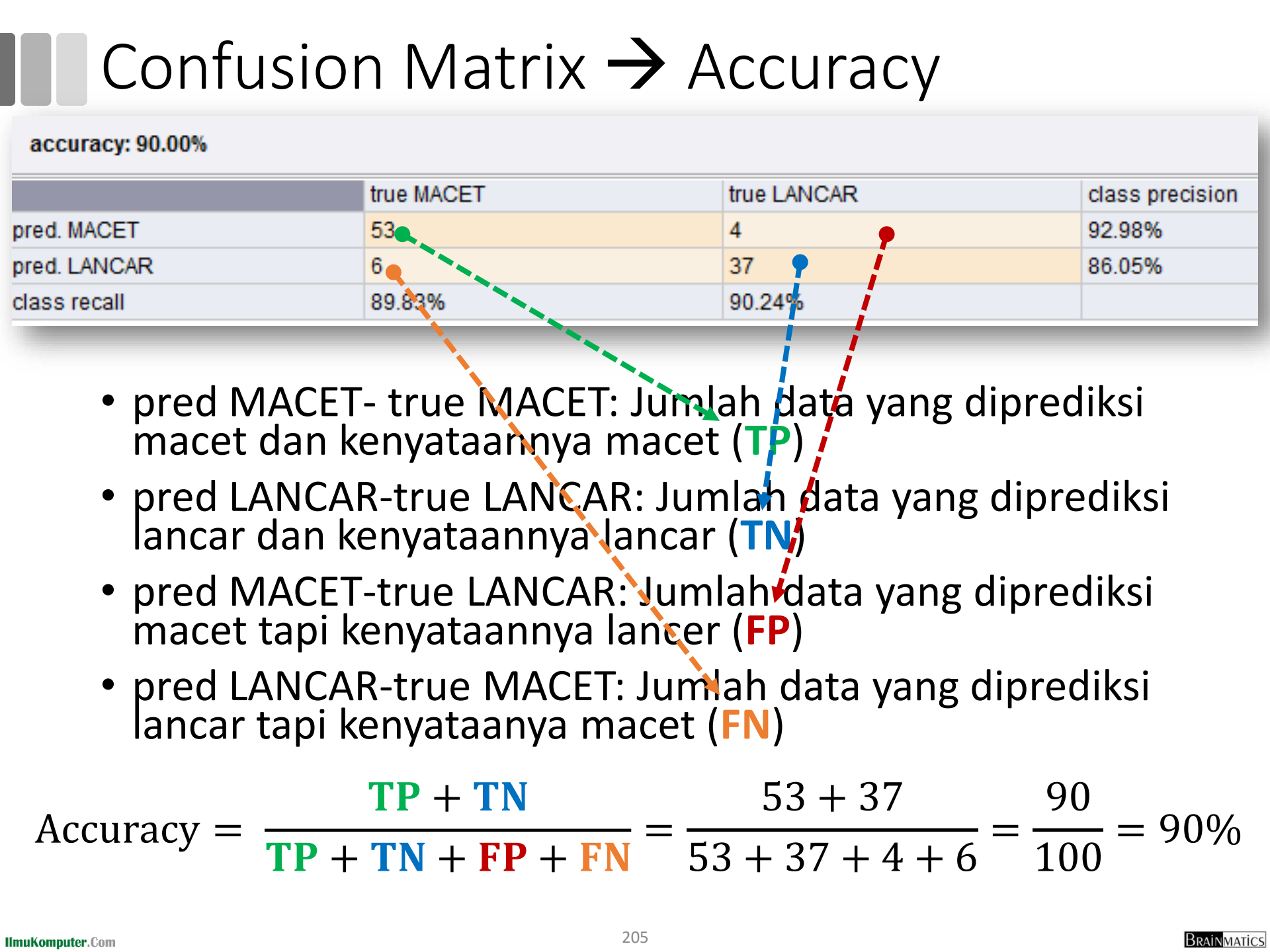
- Gunakan dataset di bawah:
  - creditapproval-training.xls: untuk membuat model
  - creditapproval-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data testing (10%) dan data training (90%)
- Data training sebagai pembentuk model, dan data testing untuk pengujian model, ukur performancenya



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 precision and recall
  - assigns ß times as much weight to recall as to precision

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

Sensitivity and Specificity

Binary classification should be both sensitive and specific as much as possible:

1. Sensitivity measures the proportion of true 'positives' that are correctly identified (True Positive Rate (TP Rate) or Recall)

$$Sensitivity = \frac{Number\ of\ 'True\ Positives'}{Number\ of\ 'True\ Positives' + Number\ of\ 'False\ Negatives'},$$

2. Specificity measures the proportion of true 'negatives' that are correctly identified (False Negative Rate (FN Rate or Precision)

$$Specificity = \frac{Number\ of\ 'True\ Negatives'}{Number\ of\ 'True\ Negatives' + Number\ of\ 'False\ Positives'},$$

PPV and NPV

We need to know the probability that the classifier will give the correct diagnosis, but the sensitivity and specificity do not give us this information

- Positive Predictive Value (PPV) is the proportion of cases with 'positive' test results that are correctly diagnosed

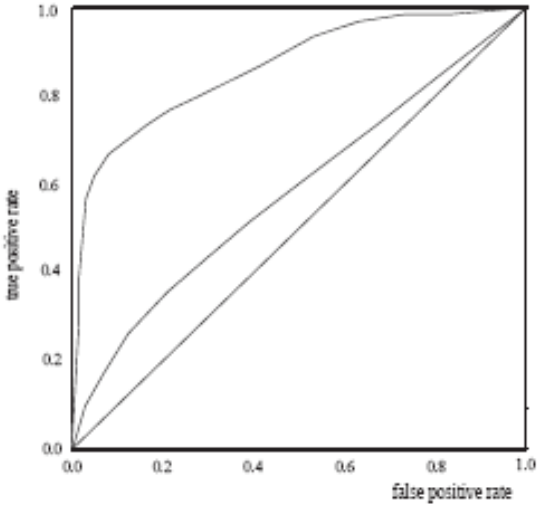
$$PPV = \frac{Number\ of\ 'True\ Positives'}{Number\ of\ 'True\ Positives' + Number\ of\ 'False\ Positives''}$$

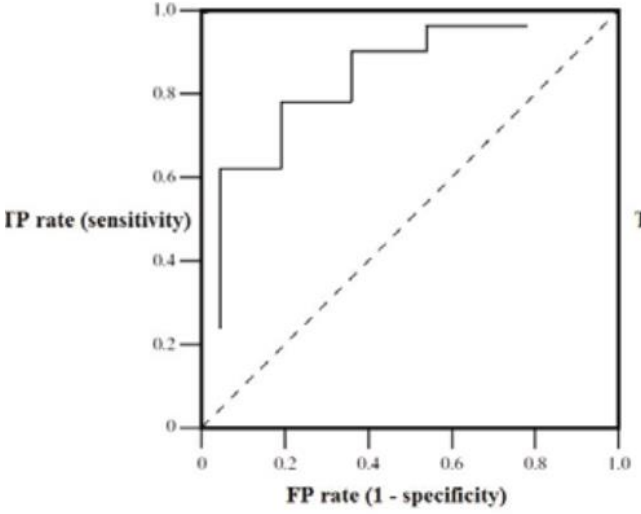
- Negative Predictive Value (NPV) is the proportion of cases with 'negative' test results that are correctly diagnosed

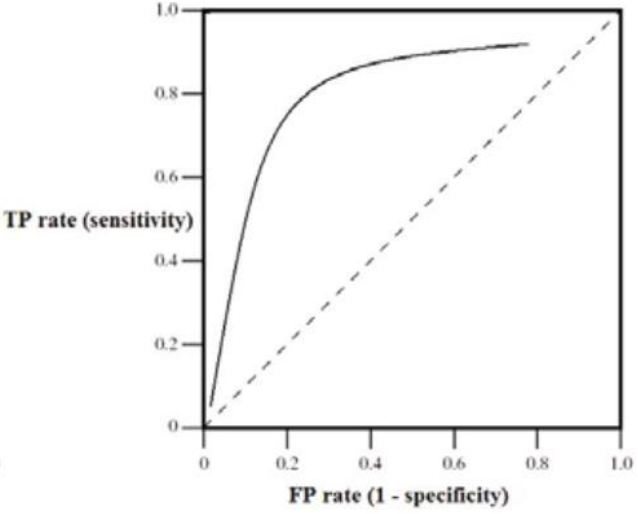
$$NPV = \frac{Number\ of\ 'True\ Negatives'}{Number\ of\ 'True\ Negatives' + Number\ of\ 'False\ Negatives'}.$$

Kurva ROC - AUC (Area Under Curve)

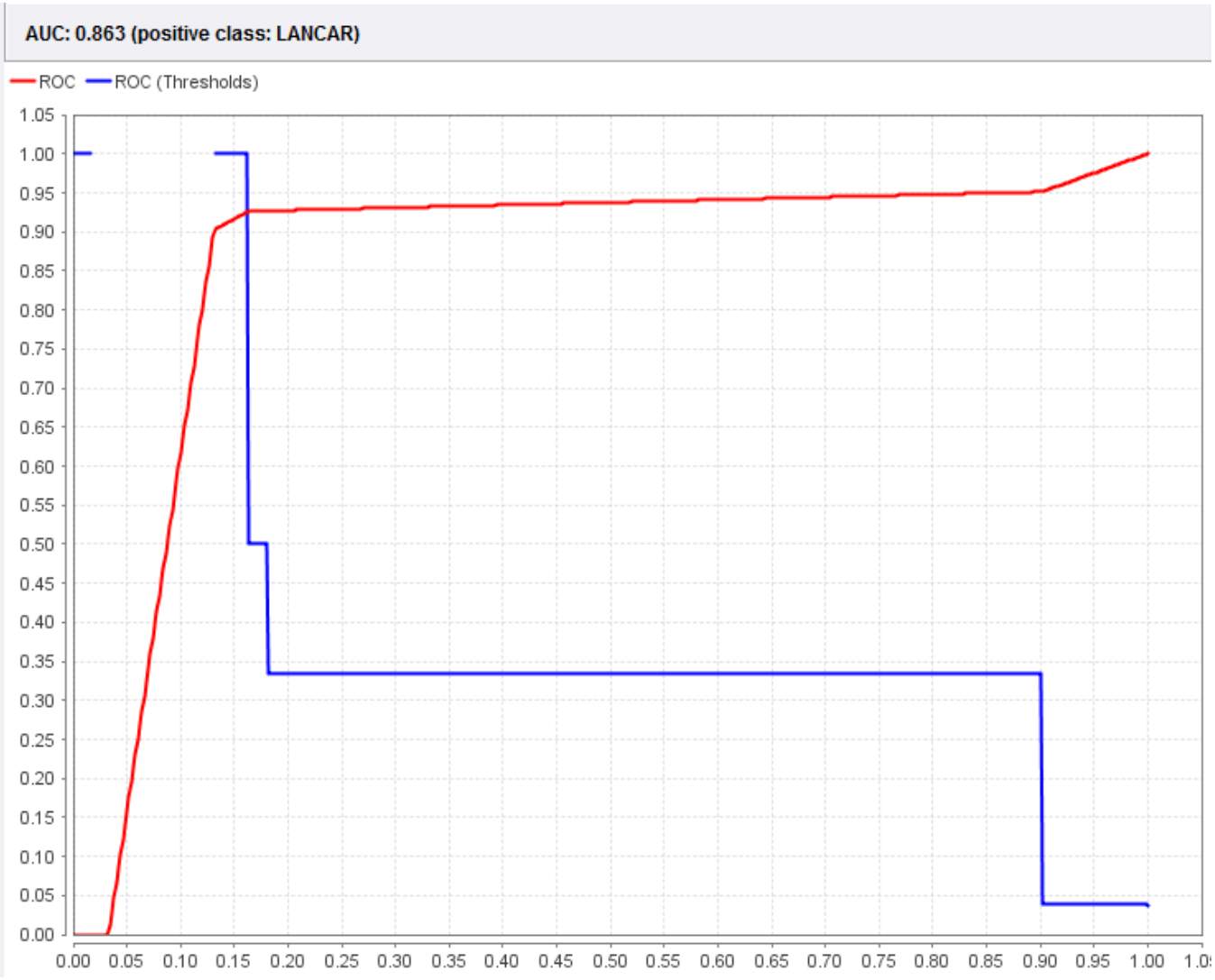
- ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models
  - Originated from signal detection theory
- ROC curves are two-dimensional graphs in which the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis
- ROC curve depicts relative trade-offs between benefits ('true positives') and costs ('false positives')
- Two types of ROC curves: discrete and continuous







Kurva ROC - AUC (Area Under Curve)



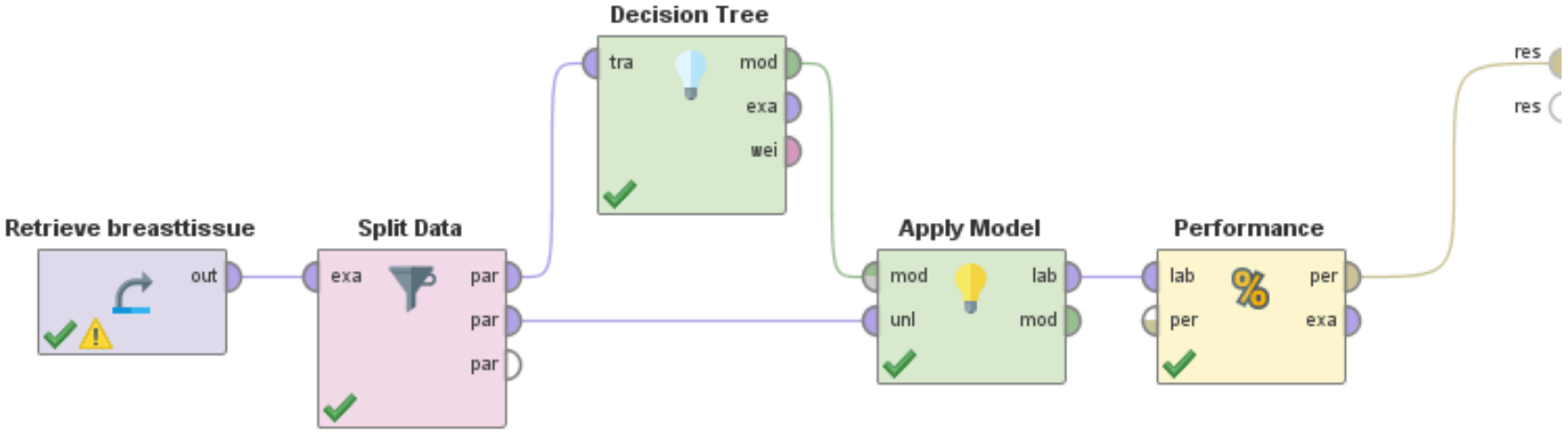
Guide for Classifying the AUC

1. 0.90 - 1.00 = excellent classification
2. 0.80 - 0.90 = good classification
3. 0.70 - 0.80 = fair classification
4. 0.60 - 0.70 = poor classification
5. 0.50 - 0.60 = failure

(Gorunescu, 2011)

Latihan: Prediksi Kanker Payudara

- Gunakan dataset: breasttissue.xls
- Split data dengan perbandingan:
  data testing (10%) dan data training (90%)
- Ukur performance
  (Accuracy dan Kappa)



Kappa Statistics

- The (Cohen's) Kappa statistics is a more vigorous measure than the 'percentage correct prediction' calculation, because Kappa considers the correct prediction that is occurring by chance
- Kappa is essentially a measure of how well the classifier performed as compared to how well it would have performed simply by chance
- A model has a high Kappa score if there is a big difference between the accuracy and the null error rate (Markham, K., 2014)
- Kappa is an important measure on classifier performance, especially on imbalanced data set

Latihan: Prediksi Harga Saham

- Gunakan dataset di bawah:
  - hargasaham-training.xls: untuk membuat model
  - hargasaham-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data testing (10%) dan data training (90%)
- Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
- Ukur performance

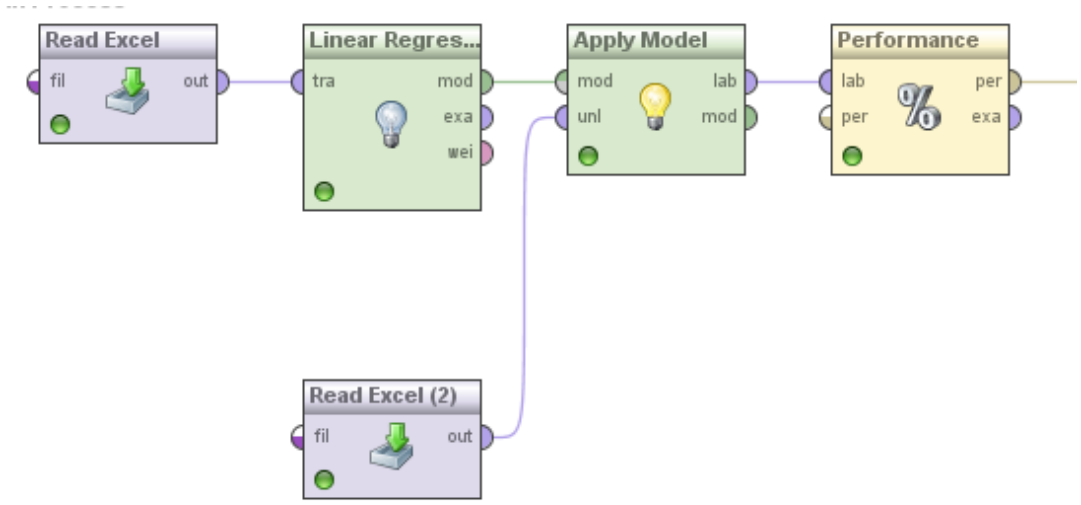


Chart style:

Scatter Multiple

x-Axis:

Date

Log scale

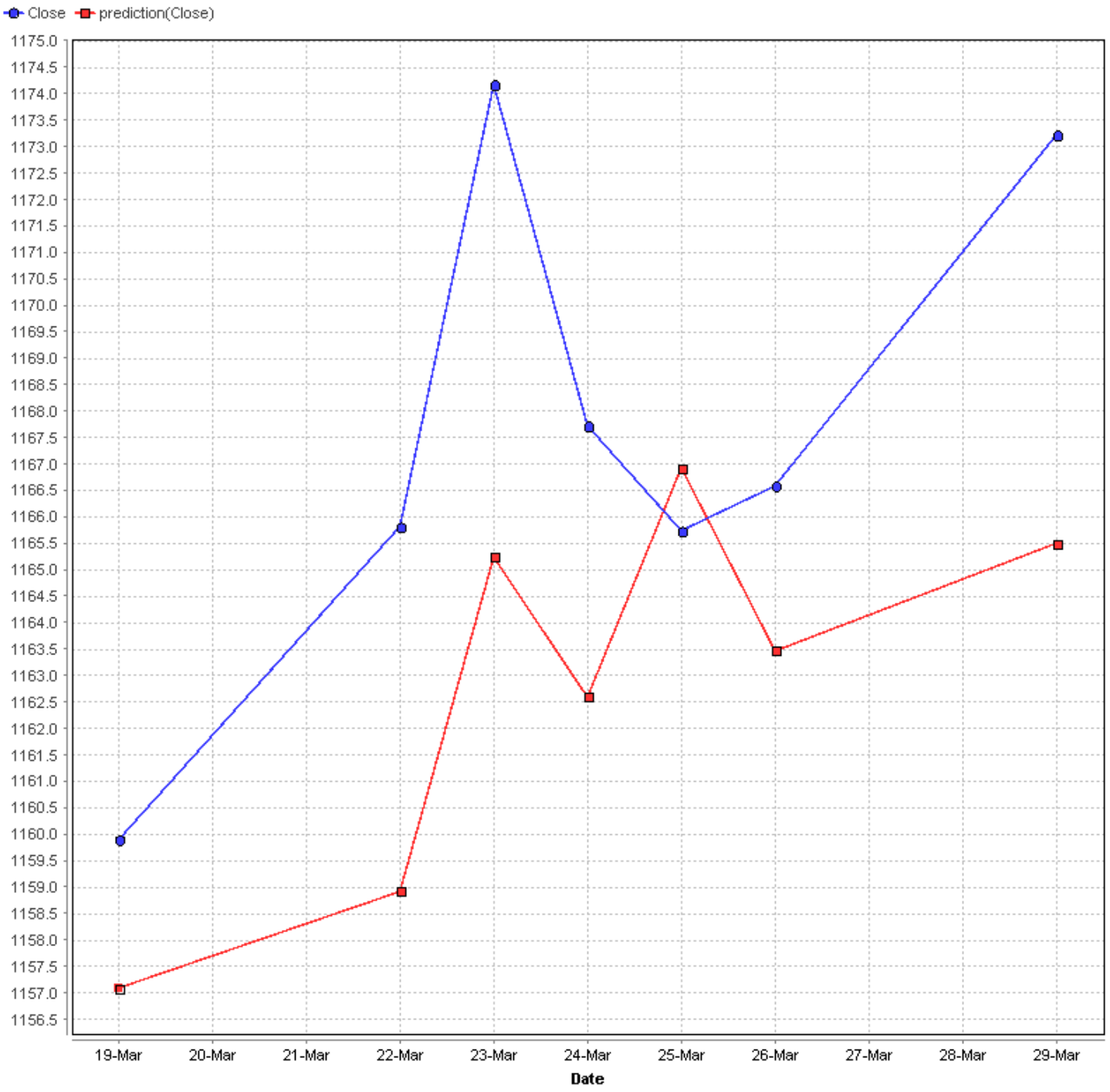
y-Axis:

Date
Open
High
Low
Volume
Close
prediction(Close)

Jitter:

Rotate labels

Points and Lines...



Root Mean Square Error

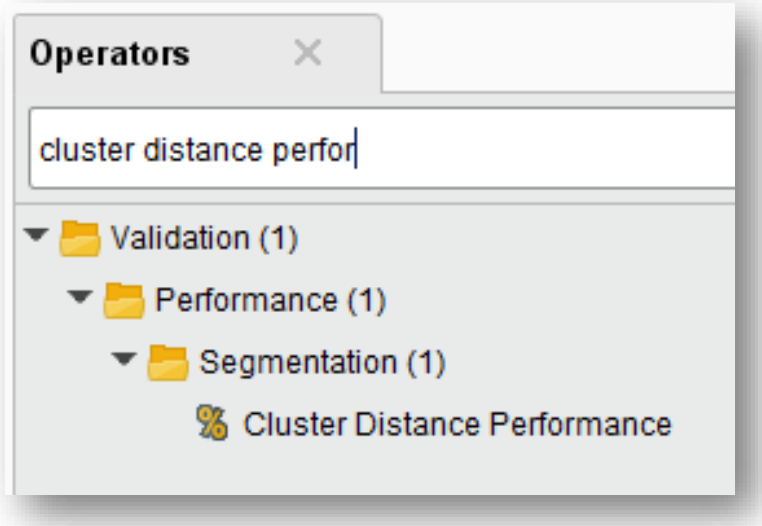
- The square root of the mean/average of the square of all of the error

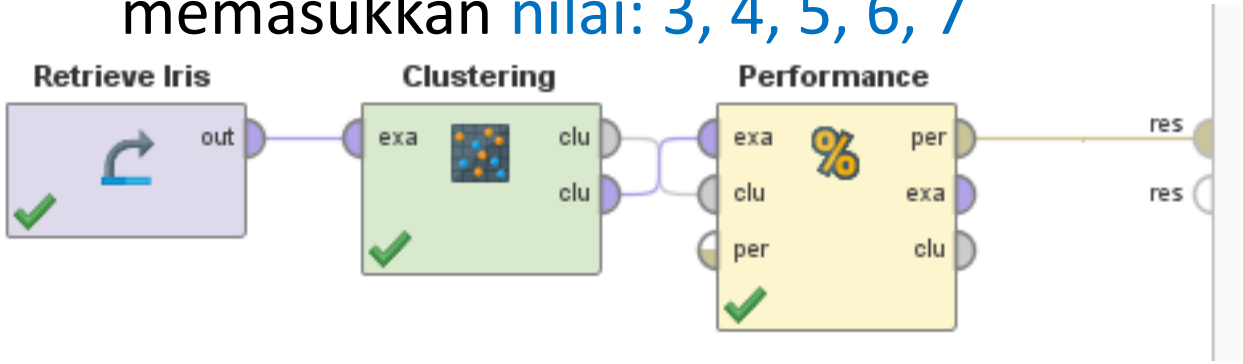
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions
- To construct the RMSE, we first need to determine the residuals
  - Residuals are the difference between the actual values and the predicted values
  - We denoted them by $\hat{y}_i - y_i$
  - where y_i is the observed value for the *i*th observation and
  - $\hat{y}_i$ is the predicted value
- They can be positive or negative as the predicted value under or over estimates the actual value
- You then use the RMSE as a measure of the spread of the *y* values about the predicted *y* value

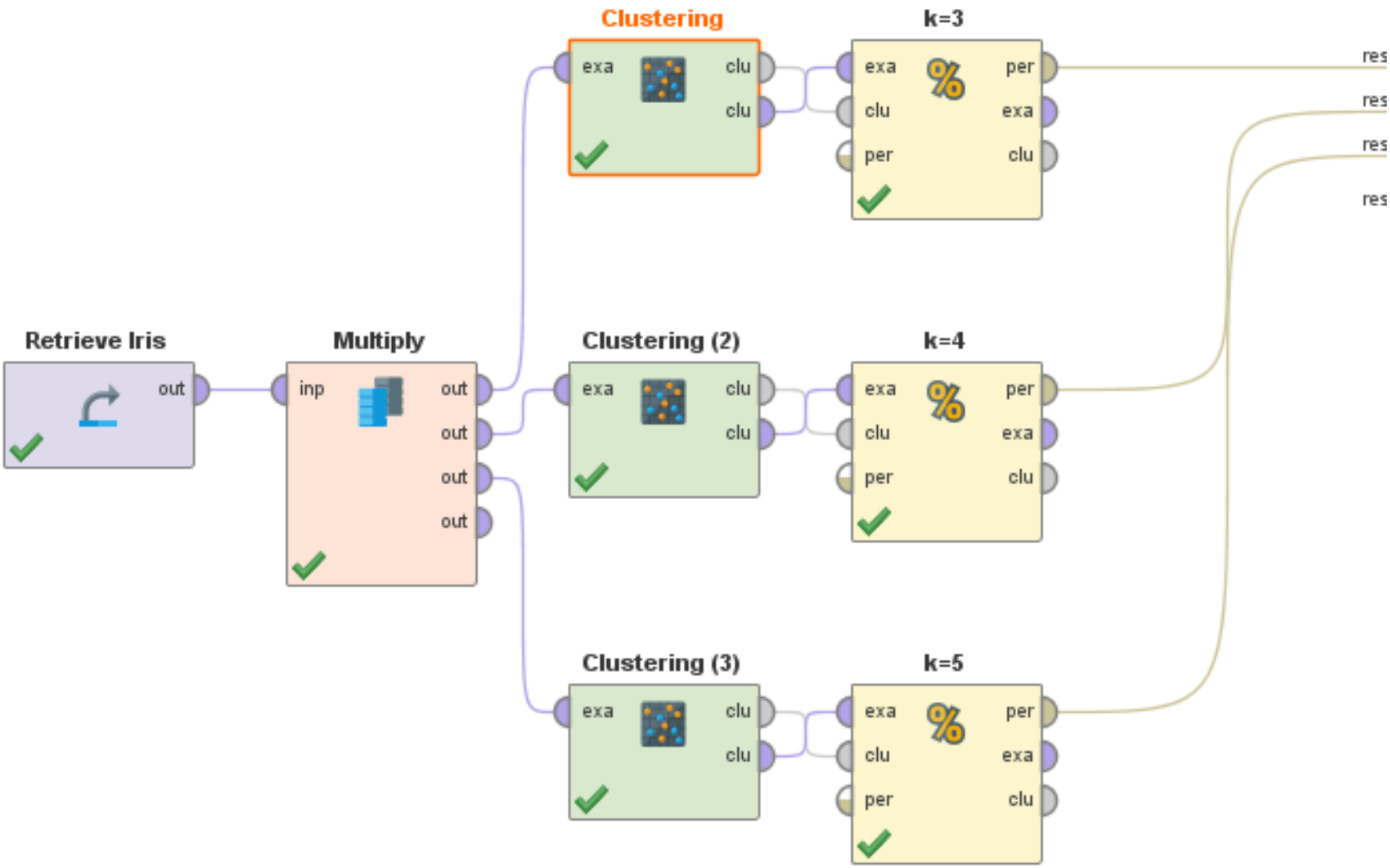
Latihan: Klastering Jenis Bunga Iris

1. Lakukan training pada data iris (ambil dari repositories rapidminer) dengan menggunakan algoritma clustering k-means
2. Ukur performance-nya dengan *Cluster Distance Performance*, cek dan analisis nilai yang keluar Davies Bouldin Indeks (DBI)
3. Lakukan pengubahan pada nilai k pada parameter k-means dengan memasukkan nilai: 3, 4, 5, 6, 7





k	DBI
3	0.666
4	0.764
5	0.806
6	0.910
7	0.999


Clustering
k=3
exa
clu
clu
exa
clu
per
per
exa
clu
res
res
res
res
Retrieve Iris
Multiply
Clustering (2)
k=4
out
inp
out
out
out
out
Clustering (3)
k=5

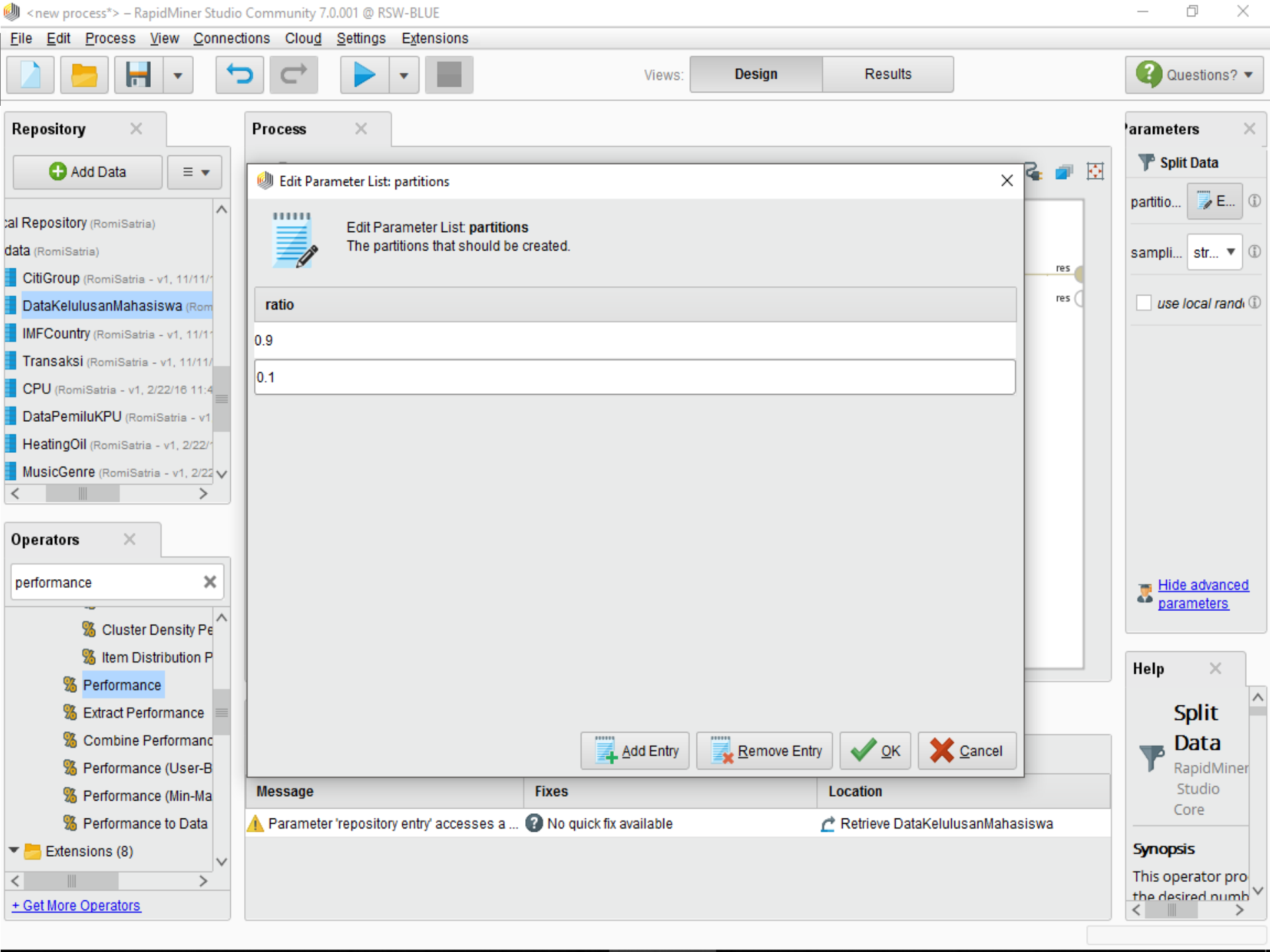
Davies–Bouldin index (DBI)

- The Davies–Bouldin index (DBI) (introduced by David L. Davies and Donald W. Bouldin in 1979) is a metric for evaluating clustering algorithms
- This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset
- As a function of the ratio of the within cluster scatter, to the between cluster separation, a lower value will mean that the clustering is better
- This affirms the idea that no cluster has to be similar to another, and hence the best clustering scheme essentially minimizes the Davies–Bouldin index
- This index thus defined is an average over all the *i* clusters, and hence a good measure of deciding how many clusters actually exists in the data is to plot it against the number of clusters it is calculated over
- The number *i* for which this value is the lowest is a good measure of the number of clusters the data could be ideally classified into

2.3.2 Pemisahan Data Otomatis dengan Operator *Split Data*

Split Data Otomatis

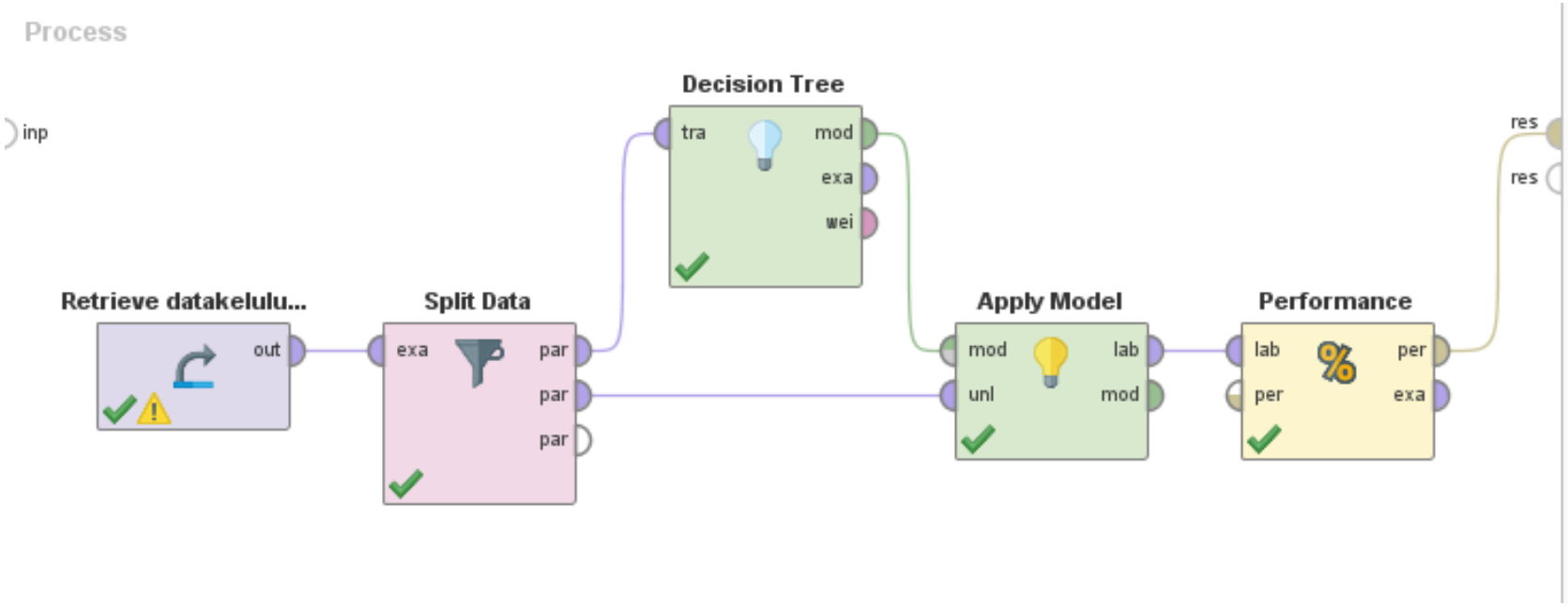
- The Split Data operator takes a dataset as its input and delivers the subsets of that dataset through its output ports
- The sampling type parameter decides how the examples should be shuffled in the resultant partitions:
  1. Linear sampling: Divides the dataset into partitions without changing the order of the examples
  2. Shuffled sampling: Builds random subsets of the dataset
  3. Stratified sampling: Builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset


<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
Process
Parameters
Split Data
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.9
0.1
Add Entry
Remove Entry
OK
Cancel
Operators
performance
Cluster Density Pe
Item Distribution P
Performance
Extract Performance
Combine Performanc
Performance (User-B
Performance (Min-Ma
Performance to Data
Extensions (8)
+ Get More Operators
Message
Fixes
Location
Parameter 'repository entry' accesses a ...
No quick fix available
Retrieve DataKelulusanMahasiswa
Hide advanced parameters
Help
Split Data
RapidMiner Studio Core
Synopsis
This operator pro

Latihan: Prediksi Kelulusan Mahasiswa

1. Dataset: datakelulusanmahasiswa.xls
2. Pisahkan data menjadi dua secara otomatis (Split Data): data testing (10%) dan data training (90%)
3. Ujicoba parameter pemisahan data baik menggunakan Linear Sampling, Shuffled Sampling dan Stratified Sampling
4. Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
5. Terapkan algoritma yang sesuai dan ukur performance dari model yang dibentuk

Proses Prediksi Kelulusan Mahasiswa



Latihan: Estimasi Konsumsi Minyak

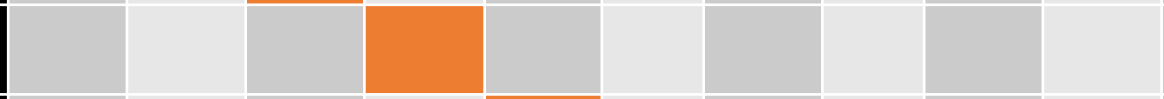
1. Dataset: HeatingOil.csv
2. Pisahkan data menjadi dua secara otomatis (Split Data): data testing (10%) dan data training (90%)
3. Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
4. Terapkan algoritma yang sesuai dan ukur performance dari model yang dibentuk

2.3.3 Pemisahan Data dan Evaluasi Model Otomatis dengan *Cross-Validation*

Metode Cross-Validation

- Metode cross-validation digunakan untuk menghindari overlapping pada data testing
- Tahapan cross-validation:
  1. Bagi data menjadi k subset yg berukuran sama
  2. Gunakan setiap subset untuk data testing dan sisanya untuk data training
- Disebut juga dengan k-fold cross-validation
- Seringkali subset dibuat stratified (bertingkat) sebelum cross-validation dilakukan, karena stratifikasi akan mengurangi variansi dari estimasi

10 Fold Cross-Validation

Eksperimen	Dataset	Akurasi
1		93%
2		91%
3		90%
4		93%
5		93%
6		91%
7		94%
8		93%
9		91%
10		90%
	Akurasi Rata-Rata	92%

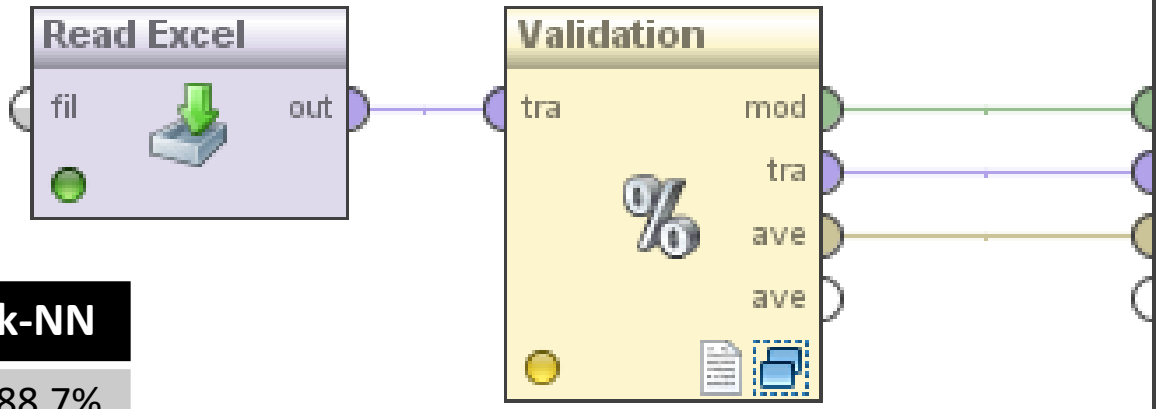
Orange: k-subset (data testing)

10 Fold Cross-Validation

- Metode evaluasi standard: stratified 10-fold cross-validation
- Mengapa 10? Hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa 10-fold cross-validation adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat
- 10-fold cross-validation akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian

Latihan: Prediksi Elektabilitas Caleg

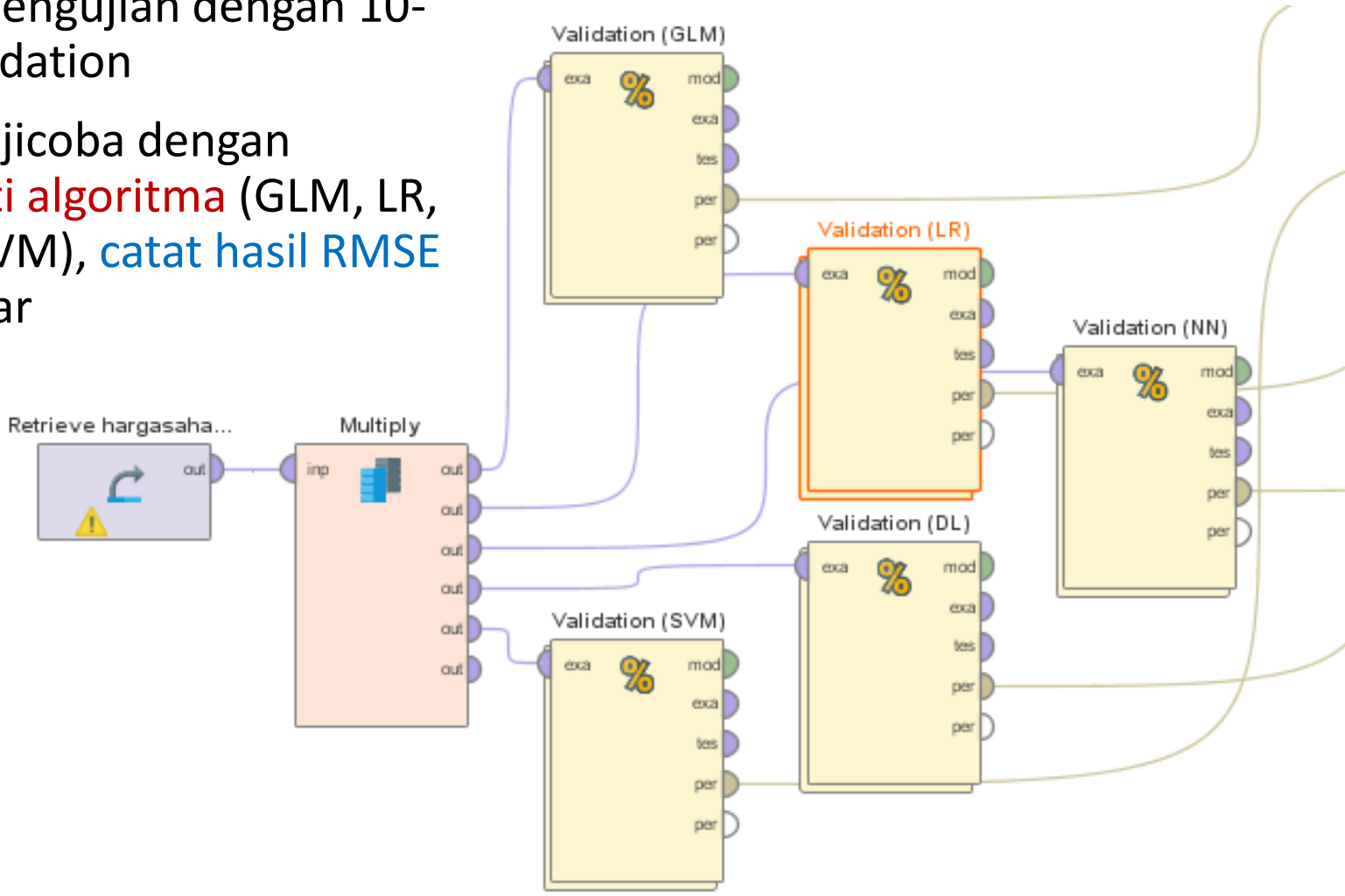
1. Lakukan training pada data pemilu (datapemilukpu.xls)
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Ukur performance-nya dengan confusion matrix dan ROC Curve
4. Lakukan ujicoba, ubah algoritma menjadi Naive Bayes, k-NN, Random Forest (RF), Logistic Regression (LogR), analisis mana algoritma yang menghasilkan model yang lebih baik (akurasi tinggi)



	C4.5	NB	k-NN
Accuracy	**92.87%**	79.34%	88.7%
AUC	**0.934**	0.849	0.5

Latihan: Komparasi Prediksi Harga Saham

- Gunakan dataset harga saham (hargasaham-training.xls)
- Lakukan pengujian dengan 10-fold X Validation
- Lakukan ujicoba dengan mengganti algoritma (GLM, LR, NN, DL, SVM), catat hasil RMSE yang keluar



2.3.4 Komparasi Algoritma Data Mining

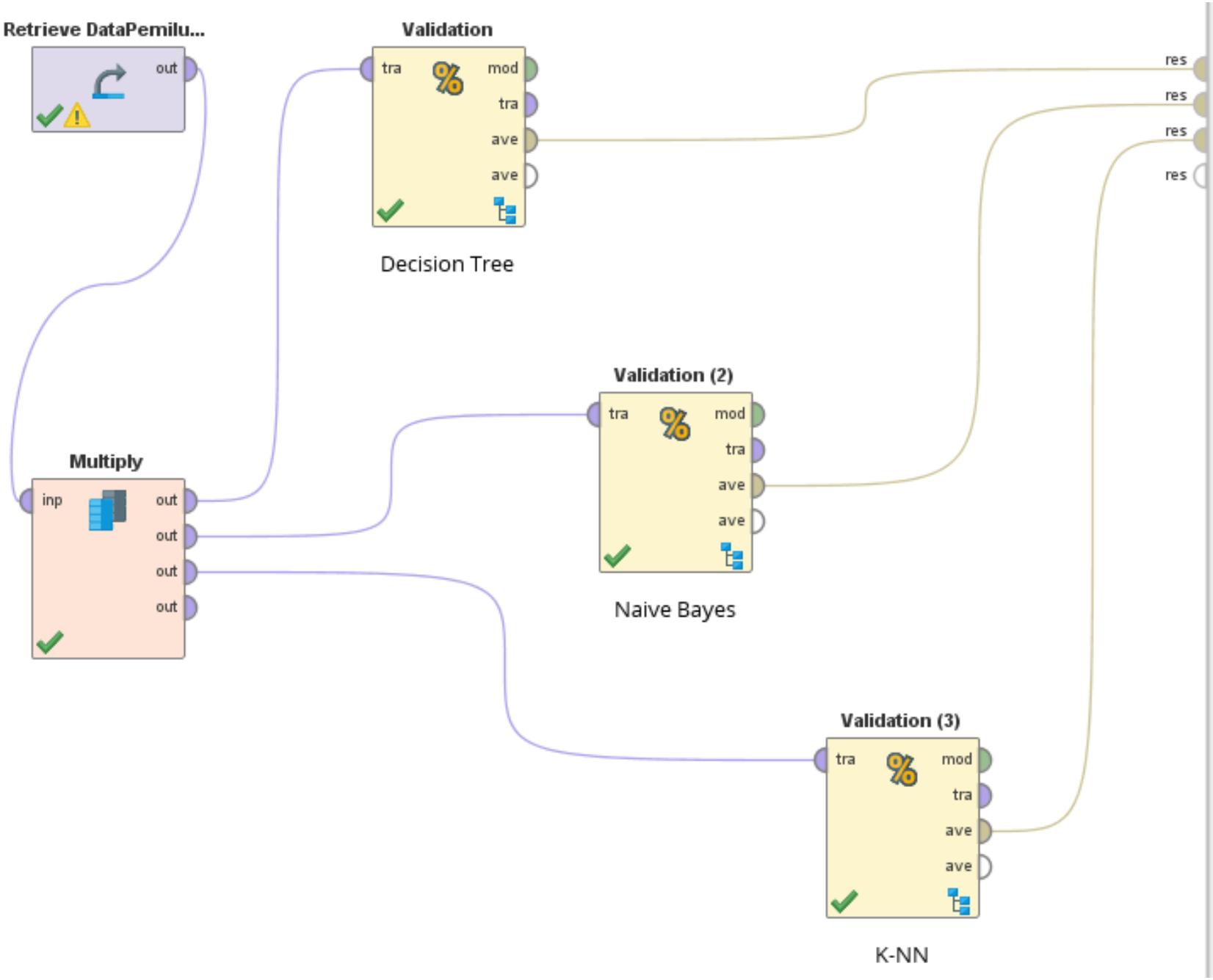
Metode Data Mining

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

Latihan: Prediksi Elektabilitas Caleg

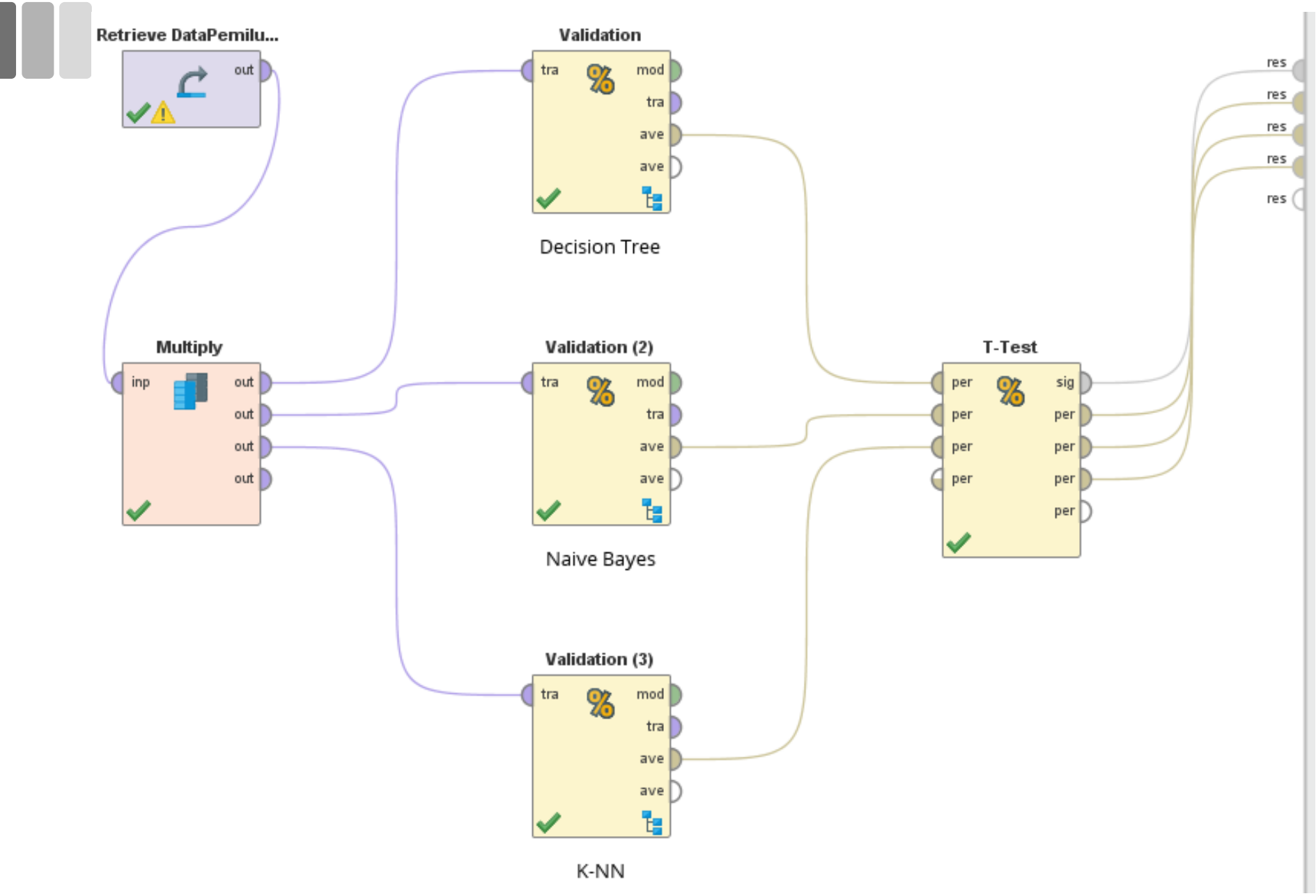
1. Lakukan training pada data pemilu (datapemilukpu.xls) dengan menggunakan algoritma
   1. Decision Tree (C4.5)
   2. Naïve Bayes (NB)
   3. K-Nearest Neighbor (K-NN)
2. Lakukan pengujian dengan menggunakan 10-fold X Validation

	DT	NB	K-NN
Accuracy	92.45%	77.46%	88.72%
AUC	0.851	0.840	0.5


Retrieve DataPemilu...
out
Validation
tra
mod
tra
ave
ave
Decision Tree
res
res
res
res
Multiply
inp
out
out
out
out
Validation (2)
tra
mod
tra
ave
ave
Naive Bayes
Validation (3)
tra
mod
tra
ave
ave
K-NN

Latihan: Prediksi Elektabilitas Caleg

1. Lakukan training pada data pemilu (datapemilukpu.xls) dengan menggunakan algoritma C4.5, NB dan K-NN
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Ukur performance-nya dengan confusion matrix dan ROC Curve
4. Uji beda dengan t-Test untuk mendapatkan model terbaik


Retrieve DataPemilu...
out
Validation
tra
mod
tra
ave
ave
Decision Tree
Multiply
inp
out
out
out
out
Validation (2)
tra
mod
tra
ave
ave
Naive Bayes
T-Test
per
per
per
per
sig
per
per
per
Validation (3)
tra
mod
tra
ave
ave
K-NN
res
res
res
res
res

Hasil Prediksi Elektabilitas Caleg

- Komparasi Accuracy dan AUC

	C4.5	NB	K-NN
Accuracy	92.45%	77.46%	88.72%
AUC	0.851	0.840	0.5

- Uji Beda (t-Test)

C4.5 NB kNN

A	B	C	D
	0.932 +/- 0.020	0.797 +/- 0.046	0.896 +/- 0.022
C4.5 0.932 +/- 0.020		0.000	0.001
NB 0.797 +/- 0.046			0.000
kNN 0.896 +/- 0.022			

Values with a colored background are smaller than alpha=0.050, which indicate a probably significant difference between the mean values

- Urutan model terbaik: 1. C4.5 2. k-NN 3. NB

Hasil Prediksi Elektabilitas Caleg

- Komparasi Accuracy dan AUC

	C4.5	NB	K-NN
Accuracy	93.41%	79.72%	91.76%
AUC	0.921	0.826	0.885

- Uji Beda (t-Test)

C4.5 NB kNN

A	B	C	D
	0.934 +/- 0.029	0.797 +/- 0.079	0.918 +/- 0.026
C4.5 0.934 +/- 0.029		0.000	0.192
NB 0.797 +/- 0.079			0.000
kNN 0.918 +/- 0.026			

Values with a white background are higher than alpha=0.050, which indicate a probably NO significant difference between the mean values

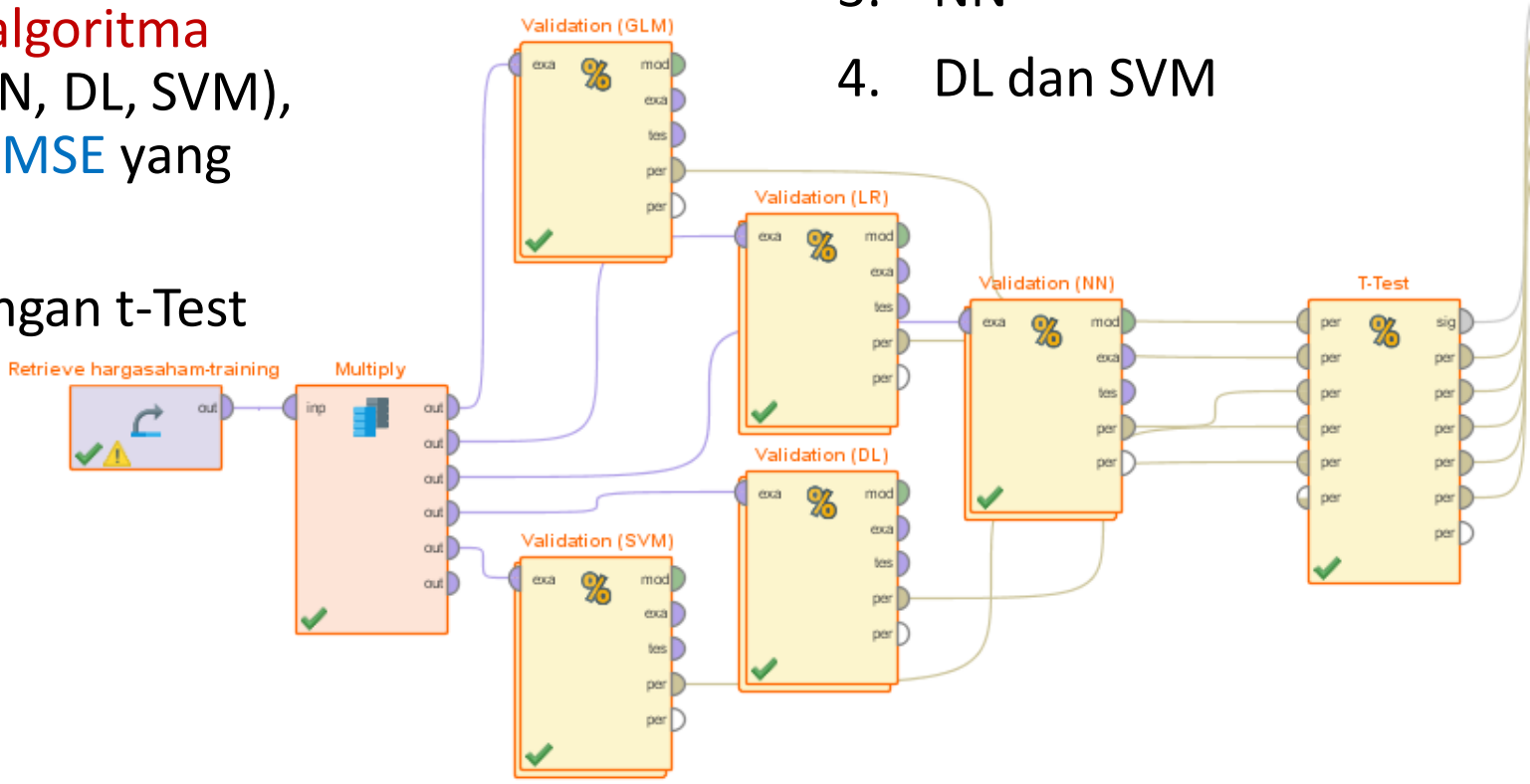
- Urutan model terbaik: 1. C4.5 1. kNN 2. NB

Latihan: Komparasi Prediksi Harga Saham

- Gunakan dataset harga saham (hargasaham-training.xls)
- Lakukan pengujian dengan 10-fold X Validation
- Lakukan ujicoba dengan mengganti algoritma (GLM, LR, NN, DL, SVM), catat hasil RMSE yang keluar
- Uji beda dengan t-Test

A	B	C	D	E	F
	22.982 +/- 1.589	10.947 +/- 2.599	7.243 +/- 0.742	16.950 +/- 6.285	15.188 +/- 1.322
22.982 +/- 1.589		0.000	0.000	0.009	0.000
10.947 +/- 2.599			0.000	0.012	0.000
7.243 +/- 0.742				0.000	0.000
16.950 +/- 6.285					0.397
15.188 +/- 1.322					

1. GLM
2. LR
3. NN
4. DL dan SVM



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

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?

Latihan: Prediksi Kelulusan Mahasiswa

1. Lakukan training pada data mahasiswa (datakelulusanmahasiswa.xls) dengan menggunakan C4.5, ID3, NB, K-NN, RF dan LogR
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Uji beda dengan t-Test untuk mendapatkan model terbaik

Hasil Prediksi Kelulusan Mahasiswa

- Komparasi Accuracy dan AUC

	C4.5	NB	K-NN	LogR
Accuracy	91.55%	82.58%	83.63%	77.47%
AUC	0.909	0.894	0.5	0.721

- Uji Beda (t-Test)

C4.5 NB kNN LogR

	A	B	C	D	E
		0.916 +/- 0.039	0.826 +/- 0.085	0.836 +/- 0.060	0.775 +/- 0.185
C4.5	0.916 +/- 0.039		0.007	0.003	0.030
NB	0.826 +/- 0.085			0.753	0.438
kNN	0.836 +/- 0.060				0.330
LogR	0.775 +/- 0.185				

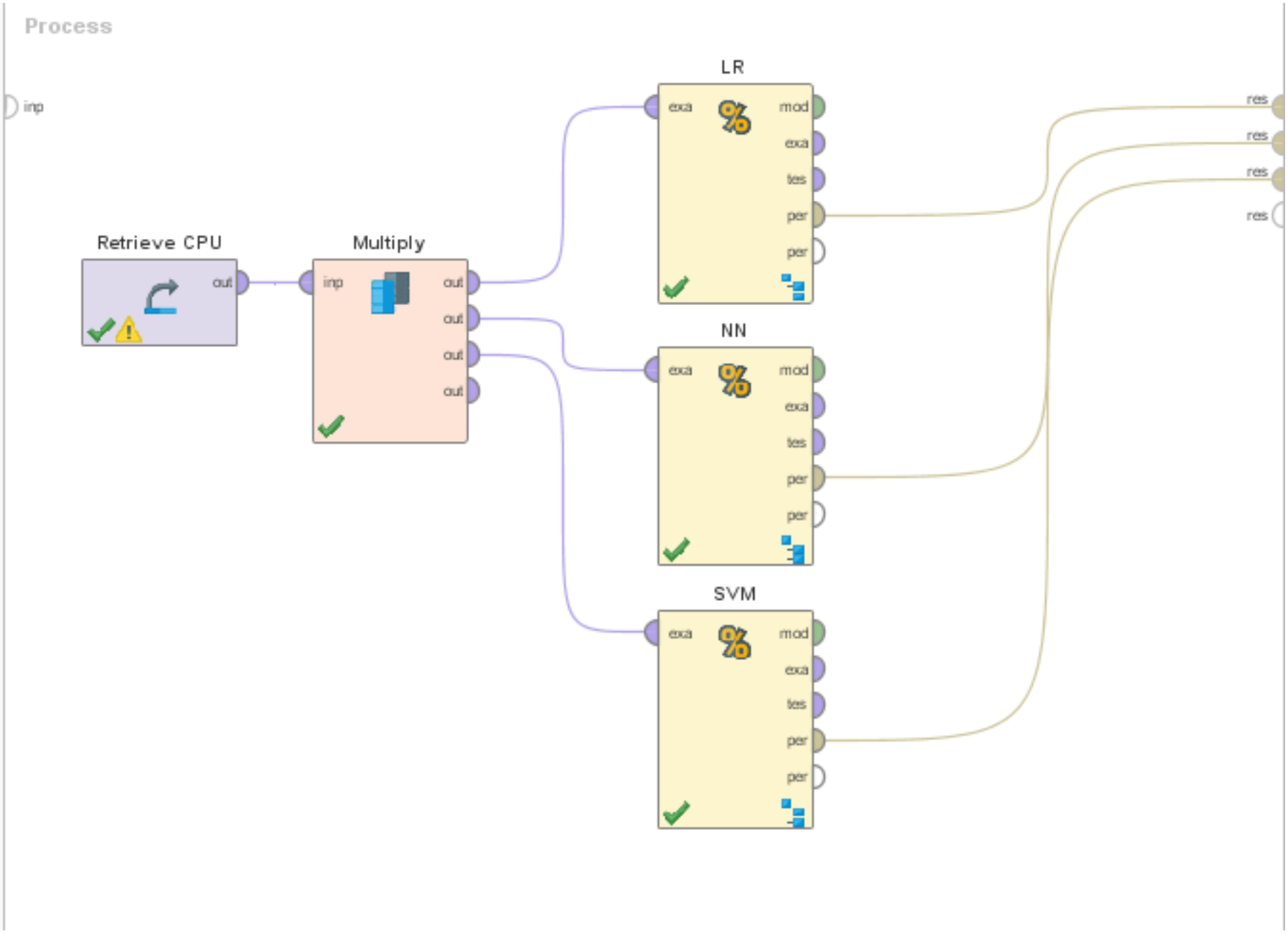
- Urutan model terbaik: 1. C4.5 2. NB 2. k-NN 2. LogR

Latihan: Estimasi Performance CPU

1. Lakukan training pada data cpu (cpu.xls) dengan menggunakan algoritma linear regression, neural network dan support vector machine
2. Lakukan pengujian dengan XValidation (*numerical*)
3. Ukur performance-nya dengan menggunakan RMSE (Root Mean Square Error)

	LR	NN	SVM
RMSE	54.676	55.192	94.676

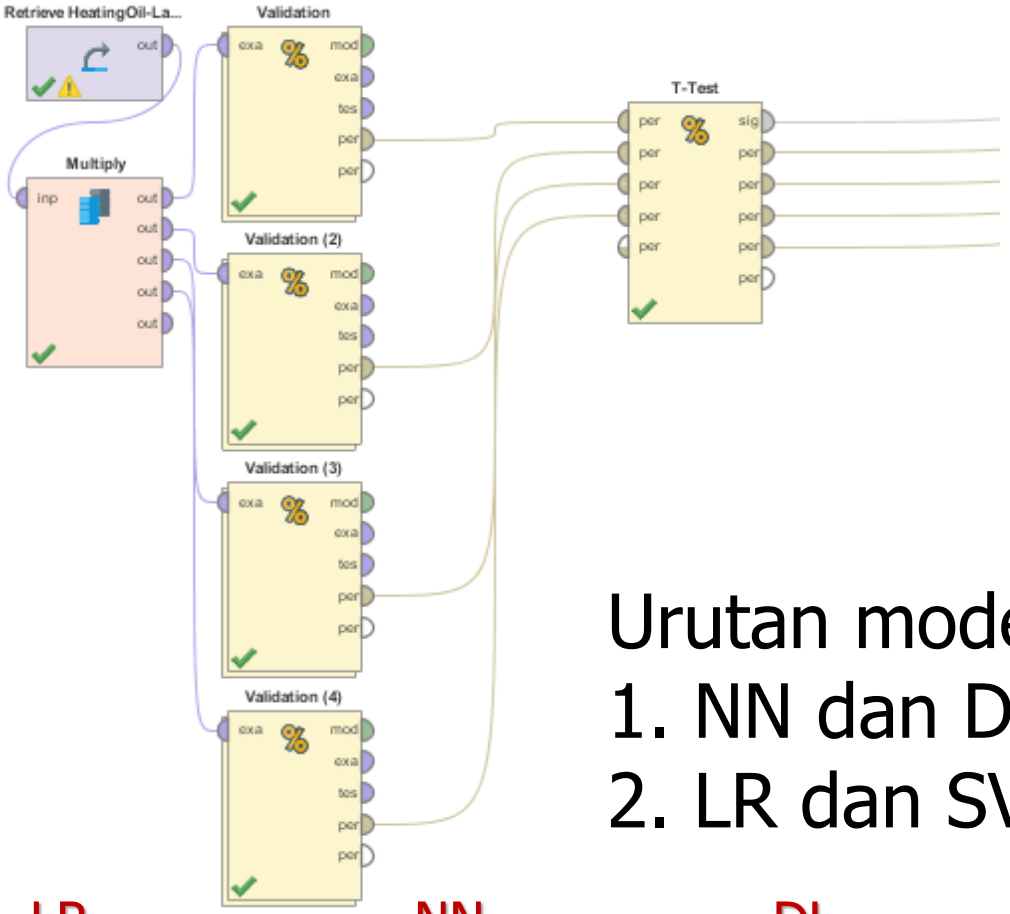
4. Urutan model terbaik: 1. LR 2. NN 3. SVM


Process
inp
Retrieve CPU
out
Multiply
inp
out
out
out
out
LR
exa
mod
exa
tes
per
per
NN
exa
mod
exa
tes
per
per
SVM
exa
mod
exa
tes
per
per
res
res
res
res

Latihan: Estimasi Konsumsi Minyak

1. Lakukan training pada data minyak pemanas (HeatingOil.csv) dengan menggunakan algoritma linear regression, neural network dan support vector machine, Deep Learning
2. Lakukan pengujian dengan XValidation (*numerical*) dan Uji beda dengan t-Test
3. Ukur performance-nya dengan menggunakan RMSE (Root Mean Square Error)

	LR	NN	SVM	DL
RMSE				



Urutan model terbaik:
1. NN dan DL
2. LR dan SVM

	LR	NN	DL	SVM
	A	B	C	D	E
LR		23.907 +/- 3.064	14.286 +/- 2.119	15.150 +/- 2.263	25.037 +/- 4.041
NN	23.907 +/- 3.064		0.000	0.000	0.490
DL	14.286 +/- 2.119			0.390	0.000
SVM	15.150 +/- 2.263				0.000
	25.037 +/- 4.041				

Latihan: Prediksi Elektabilitas Caleg

1. Lakukan training pada data pemilu (datapemilukpu.xls) dengan menggunakan algoritma Decision Tree, Naive Bayes, K-Nearest Neighbor, RandomForest, Logistic Regression
2. Lakukan pengujian dengan menggunakan XValidation
3. Ukur performance-nya dengan confusion matrix dan ROC Curve
4. Masukkan setiap hasil percobaan ke dalam file Excel

	DT	NB	K-NN	RandFor	LogReg
Accuracy	92.21%	76.89%	89.63%		
AUC	0.851	0.826	0.5		

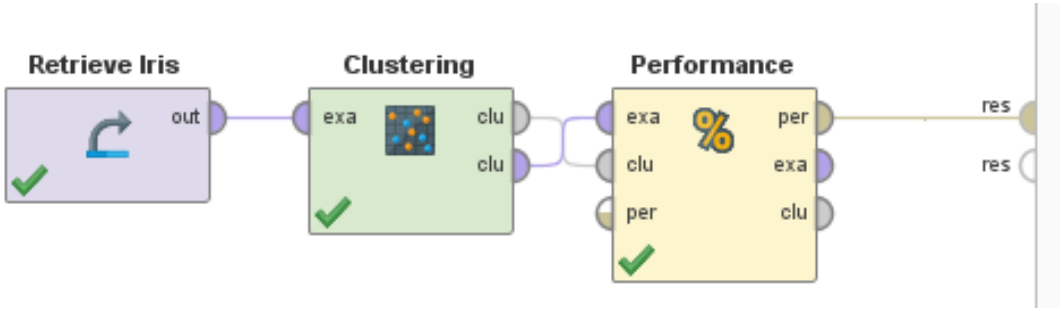
Latihan: Prediksi Harga Saham

1. Lakukan training pada data harga saham (hargasaham-training.xls) dengan neural network, linear regression, support vector machine
2. Lakukan pengujian dengan menggunakan XValidation

	LR	NN	SVM
RMSE			

Latihan: Klastering Jenis Bunga Iris

1. Lakukan training pada data iris (ambil dari repositories rapidminer) dengan menggunakan algoritma clustering k-means
2. Gunakan pilihan nilai untuk *k*, isikan dengan 3, 4, 5, 6, 7
3. Ukur performance-nya dengan *Cluster Distance Performance*, dari analisis Davies Bouldin Indeks (DBI), tentukan nilai *k* yang paling optimal



	k=3	k=4	k=5	k=6	k=7
DBI	0.666	0.764	0.806	0.910	0.99

Davies–Bouldin index (DBI)

- The Davies–Bouldin index (DBI) (introduced by David L. Davies and Donald W. Bouldin in 1979) is a metric for evaluating clustering algorithms
- This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset
- As a function of the ratio of the within cluster scatter, to the between cluster separation, a lower value will mean that the clustering is better
- This affirms the idea that no cluster has to be similar to another, and hence the best clustering scheme essentially minimizes the Davies–Bouldin index
- This index thus defined is an average over all the *i* clusters, and hence a good measure of deciding how many clusters actually exists in the data is to plot it against the number of clusters it is calculated over
- The number *i* for which this value is the lowest is a good measure of the number of clusters the data could be ideally classified into

Evaluasi 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)

Tugas: Mengolah Semua Dataset

1. Lakukan ujicoba terhadap semua dataset yang ada di folder datasets, dengan menggunakan berbagai metode data mining yang sesuai (estimasi, prediksi, klasifikasi, clustering, association)
2. Kombinasikan pengujian dengan pemecahan data training-testing, dan pengujian dengan menggunakan metode X validation
3. Ukur performance dari model yang terbentuk dengan menggunakan metode pengukuran sesuai dengan metode data mining yang dipilih
4. Jelaskan secara mendetail tahapan ujicoba yang dilakukan, kemudian lakukan analisis dan sintesis, dan buat laporan dalam bentuk slide
5. Presentasikan di depan kelas

Tugas: Mereview Paper

- Technical Paper:
  - Judul: Application and Comparison of Classification Techniques in Controlling Credit Risk
  - Author: Lan Yu, Guoqing Chen, Andy Koronios, Shiwu Zhu, and Xunhua Guo
  - Download: http://romisatriawahono.net/lecture/dm/paper/

- Baca dan pahami paper di atas dan jelaskan apa yang dilakukan peneliti pada paper tersebut:
  1. Object Penelitian
  2. Masalah Penelitian
  3. Tujuan Penelitian
  4. Metode Penelitian
  5. Hasil Penelitian

Tugas: Mereview Paper

- Technical Paper:
  - Judul: A Comparison Framework of Classification Models for Software Defect Prediction
  - Author: Romi Satria Wahono, Nanna Suryana Herman, Sabrina Ahmad
  - Publications: Adv. Sci. Lett. Vol. 20, No. 10-12, 2014
  - Download: http://romisatriawahono.net/lecture/dm/paper

- Baca dan pahami paper di atas dan jelaskan apa yang dilakukan peneliti pada paper tersebut:
  1. Object Penelitian
  2. Masalah Penelitian
  3. Tujuan Penelitian
  4. Metode Penelitian
  5. Hasil Penelitian

Tugas Mereview Paper

- Technical Paper:
  - Judul: An experimental comparison of classification algorithms for imbalanced credit scoring data sets
  - Author: Iain Brown and Christophe Mues
  - Publications: Expert Systems with Applications 39 (2012) 3446–3453
  - Download: http://romisatriawahono.net/lecture/dm/paper

- Baca dan pahami paper di atas dan jelaskan apa yang dilakukan peneliti pada paper tersebut:
  1. Object Penelitian
  2. Masalah Penelitian
  3. Tujuan Penelitian
  4. Metode Penelitian
  5. Hasil Penelitian

Tugas: Menulis Paper Penelitian

- Cari dataset yang ada di sekitar kita
- Lakukan penelitian berupa komparasi dari (minimal) 5 algoritma machine learning untuk memining knowledge dari dataset tersebut
- Gunakan uji beda (baik parametrik dan non parametric) untuk analisis dan pembuatan ranking dari algoritma machine learning
- Tulis makalah tentang penelitian yang kita buat
- Contoh-contoh makalah komparasi ada di: http://romisatriawahono.net/lecture/dm/paper/method%20comparison/
- Upload seluruh file laporan ke Card di Trello.Com
- Deadline: sehari sebelum mata kuliah berikutnya

Paper Formatting

- Ikuti template dan contoh paper dari: http://journal.ilmukomputer.org

- Isi paper:
  - Abstract: Harus berisi obyek-masalah-metode-hasil
  - Introduction: Latar belakang masalah penelitian dan struktur paper
  - Related Work: Penelitian yang berhubungan
  - Theoretical Foundation: Landasan dari berbagai teori yang digunakan
  - Proposed Method: Metode yang diusulkan
  - Experimental Results: Hasil eksperimen
  - Conclusion: Kesimpulan dan future works

Competency Check

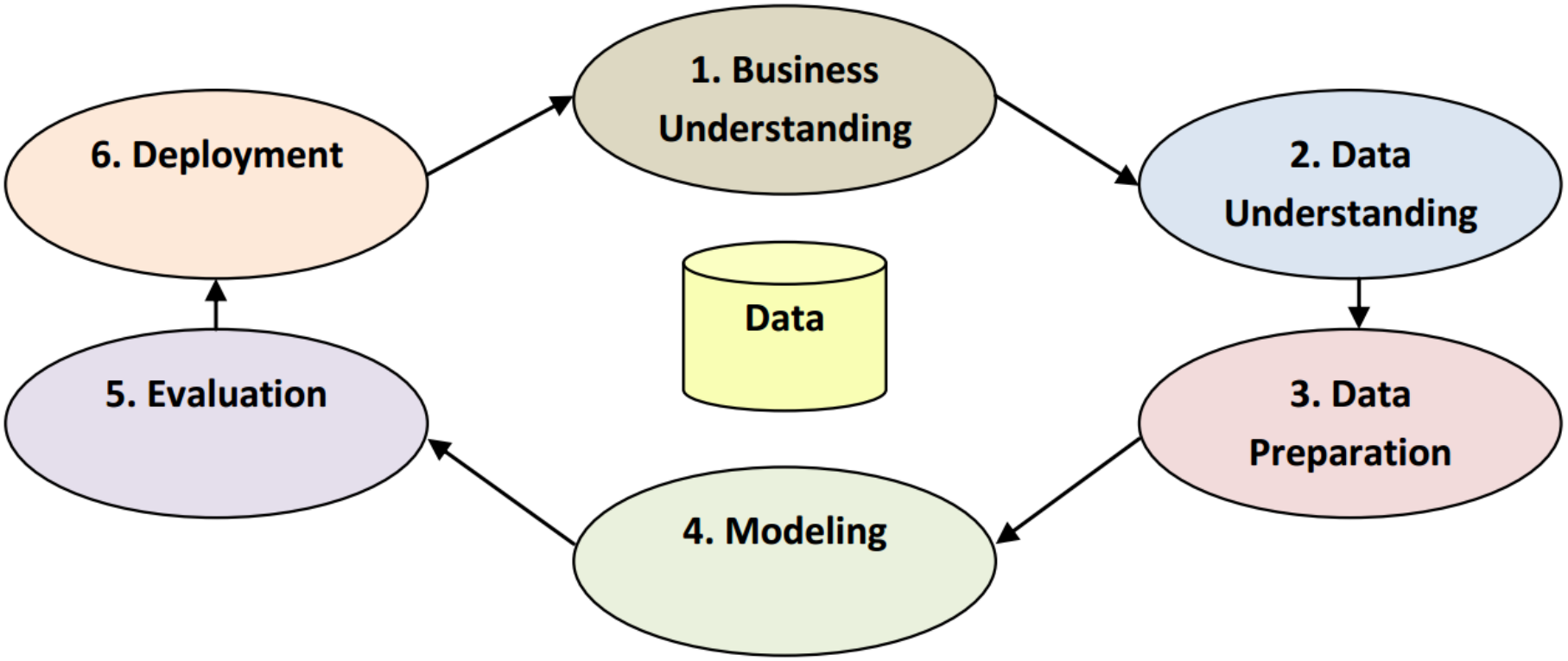
1. Dataset – Methods – Knowledge
    1. Dataset Main Golf (Klasifikasi)
    2. Dataset Iris (Klasifikasi)
    3. Dataset Iris (Klastering)
    4. Dataset CPU (Estimasi)
    5. Dataset Pemilu (Klasifikasi)
    6. Dataset Heating Oil (Association)
    7. Dataset Transaksi (Association)
    8. Dataset Harga Saham (Forecasting)
2. Dataset – Methods – Knowledge – Evaluation
    1. Manual
    2. Data Split Operator
    3. Cross Validation
3. Methods Comparison
    - Uji t-Test
4. Paper Reading
    1. Lan Yu (DeLong Pearson Test)
    2. Wahono (Friedman Test)

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

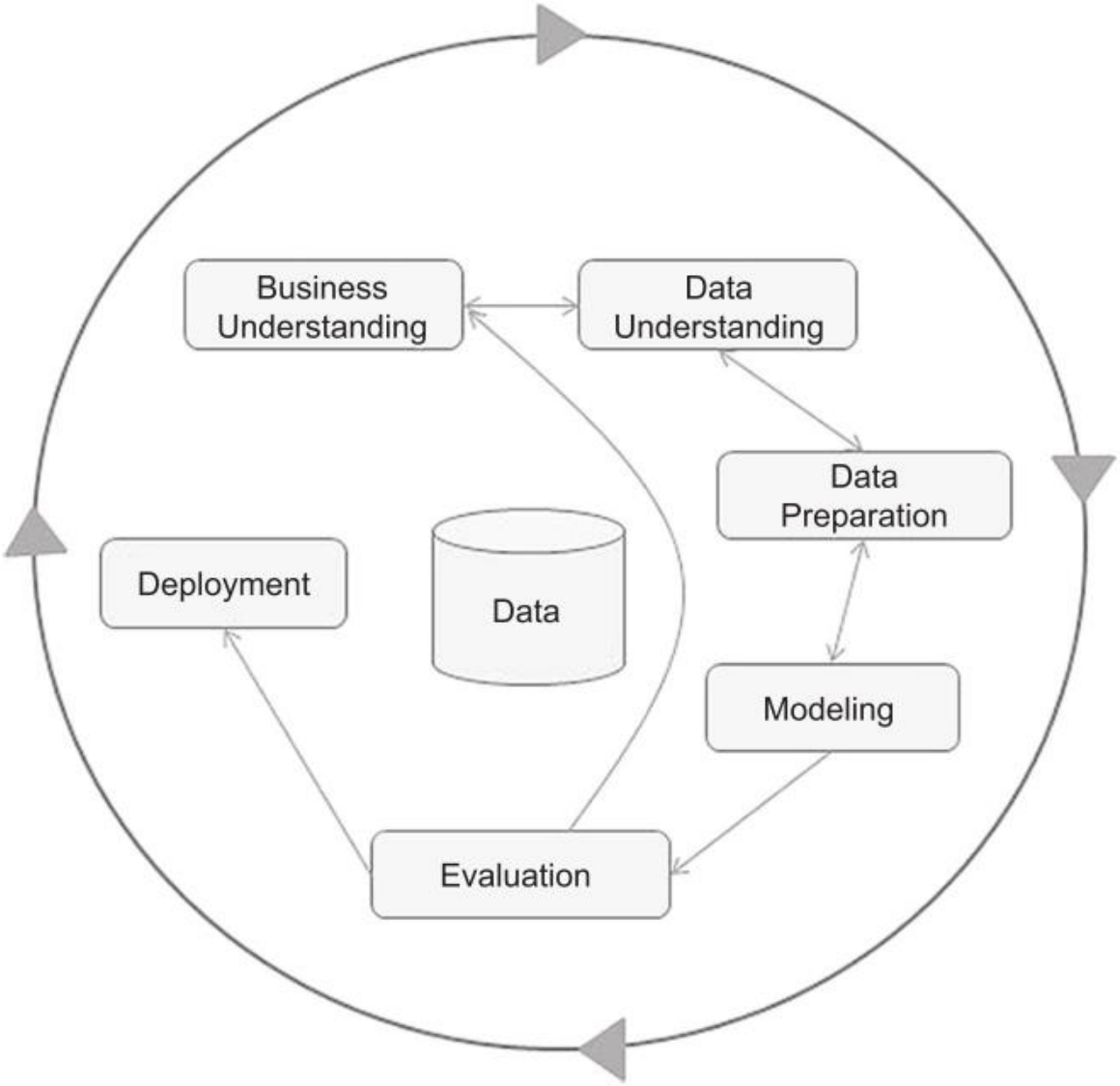
5. Evaluation

- 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



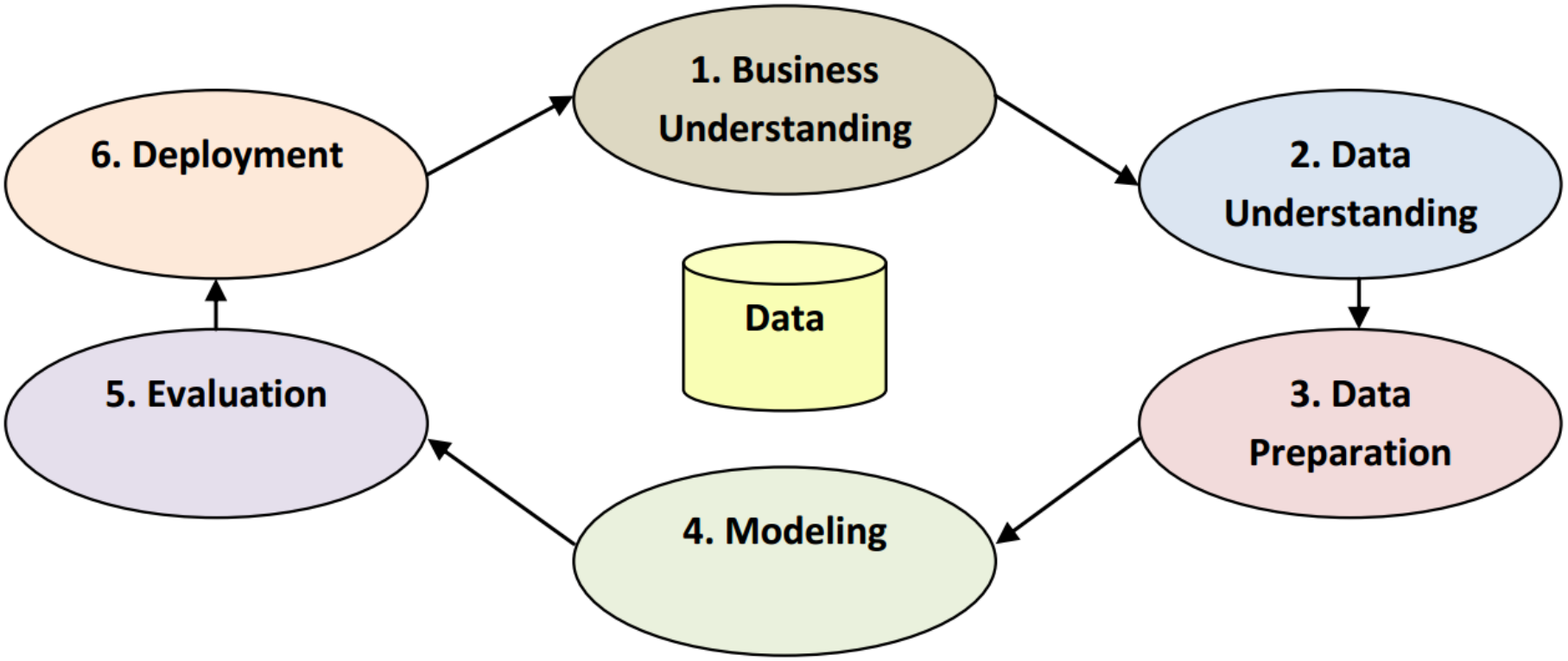
Studi Kasus CRISP-DM

Heating Oil Consumption – Correlational Methods

(Matthew North, Data Mining for the Masses 2nd Edition, 2016, ***Chapter 4 Correlational Methods, pp. 69-76****)*

Dataset: HeatingOil.csv

CRISP-DM


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

1. Business Understanding

- Problems:
  - Sarah is a regional sales manager for a nationwide supplier of fossil fuels for home heating
  - Marketing performance is very poor and decreasing, while marketing spending is increasing
  - She feels a need to understand the types of behaviors and other factors that may influence the demand for heating oil in the domestic market
  - She recognizes that there are many factors that influence heating oil consumption, and believes that by investigating the relationship between a number of those factors, she will be able to better monitor and respond to heating oil demand, and also help her to design marketing strategy in the future
- Objective:
  - To investigate the relationship between a number of factors that influence heating oil consumption

2. Data Understanding

- In order to investigate her question, Sarah has enlisted our help in creating a correlation matrix of six attributes
- Using employer's data resources which are primarily drawn from the company's billing database, we create a data set comprised of the following attributes:
  1. Insulation: This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation
  2. Temperature: This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit
  3. Heating_Oil: This is the total number of units of heating oil purchased by the owner of each home in the most recent year
  4. Num_Occupants: This is the total number of occupants living in each home
  5. Avg_Age: This is the average age of those occupants
  6. Home_Size: This is a rating, on a scale of one to eight, of the home's overall size. The higher the number, the larger the home

3. Data Preparation

Data set: HeatingOil.csv

Result History | ExampleSet (Retrieve HeatingOil)

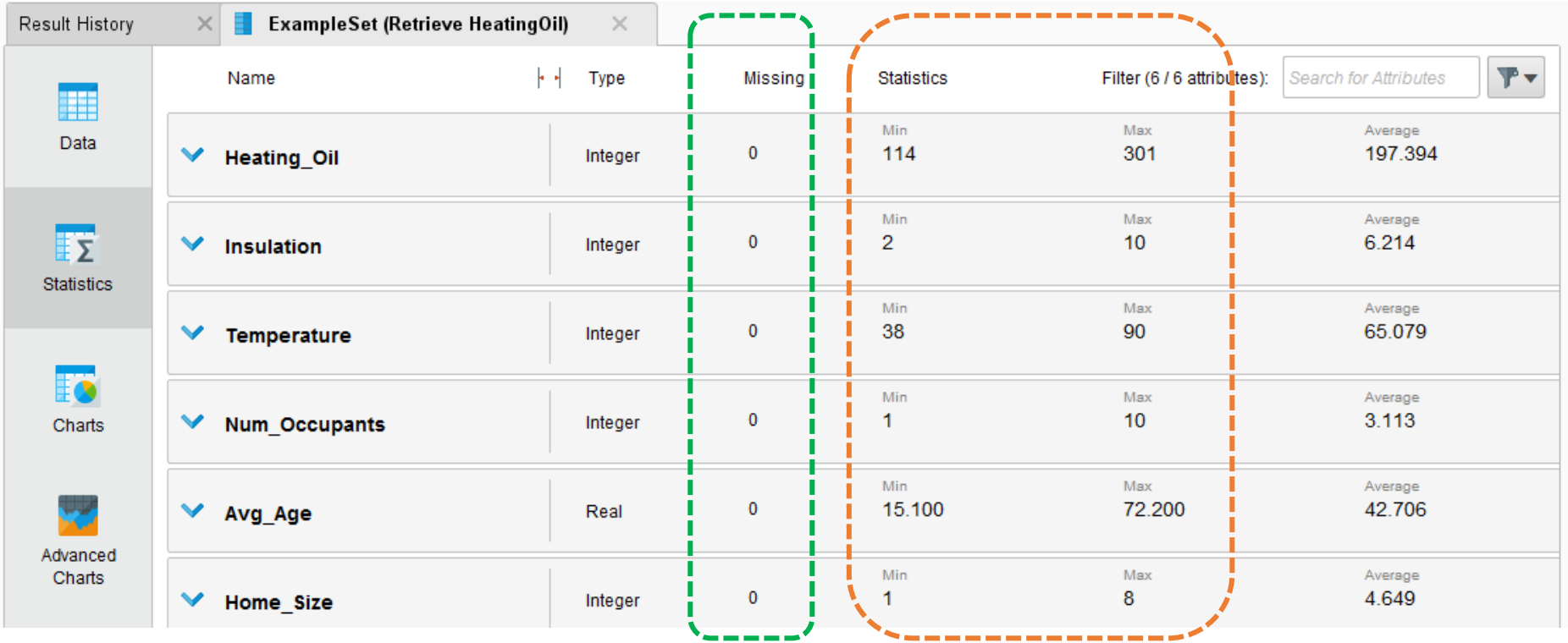
Data | Statistics | Charts | Advanced Charts | Annotations

ExampleSet (1218 examples, 0 special attributes, 6 regular attributes) Filter (1,218 / 1,218 examples):

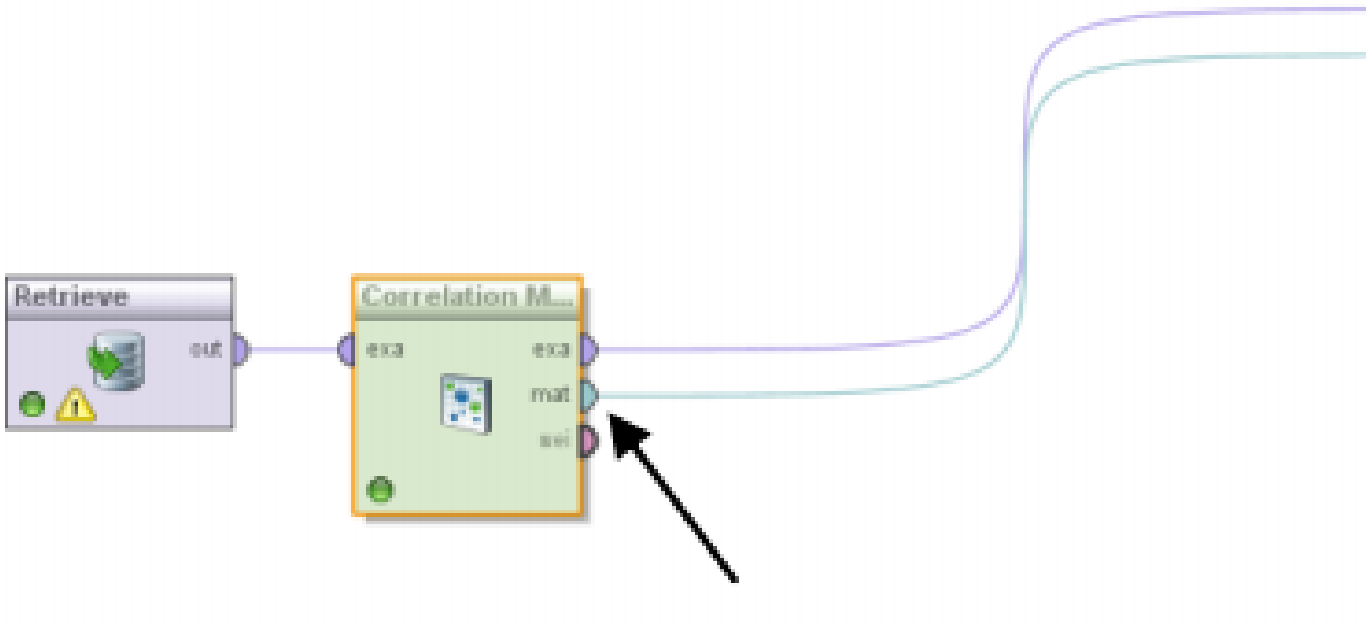
Row No.	Heating_Oil	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	132	6	74	4	23.800	4
2	263	10	43	4	56.700	4
3	145	3	81	2	28	6
4	196	9	50	4	45.100	3
5	131	2	80	5	20.800	2
6	129	5	76	3	21.500	3
7	131	5	72	4	23.500	3
8	161	6	88	2	38.200	6
9	184	5	77	3	42.500	3
10	225	10	42	3	51.100	1
11	178	6	90	2	42.100	2
12	121	3	83	1	19.800	2
13	186	10	43	5	45.100	6
14	206	8	59	2	50.100	8
15	179	4	86	5	41.400	6
16	156	4	80	3	32.800	3
17	135	4	78	4	22.800	5
18	186	4	76	1	50.500	4

3. Data Preparation

- Data set appears to be very clean with:
  - No missing values in any of the six attributes
  - No inconsistent data apparent in our ranges (Min-Max) or other descriptive statistics



4. Modeling



4. Modeling

- Hasil correlation matrix berupa tabel
- Semakin tinggi nilainya (semakin tebal warna ungu), semakin tinggi tingkat korelasinya

Attributes	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num_Occupants	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1

5. Evaluation

Positive Correlation

↑ Heating Oil use rises	↑ Insulation rating also rises		↓ Heating Oil use falls	↓ Insulation rating also falls

Whenever both attribute values move in the same direction, the correlation is positive.

Negative Correlation

↑ Temperature rises	↓ Insulation rating falls		↓ Temperature falls	↑ Insulation rating rises

Whenever attribute values move in opposite directions, the correlation is negative.

-1 ← -0.8	-0.8 ← -0.6	-0.6 ← -0.4	-0.4 ← 0	0 → 0.4	0.4 → 0.6	0.6 → 0.8	0.8 → 1.0
Very Strong Correlation	Strong Correlation	Some Correlation	No correlation	No correlation	Some correlation	Strong correlation	Very strong correlation

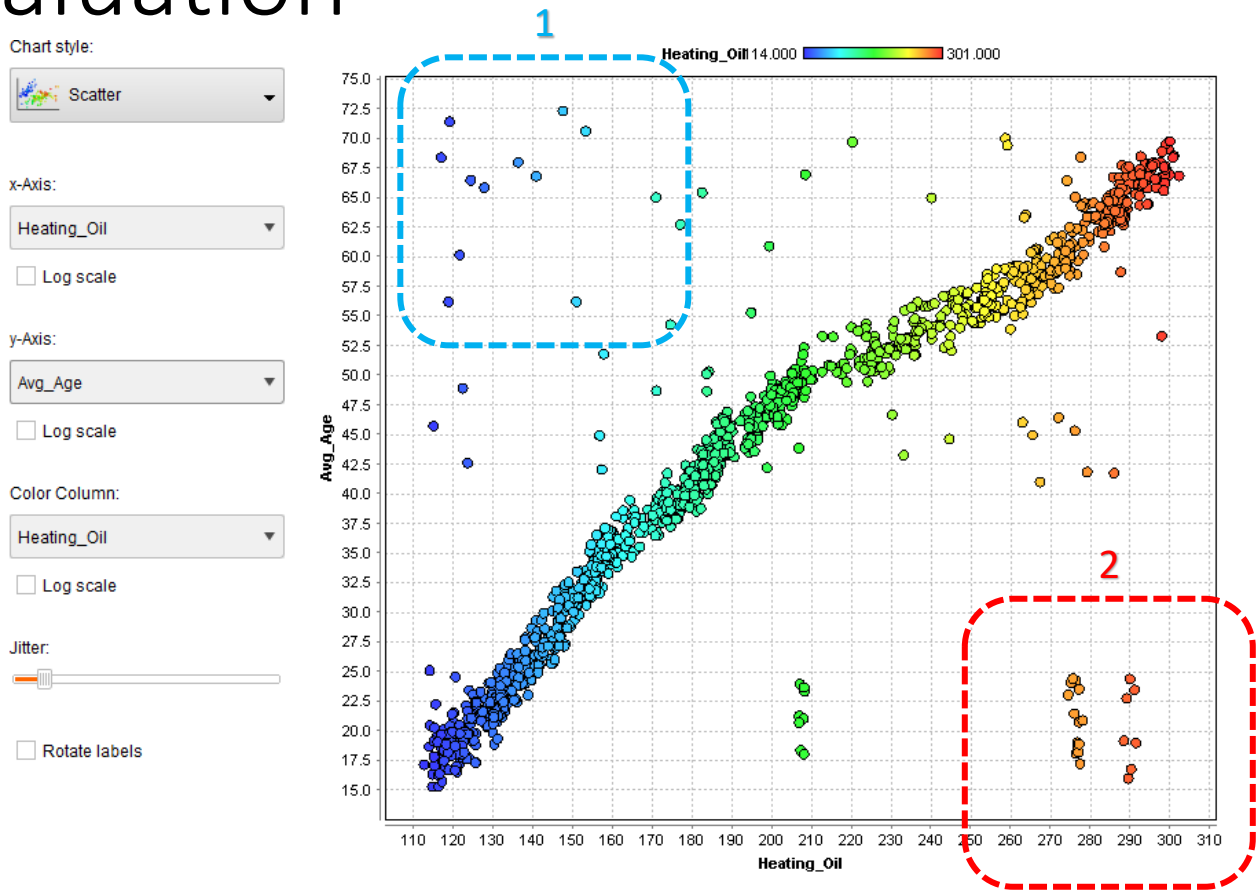
Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

5. Evaluation

- Atribut (faktor) yang paling signifikan berpengaruh (hubungan positif) pada konsumsi minyak pemanas (Heating Oil) adalah Average Age (Rata-Rata Umur) penghuni rumah
- Atribut (faktor) kedua yang paling berpengaruh adalah Temperature (hubungan negatif)
- Atribut (faktor) ketiga yang paling berpengaruh adalah Insulation (hubungan positif)
- Atribut Home Size, pengaruhnya sangat kecil, sedangkan Num_Occupant boleh dikatakan tidak ada pengaruh ke konsumsi minyak pemanas

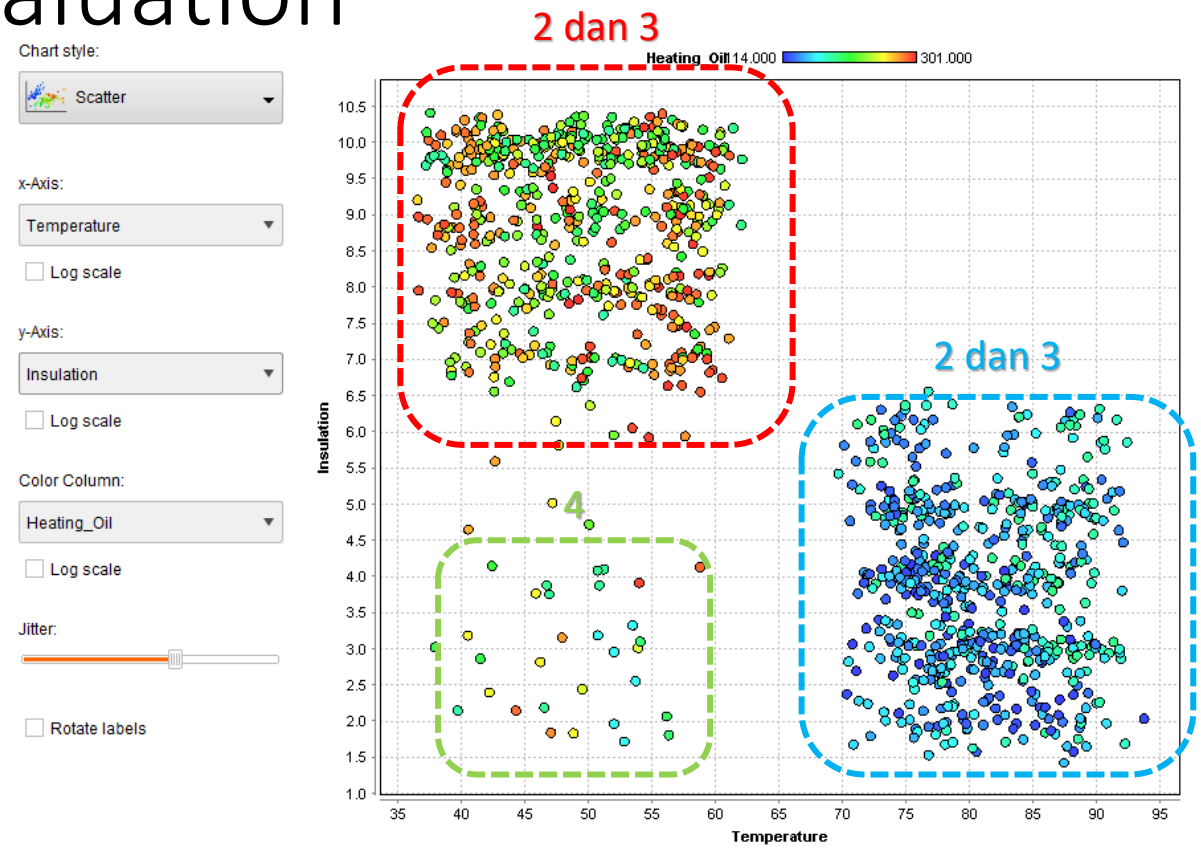
Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

5. Evaluation



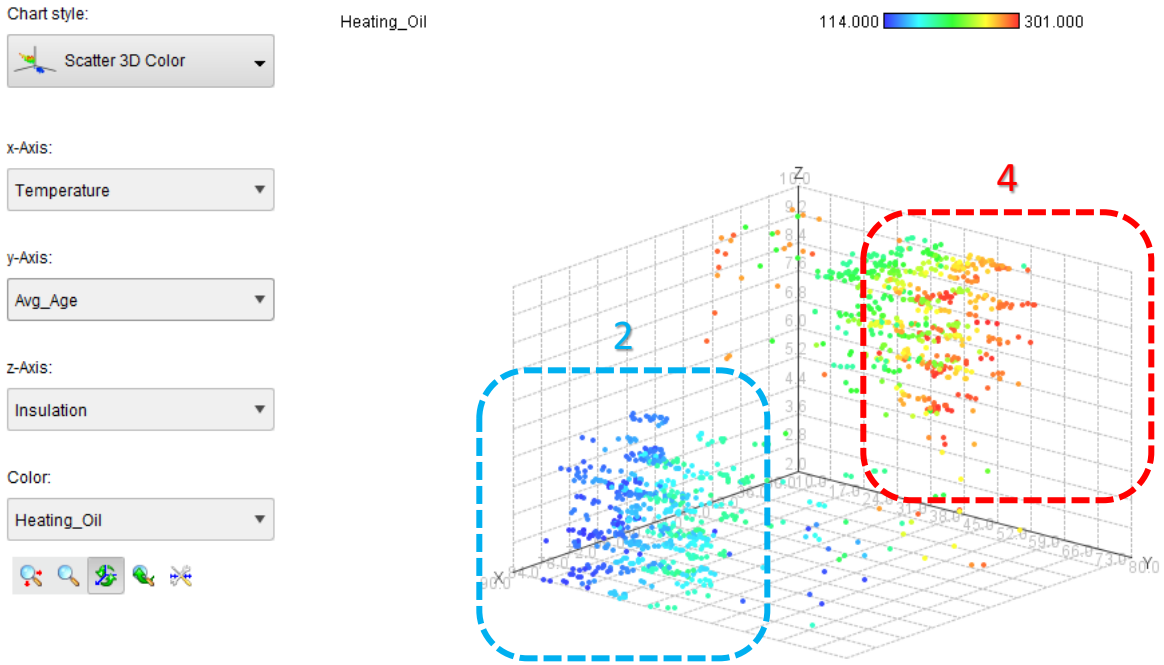
- Grafik menunjukkan bahwa konsumsi minyak memiliki korelasi positif dengan rata-rata usia
- Meskipun ada beberapa anomali juga terjadi:
  1. Ada beberapa orang yang rata-rata usia tinggi, tapi kebutuhan minyaknya rendah (warna biru muda di kolom kiri bagian atas)
  2. Ada beberapa orang yang rata-rata usia rendah, tapi kebutuhan minyaknya tinggi (warna merah di kolom kanan bagian bawah)

5. Evaluation



1. Grafik menunjukkan hubungan antara temperature dan insulation, dengan warna adalah konsumsi minyak (semakin merah kebutuhan minyak semakin tinggi)
2. Secara umum dapat dikatakan bahwa hubungan temperatur dengan insulation dan konsumsi minyak adalah negatif. Jadi temperatur semakin rendah, kebutuhan minyak semakin tinggi (kolom kiri bagian atas) ditunjukkan dengan banyak yang berwarna kuning dan merah
3. Insulation juga berhubungan negatif dengan temperatur, sehingga makin rendah temperatur, semakin butuh insulation
4. Beberapa anomali terdapat pada Insulation yang rendah nilainya, ada beberapa yang masih memerlukan minyak yang tinggi

5. Evaluation



1. Grafik tiga dimensi menunjukkan hubungan antara temperatur, rata-rata usia dan insulation
2. Warna menunjukkan kebutuhan minyak, semakin memerah maka semakin tinggi
3. Temperatur semakin tinggi semakin tidak butuh minyak (warna biru tua
4. Rata-rata usia dan insulation semakin tinggi semakin butuh minyak

6. Deployment

Dropping the Num_Occupants attribute

- While the number of people living in a home might logically seem like a variable that would influence energy usage, in our model it did not correlate in any significant way with anything else
- Sometimes there are attributes that don't turn out to be very interesting

Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

6. Deployment

Adding additional attributes to the data set

- It turned out that the number of occupants in the home didn't correlate much with other attributes, but that doesn't mean that other attributes would be equally uninteresting
- For example, what if Sarah had access to the number of furnaces and/or boilers in each home?
- Home_size was slightly correlated with Heating_Oil usage, so perhaps the number of instruments that consume heating oil in each home would tell an interesting story, or at least add to her insight

6. Deployment

Investigating the role of home insulation

- The Insulation rating attribute was fairly strongly correlated with a number of other attributes
- There may be some opportunity there to partner with a company that specializes in adding insulation to existing homes

Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

6. Deployment

Focusing the marketing efforts to the city with low temperature and high average age of citizen

- The temperature attribute was fairly strongly negative correlated with a heating oil consumption
- The average age attribute was strongest positive correlated with a heating oil consumption

Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.736	-0.774	-0.042	0.848	0.381
Insulation	0.736	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.673	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.673	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

6. Deployment

Adding greater granularity in the data set

- This data set has yielded some interesting results, but it's pretty general
- We have used average yearly temperatures and total annual number of heating oil units in this model
- But we also know that temperatures fluctuate throughout the year in most areas of the world, and thus monthly, or even weekly measures would not only be likely to show more detailed results of demand and usage over time, but the correlations between attributes would probably be more interesting
- From our model, Sarah now knows how certain attributes interact with one another, but in the day-to-day business of doing her job, she'll probably want to know about usage over time periods shorter than one year

Studi Kasus CRISP-DM

Heating Oil Consumption – Linear Regression

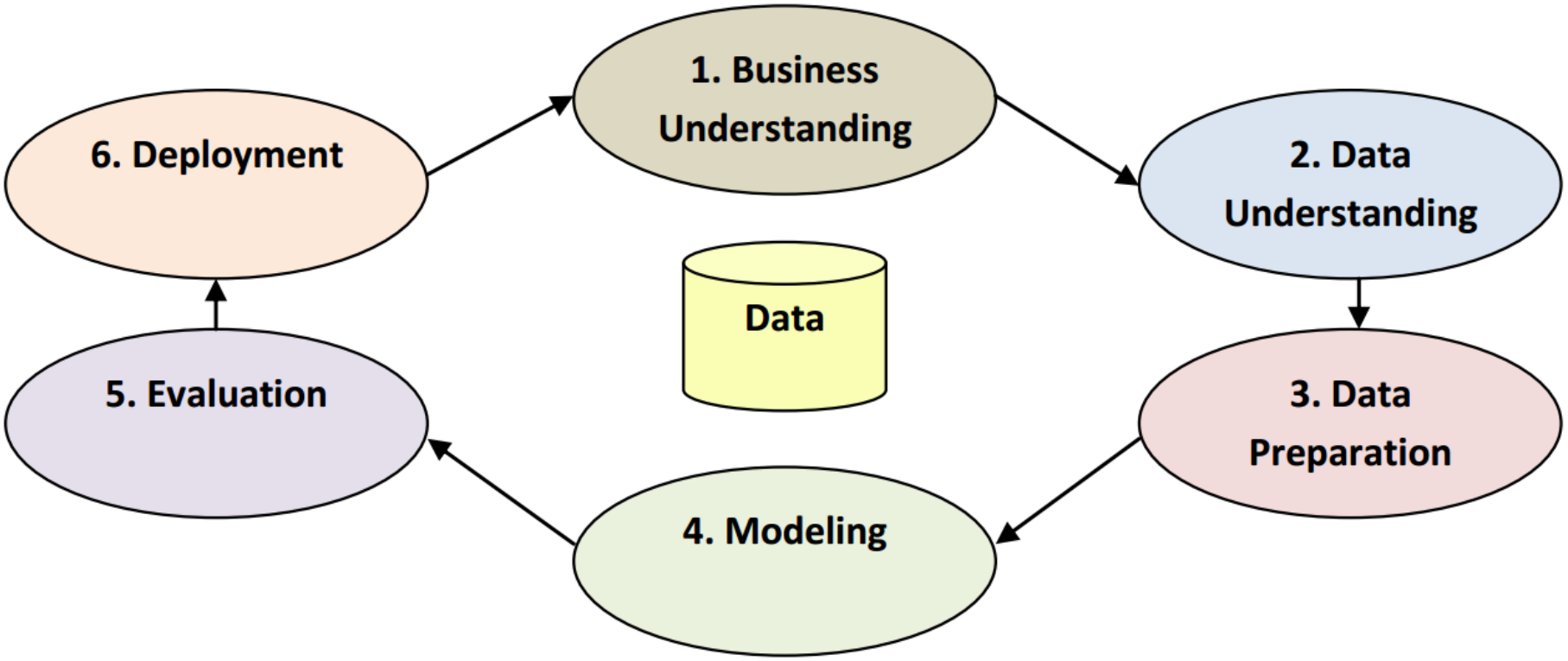
(Matthew North, Data Mining for the Masses 2nd Edition, 2016, ***Chapter 8 Linear Regression, pp. 159-171****)*

Dataset: HeatingOil.csv

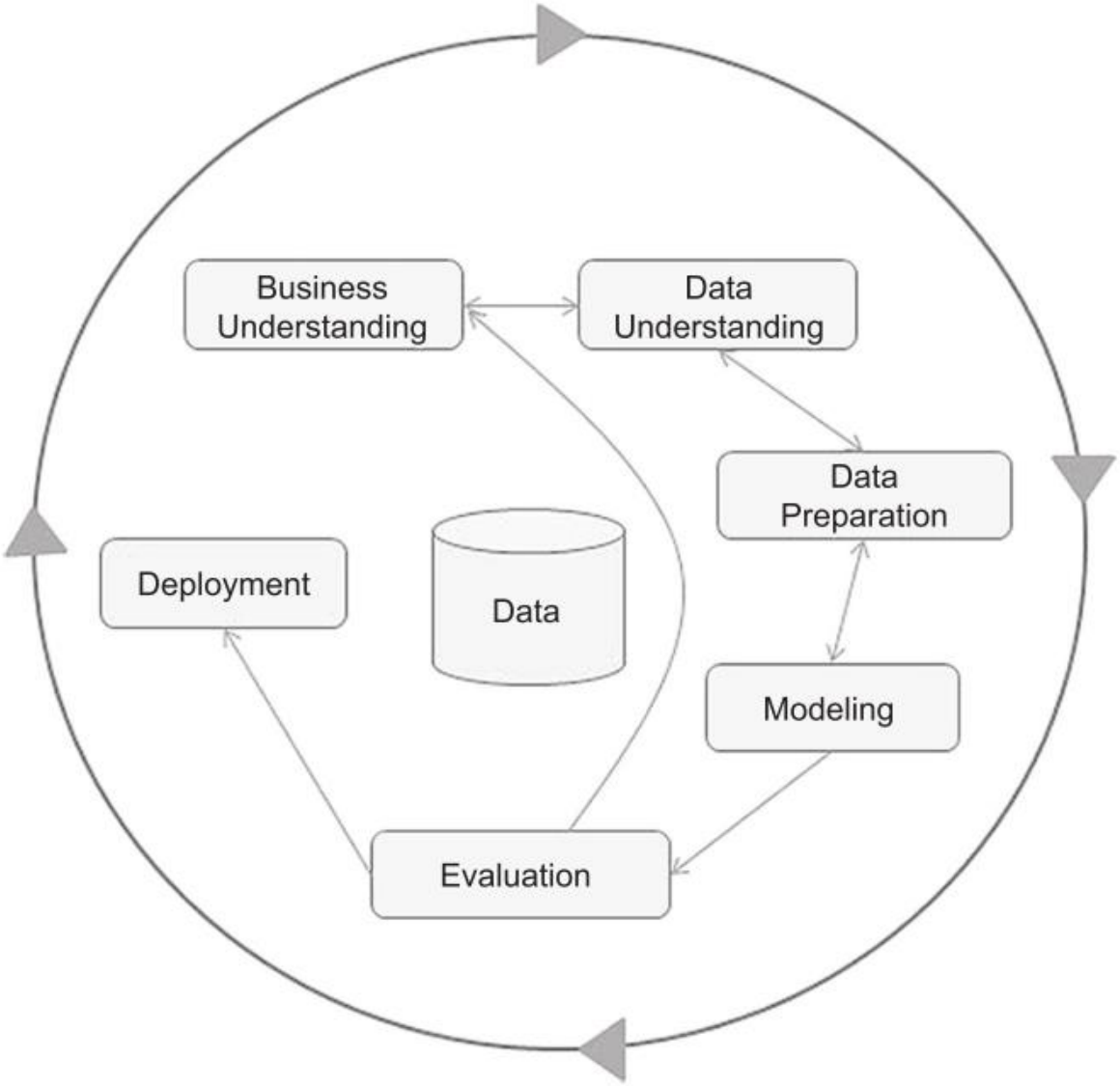
Dataset: HeatingOil-scoring.csv

http://romisatriawahono.net/lecture/dm/dataset/

CRISP-DM


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

CRISP-DM: Detail Flow



1. Business Understanding

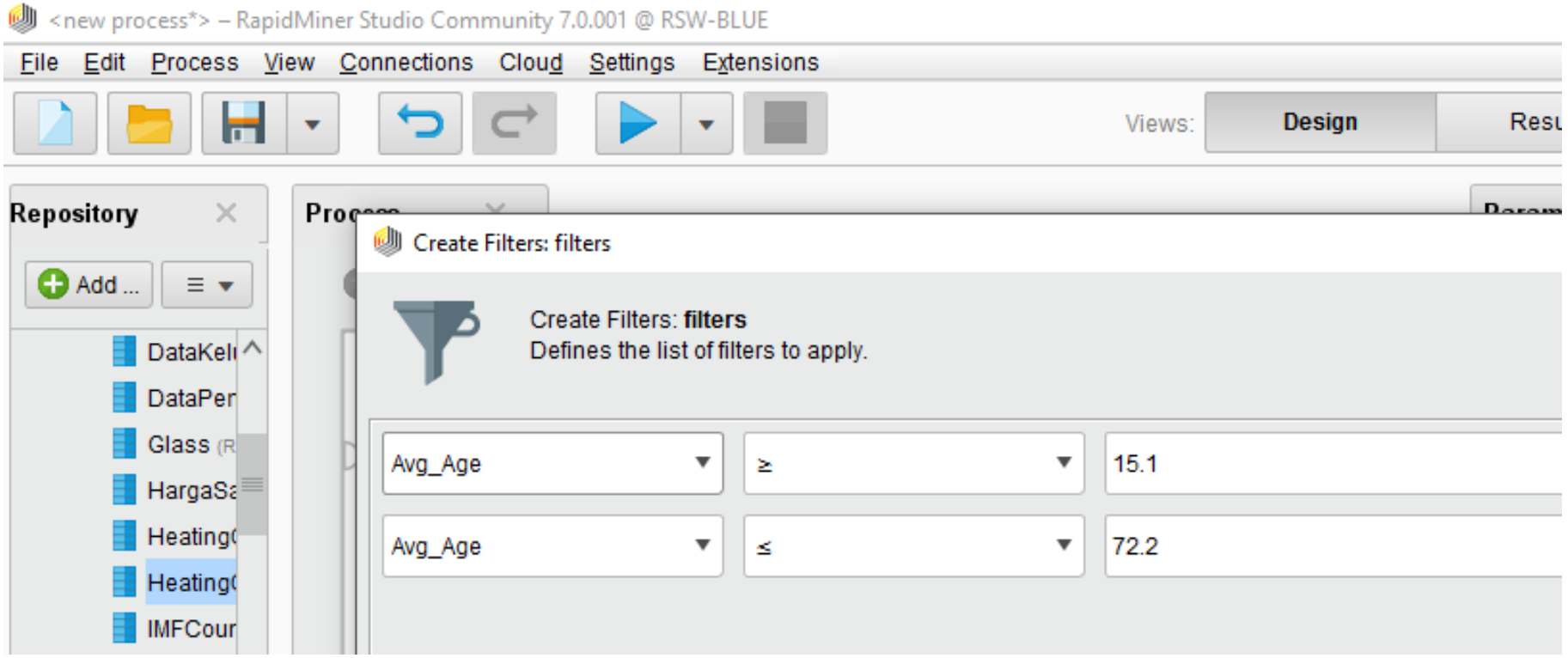
- Problems:
  - Business is booming, her sales team is signing up thousands of new clients, and she wants to be sure the company will be able to meet this new level of demand
  - Sarah's new data mining objective is pretty clear: she wants to anticipate demand for a consumable product
  - We will use a linear regression model to help her with her desired predictions. She has data, 1,218 observations that give an attribute profile for each home, along with those homes' annual heating oil consumption
  - She wants to use this data set as training data to predict the usage that 42,650 new clients will bring to her company
  - She knows that these new clients' homes are similar in nature to her existing client base, so the existing customers' usage behavior should serve as a solid gauge for predicting future usage by new customers
- Objective:
  - to predict the usage that 42,650 new clients will bring to her company

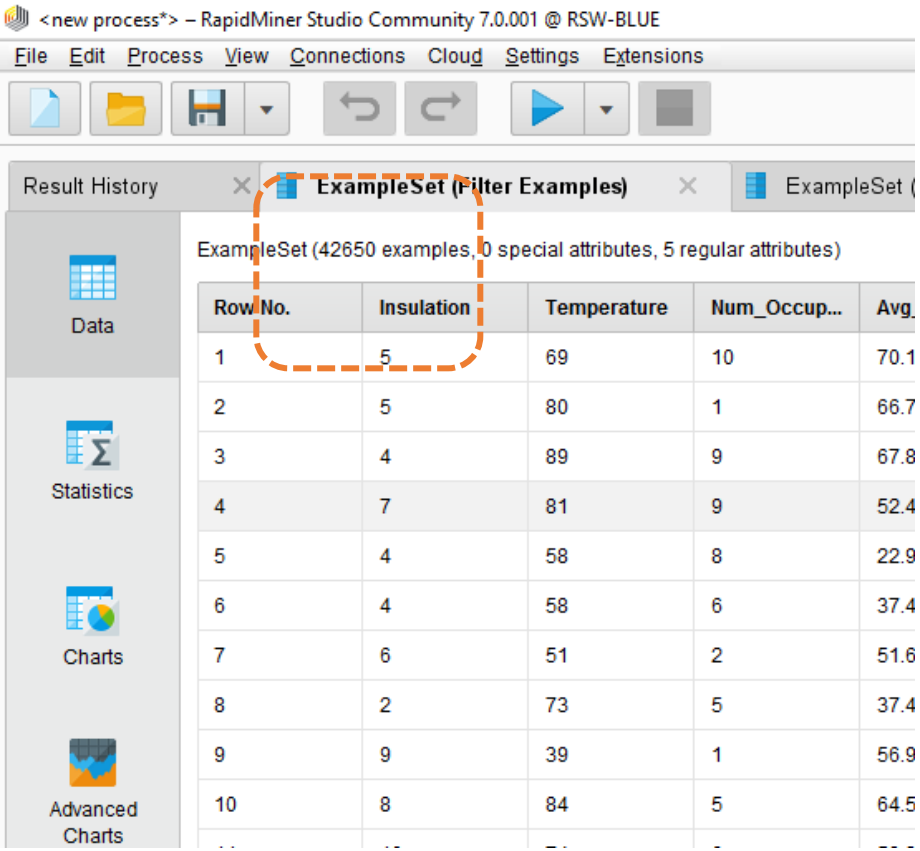
2. Data Understanding

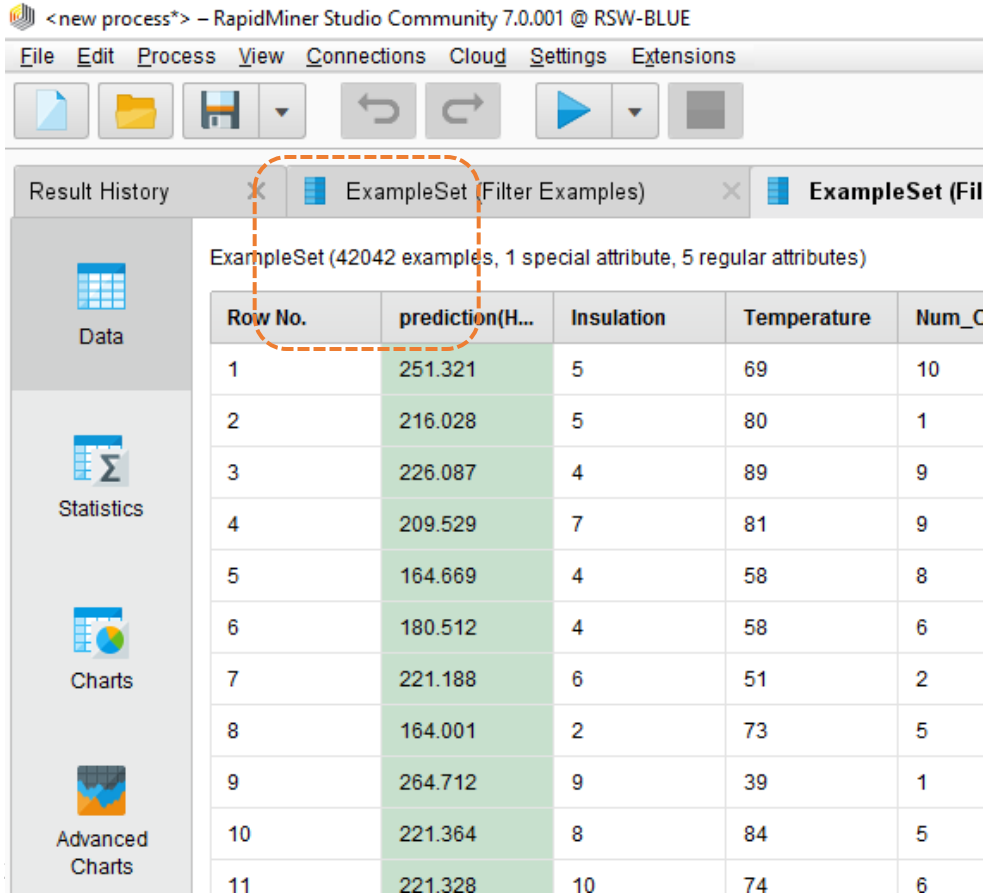
- Sarah has assembled separate Comma Separated Values file containing all of these same attributes, for her 42,650 new clients
- She has provided this data set to us to use as the scoring data set in our model
- Data set comprised of the following attributes:
  - Insulation: This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation
  - Temperature: This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit
  - Heating_Oil: This is the total number of units of heating oil purchased by the owner of each home in the most recent year
  - Num_Occupants: This is the total number of occupants living in each home
  - Avg_Age: This is the average age of those occupants
  - Home_Size: This is a rating, on a scale of one to eight, of the home's overall size. The higher the number, the larger the home

3. Data Preparation

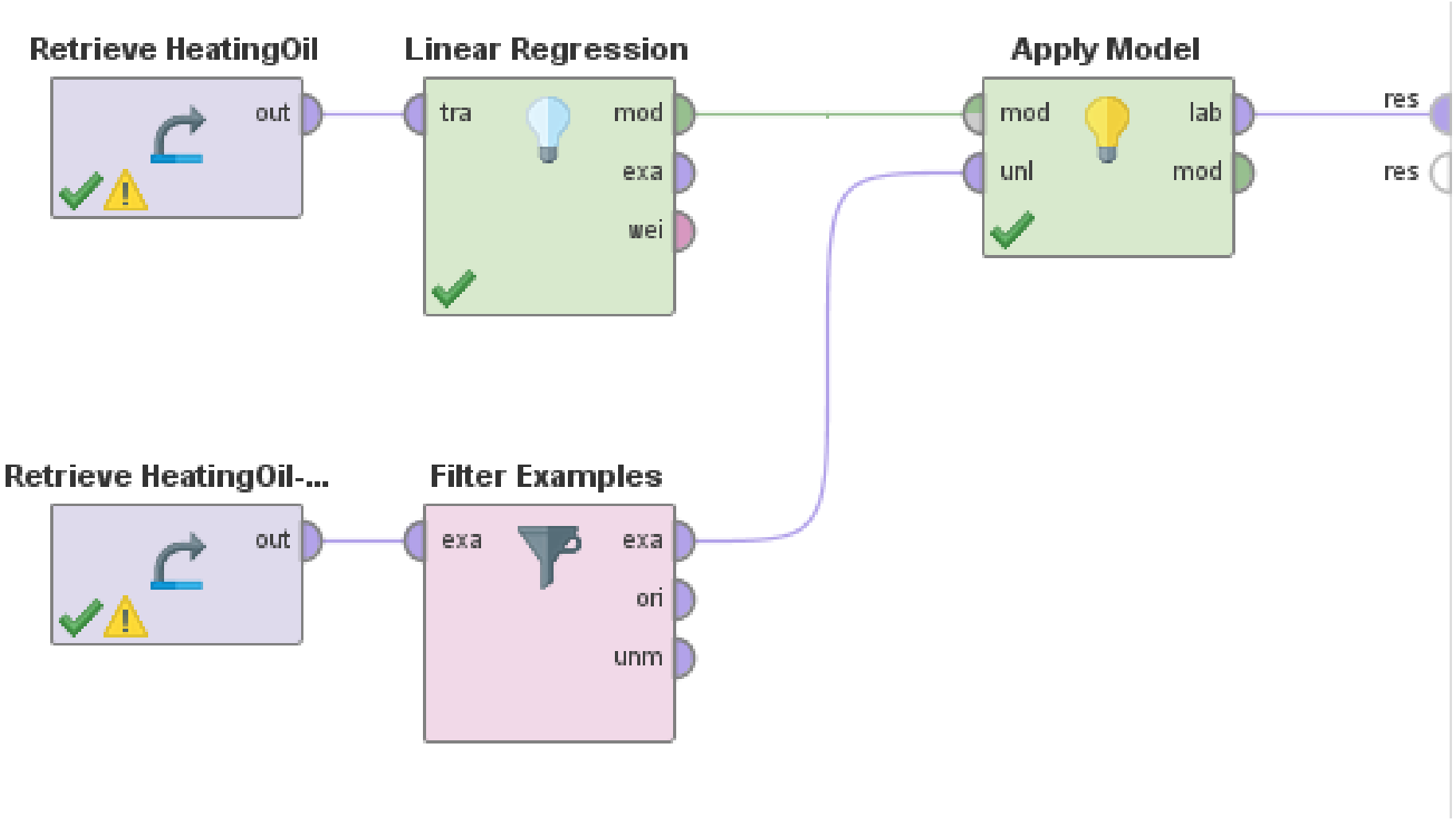
- Filter Examples: attribute value filter or custom filter
  - Avg_Age>=15.1
  - Avg_Age<=72.2
- Deleted Records= 42650-42042 = 508



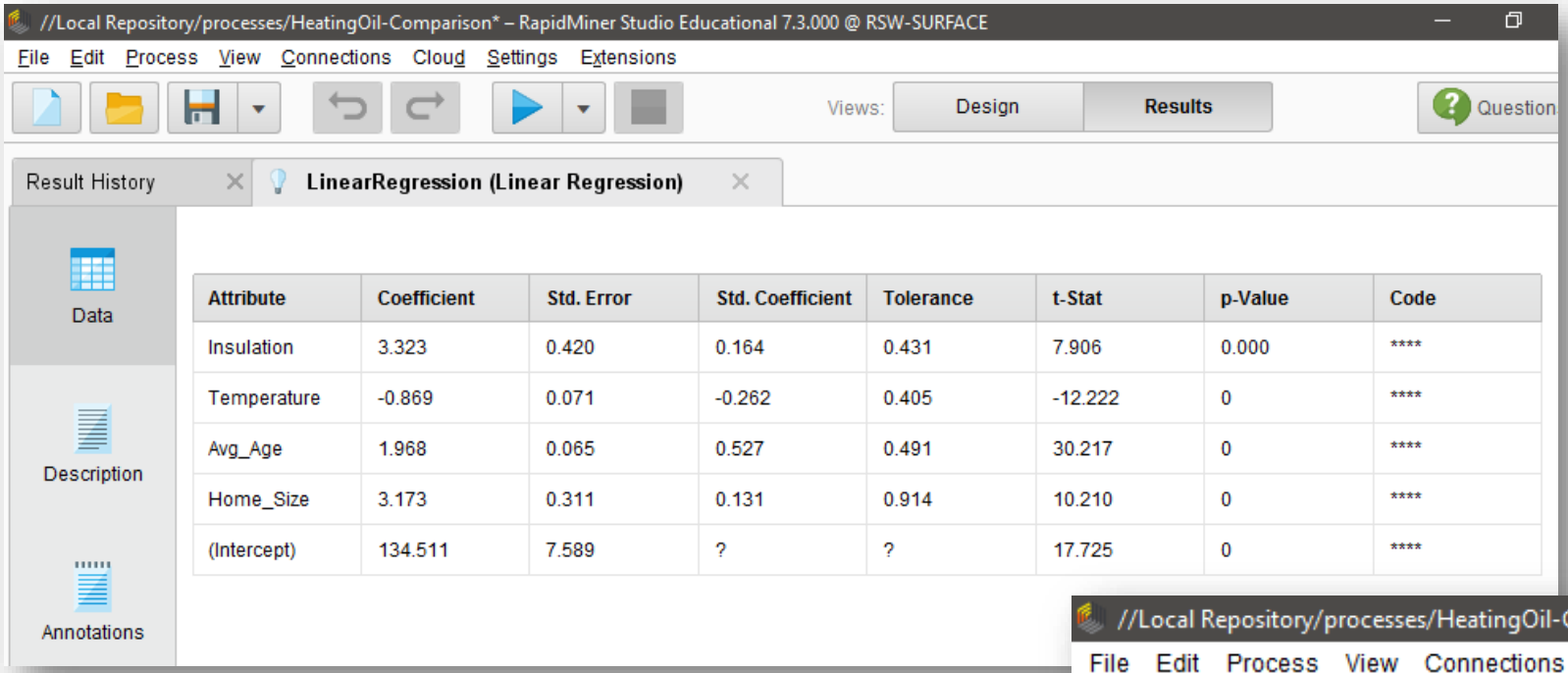




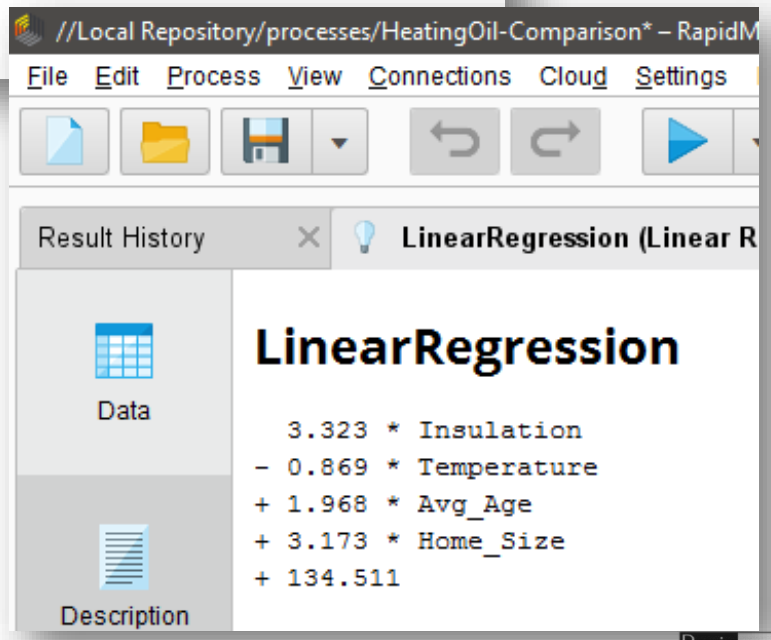
3. Modeling


Retrieve HeatingOil
out
Linear Regression
tra
mod
exa
wei
Apply Model
mod
unl
lab
mod
res
res
Retrieve HeatingOil-...
out
Filter Examples
exa
exa
ori
unm

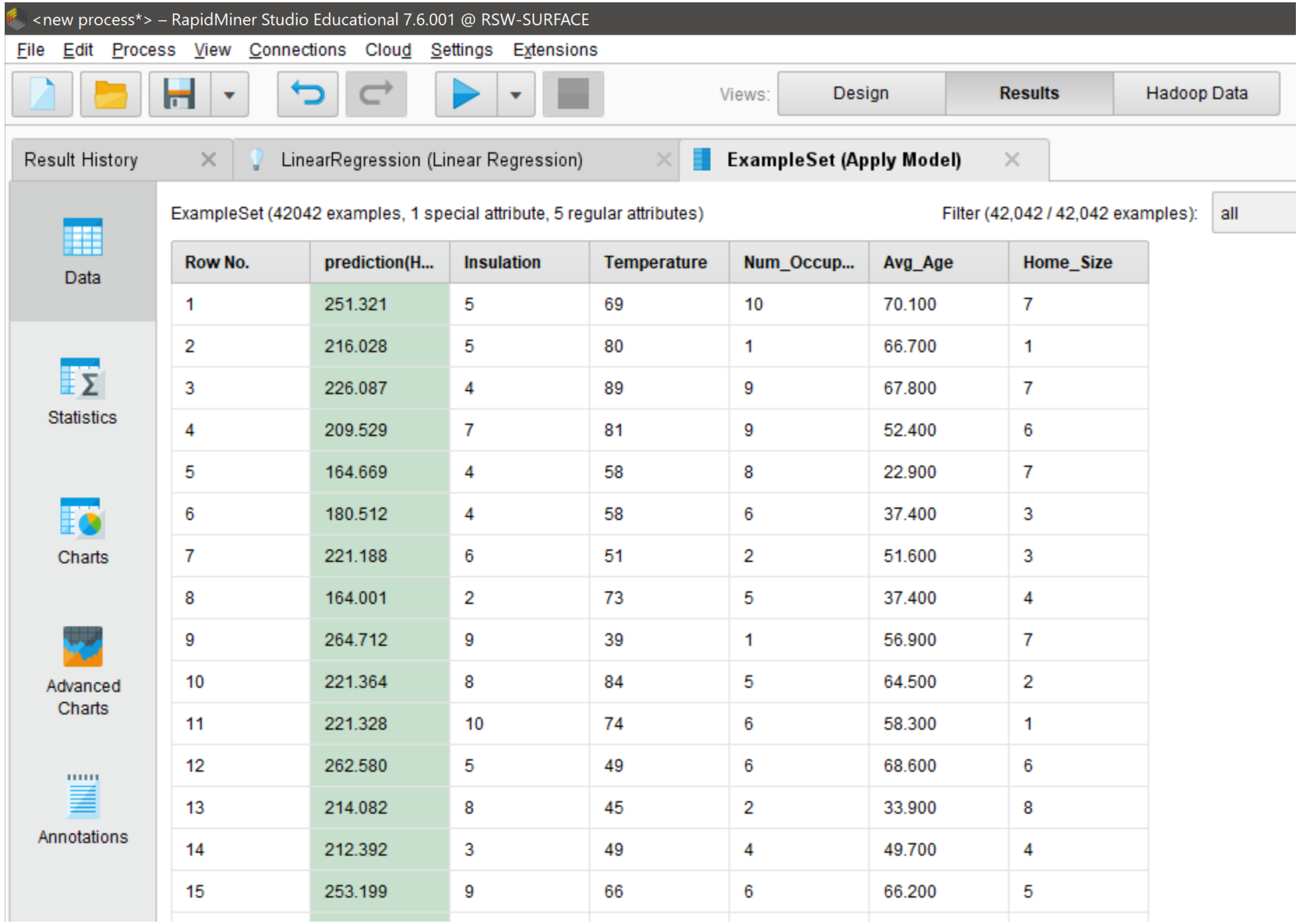
4. Evaluation – Model Regresi



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Insulation	3.323	0.420	0.164	0.431	7.906	0.000	****
Temperature	-0.869	0.071	-0.262	0.405	-12.222	0	****
Avg_Age	1.968	0.065	0.527	0.491	30.217	0	****
Home_Size	3.173	0.311	0.131	0.914	10.210	0	****
(Intercept)	134.511	7.589	?	?	17.725	0	****

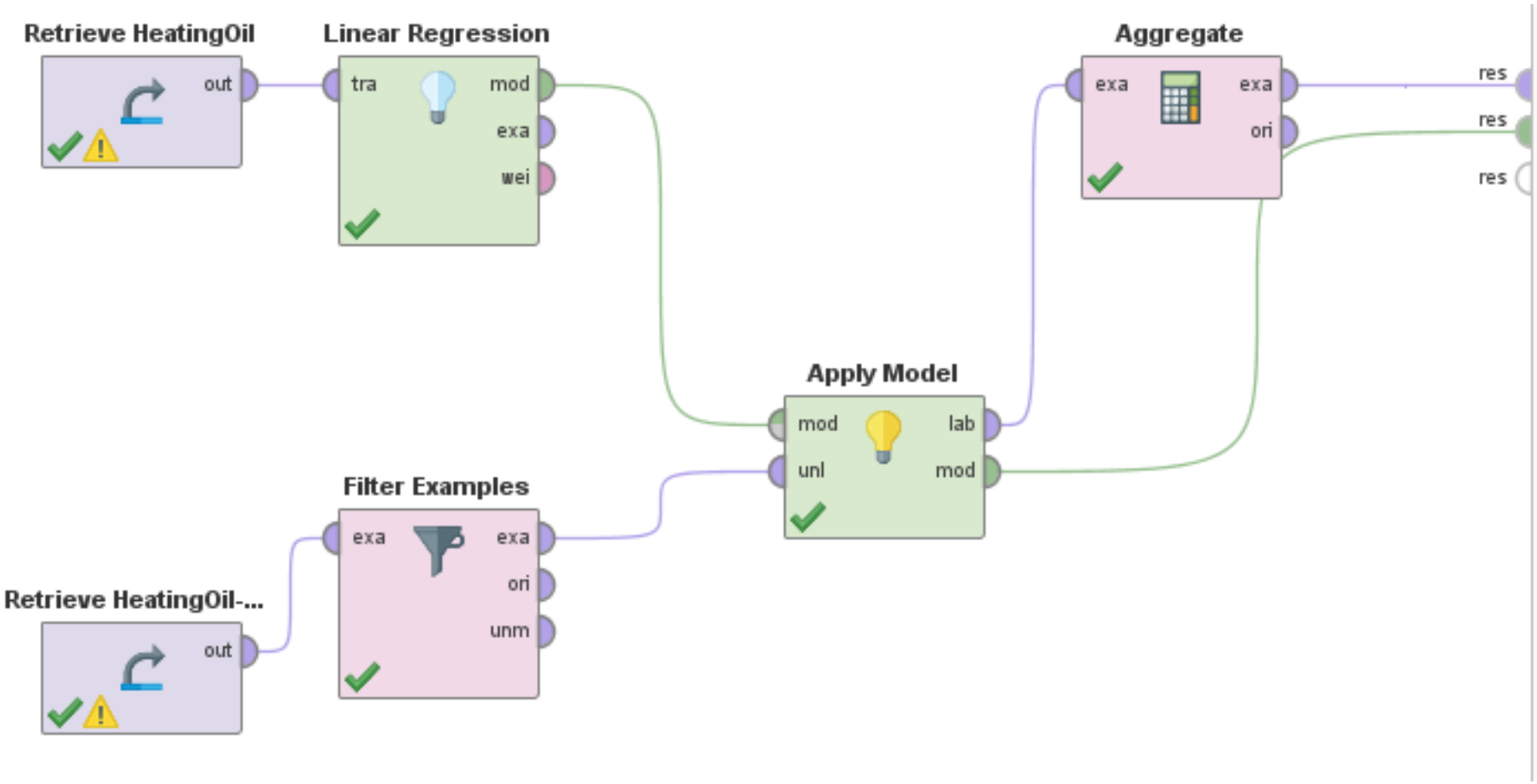


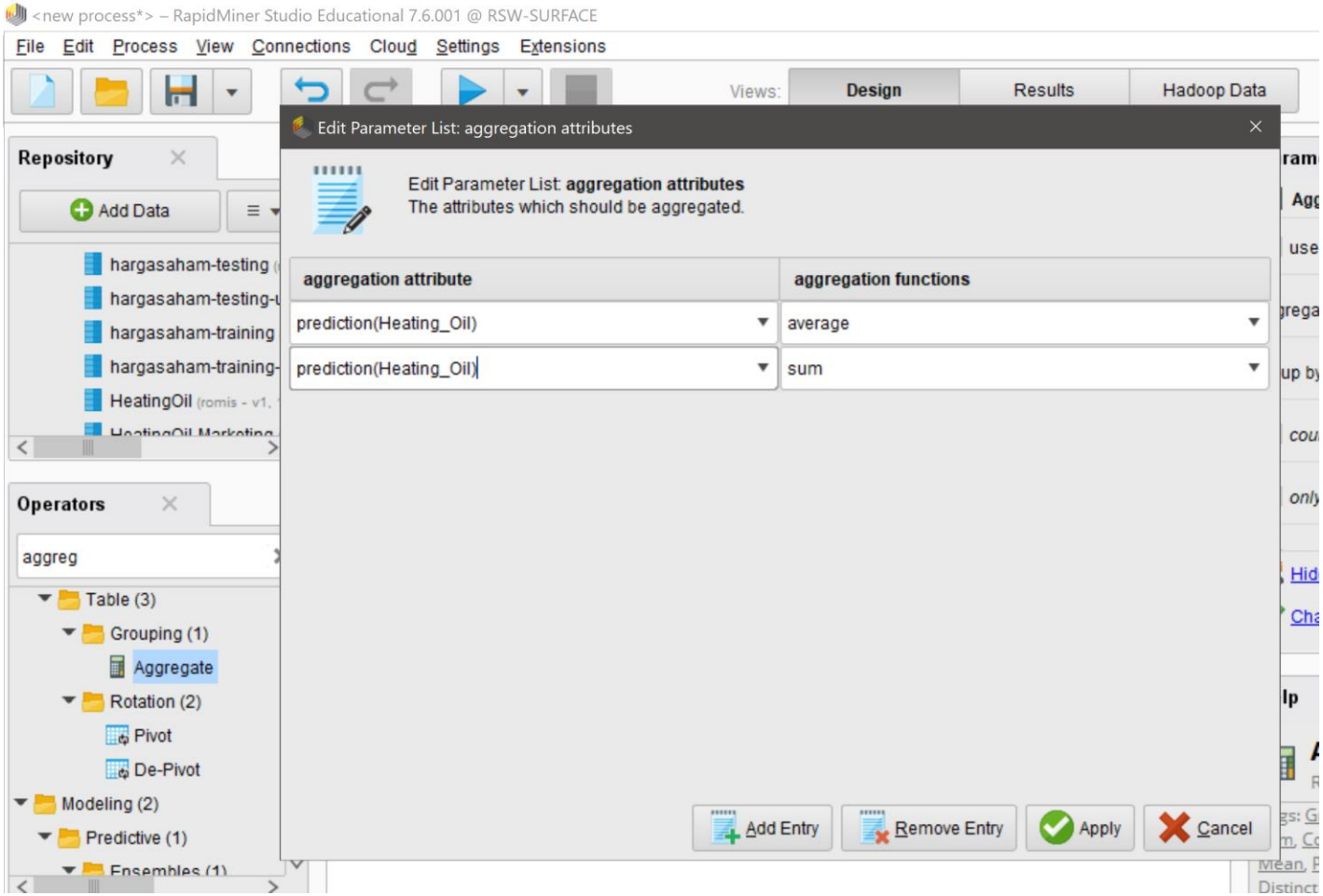
4. Evaluation – Hasil Prediksi



Row No.	prediction(H...	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	89	9	67.800	7
4	209.529	7	81	9	52.400	6
5	164.669	4	58	8	22.900	7
6	180.512	4	58	6	37.400	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4
9	264.712	9	39	1	56.900	7
10	221.364	8	84	5	64.500	2
11	221.328	10	74	6	58.300	1
12	262.580	5	49	6	68.600	6
13	214.082	8	45	2	33.900	8
14	212.392	3	49	4	49.700	4
15	253.199	9	66	6	66.200	5

5. Deployment





ExampleSet (Aggregate)

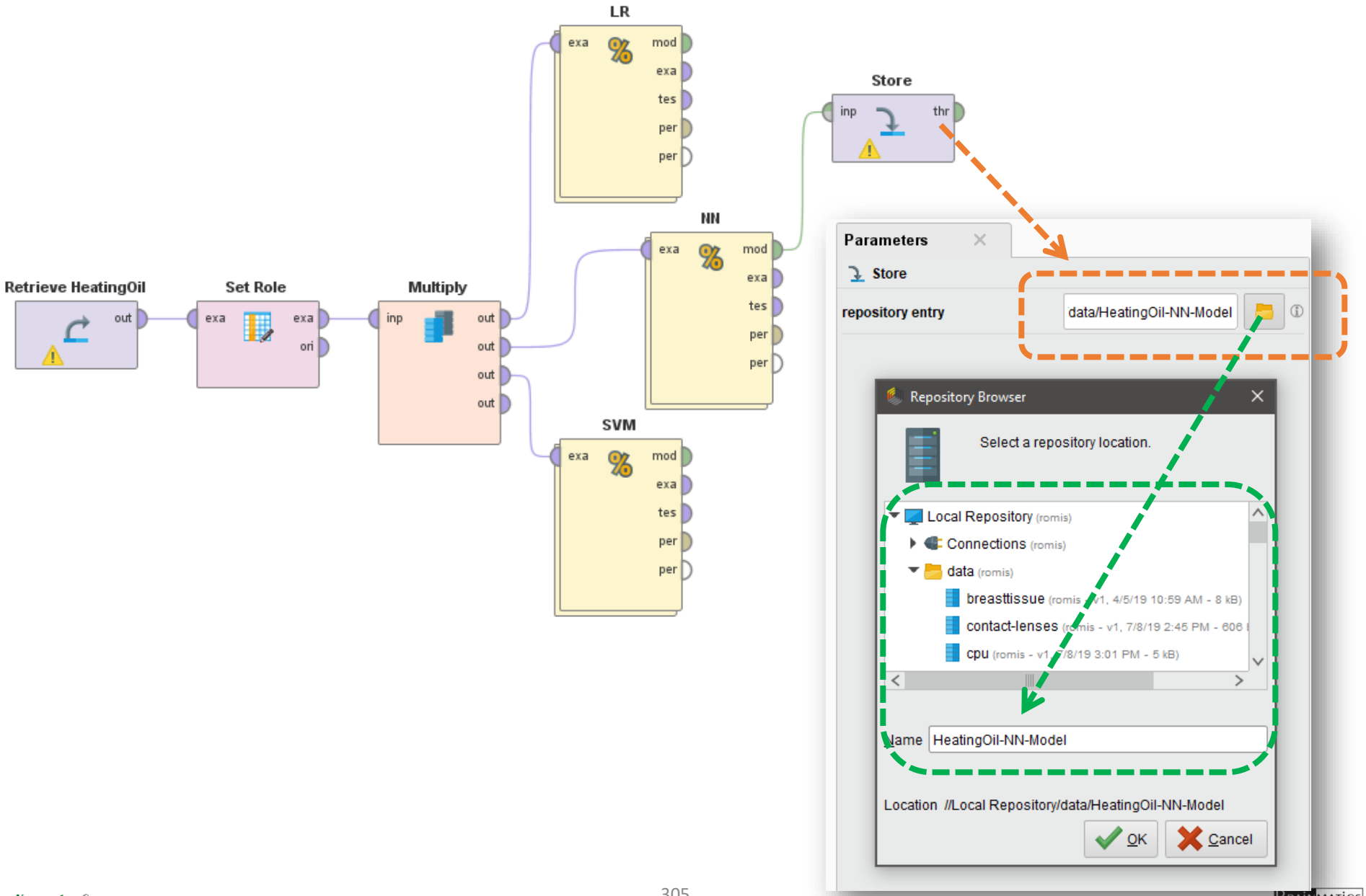
ExampleSet (1 example, 0 special attributes, 2 regular attributes)

Row No.	average(prediction(Heating_Oil))	sum(prediction(Heating_Oil))
1	199.041	8368087.536

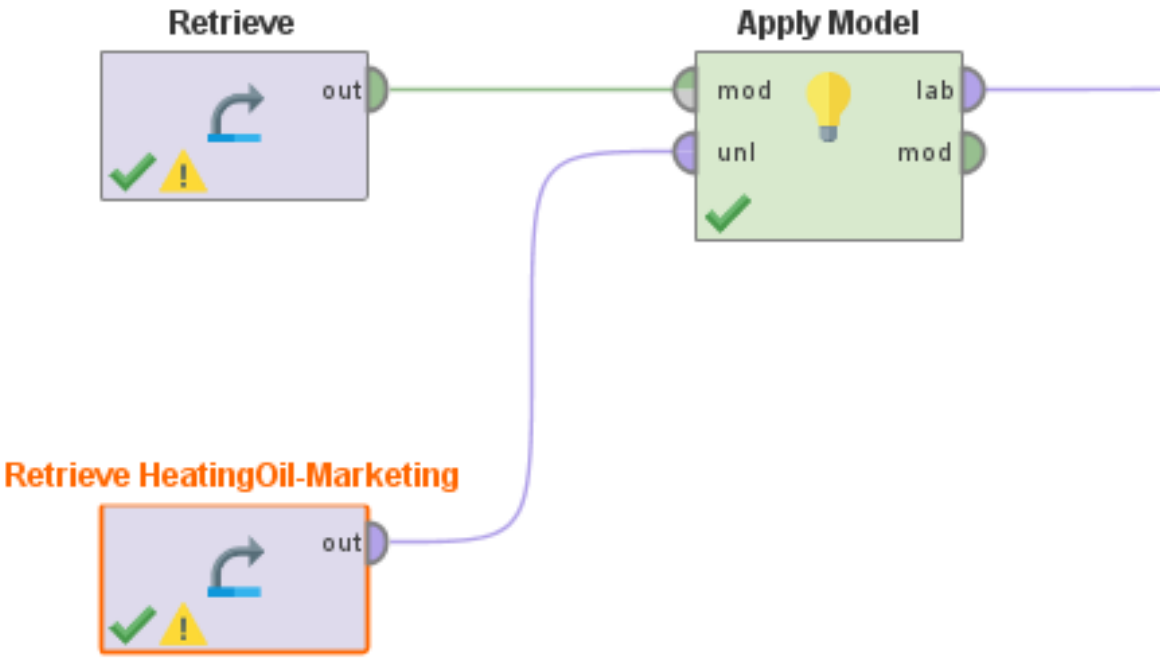
Latihan

- Karena bantuan data mining sebelumnya, Sarah akhirnya mendapatkan promosi menjadi VP marketing, yang mengelola ratusan marketer
- Sarah ingin para marketer dapat memprediksi pelanggan potensial mereka masing-masing secara mandiri. Masalahnya, data HeatingOil.csv hanya boleh diakses oleh level VP (Sarah), dan tidak diperbolehkan diakses oleh marketer secara langsung
- Sarah ingin masing-masing marketer membuat proses yang dapat mengestimasi kebutuhan konsumsi minyak dari *client* yang mereka *approach*, dengan menggunakan model yang sebelumnya dihasilkan oleh Sarah, meskipun tanpa mengakses data training (HeatingOil.csv)
- Asumsikan bahwa data HeatingOil-Marketing.csv adalah data calon pelanggan yang berhasil di *approach* oleh salah satu marketingnya
- Yang harus dilakukan Sarah adalah membuat proses untuk:
  1. Mengkomparasi algoritma yang menghasilkan model yang memiliki akurasi tertinggi (LR, NN, SVM), gunakan 10 Fold X Validation
  2. Menyimpan model terbaik ke dalam suatu file (operator **Store**)
- Yang harus dilakukan Marketer adalah membuat proses untuk:
  1. Membaca model yang dihasilkan Sarah (operator **Retrieve**)
  2. Menerapkannya di data HeatingOil-Marketing.csv yang mereka miliki
- Mari kita bantu Sarah dan Marketer membuat dua proses tersebut

Proses Komparasi Algoritma (Sarah)



Proses Pengujian Data (Marketer)



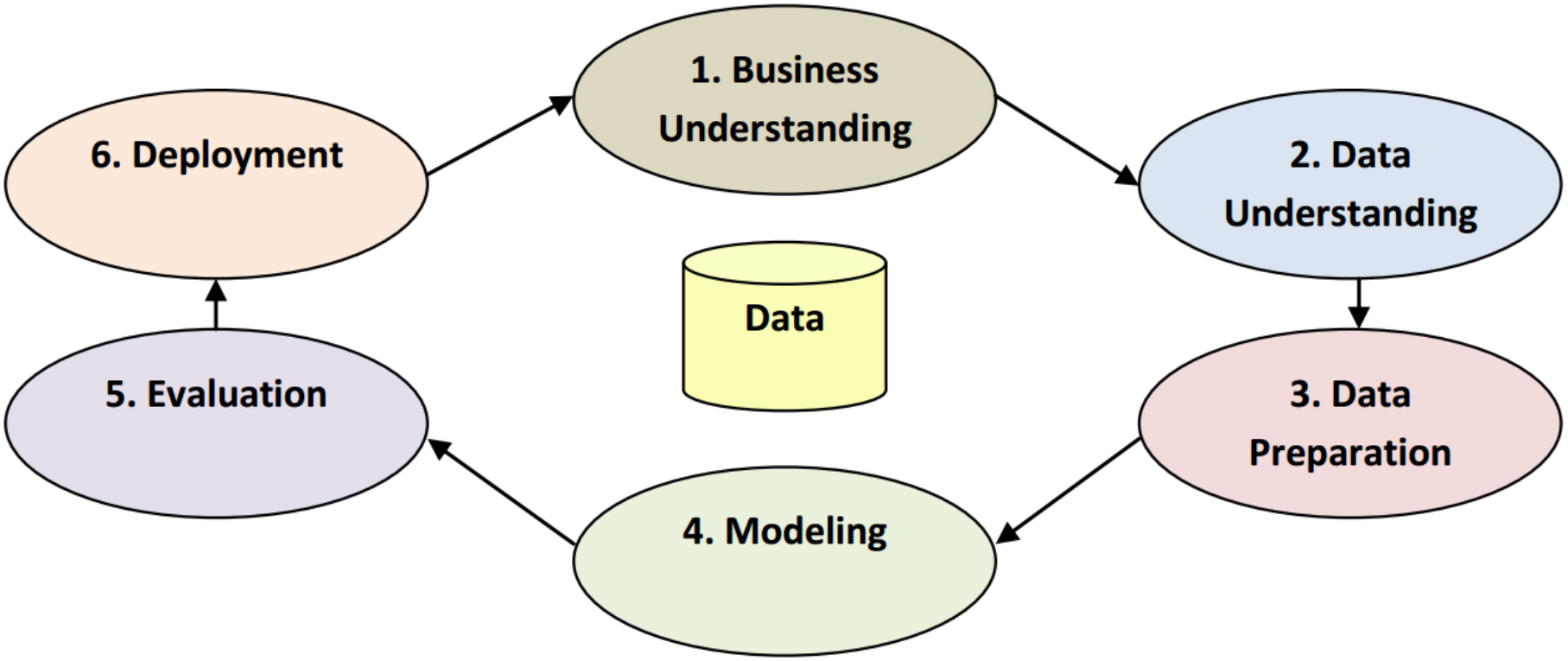
Row No.	prediction(H...	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	146.537	6	74	4	23.800	4
2	254.538	10	43	4	56.700	4
3	140.520	3	81	2	28	6
4	200.517	9	50	4	45.100	3

Studi Kasus CRISP-DM

Kelulusan Mahasiswa di Universitas Suka Belajar

Dataset: datakelulusanmahasiswa.xls

CRISP-DM


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

1. Business Understanding

- Problems:
  - Budi adalah Rektor di Universitas Suka Belajar
  - Universitas Suka Belajar memiliki masalah besar karena rasio kelulusan mahasiswa tiap angkatan sangat rendah
  - Budi ingin memahami dan membuat pola dari profile mahasiswa yang bisa lulus tepat waktu dan yang tidak lulus tepat waktu
  - Dengan pola tersebut, Budi 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. Data Understanding

- Untuk menyelesaikan masalah, Budi 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 1
    8. IPS 3: Indeks Prestasi Semester 1
    9. IPS 4: Indeks Prestasi Semester 1
    10. IPS 5: Indeks Prestasi Semester 1
    11. IPS 6: Indeks Prestasi Semester 1
    12. IPS 7: Indeks Prestasi Semester 1
    13. IPS 8: Indeks Prestasi Semester 1
    14. IPK: Indeks Prestasi Kumulatif
    15. STATUS KELULUSAN: Terlambat atau Tepat

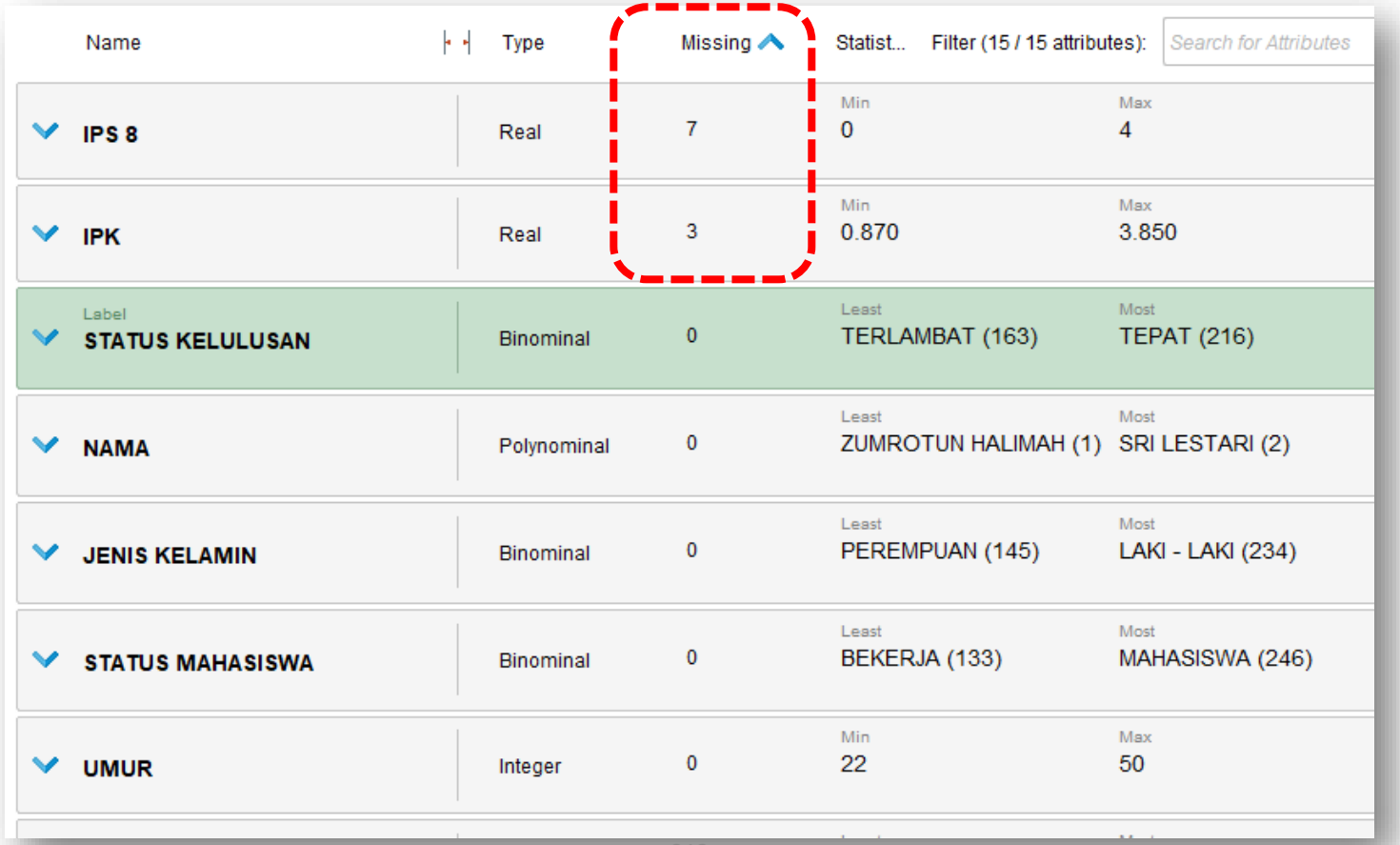
3. Data Preparation

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. Data Preparation

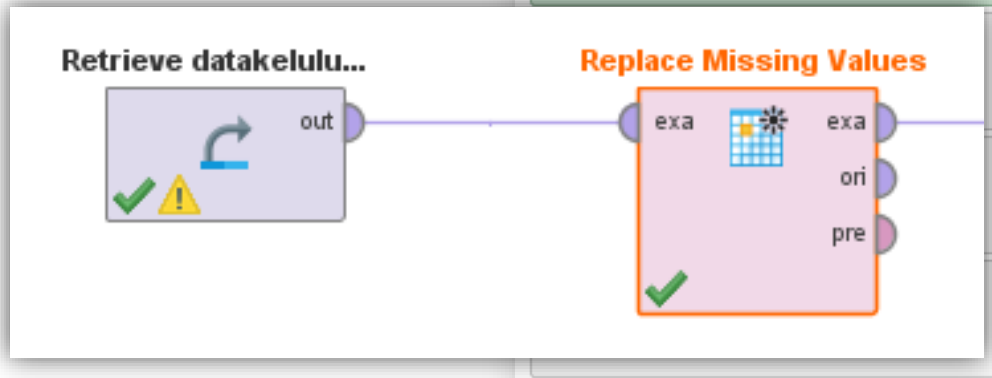
- Terdapat 379 data mahasiswa dengan 15 atribut
- Missing Value sebayak 10 data, dan tidak terdapat data noise



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. Data Preparation

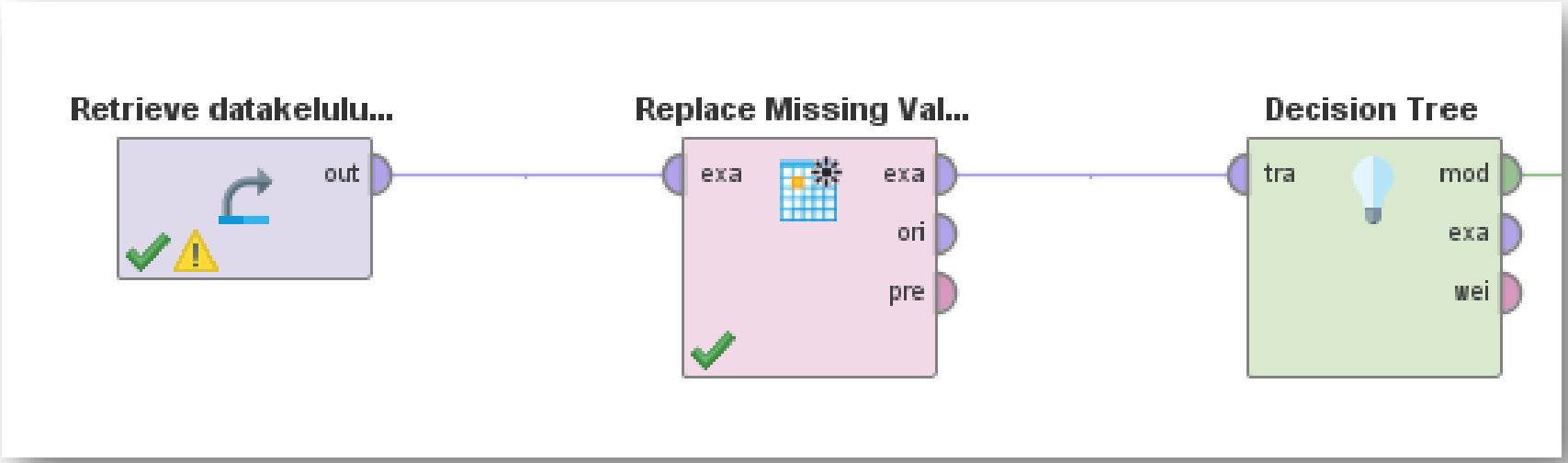
- Missing Value dipecahkan dengan menambahkan data dengan nilai rata-rata
- Hasilnya adalah data bersih tanpa missing value



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

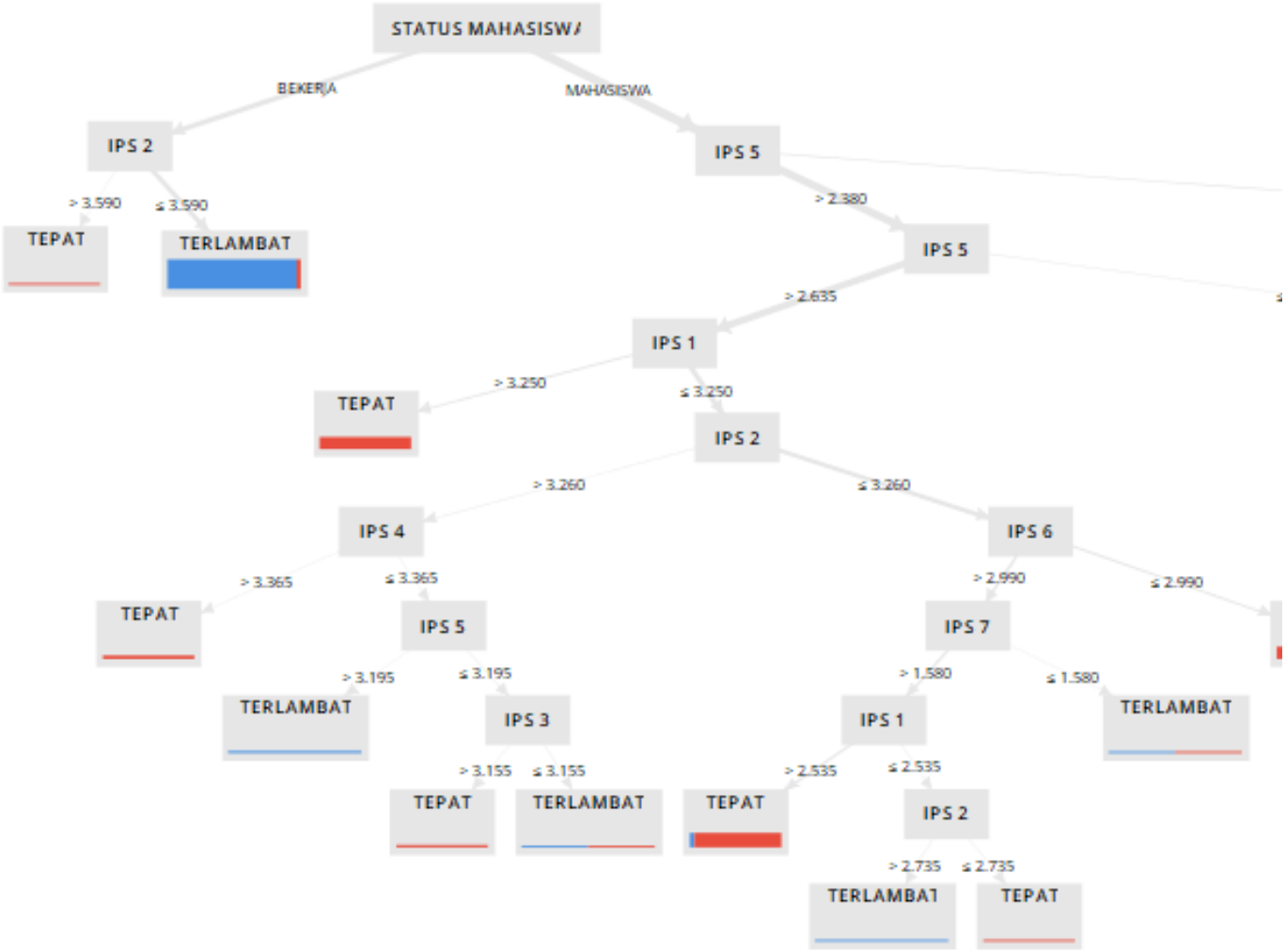
4. Modeling

- Modelkan dataset dengan Decision Tree
- Pola yang dihasilkan bisa berbentuk tree atau if-then



4. Modeling

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



5. 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}
```

5. Evaluation

- 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. Deployment

- Budi 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

- Budi membuat peraturan melarang mahasiswa bekerja paruh waktu di semester awal perkuliahan, karena beresiko tinggi di kelulusan tepat waktu

- Budi 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

- Budi memasukkan pola dan model yang terbentuk ke dalam sistem informasi akademik, dimana sistem dibuat cerdas, sehingga bisa mengirimkan email analisis pola secara otomatis ke mahasiswa sesuai profilnya

Latihan

- Pahami dan lakukan eksperimen berdasarkan seluruh studi kasus yang ada di buku Data Mining for the Masses (*Matthew North*)

- Pahami bahwa metode CRISP-DM membantu kita memahami penggunaan metode data mining yang lebih sesuai dengan kebutuhan organisasi

Tugas Menyelesaikan Masalah Organisasi

- Analisis masalah dan kebutuhan yang ada di organisasi lingkungan sekitar anda
- Kumpulkan dan review dataset yang tersedia, dan hubungkan masalah dan kebutuhan tadi dengan data yang tersedia (analisis dari 5 peran data mining)
  - Bila memungkinkan pilih beberapa peran sekaligus untuk mengolah data tersebut, misalnya: lakukan association (analisis faktor), sekaligus estimation atau clustering
- Lakukan proses CRISP-DM untuk menyelesaikan masalah yang ada di organisasi sesuai dengan data yang didapatkan
  - Pada proses data preparation, lakukan data cleaning (replace missing value, replace, filter attribute) sehingga data siap dimodelkan
  - Lakukan juga komparasi algoritma untuk memilih algoritma terbaik
- Rangkumkan dalam bentuk slide dengan contoh studi kasus Sarah yang menggunakan data mining untuk:
  - Menganalisis faktor yang berhubungan (matrix correlation)
  - Mengestimasi jumlah stok minyak (linear regression)

Studi Kasus CRISP-DM

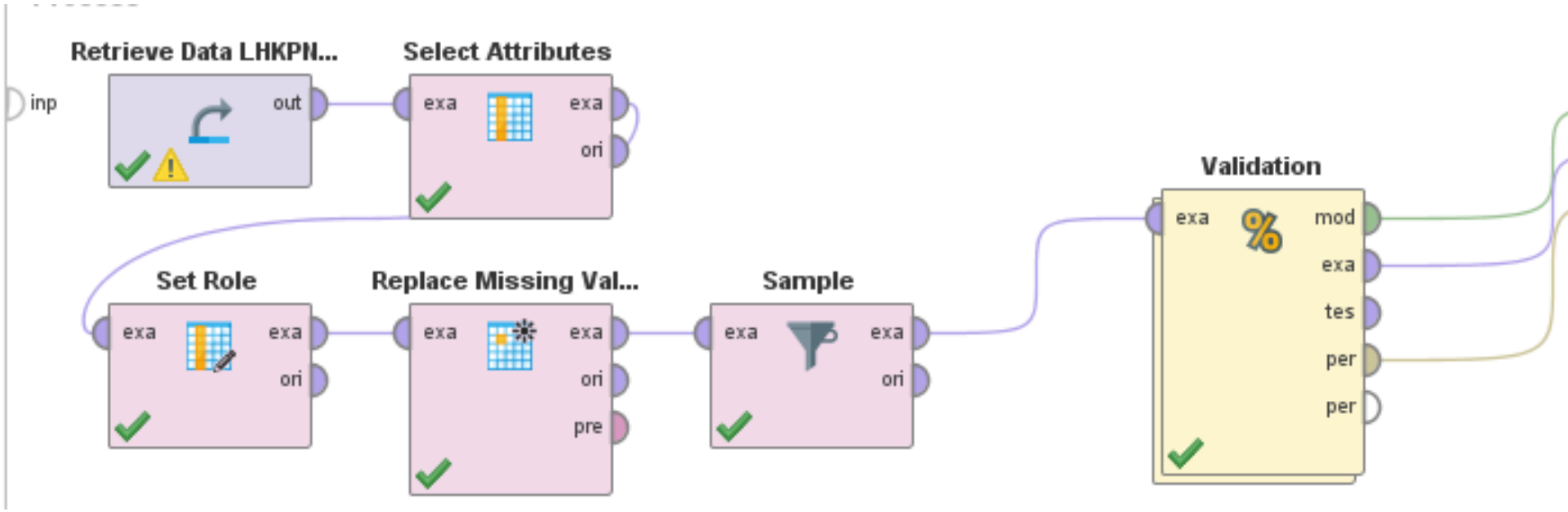
Profiling Tersangka Koruptor

Dataset: Data LHKPN KPK

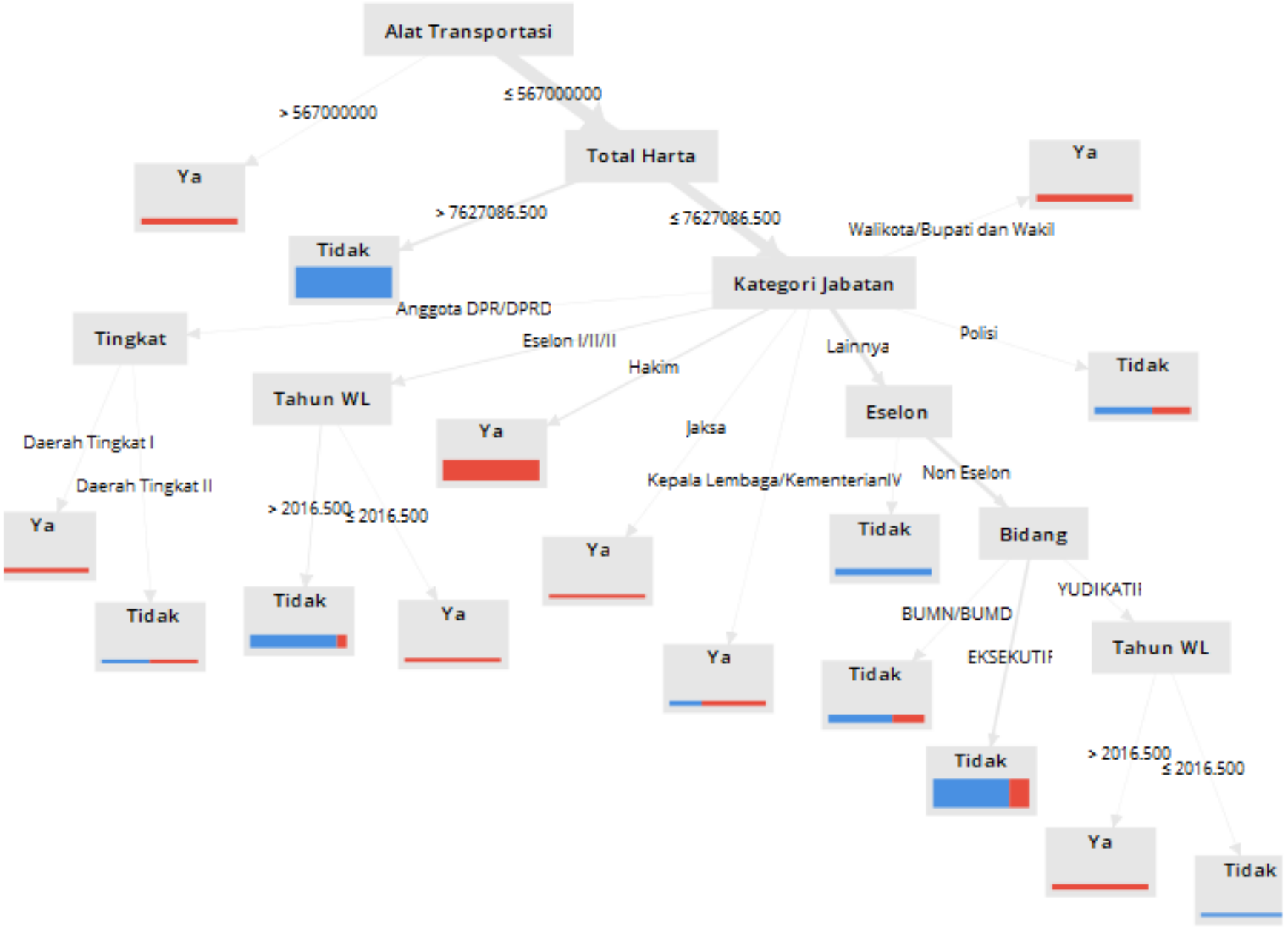
Contoh Kasus Pengolahan Data LHKPN

1. Prediksi Profil Tersangka Koruptor
   (Klasifikasi, Decision Tree)
2. Forecasting Jumlah Wajib Lapor di suatu Instansi atau suatu propinsi
   (Forecasting, Neural Network)
3. Prediksi Rekomendasi Hasil Pemeriksaan LHKPN
   (Klasifikasi, Decision Tree)

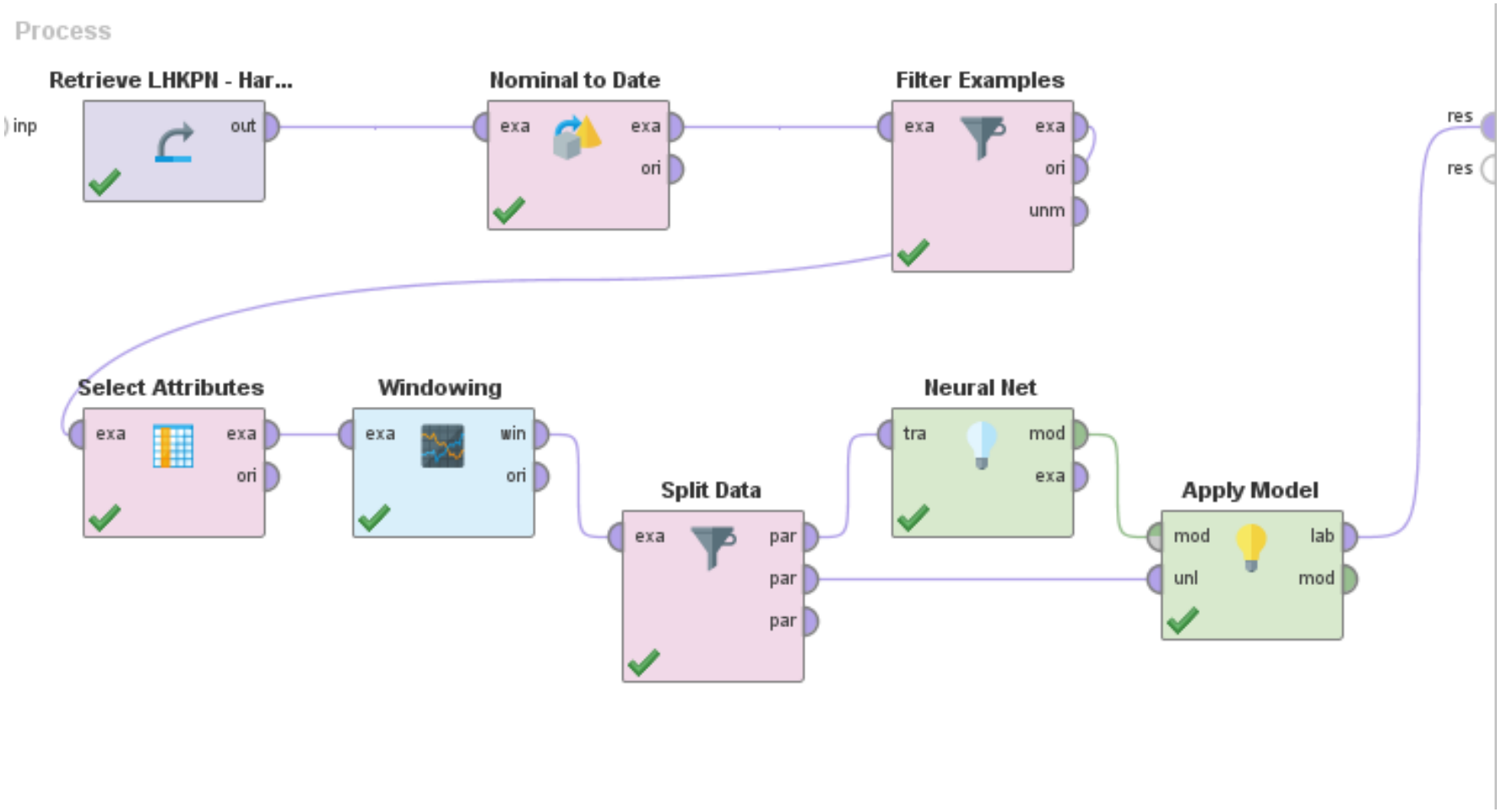
Prediksi Profil Tersangka Koruptor



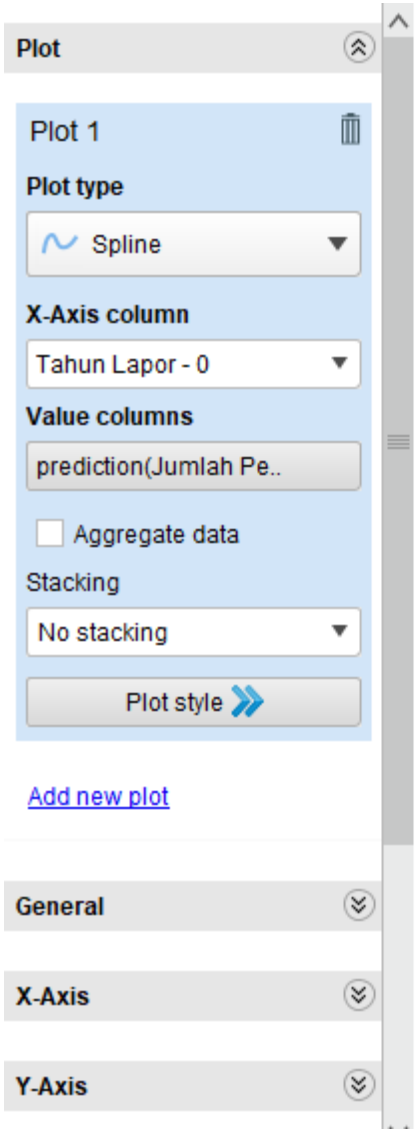
Pola Profil Tersangka Koruptor



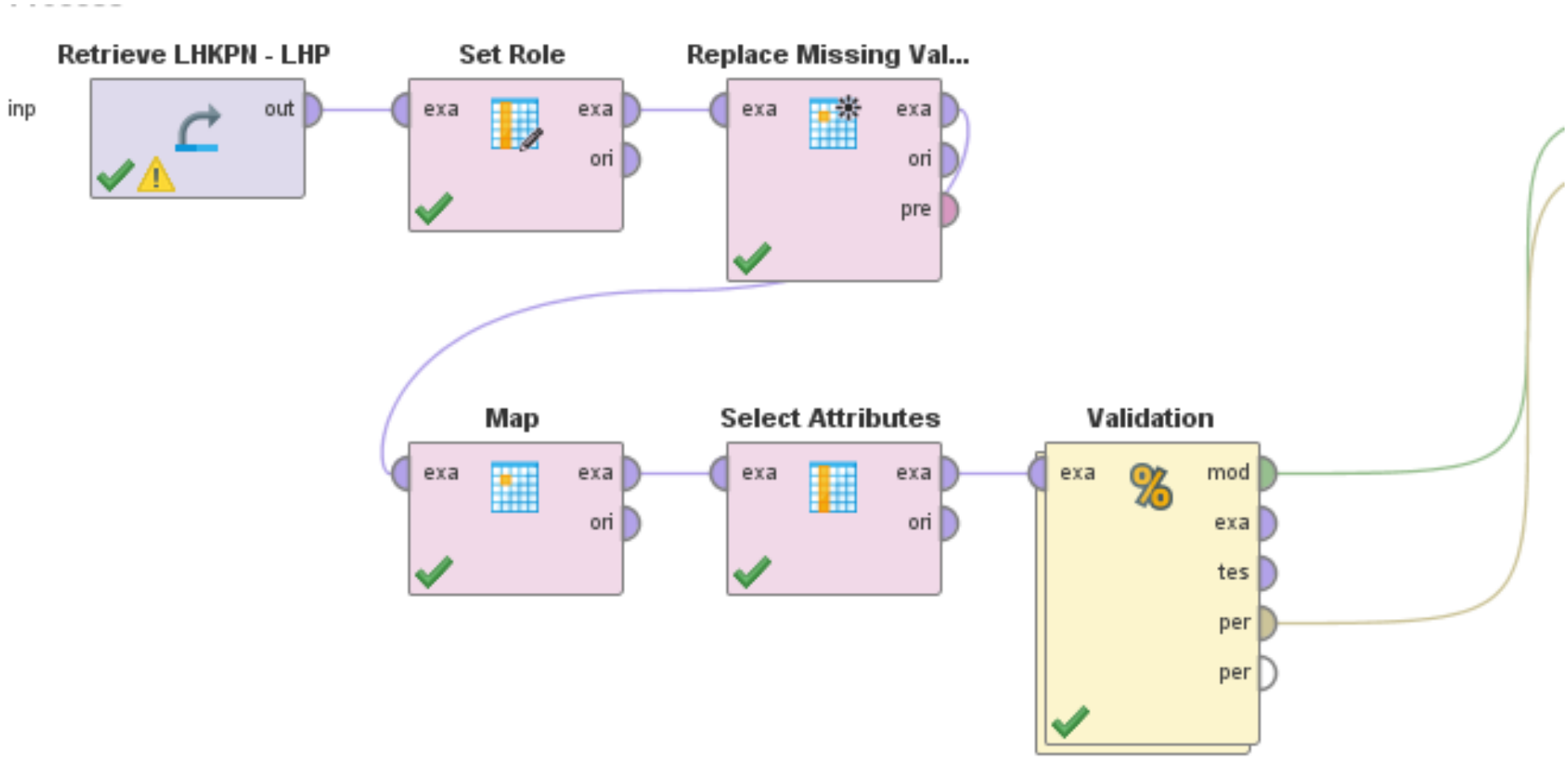
Forecasting Jumlah Wajib Lapor



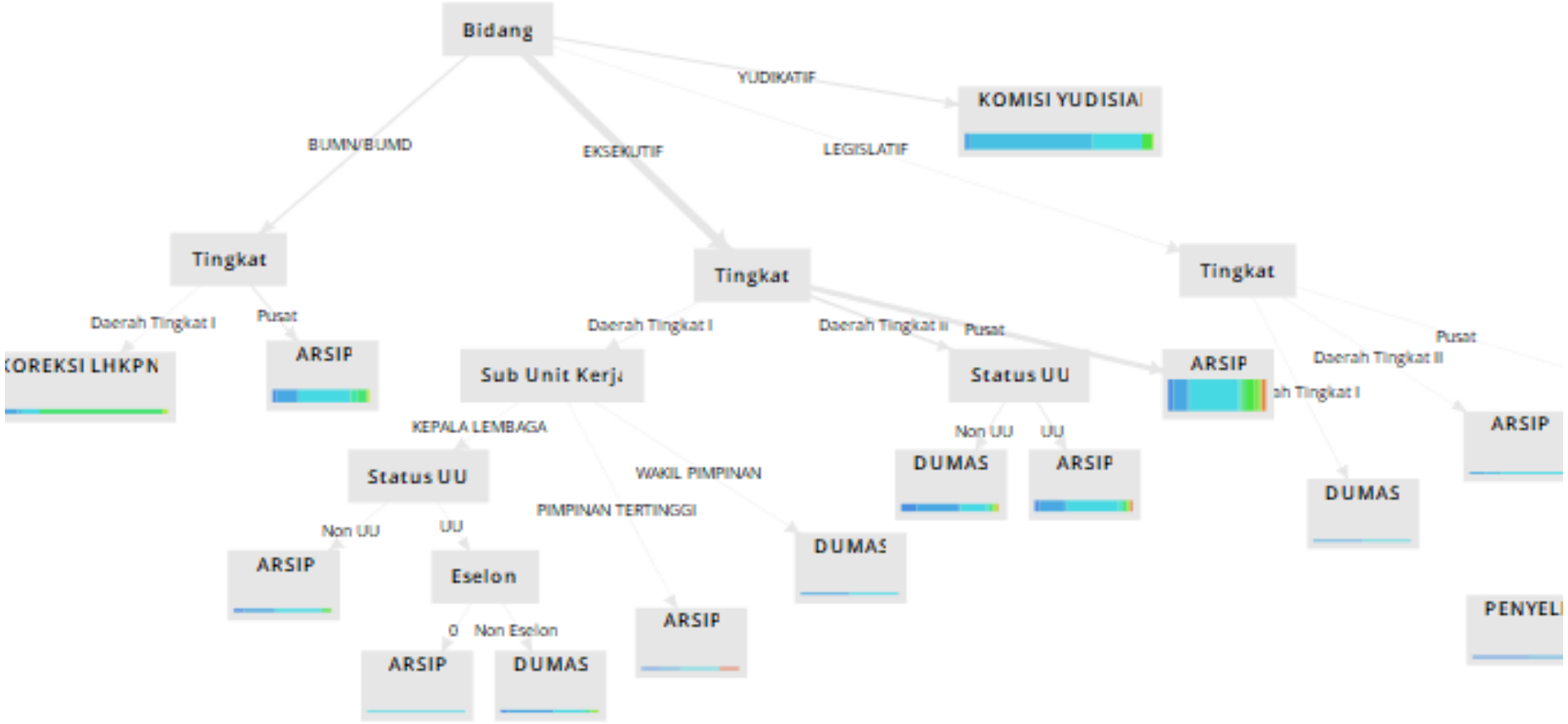
Forecasting Jumlah Wajib Lapor



Rekomendasi Hasil Pemeriksaan LHKPN



Pola Rekomendasi Hasil Pemeriksaan LHKPN



3. Persiapan Data

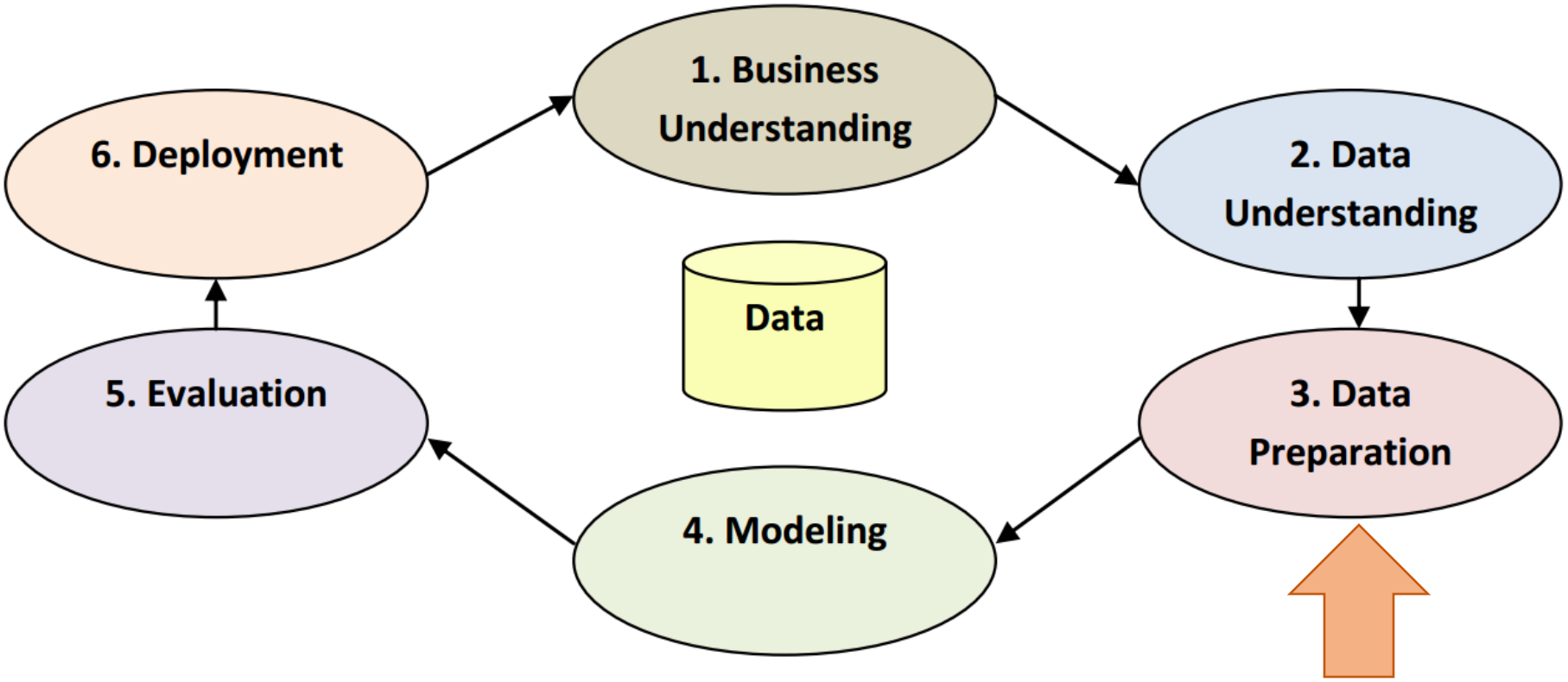
3.1 Data Cleaning

3.2 Data Reduction

3.3 Data Transformation and Data Discretization

3.4 Data Integration

CRISP-DM


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

Why Preprocess the Data?

Measures for data quality: A multidimensional view

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, …
- Consistency: some modified but some not, …
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

1. Data cleaning
   - Fill in missing values
   - Smooth noisy data
   - Identify or remove outliers
   - Resolve inconsistencies
2. Data reduction
   - Dimensionality reduction
   - Numerosity reduction
   - Data compression
3. Data transformation and data discretization
   - Normalization
   - Concept hierarchy generation
4. Data integration
   - Integration of multiple databases or files

Data Preparation Law (Data Mining Law 3)

Data preparation is more than half of every data mining process

- Maxim of data mining: most of the effort in a data mining project is spent in data acquisition and preparation, and informal estimates vary from 50 to 80 percent
- The purpose of data preparation is:
  1. To put the data into a form in which the data mining question can be asked
  2. To make it easier for the analytical techniques (such as data mining algorithms) to answer it

3.1 Data Cleaning

Data Cleaning

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., Occupation=“ ” (missing data)
- Noisy: containing noise, errors, or outliers
  - e.g., Salary=“−10” (an error)
- Inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42”, Birthday=“03/07/2010”
  - Was rating “1, 2, 3”, now rating “A, B, C”
- Discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
  - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Contoh Missing Data

- Dataset: MissingDataSet.csv

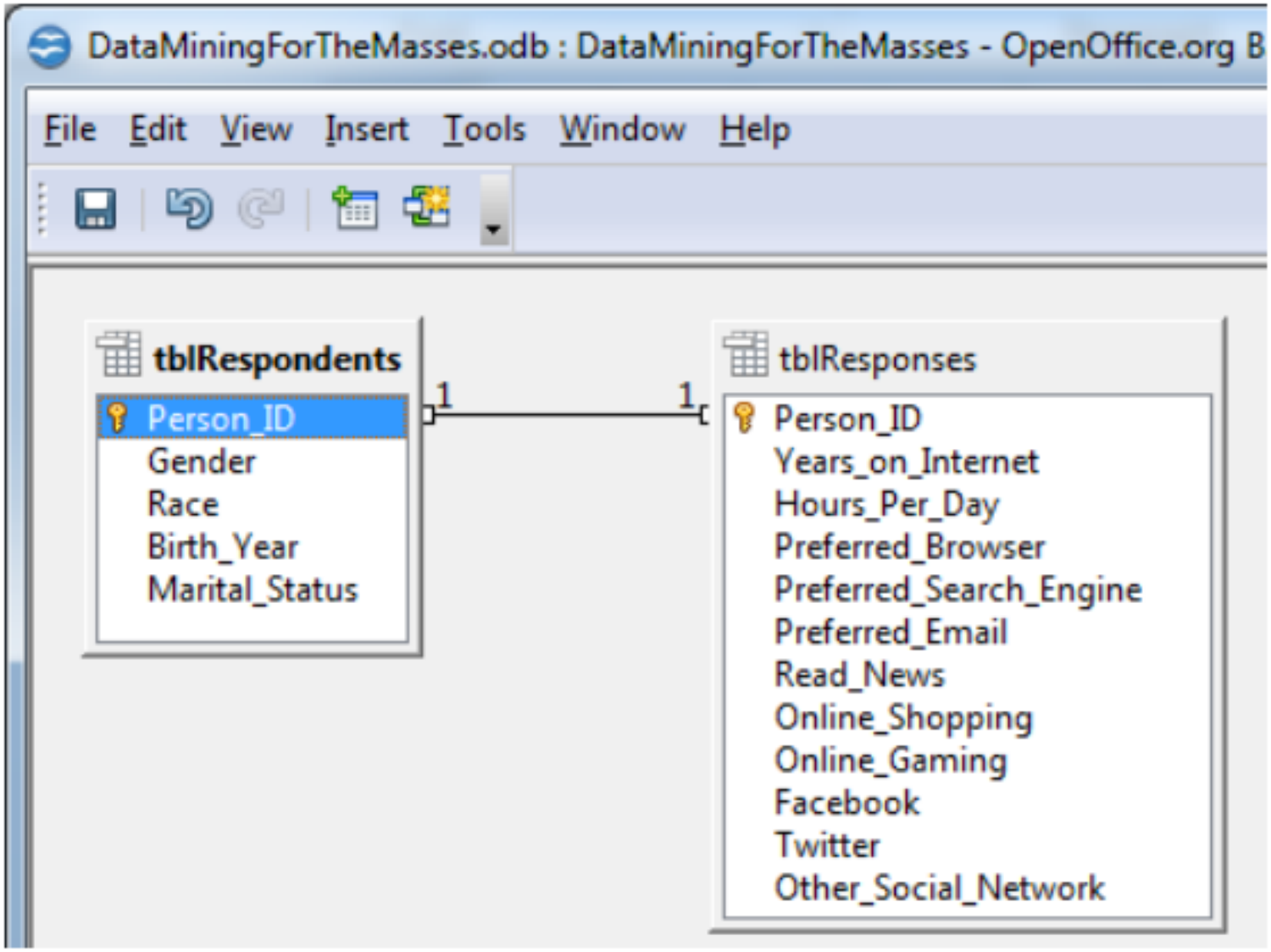
ExampleSet (11 examples, 0 special attributes, 15 regular attributes) View Filter (11 / 11): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_I...	Hours_Per...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Ga...	Facebook	Twitter	Other_Soci...
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
3	F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
5	M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y	Y	N	Y	N	?
6	M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y	N	Y	N	N	?
7	M	African Amer	1981	D	11	2	Firefox	Google	Yahoo	?	Y	?	Y	Y	LinkedIn
8	M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
9	F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?	Y	Y	N	MySpace
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

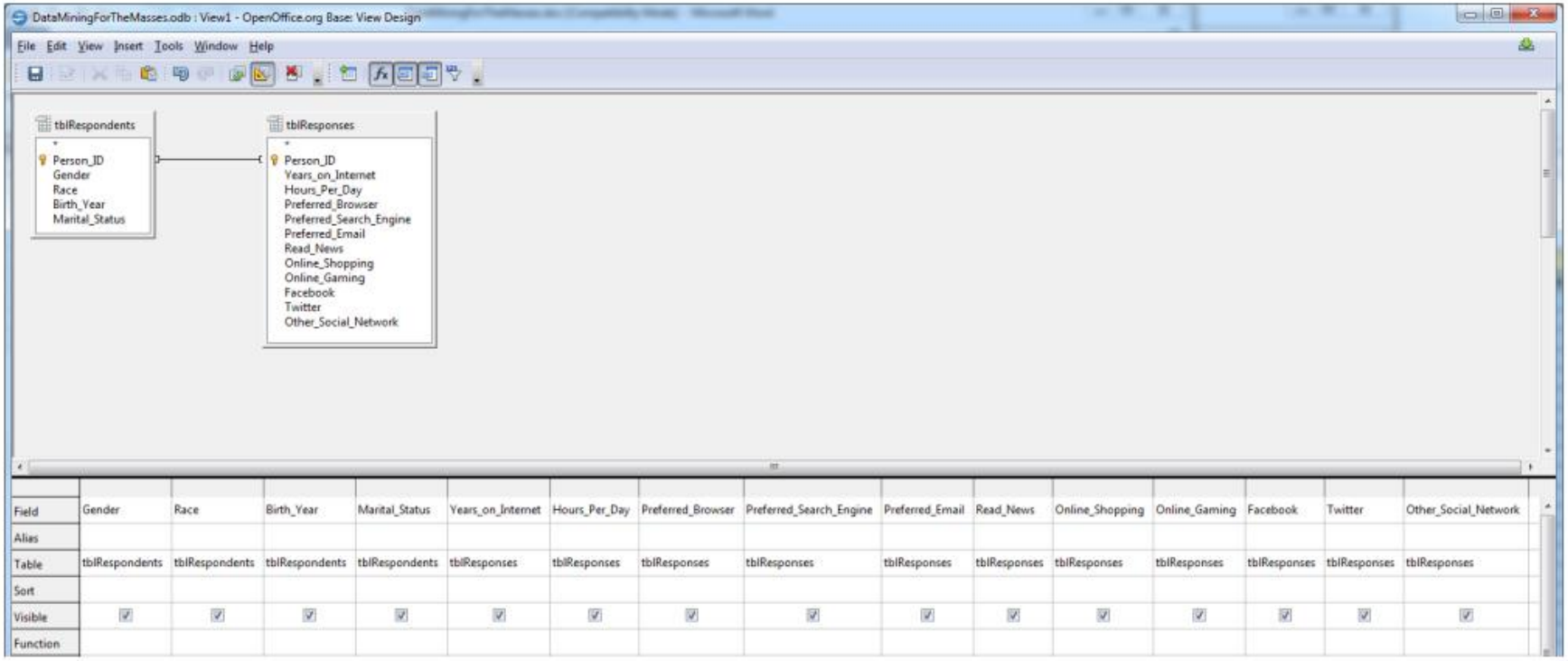
MissingDataSet.csv

- Jerry is the marketing manager for a small Internet design and advertising firm
- Jerry's boss asks him to develop a data set containing information about Internet users
- The company will use this data to determine what kinds of people are using the Internet and how the firm may be able to market their services to this group of users
- To accomplish his assignment, Jerry creates an online survey and places links to the survey on several popular Web sites
- Within two weeks, Jerry has collected enough data to begin analysis, but he finds that his data needs to be denormalized
- He also notes that some observations in the set are missing values or they appear to contain invalid values
- Jerry realizes that some additional work on the data needs to take place before analysis begins.

Relational Data



View of Data (Denormalized Data)



Field	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
Alias															
Table	tblRespondents	tblRespondents	tblRespondents	tblRespondents	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses
Sort															
Visible	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Function															

vwInternetUser - DataMiningForTheMasses - OpenOffice.org Base: Table Data View

File Edit View Insert Tools Window Help

Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y	N	
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	
M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y		Y	Y	LinkedIn
M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	
M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y	N	MySpace
F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

Contoh Missing Data

- Dataset: MissingDataSet.csv

ExampleSet (11 examples, 0 special attributes, 15 regular attributes) View Filter (11 / 11): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_I...	Hours_Per...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Ga...	Facebook	Twitter	Other_Soci...
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
3	F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
5	M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y	Y	N	Y	N	?
6	M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y	N	Y	N	N	?
7	M	African Amer	1981	D	11	2	Firefox	Google	Yahoo	?	Y	?	Y	Y	LinkedIn
8	M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
9	F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?	Y	Y	N	MySpace
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

How to Handle Missing Data?

- Ignore the tuple:
  - Usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually:
  - Tedious + infeasible?
- Fill in it automatically with
  - A global constant: e.g., “unknown”, a new class?!
  - The attribute mean
  - The attribute mean for all samples belonging to the same class: smarter
  - The most probable value: inference-based such as Bayesian formula or decision tree

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 3 Data Preparation
  1. Handling Missing Data, pp. 34-48 (*replace*)
  2. Data Reduction, pp. 48-51 (*delete/filter*)

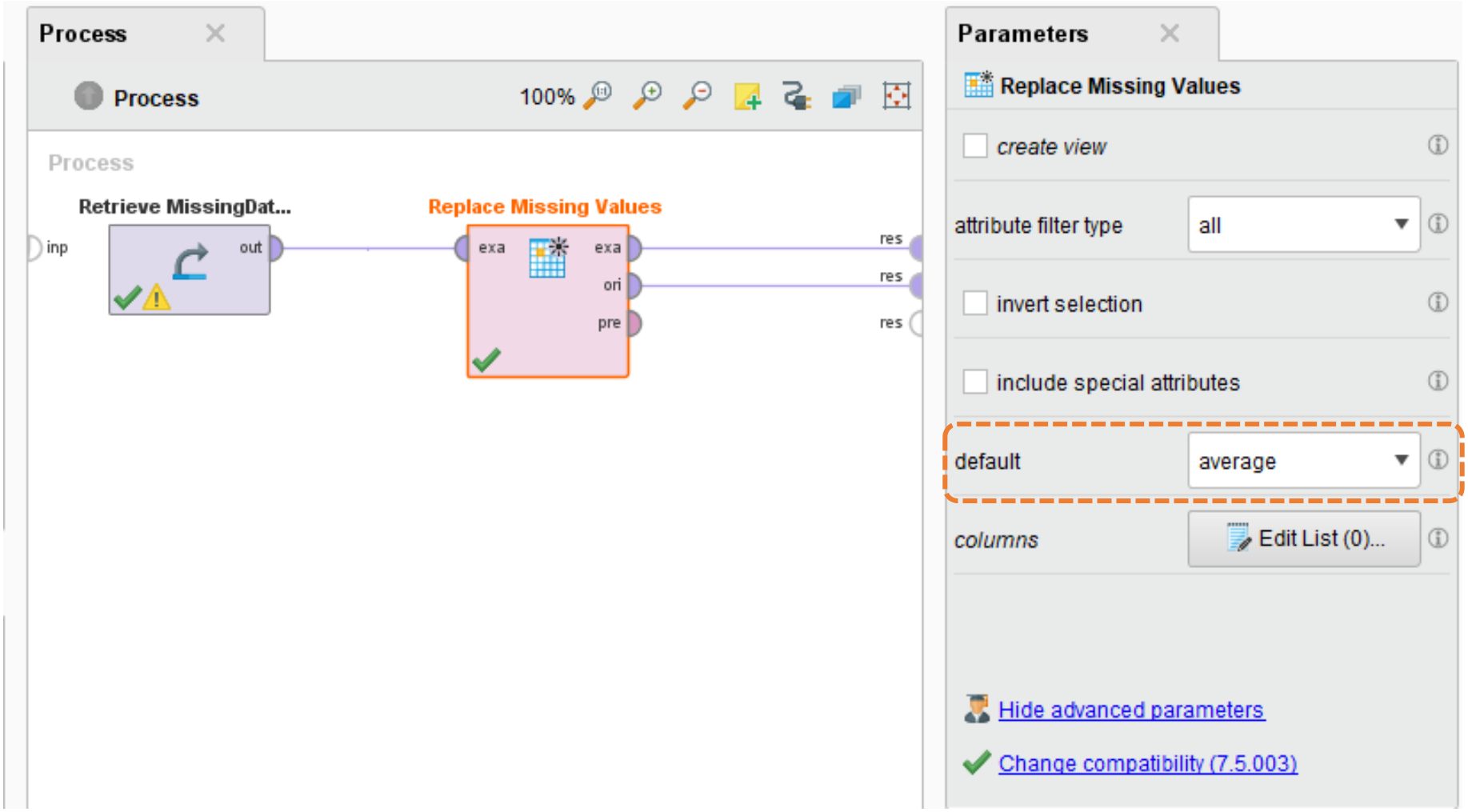
- Dataset: MissingDataSet.csv

- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut?

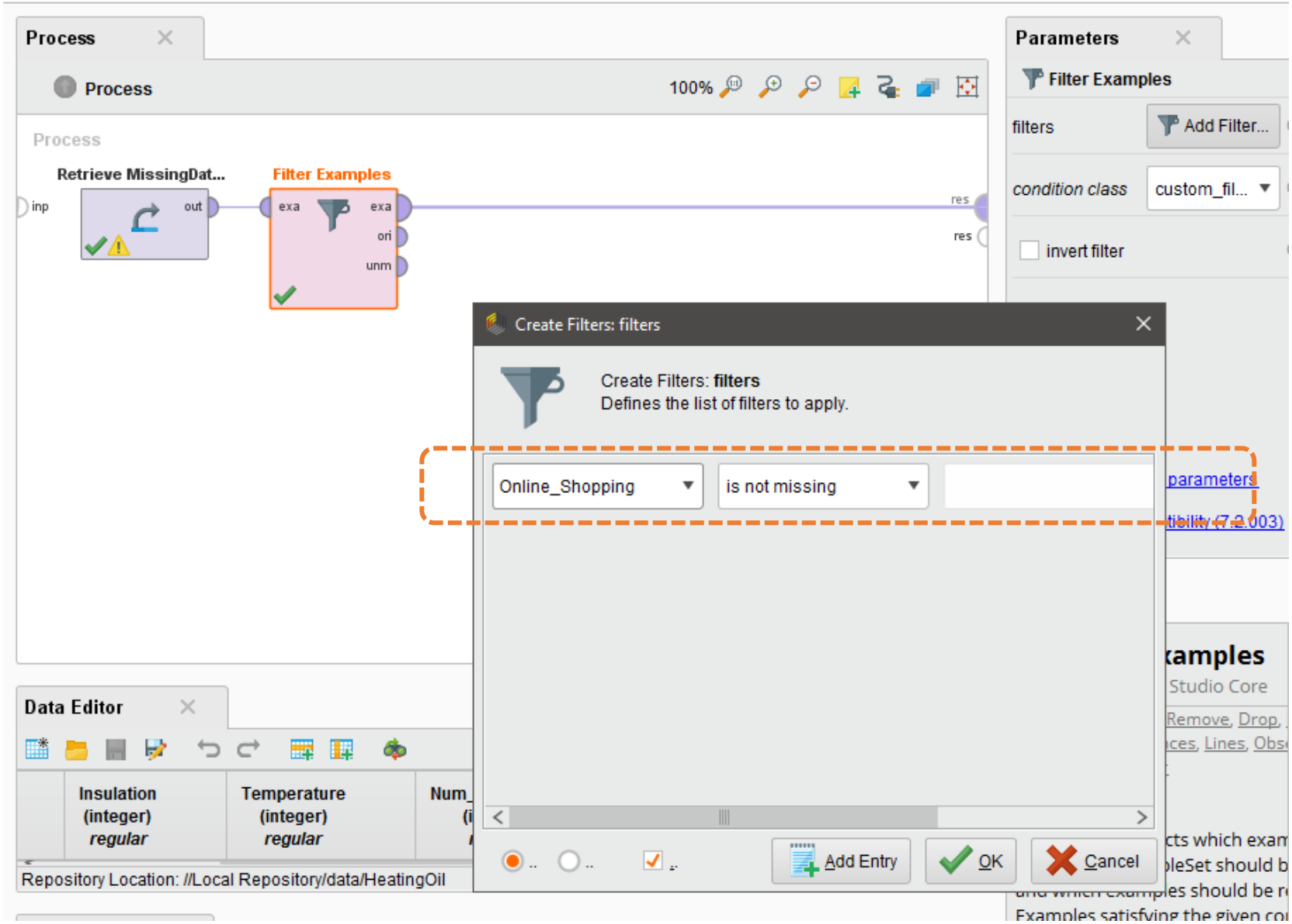
Missing Value Detection

Name	Type	Missing	Filter (15 / 15 attributes): Search for Attributes	
			Open chart	
Read_News	Polynominal	1	Least N (2)	Most Y (8)
Online_Shopping	Polynominal	2	Least N (4)	Most Y (5)
Online_Gaming	Polynominal	3	Least Y (2)	Most N (6)
Facebook	Polynominal	0	Least N (3)	Most Y (8)
Twitter	Polynominal	0	Least 99 (1)	Most N (8)
Other_Social_Network	Polynominal	7	Least MySpace (1)	Most LinkedIn (2)

Missing Value Replace



Missing Value Filtering



Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- Other data problems which require data cleaning
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data?

- Binning
  - First sort data and partition into (equal-frequency) bins
  - Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - Smooth by fitting the data into regression functions
- Clustering
  - Detect and remove outliers
- Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

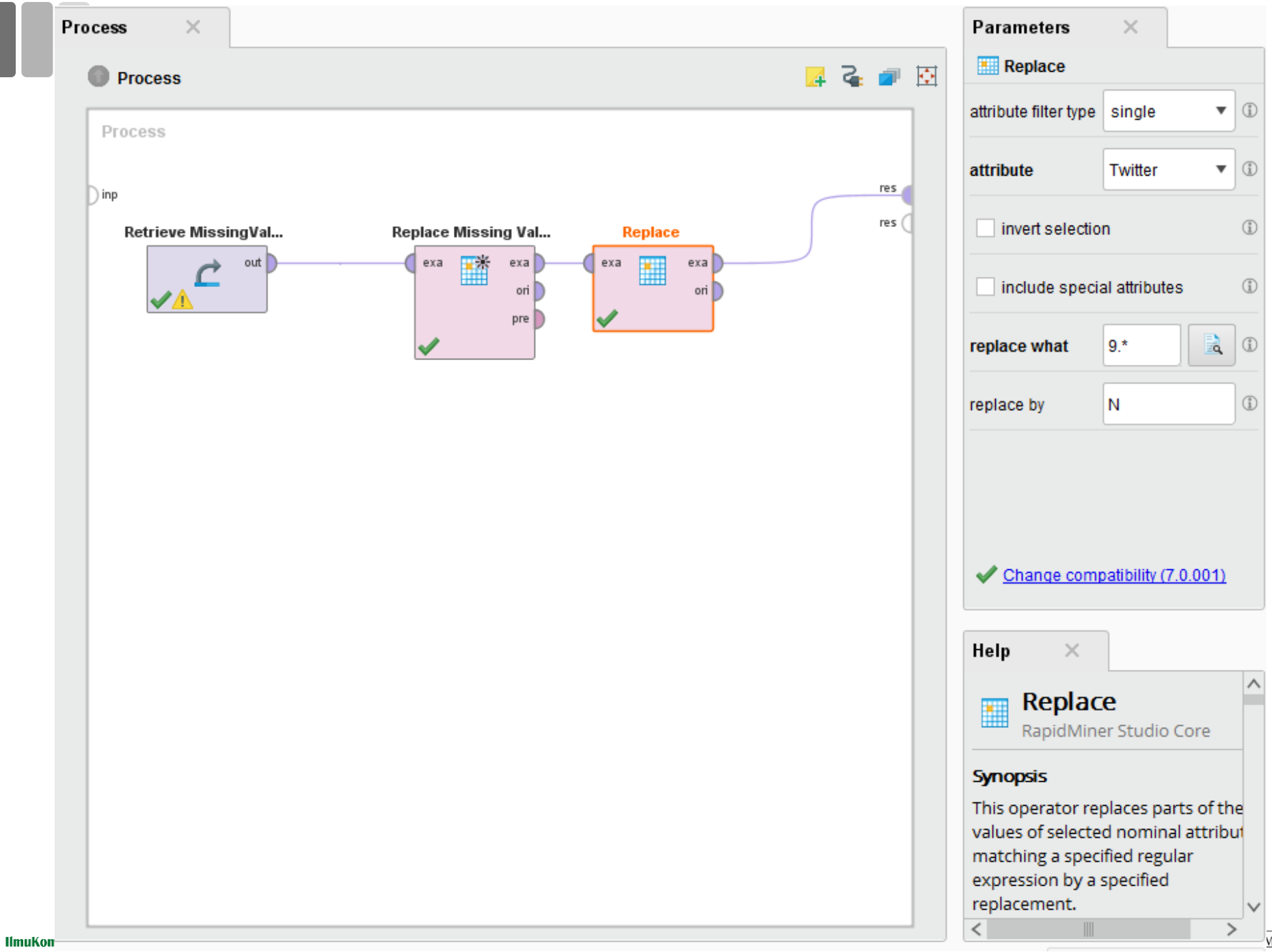
- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

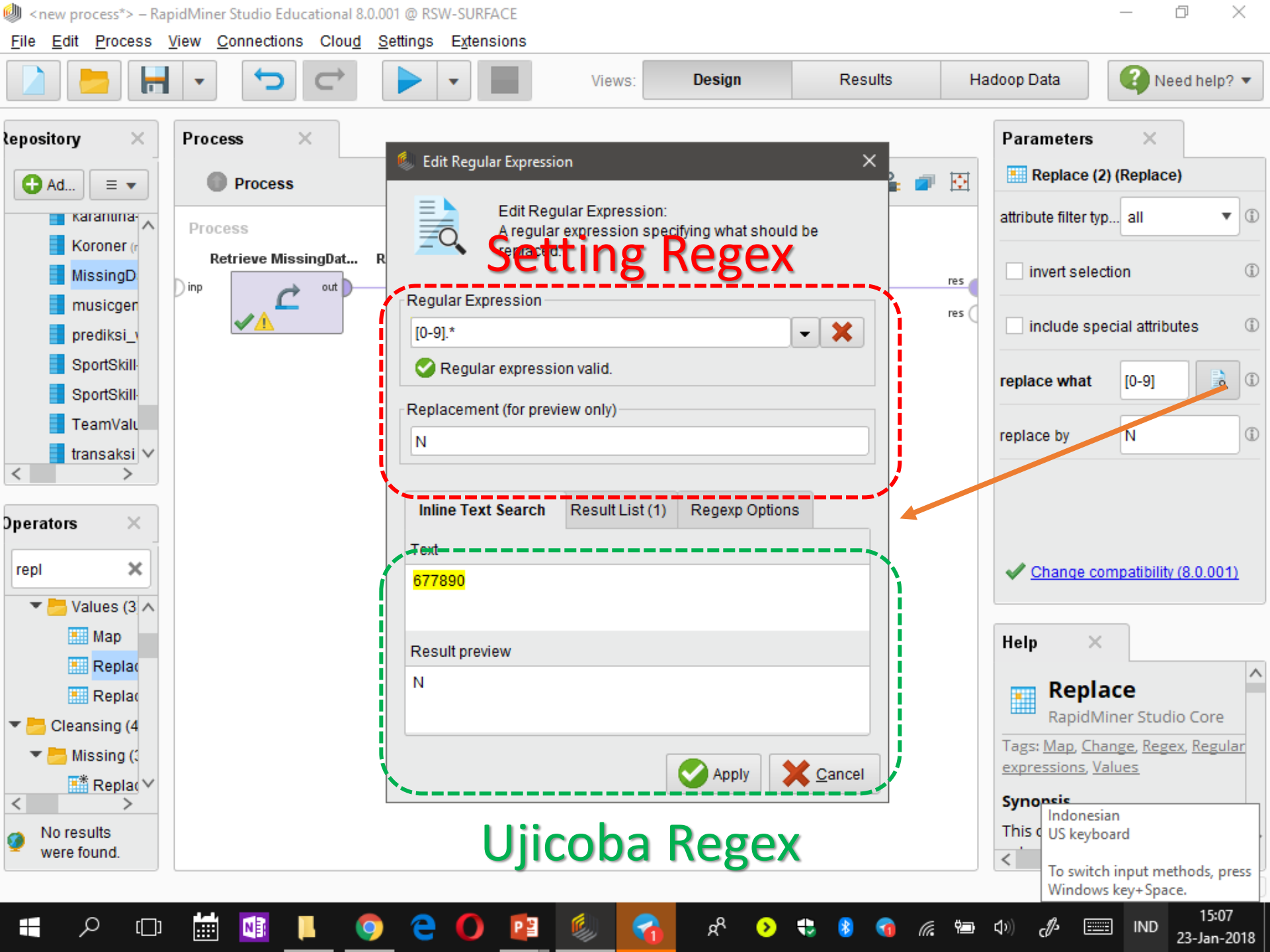
Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 3 Data Preparation, pp. 52-54 (Handling Inconsistence Data)

- Dataset: MissingDataSet.csv

- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!


Process
Process
Process
inp
Retrieve MissingVal...
out
Replace Missing Val...
exa
exa
ori
pre
Replace
exa
exa
ori
res
res
Parameters
Replace
attribute filter type
single
attribute
Twitter
invert selection
include special attributes
replace what
9.*
replace by
N
Change compatibility (7.0.001)
Help
Replace
RapidMiner Studio Core
Synopsis
This operator replaces parts of the
values of selected nominal attribut
matching a specified regular
expression by a specified
replacement.


<new process*> – RapidMiner Studio Educational 8.0.001 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results Hadoop Data Need help?
Repository
Process
Operators
Parameters
Help
Edit Regular Expression
Edit Regular Expression:
A regular expression specifying what should be replaced.
Setting Regex
Regular Expression
[0-9].*
Regular expression valid.
Replacement (for preview only)
N
Inline Text Search
Result List (1)
Regexp Options
Text
677890
Result preview
N
Apply
Cancel
Ujicoba Regex
Replace (2) (Replace)
attribute filter typ...
all
invert selection
include special attributes
replace what
[0-9]
replace by
N
Change compatibility (8.0.001)
Replace
RapidMiner Studio Core
Tags: Map, Change, Regex, Regular expressions, Values
Synopsis
No results were found.
Indonesian
US keyboard
To switch input methods, press Windows key+Space.
IND
15:07
23-Jan-2018

Latihan

- Impor data MissingDataValue-Noisy.csv
- Gunakan Regular Expression (operator Replace) untuk mengganti semua noisy data pada atribut nominal menjadi “N”

ExampleSet (Retrieve MissingDataSet-Noisy)

Turbo Prep | Auto Model | Filter (11 / 11 examples): all

Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Gam...	Facebook	Twitter	Other_Socia...
8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
2	3	Internet Explo...	Bing	Hotmail	Y	Y	N	Y	N	?
15	4	Internet Explo...	Google	Yahoo	Y	N	Y	N	N	?
11	2	Firefox	Google	Yahoo	10	Y	765	Y	Y	LinkedIn
3	3	Internet Explo...	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
12	1	Safari	Yahoo	Yahoo	Y	9	Y	Y	N	MySpace
12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

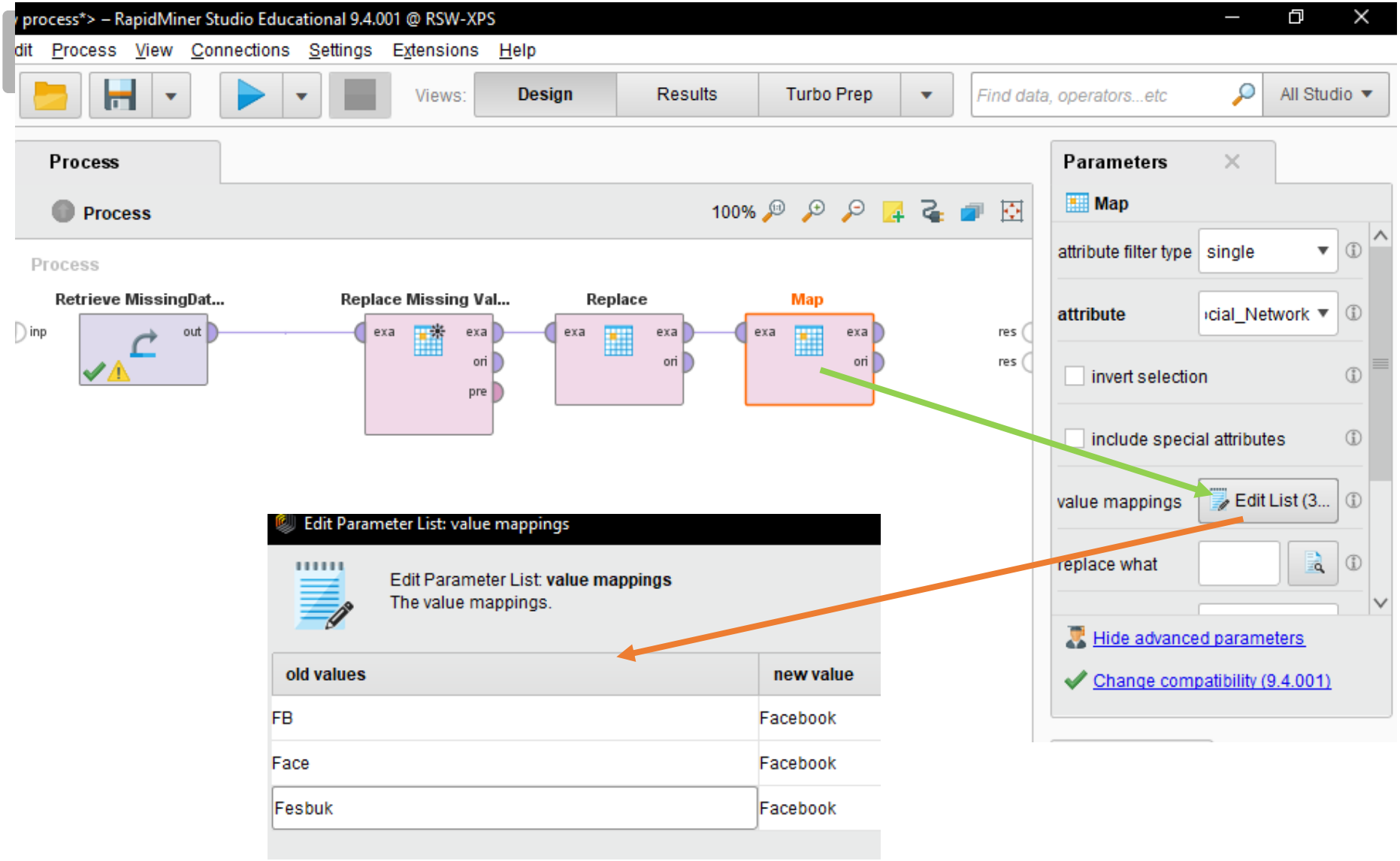
Latihan

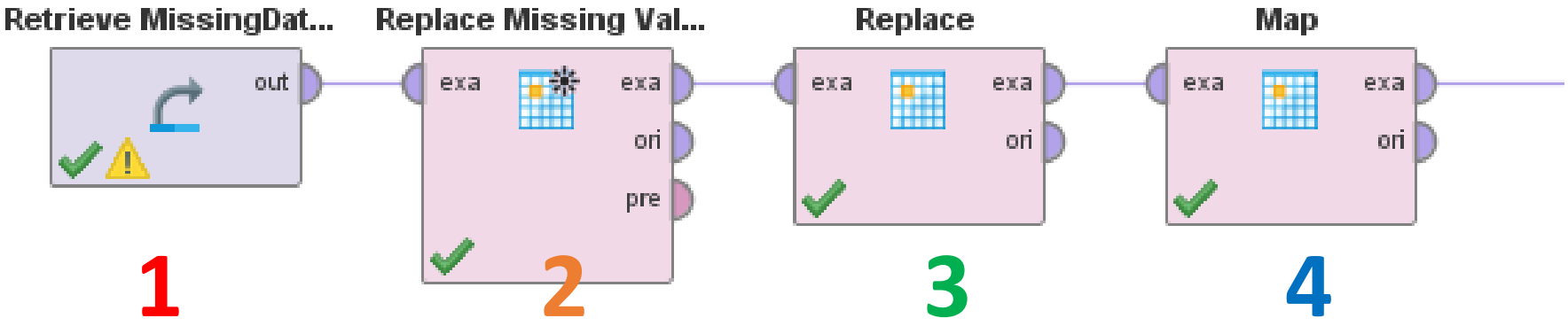
1. Impor data MissingDataValue-Noisy-Multiple.csv
2. Gunakan operator Replace Missing Value untuk mengisi data kosong
3. Gunakan Regular Expression (operator Replace) untuk mengganti semua noisy data pada atribut nominal menjadi “N”
4. Gunakan operator Map untuk mengganti semua isian Face, FB dan Fesbuk menjadi Facebook

ExampleSet (Retrieve MissingDataSet-Noisy-Multiple)

Turbo Prep | Auto Model | Filter (11 / 11 examples): all

Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Gam...	Facebook	Twitter	Other_Socia...
8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	Face
6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	FB
8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	Fesbuk
2	3	Internet Explo...	Bing	Hotmail	Y	Y	N	Y	N	?
15	4	Internet Explo...	Google	Yahoo	Y	N	Y	N	N	?
11	2	Firefox	Google	Yahoo	10	Y	765	Y	Y	LinkedIn
3	3	Internet Explo...	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
12	1	Safari	Yahoo	Yahoo	Y	9	Y	Y	N	MySpace
12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+


process*> – RapidMiner Studio Educational 9.4.001 @ RSW-XPS
dit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep
Find data, operators...etc
All Studio
Process
Process
100%
Process
Retrieve MissingDat...
Replace Missing Val...
Replace
Map
inp
out
exa
ori
pre
res
Parameters
Map
attribute filter type
single
attribute
cial_Network
invert selection
include special attributes
value mappings
Edit List (3...
replace what
Hide advanced parameters
Change compatibility (9.4.001)
Edit Parameter List: value mappings
Edit Parameter List: value mappings
The value mappings.
old values
new value
FB
Facebook
Face
Facebook
Fesbuk
Facebook



1. Impor data MissingDataValue-Noisy-Multiple.csv
2. operator Replace Missing Value untuk mengisi data kosong
3. operator Replace untuk mengganti semua noisy data pada atribut nominal menjadi “N”
4. operator Map untuk mengganti semua isian Face, FB dan Fesbuk menjadi Facebook

Studi Kasus CRISP-DM

Sport Skill – Discriminant Analysis

(Matthew North, Data Mining for the Masses 2nd Edition, 2016, ***Chapter 7 Discriminant Analysis, pp. 123-143****)*

Dataset: SportSkill-Training.csv

Dataset: SportSkill-Scoring.csv

1. Business Understanding

- **Motivation**:
  - Gill runs a sports academy designed to help high school aged athletes achieve their maximum athletic potential. He focuses on four major sports: Football, Basketball, Baseball and Hockey
  - He has found that while many high school athletes enjoy participating in a number of sports in high school, as they begin to consider playing a sport at the college level, they would prefer to specialize in one sport
  - As he's worked with athletes over the years, Gill has developed an extensive data set, and he now is wondering if he can use past performance from some of his previous clients to predict prime sports for up-and-coming high school athletes
  - By evaluating each athlete's performance across a battery of test, Gill hopes we can help him figure out for which sport each athlete has the highest aptitude
- **Objective**:
  - Ultimately, he hopes he can make a recommendation to each athlete as to the sport in which they should most likely choose to specialize

2. Data Understanding

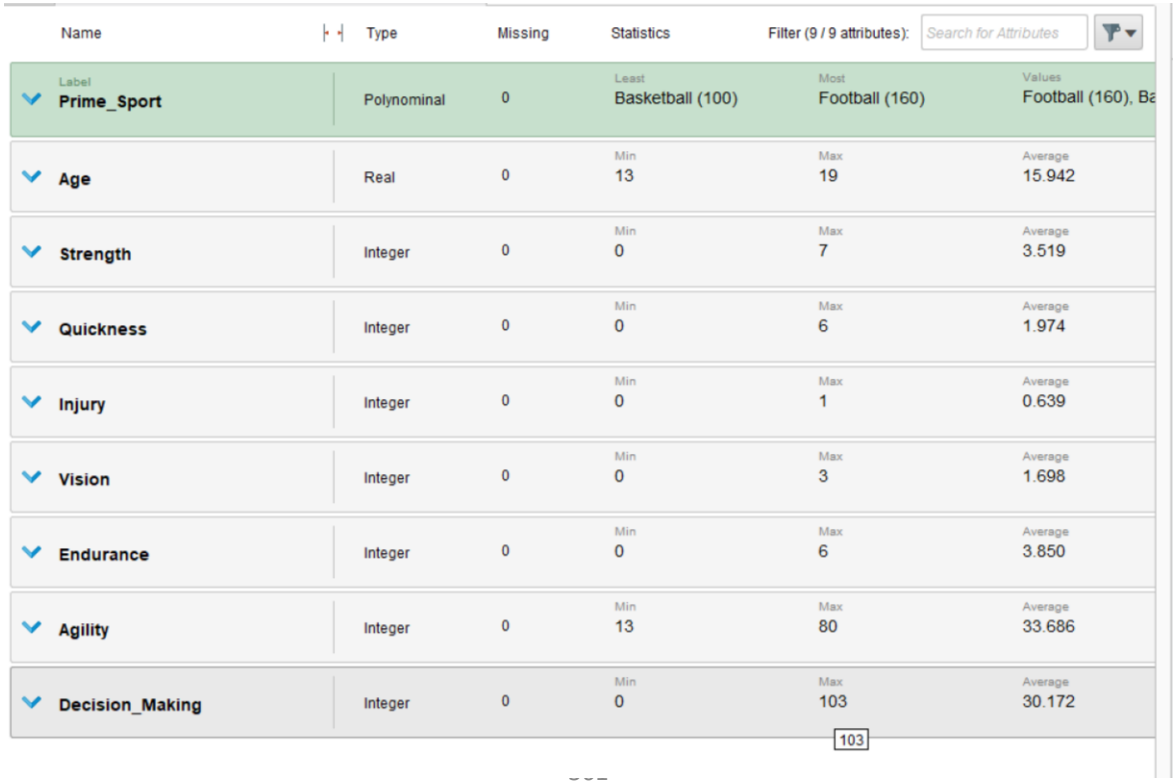
- Every athlete that has enrolled at Gill's academy over the past several years has taken a battery test, which tested for a number of athletic and personal traits
- Because the academy has been operating for some time, Gill has the benefit of knowing which of his former pupils have gone on to specialize in a single sport, and which sport it was for each of them

2. Data Understanding

- Working with Gill, we gather the results of the batteries for all former clients who have gone on to specialize
- Gill adds the sport each person specialized in, and we have a data set comprised of 493 observations containing the following attributes:
    1. Age: ....
    2. Strength: ....
    3. Quickness: ....
    4. Injury: ....
    5. Vision: ....
    6. Endurance: ....
    7. Agility: ....
    8. Decision Making: ....
    9. Prime Sport: ....

3. Data Preparation

- Filter Examples: attribute value filter
  - Decision_Making>=3
  - Decision_Making<=100
- Deleted Records= 493-482=11

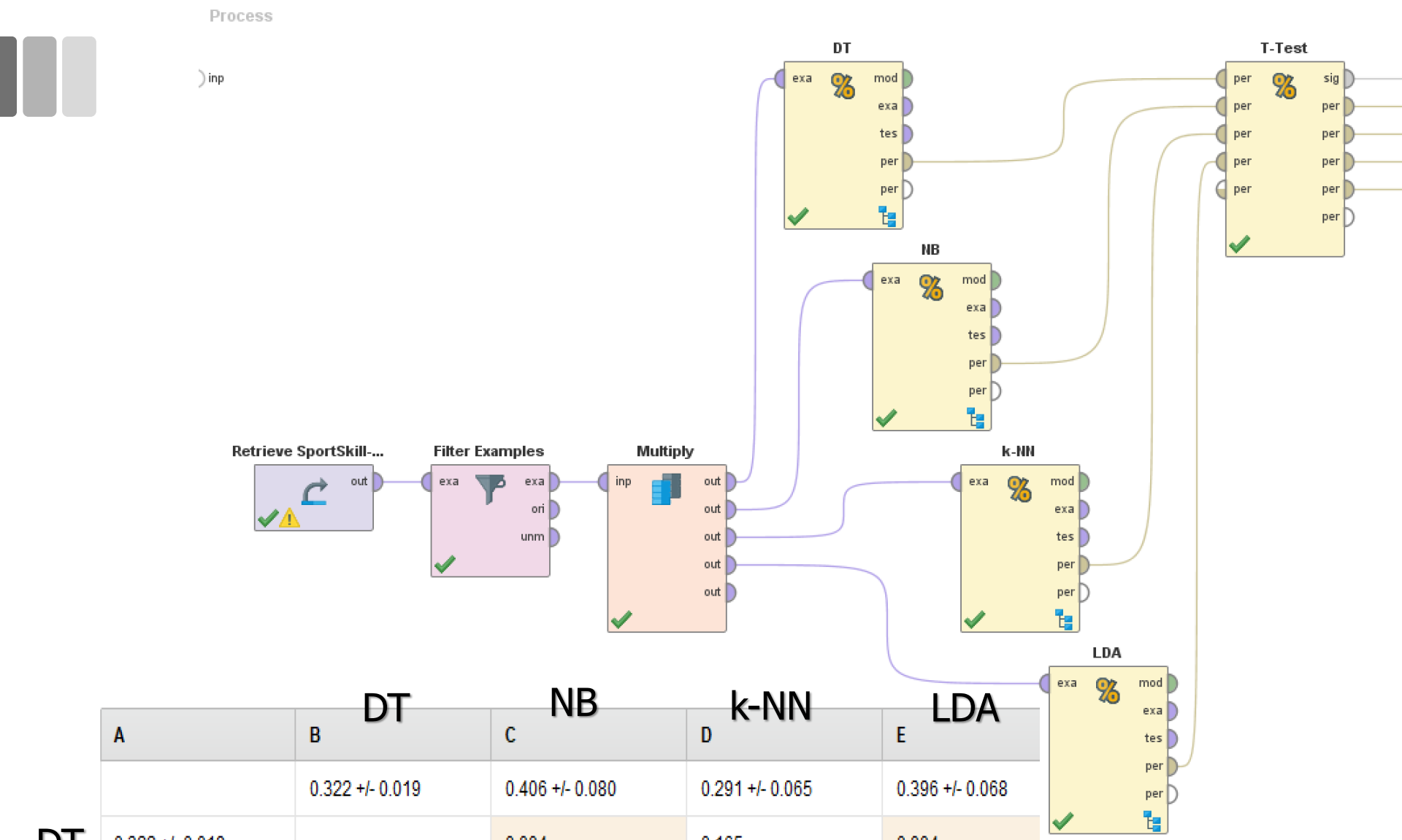


Name	Type	Missing	Statistics		Filter (9 / 9 attributes): Search for Attributes
Label **Prime_Sport**	Polynominal	0	Least Basketball (100)	Most Football (160)	Values Football (160), Ba
Age	Real	0	Min 13	Max 19	Average 15.942
Strength	Integer	0	Min 0	Max 7	Average 3.519
Quickness	Integer	0	Min 0	Max 6	Average 1.974
Injury	Integer	0	Min 0	Max 1	Average 0.639
Vision	Integer	0	Min 0	Max 3	Average 1.698
Endurance	Integer	0	Min 0	Max 6	Average 3.850
Agility	Integer	0	Min 13	Max 80	Average 33.686
Decision_Making	Integer	0	Min 0	Max 103	Average 30.172

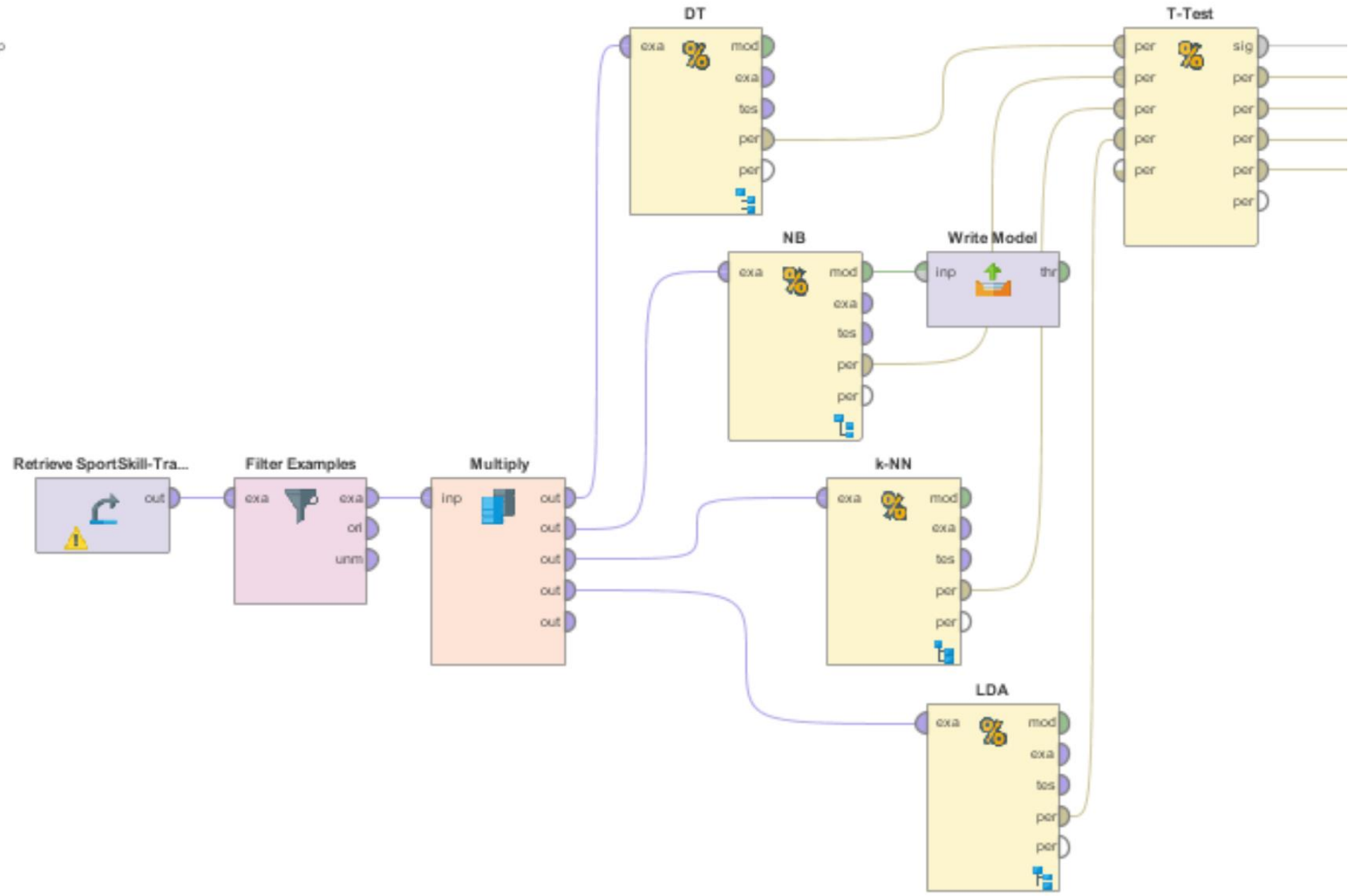
103

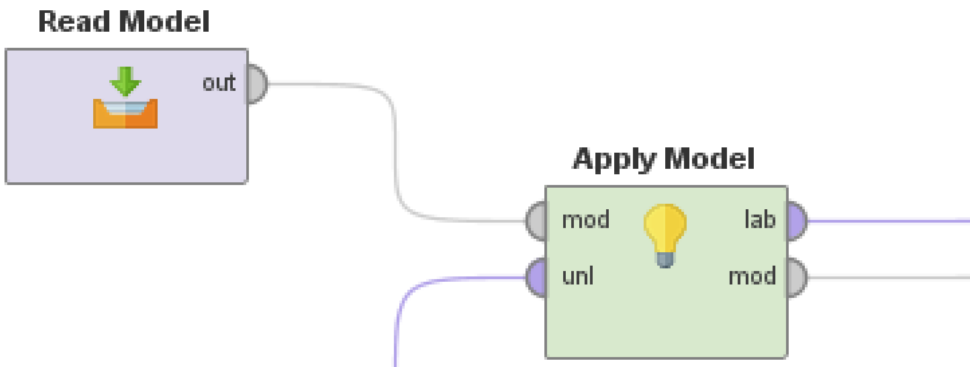
Latihan

1. Lakukan training pada data SportSkill-Training.csv dengan menggunakan C4.5, NB, K-NN dan LDA
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Uji beda dengan t-Test untuk mendapatkan model terbaik
4. Simpan model terbaik dari komparasi di atas dengan operator Write Model, dan kemudian Apply Model pada dataset SportSkill-Scoring.csv



	DT	NB	k-NN	LDA
A	B	C	D	E
	0.322 +/- 0.019	0.406 +/- 0.080	0.291 +/- 0.065	0.396 +/- 0.068
DT	0.322 +/- 0.019	0.004	0.165	0.004
NB	0.406 +/- 0.080		0.002	0.758
k-NN	0.291 +/- 0.065			0.002
LDA	0.396 +/- 0.068			


DT
exa
mod
exa
tes
per
per
T-Test
per
per
per
per
per
sig
per
per
per
per
per
NB
exa
mod
exa
tes
per
per
Write Model
inp
thr
Retrieve SportSkill-Tra...
out
Filter Examples
exa
exa
cri
unm
Multiply
inp
out
out
out
out
out
k-NN
exa
mod
exa
tes
per
per
LDA
exa
mod
exa
tes
per
per



ExampleSet (1841 examples, 1 special attribute, 8 regular attributes)

Filter (1,841 / 1,841 examples): all

Row No.	prediction(Prime_Spo...	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M...
1	Basketball	18.500	5	1	1	0	5	33	61
2	Baseball	13.300	1	2	1	3	5	18	59
3	Football	13.400	2	1	0	2	5	40	11
4	Hockey	13.600	4	1	0	0	5	28	0
5	Baseball	16.300	3	1	0	2	5	32	35
6	Football	15.700	1	1	0	2	3	43	37
7	Baseball	17	3	2	0	3	5	21	41
8	Football	16.300	3	1	0	1	1	41	29
9	Baseball	15.700	1	2	1	3	5	17	45
10	Football	16.500	3	2	0	1	3	46	40
11	Football	18.900	5	1	1	2	5	41	6
12	Football	14.600	5	2	1	0	3	35	48
13	Baseball	17.300	2	2	1	2	1	28	32
14	Football	14.900	6	1	1	3	3	42	39
15	Football	15.800	4	1	1	1	1	49	4
16	Basketball	14.200	5	1	1	0	5	24	55
17	Basketball	14.500	4	2	1	0	5	21	45
18	Football	13.300	5	1	1	3	3	42	29
19	Hockey	13.700	4	2	1	0	3	27	7
20	Baseball	17.300	0	0	0	3	5	15	43
21	Football	15	5	1	0	2	5	31	4

3.2 Data Reduction

Data Reduction Methods

- Data Reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results
- Why Data Reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis take a very long time to run on the complete dataset

- Data Reduction Methods
  1. Dimensionality Reduction
     1. Feature Extraction
     2. Feature Selection
        1. Filter Approach
        2. Wrapper Approach
        3. Embedded Approach
  2. Numerosity Reduction (Data Reduction)
     - Regression and Log-Linear Models
     - Histograms, clustering, sampling

1. Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- Dimensionality reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

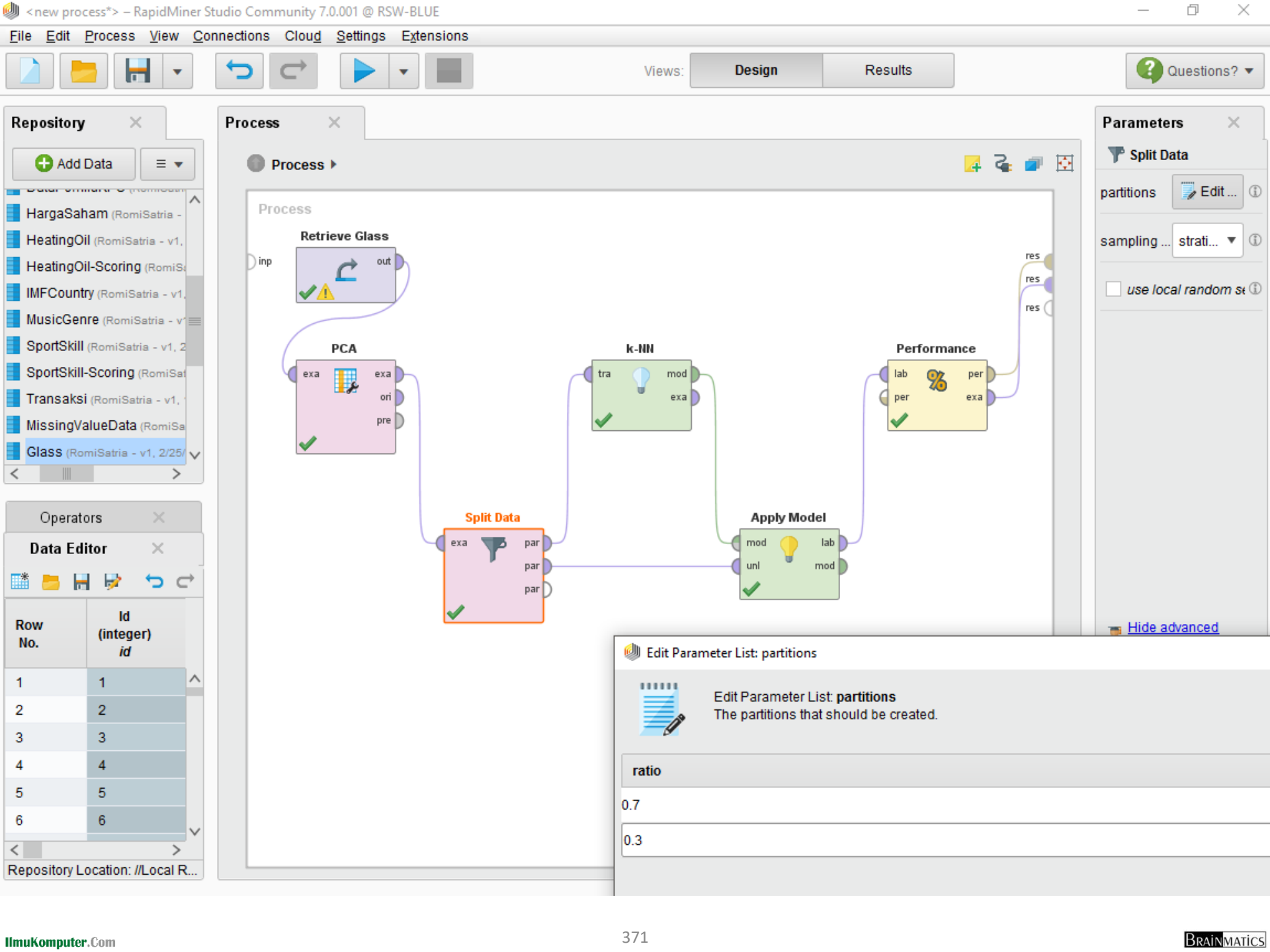
- Dimensionality Reduction Methods:
  1. Feature Extraction: Wavelet transforms, Principal Component Analysis (PCA)
  2. Feature Selection: Filter, Wrapper, Embedded

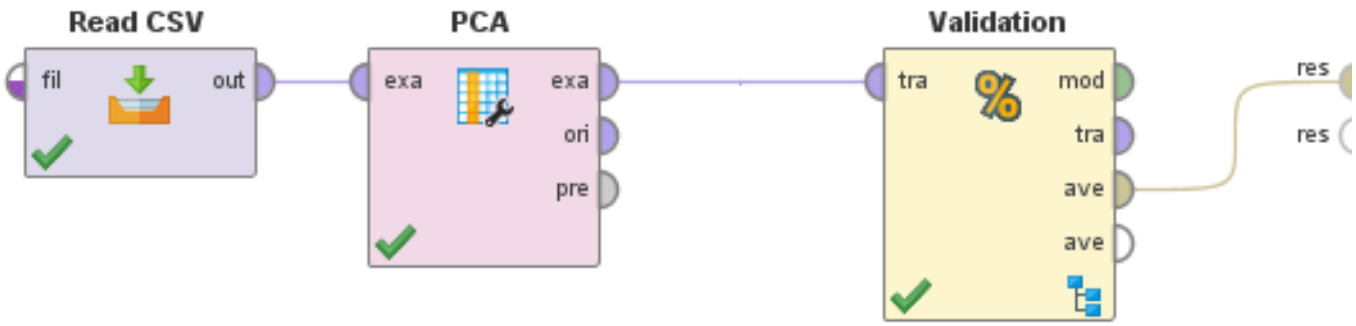
Principal Component Analysis (Steps)

- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  1. Normalize input data: Each attribute falls within the same range
  2. Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  3. Each input data (vector) is a linear combination of the *k* principal component vectors
  4. The principal components are sorted in order of decreasing "significance" or strength
  5. Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance
- Works for numeric data only

Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 4 (k-Nearest Neighbor Classification II) pp. 45-51
- Dataset: glass.data
- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!
- Bandingkan akurasi dari k-NN dan PCA+k-NN

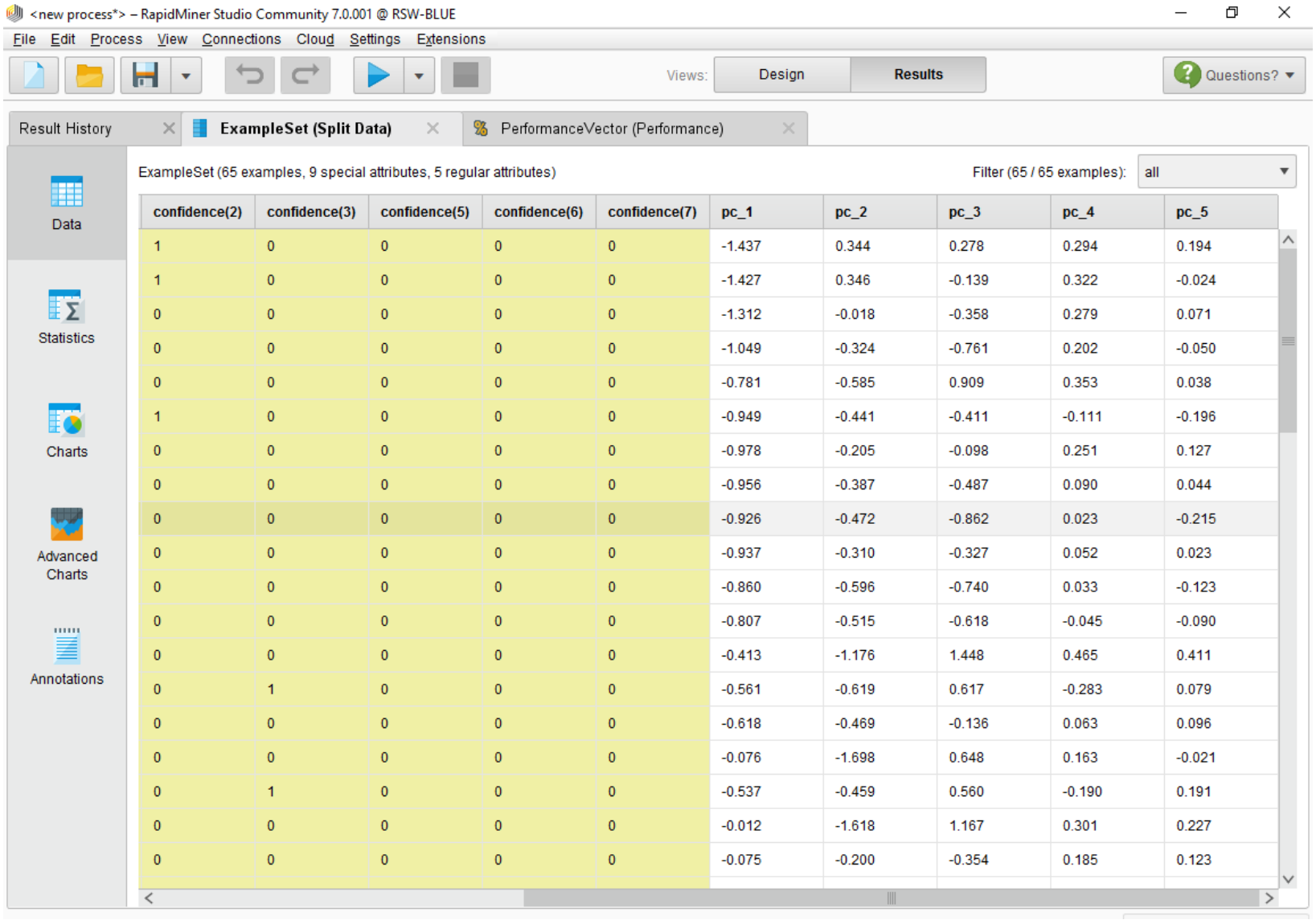

<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
HargaSaham (RomiSatria -
HeatingOil (RomiSatria - v1,
HeatingOil-Scoring (RomiS
IMFCountry (RomiSatria - v1,
MusicGenre (RomiSatria - v1
SportSkill (RomiSatria - v1, 2
SportSkill-Scoring (RomiSat
Transaksi (RomiSatria - v1,
MissingValueData (RomiSa
Glass (RomiSatria - v1, 2/25/
Operators
Data Editor
Row No.
Id (integer) id
1 1
2 2
3 3
4 4
5 5
6 6
Repository Location: //Local R...
Process
Process
Process
Retrieve Glass
inp
out
PCA
exa
exa
ori
pre
k-NN
tra
mod
exa
Performance
lab
per
per
exa
res
res
res
Split Data
exa
par
par
par
Apply Model
mod
lab
unl
mod
Parameters
Split Data
partitions
Edit ...
sampling ...
strati...
use local random se
Hide advanced
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
0.3


Read CSV
fil
out
PCA
exa
exa
ori
pre
Validation
tra
mod
tra
ave
ave
res
res

Data Awal Sebelum PCA

Type	RI	Na	Mg	Al	Si	K	Ca	Ba
1	1.521	13.640	4.490	1.100	71.780	0.060	8.750	0
1	1.518	13.890	3.600	1.360	72.730	0.480	7.830	0
1	1.516	13.530	3.550	1.540	72.990	0.390	7.780	0
1	1.518	13.210	3.690	1.290	72.610	0.570	8.220	0
1	1.517	13.270	3.620	1.240	73.080	0.550	8.070	0
1	1.516	12.790	3.610	1.620	72.970	0.640	8.070	0
1	1.517	13.300	3.600	1.140	73.090	0.580	8.170	0
1	1.518	13.150	3.610	1.050	73.240	0.570	8.240	0
1	1.519	14.040	3.580	1.370	72.080	0.560	8.300	0
1	1.518	13	3.600	1.360	72.990	0.570	8.400	0
1	1.516	12.720	3.460	1.560	73.200	0.670	8.090	0
1	1.518	12.800	3.660	1.270	73.010	0.600	8.560	0
1	1.516	12.880	3.430	1.400	73.280	0.690	8.050	0
1	1.517	12.860	3.560	1.270	73.210	0.540	8.380	0
1	1.518	12.610	3.590	1.310	73.290	0.580	8.500	0
1	1.518	12.810	3.540	1.230	73.240	0.580	8.390	0
1	1.518	12.680	3.670	1.160	73.110	0.610	8.700	0
1	1.522	14.360	3.850	0.890	71.360	0.150	9.150	0
1	1.519	13.900	3.730	1.180	72.120	0.060	8.890	0

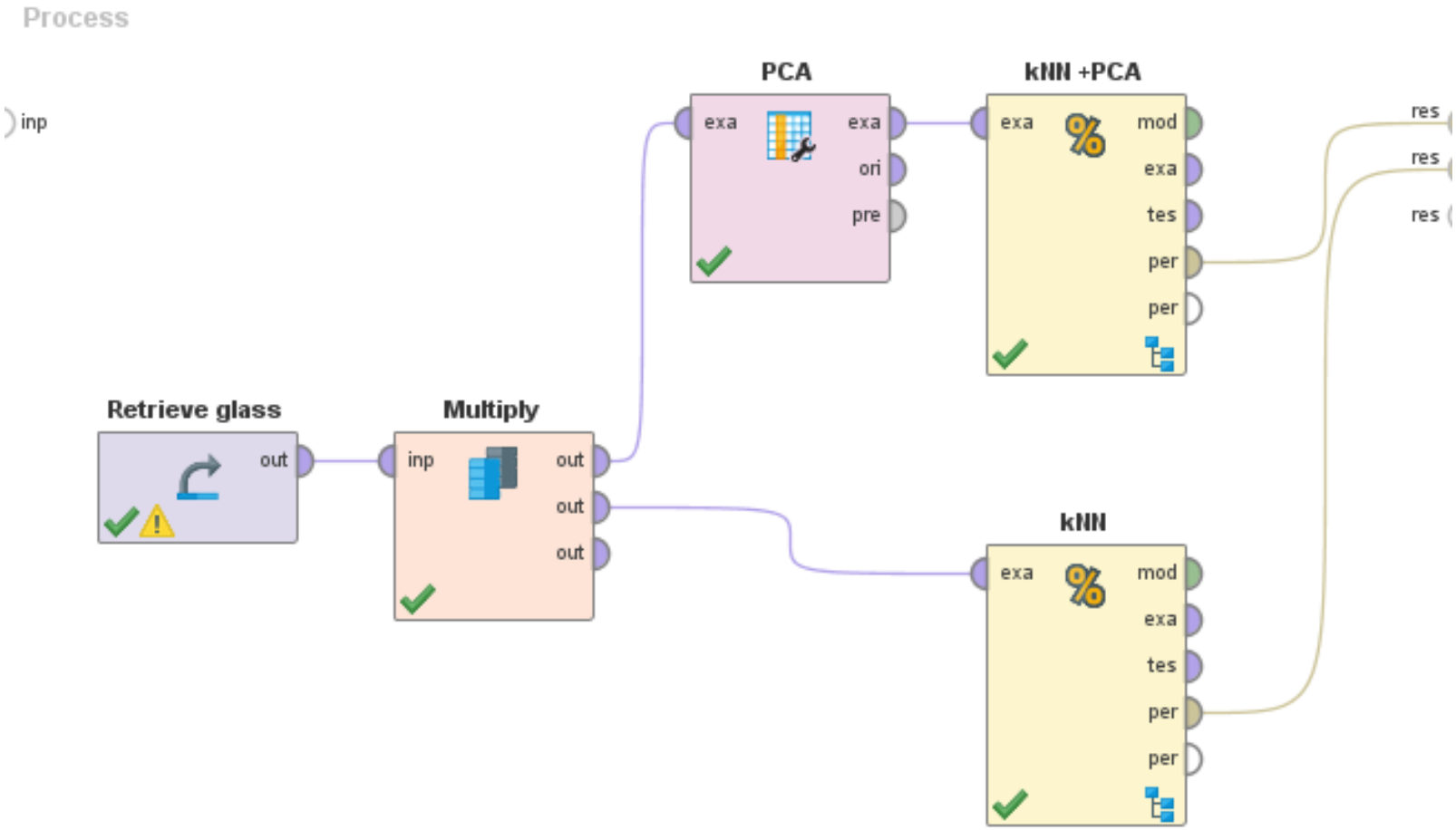
Data Setelah PCA



confidence(2)	confidence(3)	confidence(5)	confidence(6)	confidence(7)	pc_1	pc_2	pc_3	pc_4	pc_5
1	0	0	0	0	-1.437	0.344	0.278	0.294	0.194
1	0	0	0	0	-1.427	0.346	-0.139	0.322	-0.024
0	0	0	0	0	-1.312	-0.018	-0.358	0.279	0.071
0	0	0	0	0	-1.049	-0.324	-0.761	0.202	-0.050
0	0	0	0	0	-0.781	-0.585	0.909	0.353	0.038
1	0	0	0	0	-0.949	-0.441	-0.411	-0.111	-0.196
0	0	0	0	0	-0.978	-0.205	-0.098	0.251	0.127
0	0	0	0	0	-0.956	-0.387	-0.487	0.090	0.044
0	0	0	0	0	-0.926	-0.472	-0.862	0.023	-0.215
0	0	0	0	0	-0.937	-0.310	-0.327	0.052	0.023
0	0	0	0	0	-0.860	-0.596	-0.740	0.033	-0.123
0	0	0	0	0	-0.807	-0.515	-0.618	-0.045	-0.090
0	0	0	0	0	-0.413	-1.176	1.448	0.465	0.411
0	1	0	0	0	-0.561	-0.619	0.617	-0.283	0.079
0	0	0	0	0	-0.618	-0.469	-0.136	0.063	0.096
0	0	0	0	0	-0.076	-1.698	0.648	0.163	-0.021
0	1	0	0	0	-0.537	-0.459	0.560	-0.190	0.191
0	0	0	0	0	-0.012	-1.618	1.167	0.301	0.227
0	0	0	0	0	-0.075	-0.200	-0.354	0.185	0.123

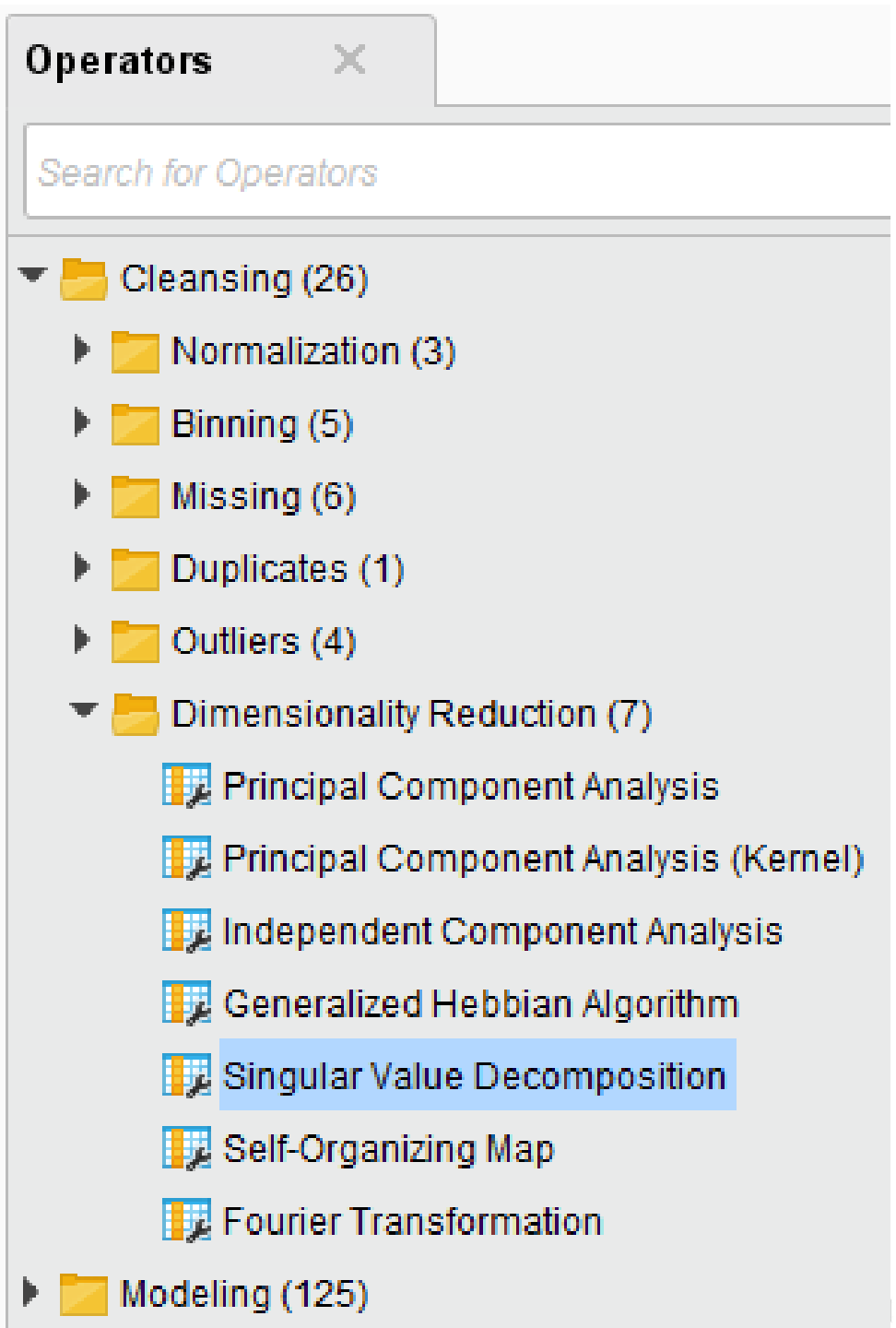
Latihan

- Susun ulang proses yang mengkomparasi model yang dihasilkan oleh k-NN dan PCA + k-NN
- Gunakan 10 Fold X Validation


Process
inp
PCA
exa
exa
ori
pre
kNN +PCA
exa
mod
exa
tes
per
per
res
res
res
Retrieve glass
out
Multiply
inp
out
out
out
kNN
exa
mod
exa
tes
per
per

Latihan

- Review operator apa saja yang bisa digunakan untuk *feature extraction*
- Ganti PCA dengan metode *feature extraction* yang lain
- Lakukan komparasi dan tentukan mana metode feature extraction terbaik untuk data Glass.data, gunakan 10-fold cross validation



Feature/Attribute Selection

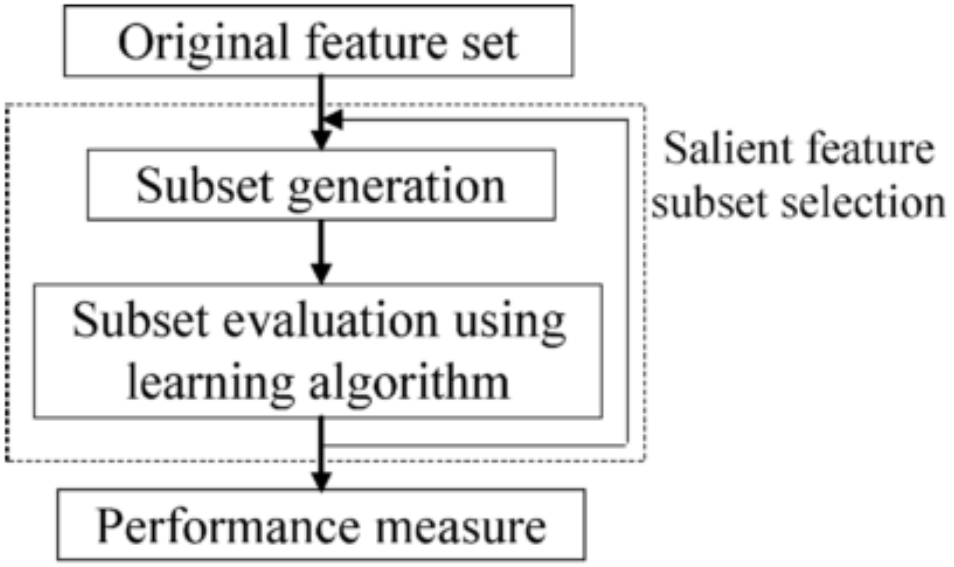
- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

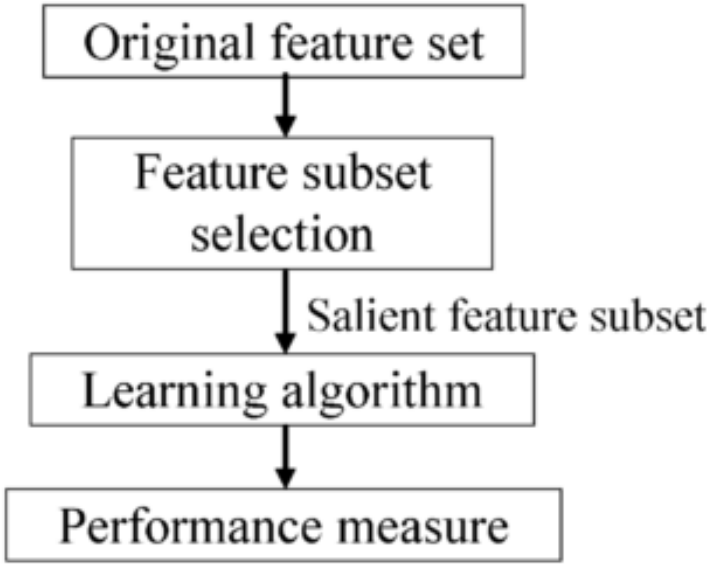
Feature Selection Approach

A number of proposed approaches for feature selection can broadly be categorized into the following three classifications: wrapper, filter, and embedded (Liu & Tu, 2004)

1. In the filter approach, statistical analysis of the feature set is required, without utilizing any learning model (Dash & Liu, 1997)
2. In the wrapper approach, a predetermined learning model is assumed, wherein features are selected that justify the learning performance of the particular learning model (Guyon & Elisseeff, 2003)
3. The embedded approach attempts to utilize the complementary strengths of the wrapper and filter approaches (Huang, Cai, & Xu, 2007)

Wrapper Approach vs Filter Approach





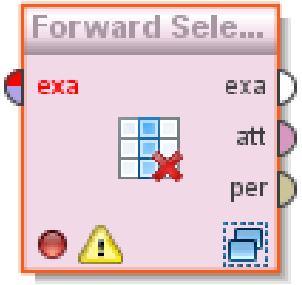
Wrapper Approach

Filter Approach

Feature Selection Approach

1. Filter Approach:
   - information gain
   - chi square
   - log likehood rasio
   - etc

2. Wrapper Approach:
   - forward selection
   - backward elimination
   - randomized hill climbing
   - etc



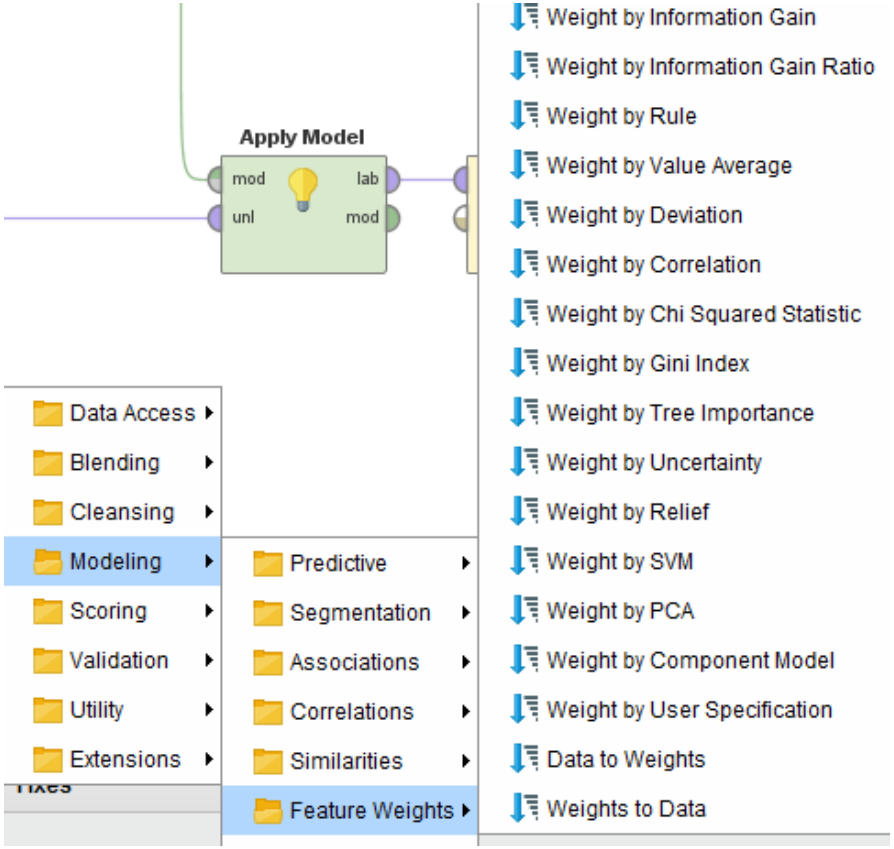
3. Embedded Approach:
   - decision tree
   - weighted naïve bayes
   - etc

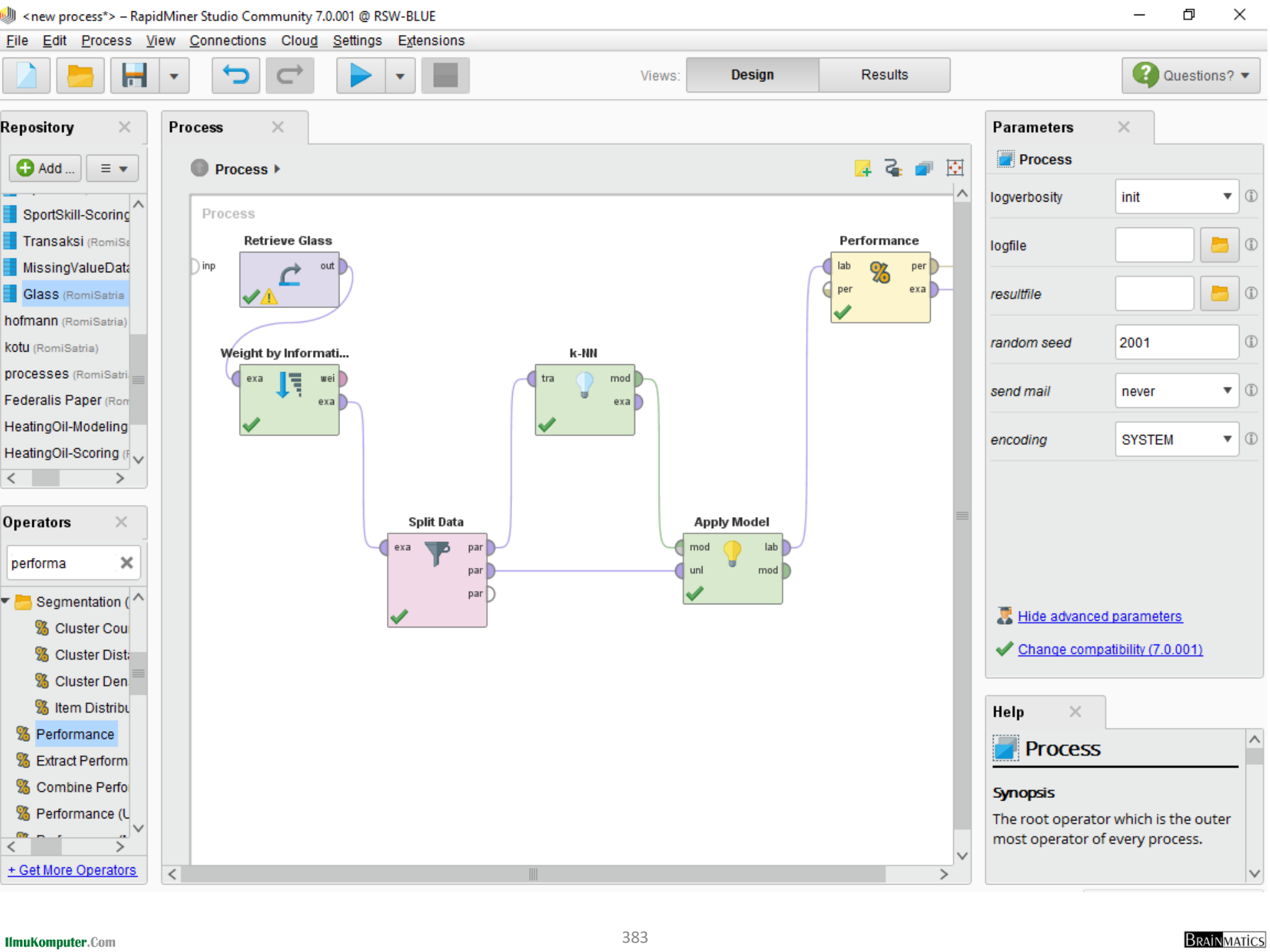
Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 4 (k-Nearest Neighbor Classification II)

- Ganti PCA dengan metode feature selection (filter), misalnya:
  - Information Gain
  - Chi Squared
  - etc

- Cek di RapidMiner, operator apa saja yang bisa digunakan untuk mengurangi atau membobot atribute dari dataset!




<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add ...
SportSkill-Scoring
Transaksi (RomiSa
MissingValueData
Glass (RomiSatria
hofmann (RomiSatria)
kotu (RomiSatria)
processes (RomiSatri
Federalis Paper (Ron
HeatingOil-Modeling
HeatingOil-Scoring
Operators
performa
Segmentation (
Cluster Cou
Cluster Dist
Cluster Den
Item Distribu
Performance
Extract Perform
Combine Perfo
Performance (U
+ Get More Operators
Process
Process
Process
Retrieve Glass
inp
out
Weight by Informati...
exa
wei
exa
k-NN
tra
mod
exa
Split Data
exa
par
par
par
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Hide advanced parameters
Change compatibility (7.0.001)
Help
Process
Synopsis
The root operator which is the outer most operator of every process.

<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Repository
Add Data
Samples
DB
Local Repository (romis)
data (romis)
Operators
Search for Operators
Cleansing (26)
Normalization (3)
Binning (5)
Missing (6)
Duplicates (1)
Outliers (4)
Dimensionality Reductior
Principal Component
Principal Component
Independent Compon
Generalized Hebbian
Singular Value Decom
Self-Organizing Map
Fourier Transformatio
Process
Process
Process
inp
Read CSV
fil
out
Weight by Informati...
exa
wei
exa
Validation
tra
mod
tra
ave
ave
res
res
Problems
No problems found
Message
Fixes
Location

Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 4 (k-Nearest Neighbor Classification II)

- Ganti PCA dengan metode feature selection (wrapper), misalnya:
  - Backward Elimination
  - Forward Selection
  - etc

- Ganti metode validasi dengan 10-Fold X Validation
- Bandingkan akurasi dari k-NN dan BE+k-NN or FS+k-NN

Repository

Add Data

SportSkill-Scoring (RomiSatria - v1, 2/24/
Transaksi (RomiSatria - v1, 11/11/15 2:10
MissingValueData (RomiSatria - v1, 2/25/
Glass (RomiSatria - v1, 2/25/16 10:38 AM
hofmann (RomiSatria)
kotu (RomiSatria)

Operators

Search for Operators

Parameters (5)
Feature Selection (6)
Forward Selection
Backward Elimination
Optimize Selection
Optimize Selection (Brute
Optimize Selection (Weigh
Optimize Selection (Evolut
Feature Generation (5)
Feature Weighting (4)
Scoring (10)
Validation (30)
Utility (87)

+ Get More Operators

Process

Process

Process

Retrieve Glass

inp out

Backward Elimination

exa exa att per

res res

Process

Process Backward Elimination

Backward Elimination

Validation

exa tra tra mod ave ave

A cross-validation evaluating a decision tree model.

Parameters

Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Hide advanced parameters

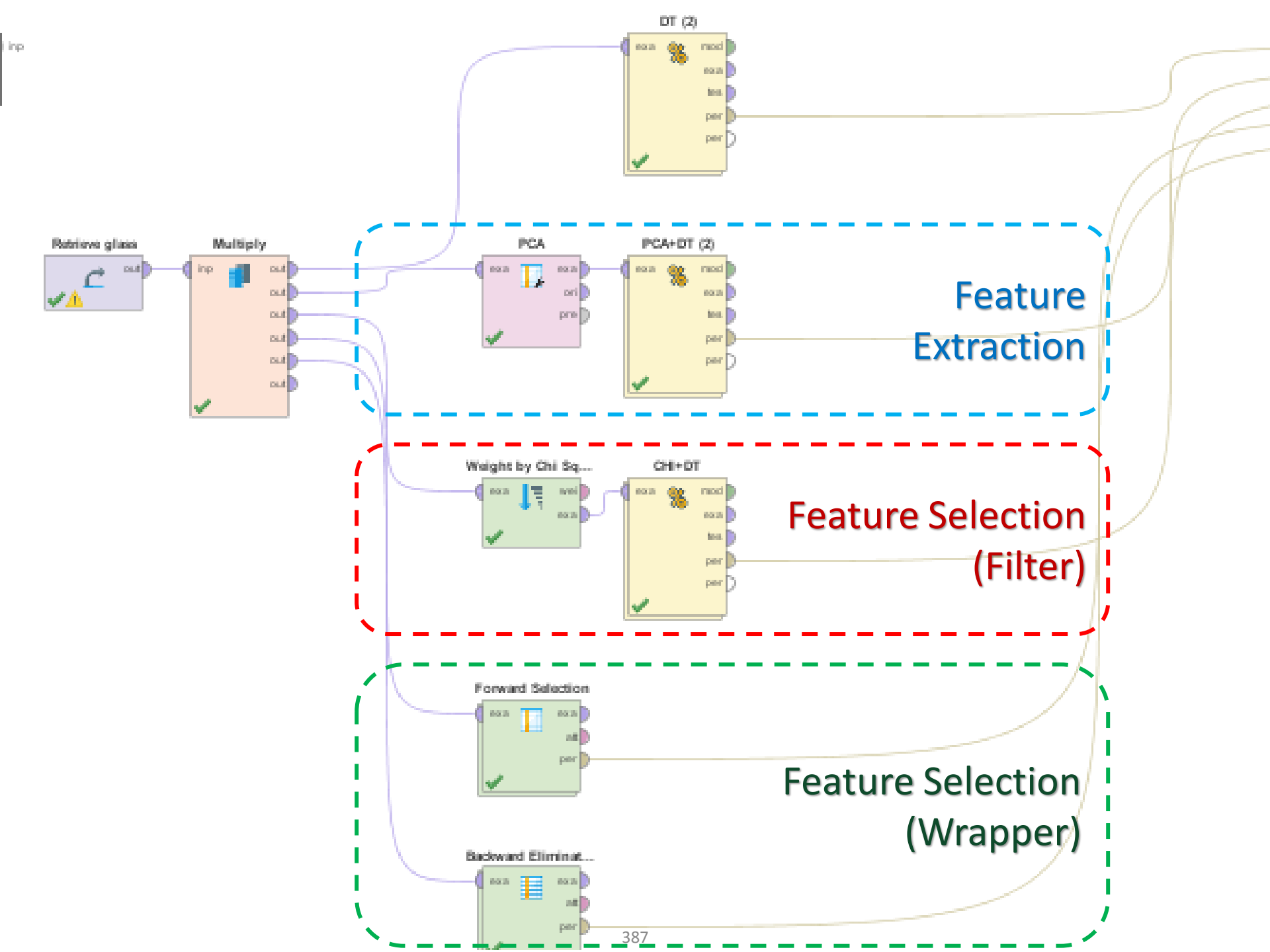
Change compatibility (7.0.001)

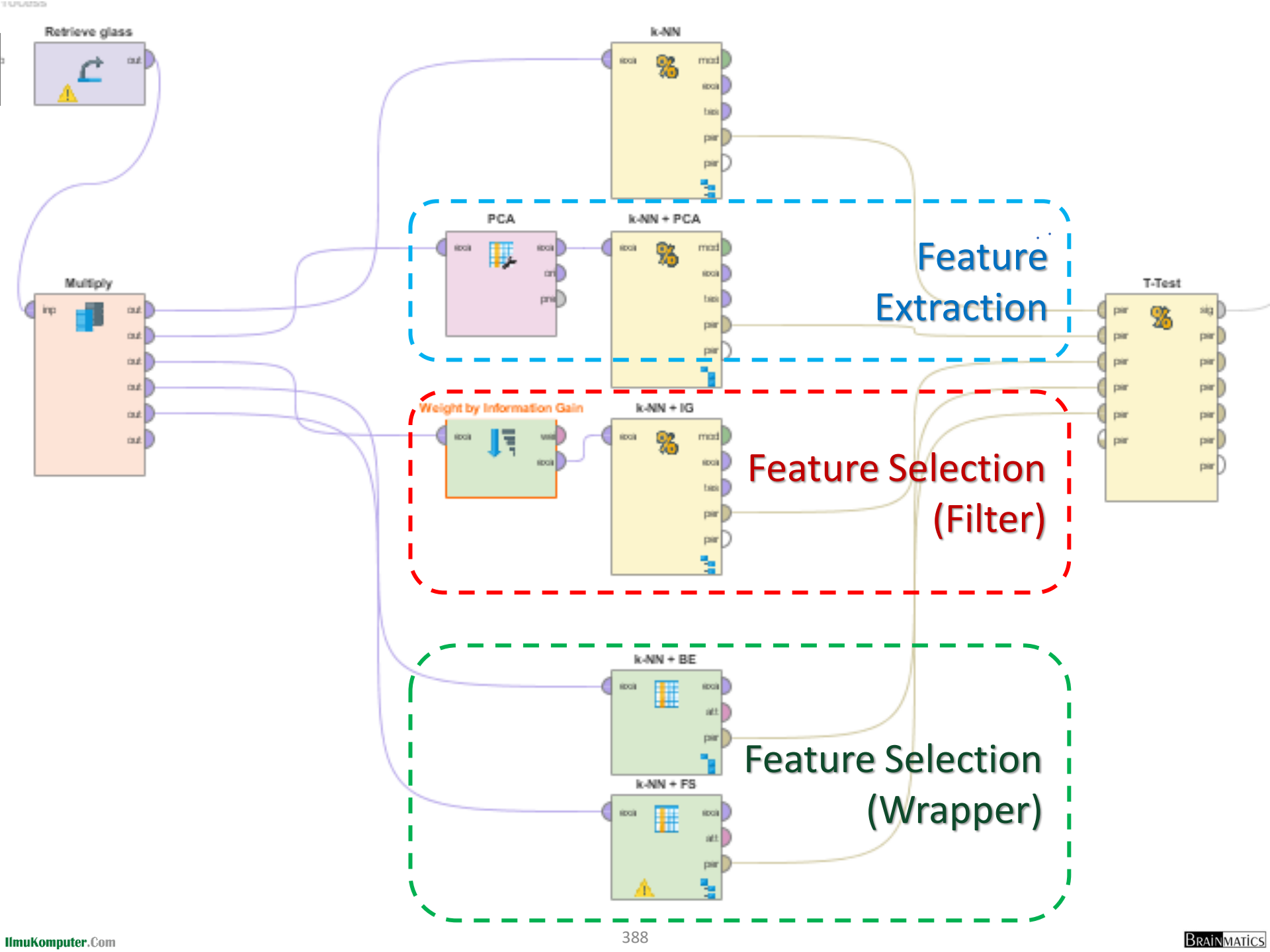
Help

Process

Synopsis

The root operator which is the outer most operator of every process.


DT (2)
Retrieve glass
Multiply
PCA
PCA+DT (2)
Feature Extraction
Weight by Chi Sq...
Chi+DT
Feature Selection (Filter)
Forward Selection
Backward Eliminat...
Feature Selection (Wrapper)


Retrieve glass
k-NN
PCA
k-NN + PCA
Feature Extraction
Multiply
T-Test
Weight by Information Gain
k-NN + IG
Feature Selection (Filter)
k-NN + BE
k-NN + FS
Feature Selection (Wrapper)

Hasil Komparasi Akurasi dan Signifikansi t-Test

	k-NN	k-NN+PCA	k-NN+ICA	k-NN+IG	k-NN+IGR	kNN + FS	k-NN+BE
Accuracy							
AUC							

Latihan: Prediksi Kelulusan Mahasiswa

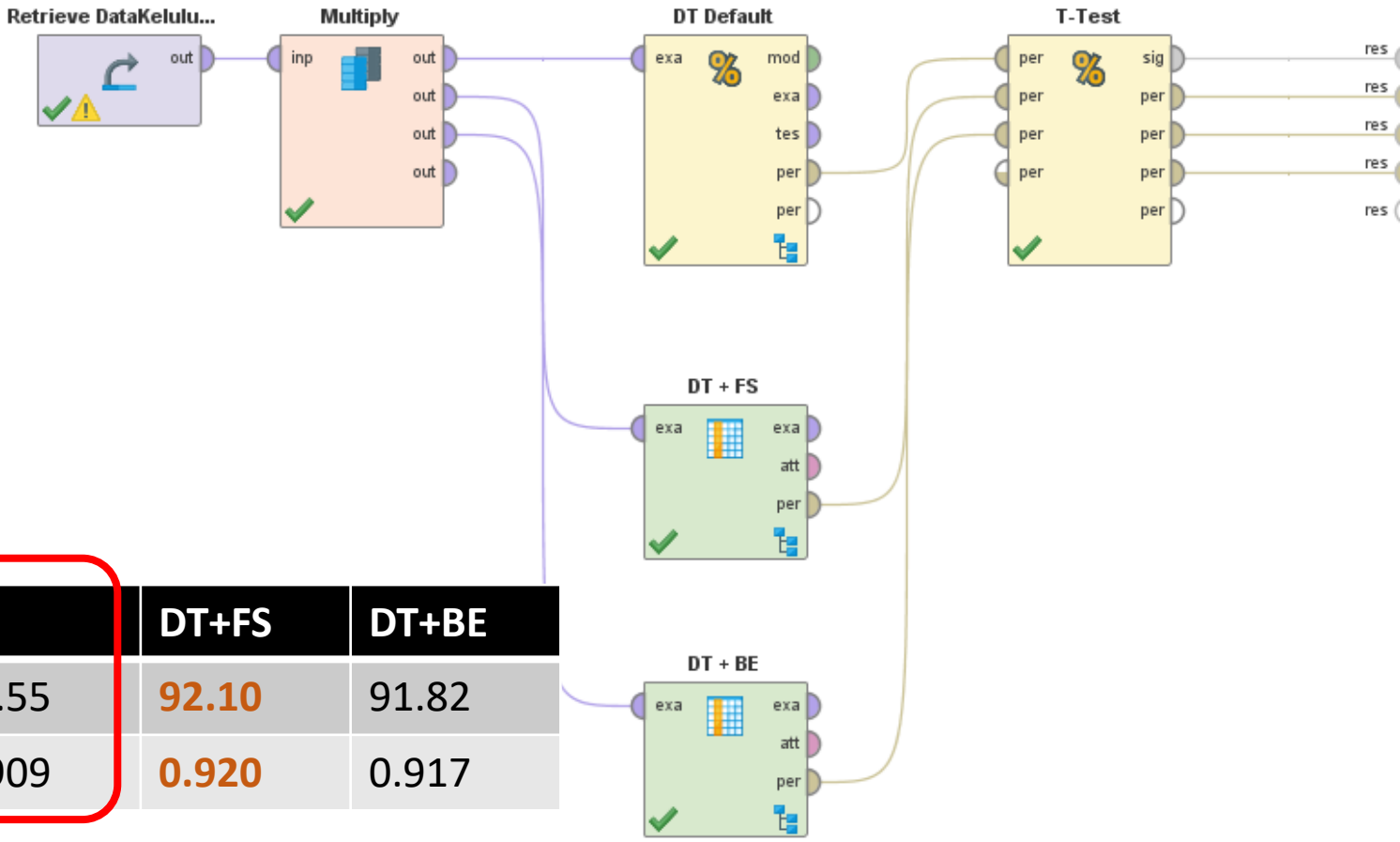
1. Lakukan training pada data mahasiswa (datakelulusanmahasiswa.xls) dengan menggunakan 3 algoritma klasifikasi (DT, NB, k-NN)
2. Analisis dan komparasi, mana algoritma klasifikasi yang menghasilkan model paling akurat (AK)
3. Lakukan feature selection dengan Information Gain (Filter), Forward Selection, Backward Elimination (Wrapper) untuk model yang paling akurat
4. Analisis dan komparasi, mana algoritma feature selection yang menghasilkan model paling akurat
5. Lakukan pengujian dengan menggunakan 10-fold X Validation

	AK	AK+IG	AK+FS	AK+BE
Accuracy	91.55		**92.10**	91.82
AUC	0.909		**0.920**	0.917

Latihan: Prediksi Kelulusan Mahasiswa

1. Lakukan training pada data mahasiswa (datakelulusanmahasiswa.xls) dengan menggunakan 4 algoritma klasifikasi (DT
2. Lakukan feature selection dengan Forward Selection untuk algoritma DT (DT+FS)
3. Lakukan feature selection dengan Backward Elimination untuk algoritma DT (DT+BE)
4. Lakukan pengujian dengan menggunakan 10-fold X Validation
5. Uji beda dengan t-Test untuk mendapatkan model terbaik (DT vs DT+FS vs DT+BE)

	DT	DT+FS	DT+BE
Accuracy	91.55	**92.10**	91.82
AUC	0.909	**0.920**	0.917



	DT	DT+FS	DT+BE
Accuracy	91.55	92.10	91.82
AUC	0.909	0.920	0.917

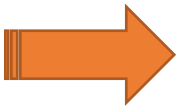
A	B	C	D
	0.916 +/- 0.039	0.921 +/- 0.050	0.918 +/- 0.032
0.916 +/- 0.039		0.787	0.867
0.921 +/- 0.050			0.884
0.918 +/- 0.032			

no significant difference

Latihan: Prediksi Elektabilitas Pemilu

1. Lakukan komparasi algoritma pada data pemilu (datapemilukpu.xls), sehingga didapatkan algoritma terbaik
2. Ambil algoritma terbaik dari langkah 1, kemudian lakukan feature selection dengan Forward Selection dan Backward Elimination
3. Tentukan kombinasi algoritma dan feature selection apa yang memiliki performa terbaik
4. Lakukan pengujian dengan menggunakan 10-fold X Validation
5. Uji beda dengan t-Test untuk mendapatkan model terbaik

	DT	NB	K-NN
Accuracy			
AUC			



	A	A + FS	A + BE
Accuracy			
AUC			

Latihan: Prediksi Kelulusan Mahasiswa

1. Lakukan training pada data mahasiswa (datakelulusanmahasiswa.xls) dengan menggunakan DT, NB, K-NN
2. Lakukan dimension reduction dengan Forward Selection untuk ketiga algoritma di atas
3. Lakukan pengujian dengan menggunakan 10-fold X Validation
4. Uji beda dengan t-Test untuk mendapatkan model terbaik

	DT	NB	K-NN	DT+FS	NB+FS	K-NN+FS
Accuracy						
AUC						

No Free Lunch Theory (Data Mining Law 4)

There is No Free Lunch for the Data Miner (NFL-DM)

The right model for a given application can only be discovered by experiment

- Axiom of machine learning: if we knew enough about a problem space, we could choose or design an algorithm to find optimal solutions in that problem space with maximal efficiency
- Arguments for the superiority of one algorithm over others in data mining rest on the idea that data mining problem spaces have one particular set of properties, or that these properties can be discovered by analysis and built into the algorithm
- However, these views arise from the erroneous idea that, in data mining, the data miner formulates the problem and the algorithm finds the solution
- In fact, the data miner both formulates the problem and finds the solution – the algorithm is merely a tool which the data miner uses to assist with certain steps in this process

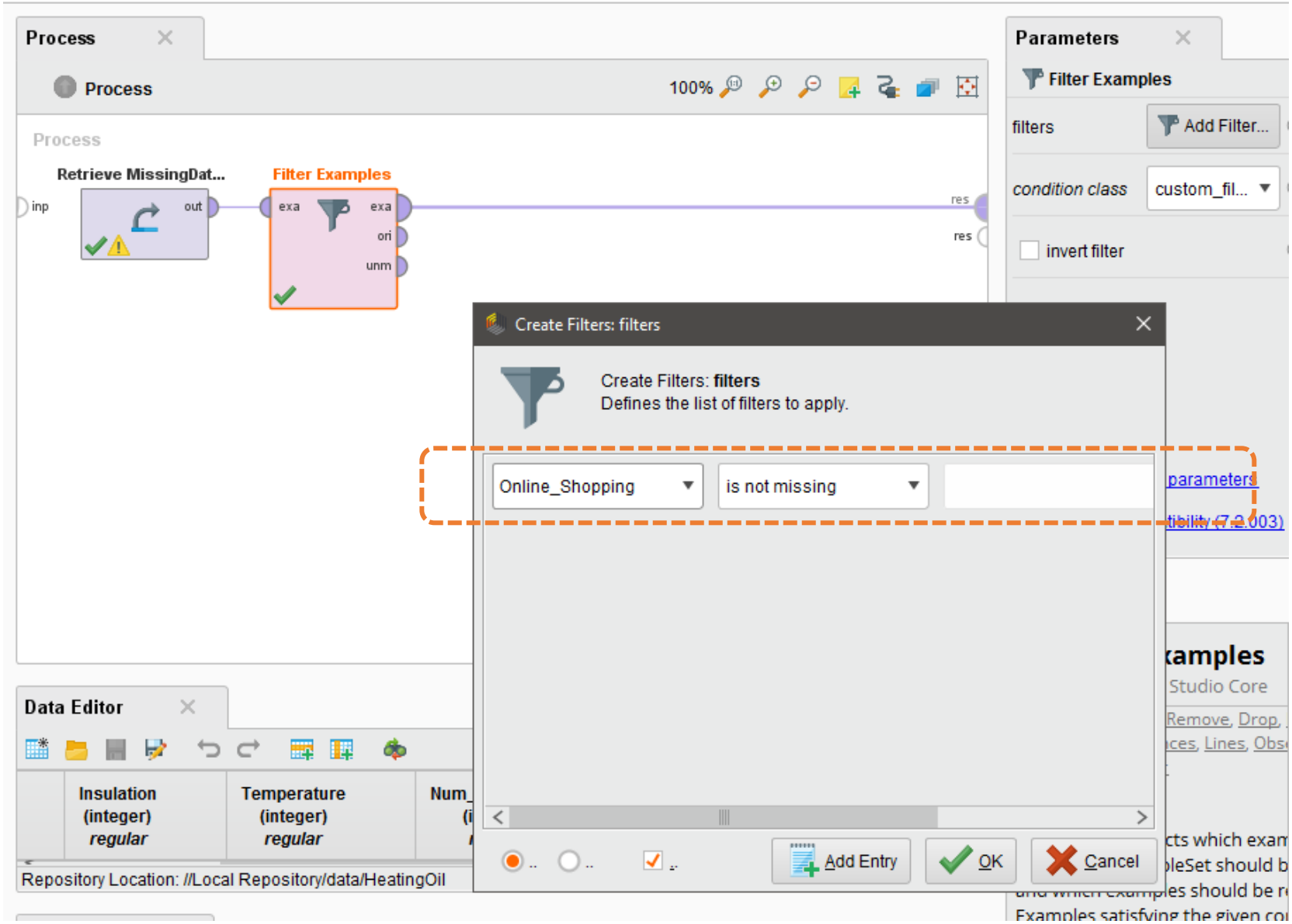
2. Numerosity Reduction

Reduce data volume by choosing alternative, smaller forms of data representation

1. Parametric methods (e.g., regression)
   - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
   - Ex.: Log-linear models—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces

2. Non-parametric methods
   - Do not assume models
   - Major families: histograms, clustering, sampling, …

Numerosity Reduction

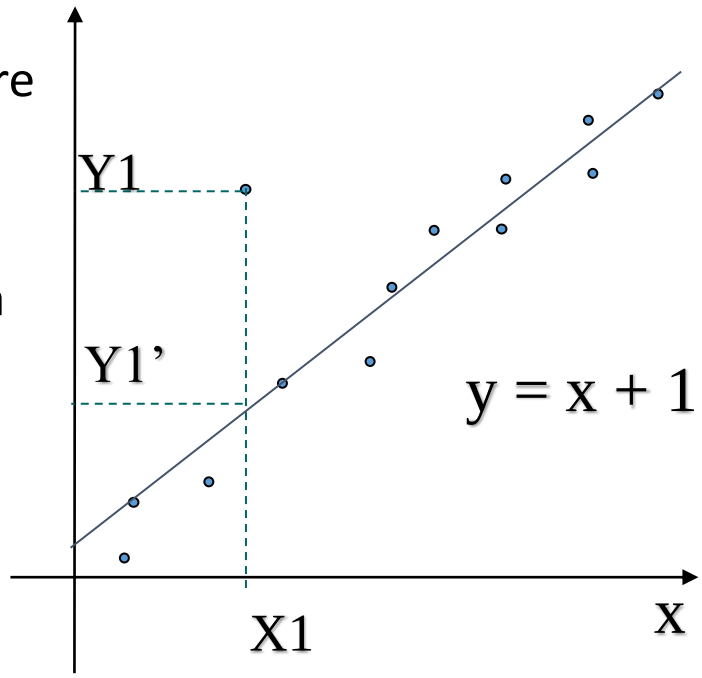


Parametric Data Reduction: Regression and Log-Linear Models

- Linear regression
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression
  - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model
  - Approximates discrete multidimensional probability distributions

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

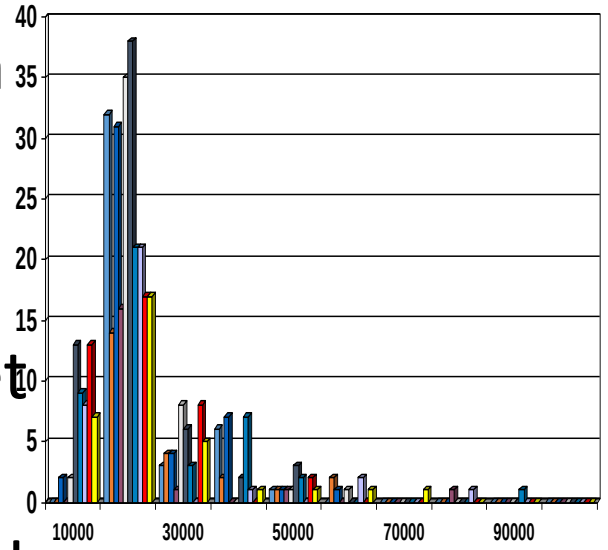


Regress Analysis and Log-Linear Models

- Linear regression: $Y = w X + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of Y_1, Y_2, …, X_1, X_2, ….
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - Approximate discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduction and data smoothing

Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

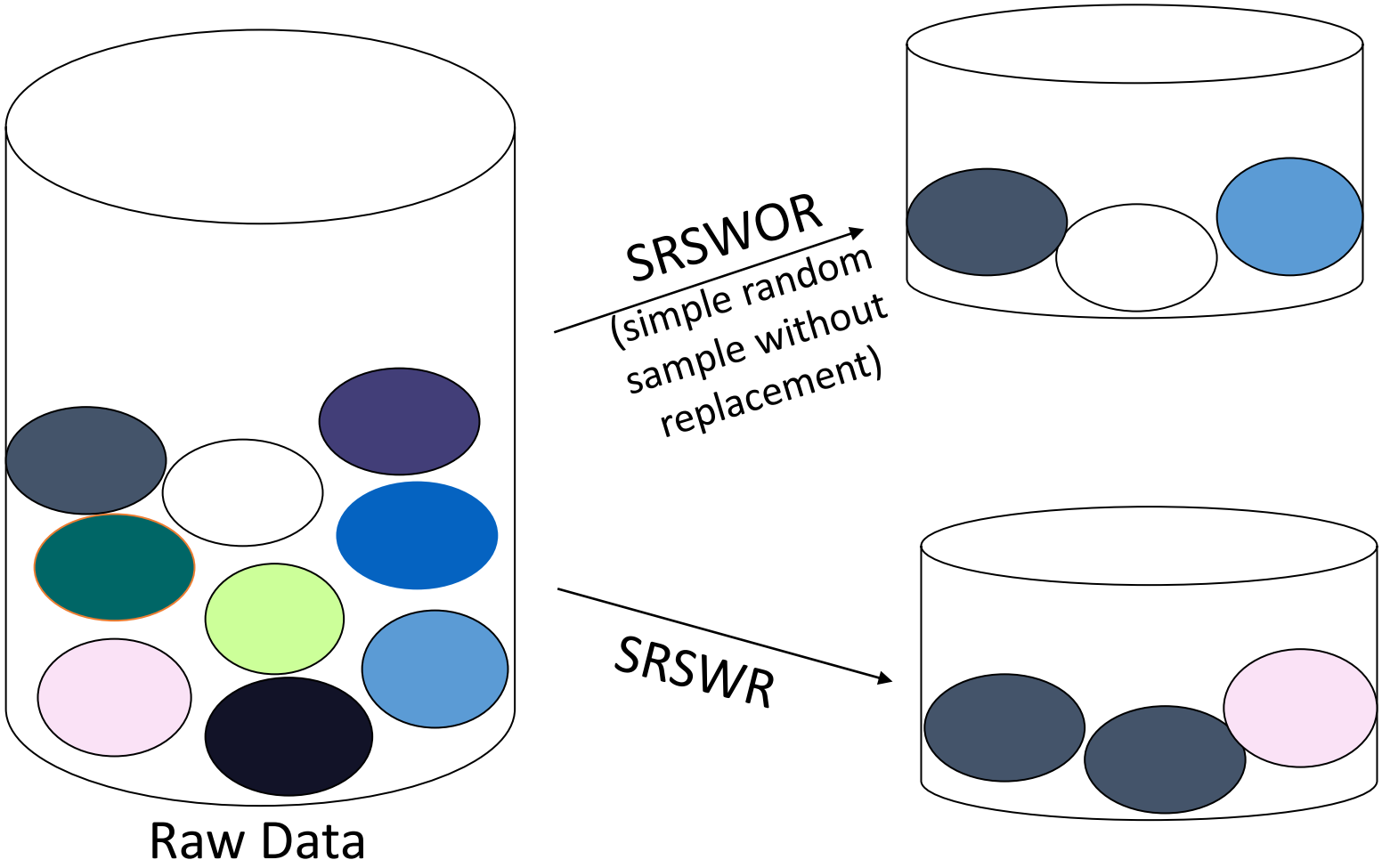
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

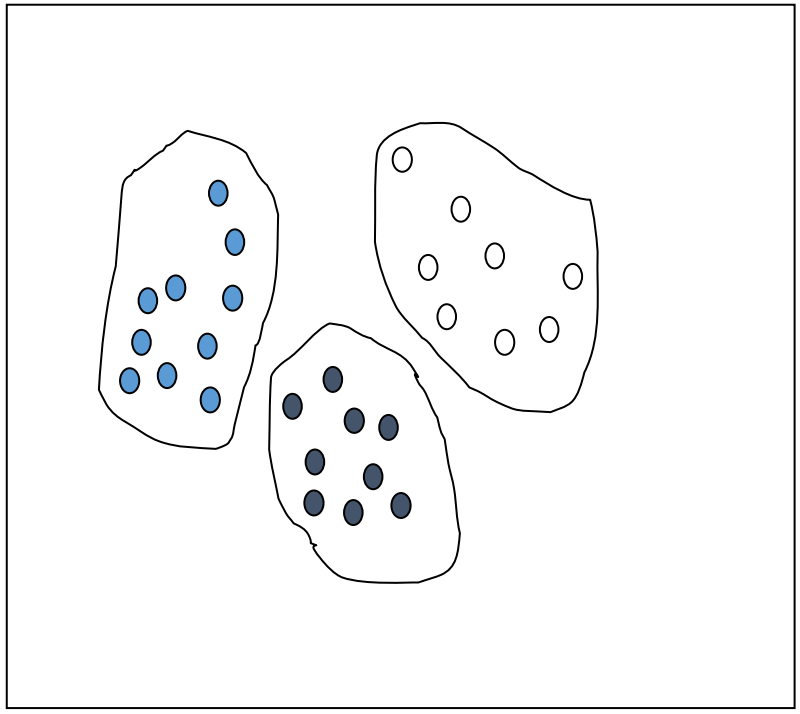
- Simple random sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - Once an object is selected, it is removed from the population
- Sampling with replacement
  - A selected object is not removed from the population
- Stratified sampling
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Sampling: With or without Replacement

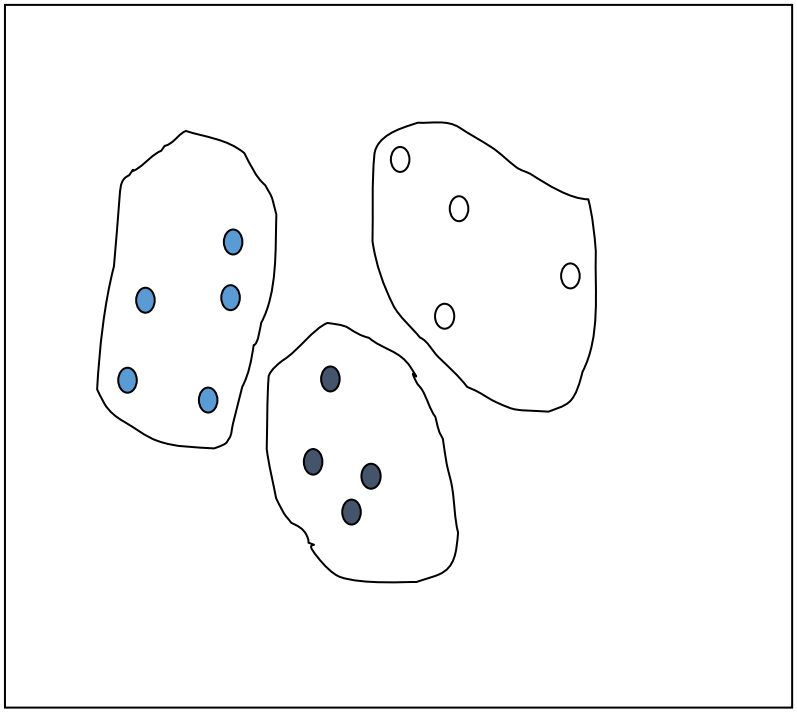


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Stratified Sampling

- Stratification is the process of dividing members of the population into homogeneous subgroups before sampling
- Suppose that in a company there are the following staff:
  - Male, full-time: 90
  - Male, part-time: 18
  - Female, full-time: 9
  - Female, part-time: 63
  - Total: 180
- We are asked to take a sample of 40 staff, stratified according to the above categories
- An easy way to calculate the percentage is to multiply each group size by the sample size and divide by the total population:
  - Male, full-time = 90 × (40 ÷ 180) = 20
  - Male, part-time = 18 × (40 ÷ 180) = 4
  - Female, full-time = 9 × (40 ÷ 180) = 2
  - Female, part-time = 63 × (40 ÷ 180) = 14

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 7 Discriminant Analysis, pp. 125-143

- Datasets:
  - SportSkill-Training.csv
  - SportSkill-Scoring.csv

- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!

3.3 Data Transformation and Data Discretization

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values
  - Each old value can be identified with one of the new values
- Methods:
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

Discretization

- Three types of attributes
  - Nominal —values from an unordered set, e.g., color, profession
  - Ordinal —values from an ordered set, e.g., military or academic rank
  - Numeric —real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

Typical methods: All the methods can be applied recursively

- Binning: Top-down split, unsupervised
- Histogram analysis: Top-down split, unsupervised
- Clustering analysis: Unsupervised, top-down split or bottom-up merge
- Decision-tree analysis: Supervised, top-down split
- Correlation (e.g., χ^2) analysis: Unsupervised, bottom-up merge

Simple Discretization: Binning

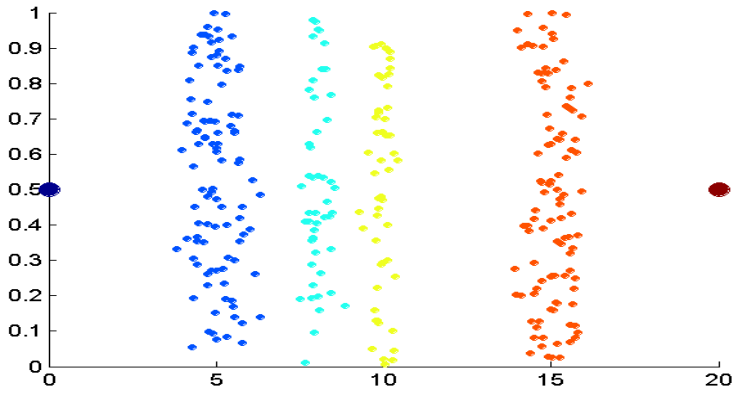
- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

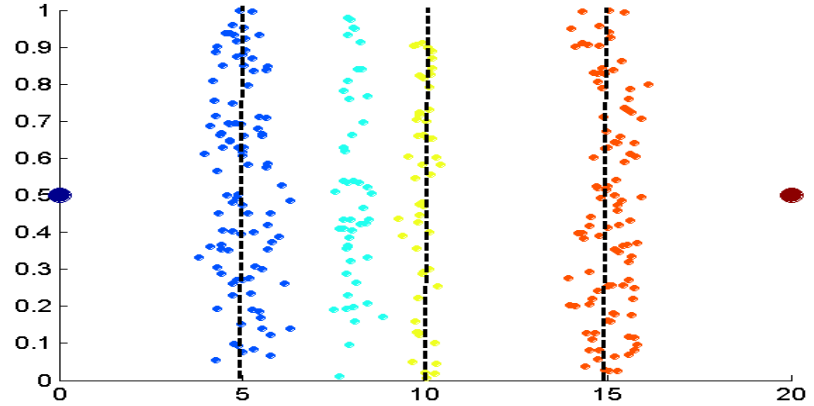
Binning Methods for Data Smoothing

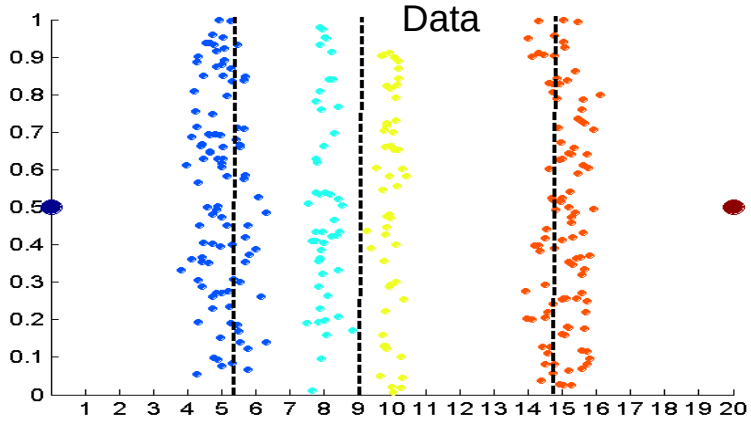
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

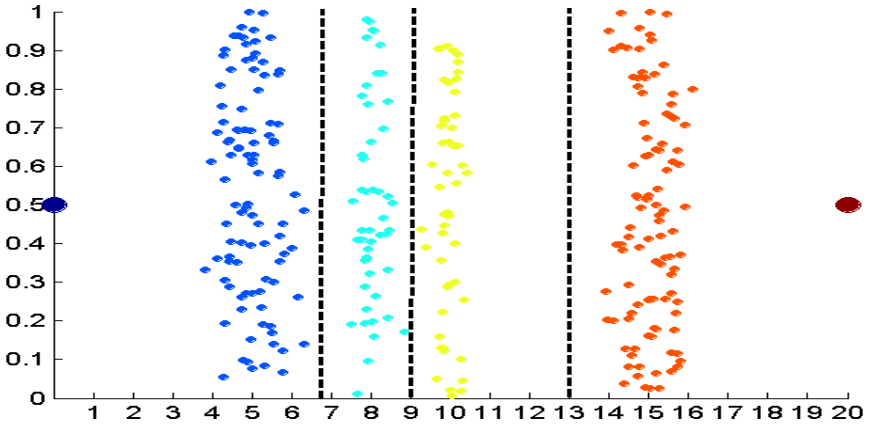
- Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Discretization Without Using Class Labels (Binning vs. Clustering)









Equal frequency (binning)

K-means clustering leads to better results

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using *entropy* to determine split point (discretization point)
  - Top-down, recursive split
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 5 (Naïve Bayes Classification I)
- Dataset: crx.data
- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!
- Bandingkan akurasi model apabila tidak menggunakan filter dan diskretisasi
- Bandingkan pula apabila digunakan feature selection (wrapper) dengan Backward Elimination

<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
SportSkill-Scoring (Ror
Transaksi (RomiSatria -
MissingValueData (Ro
Glass (RomiSatria - v1,
eReader-Training (Ron
eReader-Scoring (Rom
CRX (RomiSatria - v1, 2
Operators
discret
Cleansing (5)
Binning (5)
Discretize by
Discretize by
Discretize by
Discretize by
Discretize by
Extensions (4)
Weka (1)
Modeling (1)
Predictive
W-Reg
Series (3)
+ Get More Operators
Process
Process
Process
inp
Retrieve CRX
out
Filter Examples
exa
exa
ori
unm
Discretize
exa
exa
ori
pre
Validation
tra
mod
tra
ave
ave
res
res
A cross-validation evaluating a decision tree model.
Parameters
Process
logverbosity init
logfile
resultfile
random seed 2001
send mail never
Hide advanced parameters
Change compatibility (7.0.001)
Help
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
15:08
25-Feb-16

Hasil

	NB	NB+ Filter	NB+ Discretization	NB+ Filter+ Discretization	NB+ Filter+ Discretization + Backward Elimination
Accuracy				85.79	86.26
AUC					

3.4 Data Integration

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema Integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity Identification Problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and Resolving Data Value Conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases
  - Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Correlation Analysis (Numeric Data)

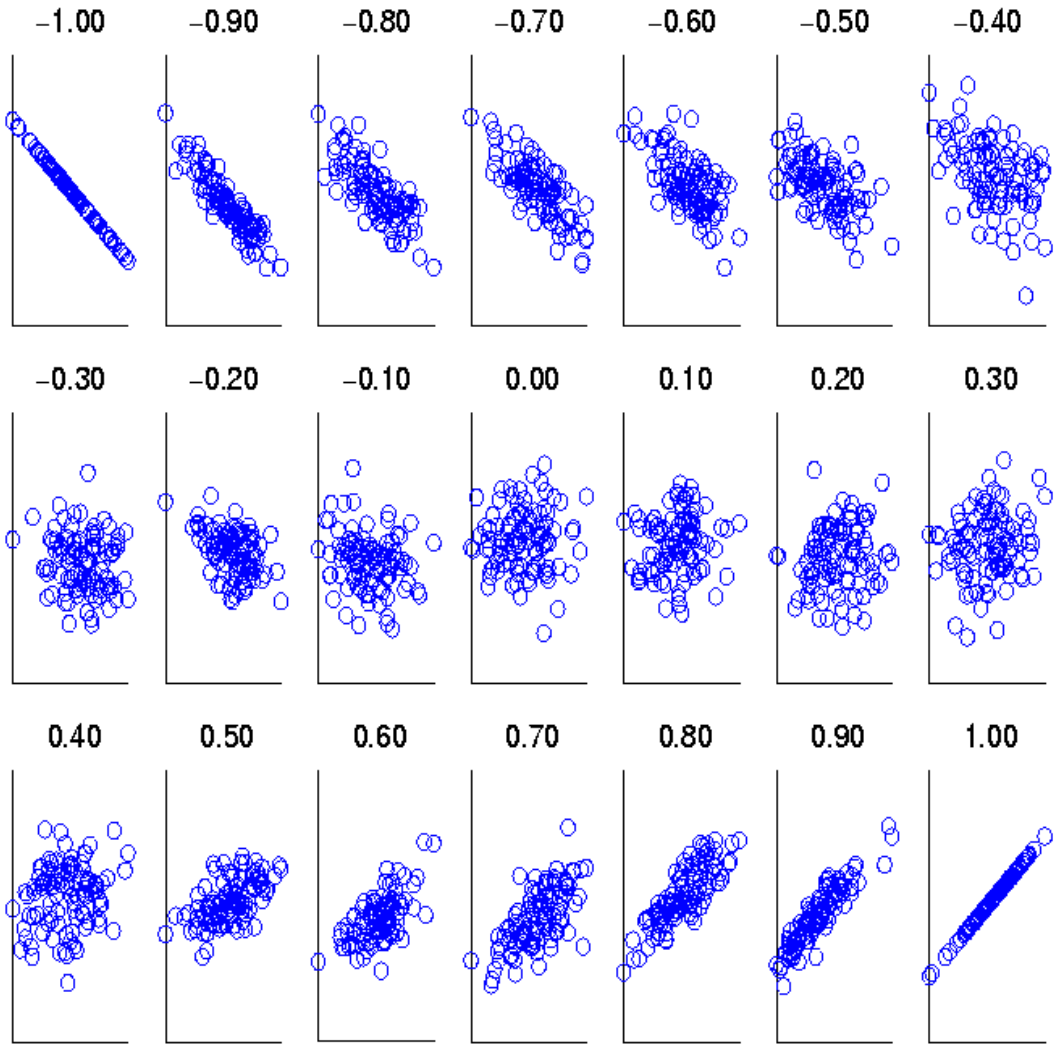
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Covariance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0

Rangkuman

1. Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability
2. Data cleaning: e.g. missing/noisy values, outliers
3. Data reduction
   - Dimensionality reduction
   - Numerosity reduction
4. Data transformation and data discretization
   - Normalization
5. Data integration from multiple sources:
   - Entity identification problem
   - Remove redundancies
   - Detect inconsistencies

Tugas Membuat Tulisan Ilmiah

- Buat tulisan ilmiah dari slide (ppt) yang sudah dibuat, dengan menggunakan template di *http://journal.ilmukomputer.org*
- Struktur Paper mengikuti format di bawah:
  1. Pendahuluan
     - Latar belakang masalah dan tujuan
  2. Penelitian Yang Berhubungan
     - Penelitian lain yang melakukan hal yang mirip dengan yang kita lakukan
  3. Metode Penelitian
     - Cara kita menganalisis data, jelaskan bahwa kita menggunakan CRISP-DM
  4. Hasil dan Pembahasan
     - 4.1 Business Understanding
     - 4.2 Data Understanding
     - 4.3 Data Preparation
     - 4.4 Modeling
     - 4.5 Evaluation
     - 4.6 Deployment
  5. Kesimpulan
     - Kesimpulan harus sesuai dengan tujuan
  6. Daftar Referensi
     - Masukan daftar referensi yang digunakan

Tugas Menyelesaikan Masalah Organisasi

- Analisis masalah dan kebutuhan yang ada di organisasi lingkungan sekitar anda
- Kumpulkan dan review dataset yang tersedia, dan hubungkan masalah dan kebutuhan tadi dengan data yang tersedia (analisis dari 5 peran data mining)
  - Bila memungkinkan pilih beberapa peran sekaligus untuk mengolah data tersebut, misalnya: lakukan association (analisis faktor), sekaligus estimation atau clustering
- Lakukan proses CRISP-DM untuk menyelesaikan masalah yang ada di organisasi sesuai dengan data yang didapatkan
  - Pada proses data preparation, lakukan data cleaning (replace missing value, replace, filter attribute) sehingga data siap dimodelkan
  - Lakukan juga komparasi algoritma dan *feature selection* untuk memilih pola dan model terbaik
  - Rangkumkan evaluasi dari pola/model/knowledge yang dihasilkan dan relasikan hasil evaluasi dengan deployment yang dilakukan
- Rangkumkan dalam bentuk slide dengan contoh studi kasus Sarah untuk membantu bidang marketing

4. Algoritma Data Mining

4.1 Algoritma Klasifikasi

4.2 Algoritma Klastering

4.3 Algoritma Asosiasi

4.4 Algoritma Estimasi dan Forecasting

4.1 Algoritma Klasifikasi

4.1.1 Decision Tree

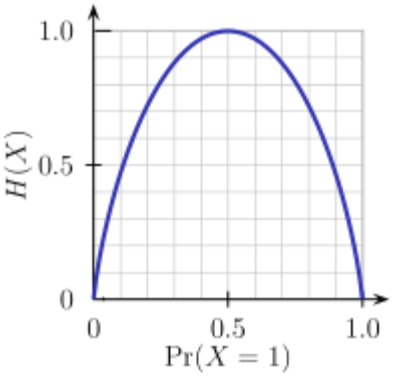
Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  1. Tree is constructed in a top-down recursive divide-and-conquer manner
  2. At start, all the training examples are at the root
  3. Attributes are categorical (if continuous-valued, they are discretized in advance)
  4. Examples are partitioned recursively based on selected attributes
  5. Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain, gain ratio, gini index)

- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Brief Review of Entropy

- Entropy (Information Theory)
  - A measure of uncertainty associated with a random variable
  - Calculation: For a discrete random variable *Y* taking *m* distinct values $\{y_1, \ldots, y_m\}$,
    - $H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)$, where $p_i = P(Y = y_i)$
  - Interpretation:
    - Higher entropy => higher uncertainty
    - Lower entropy => lower uncertainty
- Conditional Entropy
  - $H(Y|X) = \sum_x p(x) H(Y|X = x)$



m = 2

Attribute Selection Measure: Information Gain (ID3)

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_i, D|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

Computing Information-Gain for Continuous-Valued Attributes

- Let attribute *A* be a continuous-valued attribute
- Must determine the best split point for *A*
  - Sort the value *A* in increasing order
  - Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*
    - $(a_i+a_{i+1})/2$ is the midpoint between the values of a_i and a_{i+1}
  - The point with the *minimum expected information requirement* for *A* is selected as the split-point for *A*
- Split:
  - D_1 is the set of tuples in *D* satisfying *A* ≤ split-point, and D_2 is the set of tuples in *D* satisfying *A* > split-point

Tahapan Algoritma Decision Tree (ID3)

1. Siapkan data training
2. Pilih atribut sebagai akar

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

3. Buat cabang untuk tiap-tiap nilai
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yg sama

1. Siapkan data training

No	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Cloudy	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	Yes
7	Cloudy	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Cloudy	Mild	High	TRUE	Yes
13	Cloudy	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

2. Pilih atribut sebagai akar

- Untuk memilih atribut akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Untuk mendapatkan nilai Gain, harus ditentukan terlebih dahulu nilai Entropy

- Rumus Entropy: $$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

  - *S* = Himpunan Kasus
  - *n* = Jumlah Partisi *S*
  - *pi* = Proporsi dari *Si* terhadap *S*

- Rumus Gain: $$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

  - *S* = Himpunan Kasus
  - *A* = Atribut
  - *n* = Jumlah Partisi Atribut *A*
  - | Si | = Jumlah Kasus pada partisi ke-*i*
  - | S | = Jumlah Kasus dalam *S*

Perhitungan Entropy dan Gain Akar

NODE			Jml Kasus (S)	Tidak (S_1)	Ya (S_2)	Entropy	Gain
1	TOTAL						
	OUTLOOK						
		CLOUDY					
		RAINY					
		SUNNY					
	TEMPERATURE						
		COOL					
		HOT					
		MILD					
	HUMIDITY						
		HIGH					
		NORMAL					
	WINDY						
		FALSE					
		TRUE					

Penghitungan Entropy Akar

- Entropy Total

$$Entropy(Total) = (-\frac{4}{14} * \log_2(\frac{4}{14})) + (-\frac{10}{14} * \log_2(\frac{10}{14}))$$

$$Entropy(Total) = 0.863120569$$

- Entropy (Outlook)

$$Entropy(Cloudy) = (-\frac{0}{4} * \log_2(\frac{0}{4})) + (-\frac{4}{4} * \log_2(\frac{4}{4})) = 0.000000000$$

$$Entropy(Rainy) = (-\frac{1}{5} * \log_2(\frac{1}{5})) + (-\frac{4}{5} * \log_2(\frac{4}{5})) = 0.721928095$$

$$Entropy(Sunny) = (-\frac{3}{5} * \log_2(\frac{3}{5})) + (-\frac{2}{5} * \log_2(\frac{2}{5})) = 0.970950594$$

- Entropy (Temperature)

$$Entropy(Cool) = (-\frac{0}{4} * \log_2(\frac{0}{4})) + (-\frac{4}{4} * \log_2(\frac{4}{4})) = 0.000000000$$

$$Entropy(Hot) = (-\frac{2}{4} * \log_2(\frac{2}{4})) + (-\frac{2}{4} * \log_2(\frac{2}{4})) = 1.000000000$$

$$Entropy(Mild) = (-\frac{2}{6} * \log_2(\frac{2}{6})) + (-\frac{4}{6} * \log_2(\frac{4}{6})) = 0.918295834$$

- Entropy (Humidity)

$$Entropy(High) = (-\frac{4}{7} * \log_2(\frac{4}{7})) + (-\frac{3}{7} * \log_2(\frac{3}{7})) = 0.985228136$$

$$Entropy(Normal) = (-\frac{0}{7} * \log_2(\frac{0}{7})) + (-\frac{7}{7} * \log_2(\frac{7}{7})) = 0.000000000$$

- Entropy (Windy)

$$Entropy(False) = (-\frac{2}{8} * \log_2(\frac{2}{8})) + (-\frac{6}{8} * \log_2(\frac{6}{8})) = 0.811278124$$

$$Entropy(True) = (-\frac{4}{6} * \log_2(\frac{4}{6})) + (-\frac{2}{6} * \log_2(\frac{2}{6})) = 0.918295834$$

Penghitungan Entropy Akar

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1	**TOTAL**		14	10	4	0,86312	
	OUTLOOK						
		CLOUDY	4	4	0	0	
		RAINY	5	4	1	0,72193	
		SUNNY	5	2	3	0,97095	
	TEMPERATURE						
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,91830	
	HUMADITY						
		HIGH	7	4	3	0,98523	
		NORMAL	7	7	0	0	
	WINDY						
		FALSE	8	2	6	0,81128	
		TRUE	6	4	2	0,91830	

Penghitungan Gain Akar

$$Gain(Total, Outlook) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Outlook_i|}{|Total|} * Entropy\ (Outlook_i)$$

$$Gain(Total, Outlook) = 0.863120569 - \left(\left(\frac{4}{14} * 0.000000000\right) + \left(\frac{5}{14} * 0.721928095\right) + \left(\frac{5}{14} * 0.970950594\right)\right)$$

$$Gain(Total, Outlook) = 0.258521037$$

$$Gain(Total, Temperature) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Temperature_i|}{|Total|} * Entropy\ (Temperature_i)$$

$$Gain(Total, Temperature) = 0.863120569 - \left(\left(\frac{4}{14} * 0.000000000\right) + \left(\frac{4}{14} * 1.000000000\right) + \left(\frac{6}{14} * 0.918295834\right)\right)$$

$$Gain(Total, Temperature) = 0.183850925$$

$$Gain(Total, Humidity) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Humidity_i|}{|Total|} * Entropy\ (Humidity_i)$$

$$Gain(Total, Humidity) = 0.863120569 - \left(\left(\frac{7}{14} * 0.985228136\right) + \left(\frac{7}{14} * 0.000000000\right)\right)$$

$$Gain(Total, Humidity) = 0.370506501$$

$$Gain(Total, Windy) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Windy_i|}{|Total|} * Entropy\ (Windy_i)$$

$$Gain(Total, Windy) = 0.863120569 - \left(\left(\frac{8}{14} * 0.811278124\right) + \left(\frac{6}{14} * 0.918295834\right)\right)$$

$$Gain(Total, Windy) = 0.005977711$$

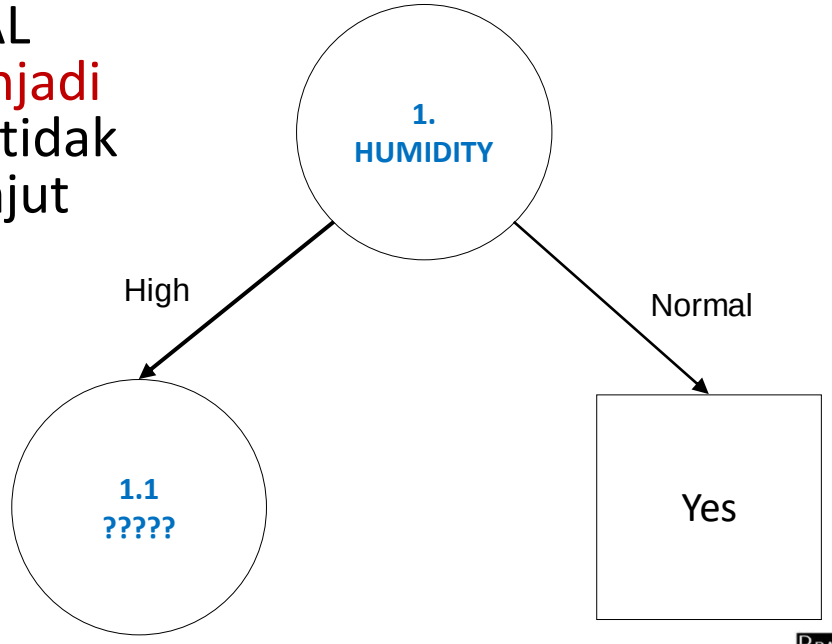
Penghitungan Gain Akar

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1	**TOTAL**		14	10	4	0,86312	
	OUTLOOK						0,25852
		CLOUDY	4	4	0	0	
		RAINY	5	4	1	0,72193	
		SUNNY	5	2	3	0,97095	
	TEMPERATURE						0,18385
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0,91830	
	HUMADITY						0,37051
		HIGH	7	4	3	0,98523	
		NORMAL	7	7	0	0	
	WINDY						0,00598
		FALSE	8	2	6	0,81128	
		TRUE	6	4	2	0,91830	

Gain Tertinggi Sebagai Akar

- Dari hasil pada Node 1, dapat diketahui bahwa atribut dengan Gain tertinggi adalah HUMIDITY yaitu sebesar 0.37051
  - Dengan demikian HUMIDITY dapat menjadi node akar

- Ada 2 nilai atribut dari HUMIDITY yaitu HIGH dan NORMAL. Dari kedua nilai atribut tersebut, nilai atribut NORMAL sudah mengklasifikasikan kasus menjadi 1 yaitu keputusan-nya Yes, sehingga tidak perlu dilakukan perhitungan lebih lanjut
  - Tetapi untuk nilai atribut HIGH masih perlu dilakukan perhitungan lagi



2. Buat cabang untuk tiap-tiap nilai

- Untuk memudahkan, dataset di filter dengan mengambil data yang memiliki kelembaban HUMADITY=HIGH untuk membuat table Node 1.1

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Cloudy	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Sunny	Mild	High	FALSE	No
Cloudy	Mild	High	TRUE	Yes
Rainy	Mild	High	TRUE	No

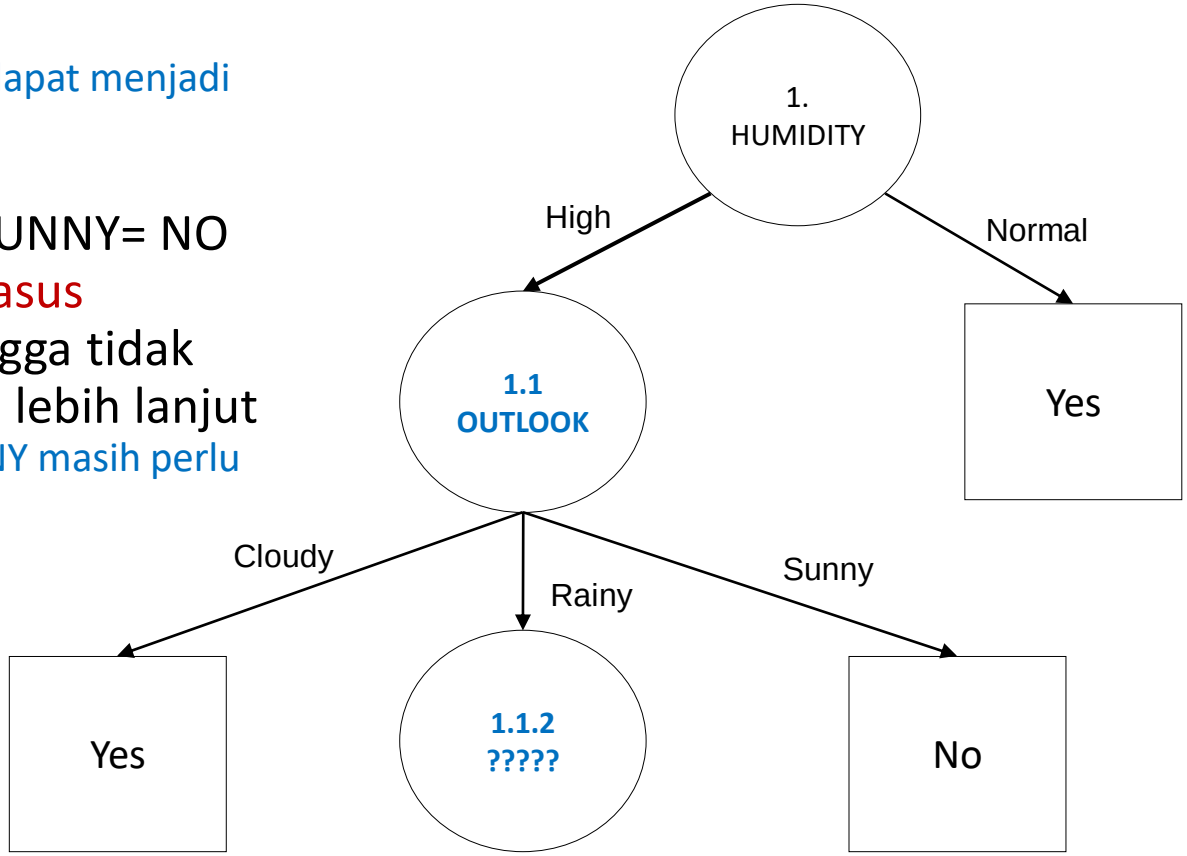
Perhitungan Entropi Dan Gain Cabang

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1.1	**HUMADITY**		7	3	4	0,98523	
	OUTLOOK						0,69951
		CLOUDY	2	2	0	0	
		RAINY	2	1	1	1	
		SUNNY	3	0	3	0	
	TEMPERATURE						0,02024
		COOL	0	0	0	0	
		HOT	3	1	2	0,91830	
		MILD	4	2	2	1	
	WINDY						0,02024
		FALSE	4	2	2	1	
		TRUE	3	1	2	0,91830	

Gain Tertinggi Sebagai Node 1.1

- Dari hasil pada Tabel Node 1.1, dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK yaitu sebesar 0.69951
  - Dengan demikian OUTLOOK dapat menjadi node kedua

- Artibut CLOUDY = YES dan SUNNY= NO sudah mengklasifikasikan kasus menjadi 1 keputusan, sehingga tidak perlu dilakukan perhitungan lebih lanjut
  - Tetapi untuk nilai atribut RAINY masih perlu dilakukan perhitungan lagi



3. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yg sama

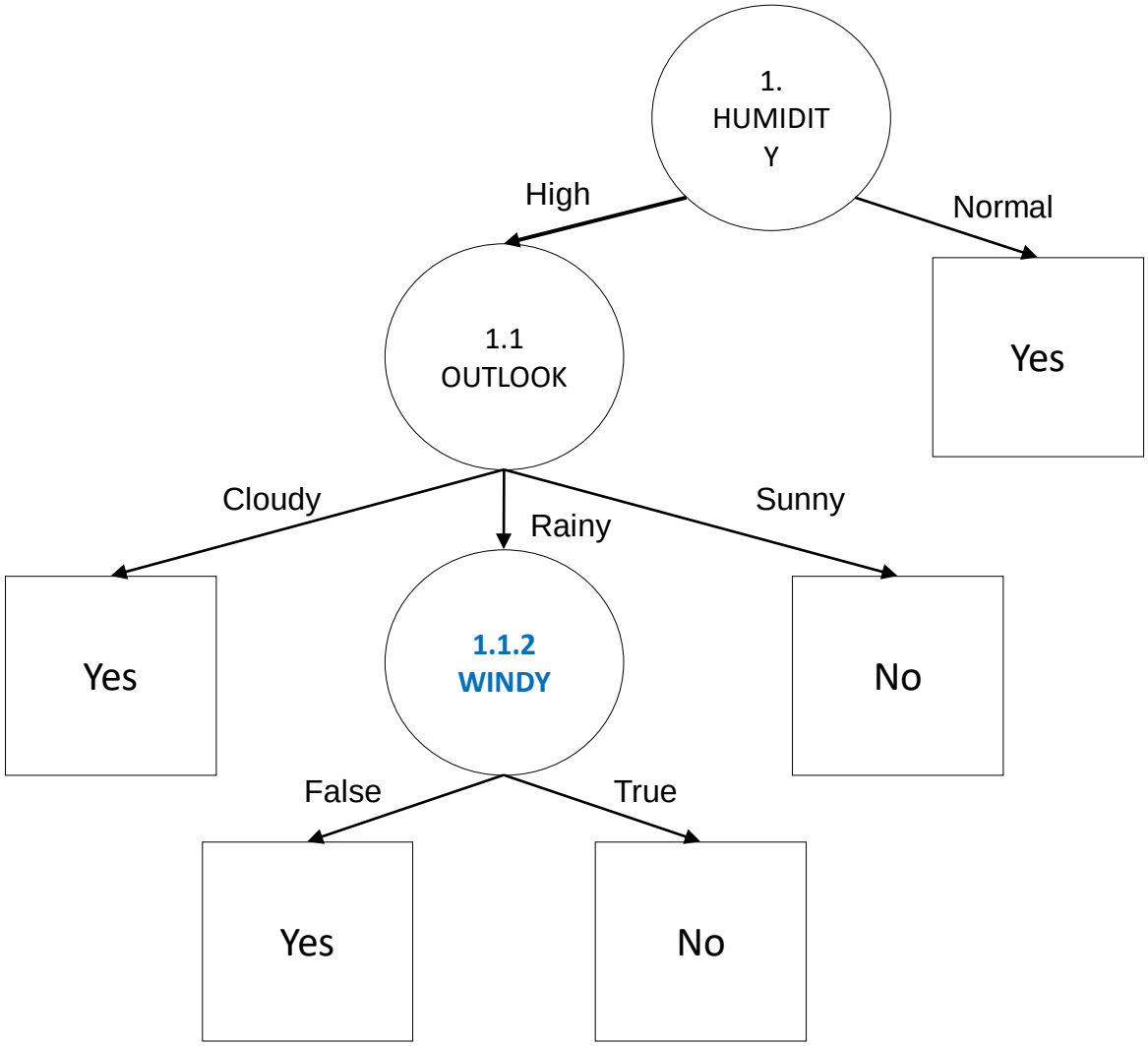
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Rainy	Mild	High	FALSE	Yes
Rainy	Mild	High	TRUE	No

NODE	ATRIBUT		JML KASUS (S)	YA (Si)	TIDAK (Si)	ENTROPY	GAIN
1.2	**HUMADITY HIGH & OUTLOOK RAINY**		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	1	0	0	
		TRUE	1	0	1	0	

Gain Tertinggi Sebagai Node 1.1.2

- Dari tabel, Gain Tertinggi adalah WINDY dan menjadi node cabang dari atribut RAINY

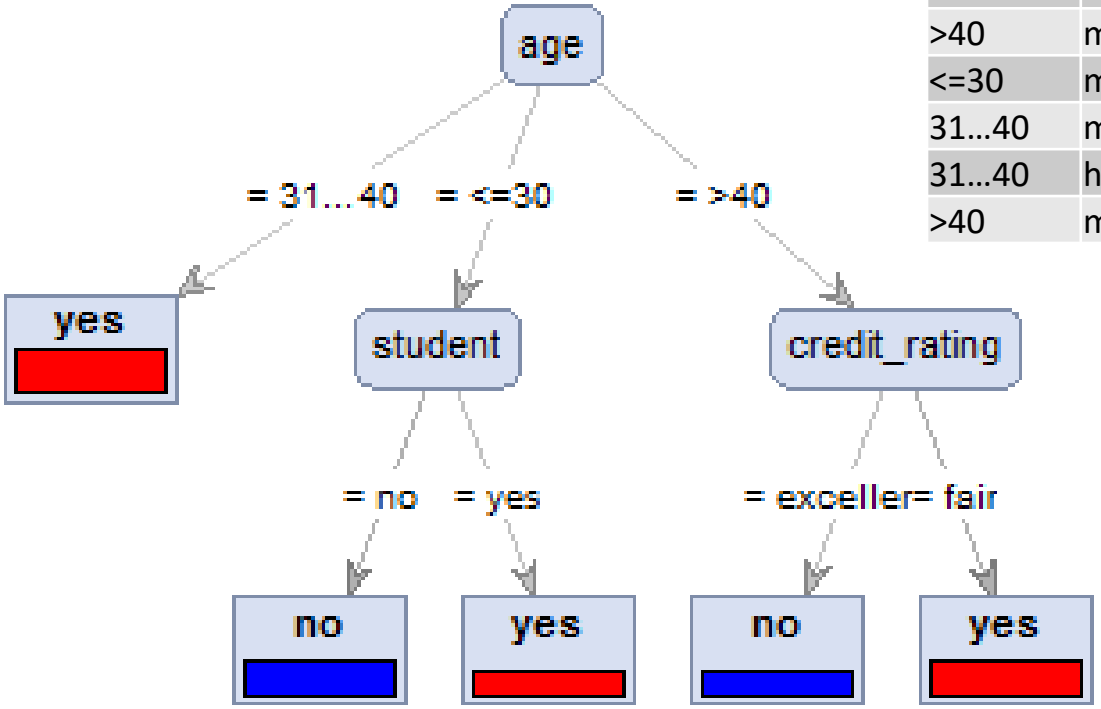
- Karena semua kasus sudah masuk dalam kelas
  - Jadi, pohon keputusan pada Gambar merupakan pohon keputusan terakhir yang terbentuk



Decision Tree Induction: An Example

- Training data set: Buys_computer

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Gain Ratio for Attribute Selection (C4.5)

- Information gain measure is biased towards attributes with a large number of values
- C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

  - GainRatio(A) = Gain(A)/SplitInfo(A)

- Ex. $SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) = 1.557$

  - gain_ratio(income) = 0.029/1.557 = 0.019

- The attribute with the maximum gain ratio is selected as the splitting attribute

Gini Index (CART)

- If a data set *D* contains examples from *n* classes, gini index, *gini*(*D*) is defined as

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class *j* in *D*

- If a data set *D* is split on *A* into two subsets D_1 and D_2, the gini index *gini*(*D*) is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node (*need to enumerate all the possible splitting points for each attribute*)

Computation of Gini Index

- Ex. D has 9 tuples in buys_computer = “yes” and 5 in “no”

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Suppose the attribute income partitions *D* into 10 in D_1: *{low, medium}* and 4 in D_2

$$gini_{income \in \{low, medium\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

$Gini_{\{low,high\}}$ is 0.458; $Gini_{\{medium,high\}}$ is 0.450. Thus, split on the {low,medium} (and {high}) since it has the lowest Gini index

- All attributes are assumed continuous-valued
- May need other tools, e.g., clustering, to get the possible split values
- Can be modified for categorical attributes

Comparing Attribute Selection Measures

The three measures, in general, return good results but

- Information gain:
  - biased towards multivalued attributes
- Gain ratio:
  - tends to prefer unbalanced splits in which one partition is much smaller than the others
- Gini index:
  - biased to multivalued attributes
  - has difficulty when # of classes is large
  - tends to favor tests that result in equal-sized partitions and purity in both partitions

Other Attribute Selection Measures

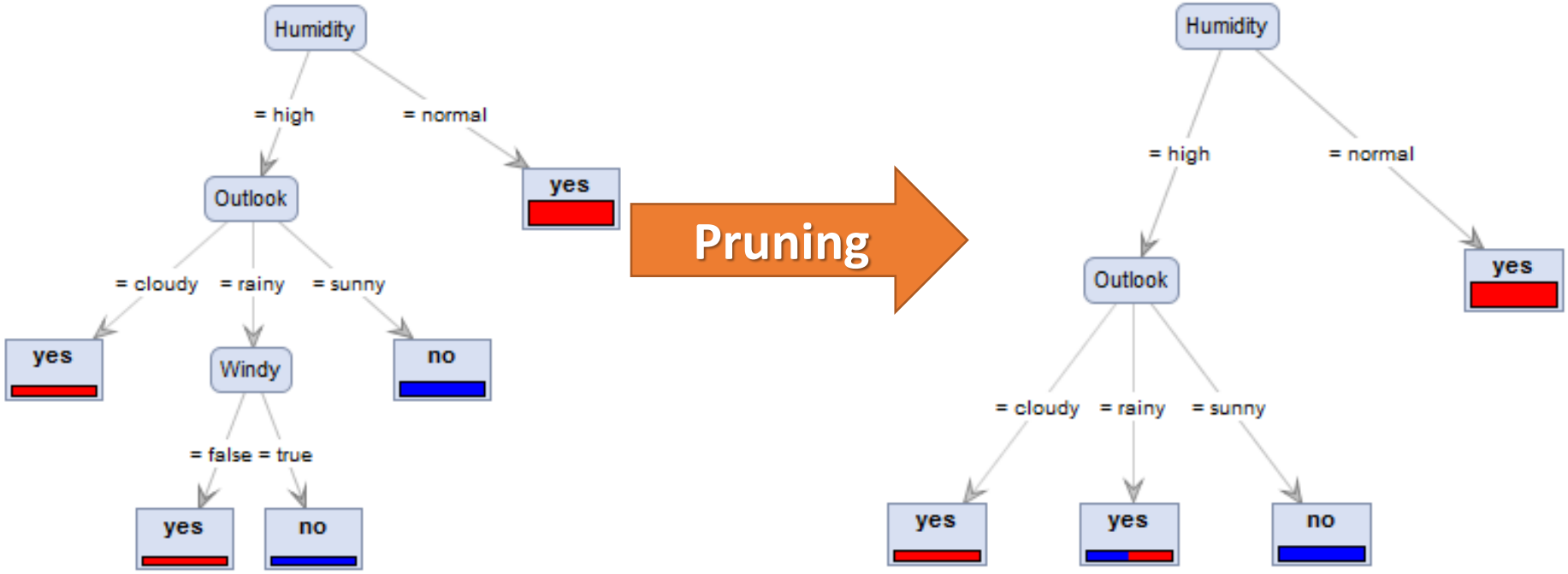
- CHAID: a popular decision tree algorithm, measure based on χ^2 test for independence
- C-SEP: performs better than info. gain and gini index in certain cases
- G-statistic: has a close approximation to χ^2 distribution
- MDL (Minimal Description Length) principle (i.e., the simplest solution is preferred):
  - The best tree as the one that requires the fewest # of bits to both (1) encode the tree, and (2) encode the exceptions to the tree
- Multivariate splits (partition based on multiple variable combinations)
  - CART: finds multivariate splits based on a linear comb. of attrs.
- Which attribute selection measure is the best?
  - Most give good results, none is significantly superior than others

Overfitting and Tree Pruning

- Overfitting: An induced tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Poor accuracy for unseen samples

- Two approaches to avoid overfitting
  1. Prepruning: *Halt tree construction early* - do not split a node if this would result in the goodness measure falling below a threshold
     - Difficult to choose an appropriate threshold
  2. Postpruning: *Remove branches from a "fully grown" tree* -get a sequence of progressively pruned trees
     - Use a set of data different from the training data to decide which is the "best pruned tree"

Row No.	Play	Outlook	Temperature	Humidity	Windy
1	no	sunny	hot	high	false
2	no	sunny	hot	high	true
3	yes	cloudy	hot	high	false
4	yes	rainy	mild	high	false
5	yes	rainy	cool	normal	false
6	yes	rainy	cool	normal	true
7	yes	cloudy	cool	normal	true
8	no	sunny	mild	high	false
9	yes	sunny	cool	normal	false
10	yes	rainy	mild	normal	false
11	yes	sunny	mild	normal	true
12	yes	cloudy	mild	high	true
13	yes	cloudy	hot	normal	false
14	no	rainy	mild	high	true



Why is decision tree induction popular?

- Relatively faster learning speed (than other classification methods)
- Convertible to simple and easy to understand classification rules
- Can use SQL queries for accessing databases
- Comparable classification accuracy with other methods

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 10 (Decision Tree), p 195-217

- Datasets:
  - eReaderAdoption-Training.csv
  - eReaderAdoption-Scoring.csv

- Analisis peran metode prunning pada decision tree dan hubungannya dengan nilai confidence
- Analisis jenis decision tree apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut

1 Business Understanding

Motivation:

- Richard works for a large online retailer
- His company is launching a tablet soon, and he want to maximize the effectiveness of his marketing
- They have a large number of customers, many of whom have purchased digital devices and other services previously
- Richard has noticed that certain types of people were anxious to get new devices as soon as they became available, while other folks seemed content to wait to buy their electronic gadgets later
- He's wondering what makes some people motivated to buy something as soon as it comes out, while others are less driven to have the product right away

Objectives:

- To mine the customers' consumer behaviors on the web site, in order to figure out which customers will buy the new tablet early, which ones will buy next, and which ones will buy later on

Latihan

- Lakukan training pada data eReader Adoption (eReader-Training.csv) dengan menggunakan DT dengan 3 alternative criterion (Gain Ratio, Information Gain dan Gini Index)
- Ujicoba masing-masing split criterion baik menggunakan prunning atau tidak
- Lakukan pengujian dengan menggunakan 10-fold X Validation
- Dari model terbaik, tentukan faktor (atribut) apa saja yang berpengaruh pada tingkat adopsi eReader

	DTGR	DTIG	DTGI	DTGR+Pr	DTIG+Pr	DTGI+Pr
Accuracy				58.39	51.01	31.01

Latihan

- Lakukan feature selection dengan Forward Selection untuk ketiga algoritma di atas
- Lakukan pengujian dengan menggunakan 10-fold X Validation
- Dari model terbaik, tentukan faktor (atribut) apa saja yang berpengaruh pada tingkat adopsi eReader

	DTGR	DTIG	DTGI	DTGR+FS	DTIG+FS	DTGI+FS
Accuracy	58.39	51.01	31.01	61.41	56.73	31.01

Data

Charts

Annotations

attribute	weight ↓
Age	1
Website_Activity	1
Browsed_Electronics_12Mo	1
Bought_Digital_Media_18Mo	1
Bought_Digital_Books	1
Payment_Method	1
Gender	0
Marital_Status	0
Bought_Electronics_12Mo	0

4.1.2 Bayesian Classification

Bayesian Classification: Why?

- A statistical classifier: performs probabilistic prediction, i.e., predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

Bayes' Theorem: Basics

- Total probability Theorem: $$P(B) = \sum_{i=1}^{M} P(B|A_i)P(A_i)$$

- Bayes' Theorem: $$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

  - Let **X** be a data sample ("*evidence*"): class label is unknown
  - Let H be a *hypothesis* that X belongs to class C
  - Classification is to determine P(H|**X**), (i.e., *posteriori probability):* the probability that the hypothesis holds given the observed data sample **X**
  - P(H) (*prior probability*): the initial probability
    - E.g., **X** will buy computer, regardless of age, income, …
  - P(**X**): probability that sample data is observed
  - P(**X**|H) (likelihood): the probability of observing the sample **X**, given that the hypothesis holds
    - E.g., Given that **X** will buy computer, the prob. that X is 31..40, medium income

Prediction Based on Bayes' Theorem

- Given training data **X**, *posteriori probability of a hypothesis* H, P(H|**X**), follows the Bayes' theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- Informally, this can be viewed as

  posteriori = likelihood x prior/evidence

- Predicts **X** belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes

- Practical difficulty: It requires initial knowledge of many probabilities, involving significant computational cost

Classification is to Derive the Maximum Posteriori

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x_1, x_2, …, x_n)
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

- This greatly reduces the computation cost: Only counts the class distribution
- If A_k is categorical, $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i,D}|$ (# of tuples of C_i in D)
- If A_k is continous-valued, $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and $P(x_k|C_i)$ is

$$P(\mathbf{X}|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

Naïve Bayes Classifier: Training Dataset

Class:

C_1:buys_computer = 'yes'

C_2:buys_computer = 'no'

Data to be classified:

X = (age <=30,
income = medium,
student = yes,
credit_rating = fair)

X → buy computer?

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	No	excellent	yes
31…40	high	Yes	fair	yes
>40	medium	No	excellent	no

Naïve Bayes Classifier: An Example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643

  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class

  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222

  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6

  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444

  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4

  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667

  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2

  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667

  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

  $P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

  P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

  $P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028

  P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

Tahapan Algoritma Naïve Bayes

1. Baca Data Training
2. Hitung jumlah class
3. Hitung jumlah kasus yang sama dengan class yang sama
4. Kalikan semua nilai hasil sesuai dengan data X yang dicari class-nya

1. Baca Data Training

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

Teorema Bayes

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})} = P(\mathbf{X} \mid H) \times P(H) / P(\mathbf{X})$$

- ***X*** → Data dengan class yang belum diketahui
- ***H*** → Hipotesis data *X* yang merupakan suatu class yang lebih spesifik
- ***P (H|X)*** → Probabilitas hipotesis *H* berdasarkan kondisi *X* (*posteriori probability*)
- ***P (H)*** → Probabilitas hipotesis *H* (*prior probability*)
- ***P (X|H)*** → Probabilitas *X* berdasarkan kondisi pada hipotesis *H*
- ***P (X)*** → Probabilitas *X*

2. Hitung jumlah class/label

- Terdapat 2 class dari data training tersebut, yaitu:
  - C1 (Class 1) →Play = yes → 9 record
  - C2 (Class 2) →Play = no → 5 record
  - Total = 14 record

- Maka:
  - P (C1) = 9/14 = 0.642857143
  - P (C2) = 5/14 = 0.357142857

- Pertanyaan:
  - Data *X = (outlook=rainy, temperature=cool, humidity=high, windy=true)*
  - Main golf atau tidak?

3. Hitung jumlah kasus yang sama dengan class yang sama

- Untuk *P(Ci)* yaitu *P(C1)* dan *P(C2)* sudah diketahui hasilnya di langkah sebelumnya.
- Selanjutnya Hitung *P(X|Ci)* untuk *i* = 1 dan 2
  - *P(outlook="sunny"|play="yes")*=2/9=0.222222222
  - *P(outlook="sunny"|play="no")*=3/5=0.6
  - *P(outlook="overcast"|play="yes")*=4/9=0.444444444
  - *P(outlook="overcast"|play="no")*=0/5=0
  - *P(outlook="rainy"|play="yes")*=3/9=0.333333333
  - *P(outlook="rainy"|play="no")*=2/5=0.4

3. Hitung jumlah kasus yang sama dengan class yang sama

- Jika semua atribut dihitung, maka didapat hasil akhirnya seperti berikut ini:

Atribute	Parameter	No	Yes
Outlook	value=sunny	0.6	0.2222222222222222
Outlook	value=cloudy	0.0	0.4444444444444444
Outlook	value=rainy	0.4	0.3333333333333333
Temperature	value=hot	0.4	0.2222222222222222
Temperature	value=mild	0.4	0.4444444444444444
Temperature	value=cool	0.2	0.3333333333333333
Humidity	value=high	0.8	0.3333333333333333
Humidity	value=normal	0.2	0.6666666666666666
Windy	value=false	0.4	0.6666666666666666
Windy	value=true	0.6	0.3333333333333333

4. Kalikan semua nilai hasil sesuai dengan data X yang dicari class-nya

- Pertanyaan:
  - Data X = (outlook=rainy, temperature=cool, humidity=high, windy=true)
  - Main Golf atau tidak?
- Kalikan semua nilai hasil dari data X
  - P(X|play="yes") = 0.333333333* 0.333333333* 0.333333333*0.333333333 = 0.012345679
  - P(X|play="no") = 0.4*0.2*0.8*0.6=0.0384
  - P(X|play="yes")*P(C1) = 0.012345679*0.642857143 = 0.007936508
  - P(X|play="no")*P(C2) = 0.0384*0.357142857 = **0.013714286**
- Nilai "no" lebih besar dari nilai "yes" maka class dari data X tersebut adalah "**No**"

Avoiding the Zero-Probability Problem

- Naïve Bayesian prediction requires each conditional prob. be non-zero. Otherwise, the predicted prob. will be zero

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Use Laplacian correction (or Laplacian estimator)
  - *Adding 1 to each case*

    Prob(income = low) = 1/1003

    Prob(income = medium) = 991/1003

    Prob(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts

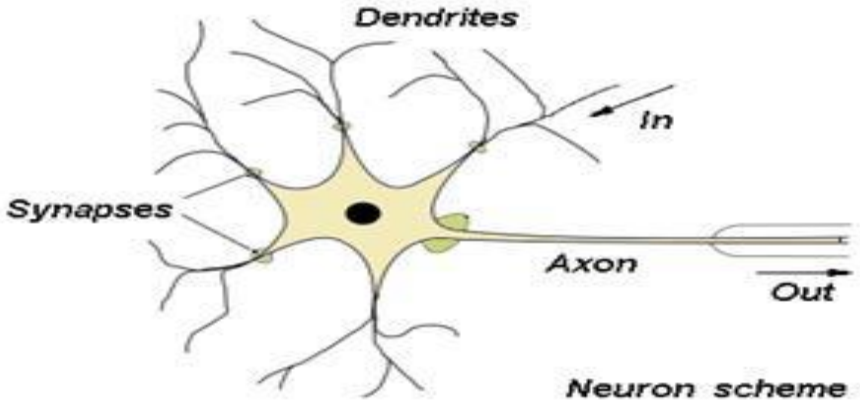
Naïve Bayes Classifier: Comments

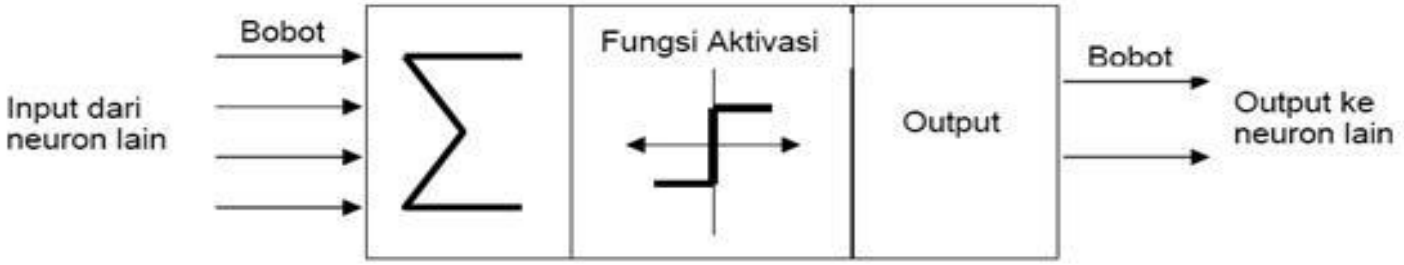
- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables, e.g.:
    - Hospitals Patients Profile: age, family history, etc.
    - Symptoms: fever, cough etc.,
    - Disease: lung cancer, diabetes, etc.
  - Dependencies among these cannot be modeled by Naïve Bayes Classifier
- How to deal with these dependencies? Bayesian Belief Networks

4.1.3 Neural Network

Neural Network

- Neural Network adalah suatu model yang dibuat untuk meniru fungsi belajar yang dimiliki otak manusia atau jaringan dari sekelompok unit pemroses kecil yang dimodelkan berdasarkan jaringan saraf manusia





Neural Network

- Model Perceptron adalah model jaringan yang terdiri dari beberapa unit masukan (ditambah dengan sebuah bias), dan memiliki sebuah unit keluaran
- Fungsi aktivasi bukan hanya merupakan fungsi biner (0,1) melainkan bipolar (1,0,-1)
- Untuk suatu harga *threshold* θ yang ditentukan:

$$F(net) = \begin{cases} 1 & \text{Jika net} > \theta \\ 0 & \text{Jika} -\theta \leq net \leq \theta \\ -1 & \text{Jika net} < -\theta \end{cases}$$

Fungsi Aktivasi

Macam fungsi aktivasi yang dipakai untuk mengaktifkan *net* diberbagai jenis neural network:

1. Aktivasi linear, Rumus: $y = sign(v) = v$

2. Aktivasi step, Rumus:

$$y = \text{sign}(v) = \begin{cases} 1 & \text{Jika } v \geq T \\ -1 & \text{Jika } v < T \end{cases}$$

3. Aktivasi sigmoid biner, Rumus:

$$y = \frac{1}{1 + e^{-v}}$$

4. Aktivasi sigmoid bipolar, Rumus:

$$y = \frac{2}{1 + e^{-v}} - 1$$

Tahapan Algoritma Perceptron

1. Inisialisasi semua bobot dan bias (umumnya *wi = b = 0*)
2. Selama ada element vektor masukan yang respon unit keluarannya tidak sama dengan target, lakukan:
   2.1 Set aktivasi unit masukan *xi = Si (i = 1,...,n)*
   2.2 Hitung respon unit keluaran: $net = \sum_i xiwi + b$

$$F(net) = \begin{cases} 1 & \text{Jika net} > \theta \\ 0 & \text{Jika} -\theta \leq net \leq \theta \\ -1 & \text{Jika net} < -\theta \end{cases}$$

   2.3 Perbaiki bobot pola yang mengadung kesalahan menurut persamaan:

   Wi (baru) = wi (lama) + Δw (i = 1,...,n) dengan $\Delta w = \alpha\, t\, xi$
   B (baru) = b(lama) + Δ b dengan $\Delta b = \alpha\, t$

   Dimana: α = Laju pembelajaran (Learning rate) yang ditentukan
   θ = Threshold yang ditentukan
   t = Target

   2.4 Ulangi iterasi sampai perubahan bobot (*Δwn = 0*) tidak ada

Studi Kasus

- Diketahui sebuah dataset kelulusan berdasarkan IPK untuk program S1:

Status	IPK	Semester
Lulus	2.9	1
Tidak Lulus	2.8	3
Tidak Lulus	2.3	5
Tidak lulus	2.7	6

- Jika ada mahasiswa IPK 2.85 dan masih semester 1, maka masuk ke kedalam manakah status tersebut ?

1: Inisialisasi Bobot

- Inisialisasi Bobot dan bias awal: b = 0 dan bias = 1

t	X1	X2
1	2,9	1
-1	2.8	3
-1	2.3	5
-1	2,7	6

2.1: Set aktivasi unit masukan

- Treshold (batasan), $\theta = 0$, yang artinya :

$$F(net) = \begin{cases} 1 & \text{Jika } net > 0 \\ 0 & \text{Jika } net = 0 \\ -1 & \text{Jika } net < 0 \end{cases}$$

2.2 - 2.3 Hitung Respon dan Perbaiki Bobot

- Hitung Response Keluaran iterasi 1
- Perbaiki bobot pola yang mengandung kesalahan

MASUKAN			TARGET		y=	PERUBAHAN BOBOT			BOBOT BARU		
X1	X2	1	t	NET	f(NET)	ΔW1	ΔW2	Δb	W1	W2	b
INISIALISASI									0	0	0
2,9	1	1	1	0	0	2,9	1	1	2,9	7	1
2,8	3	1	-1	8,12	1	-2,8	-3	-1	0,1	4	0
2,3	5	1	-1	0,23	1	-2,3	-5	-1	-2,2	-1	-1
2,7	6	1	-1	-5,94	-1	0	0	0	-2,2	-1	-1

2.4 Ulangi iterasi sampai perubahan bobot (Δwn = 0) tidak ada (Iterasi 2)

- Hitung Response Keluaran iterasi 2
- Perbaiki bobot pola yang mengandung kesalahan

MASUKAN			TARGET		y=	PERUBAHAN BOBOT			BOBOT BARU		
X1	X2	1	t	NET	f(NET)	ΔW1	ΔW2	Δb	W1	W2	b
INISIALISASI									-2,2	-1	-1
2,9	1	1	1	-8,38	-1	2,9	1	1	0,7	0	0
2,8	3	1	-1	1,96	1	-2,8	-3	-1	-2,1	-3	-1
2,3	5	1	-1	-20,83	-1	0	0	0	-2,1	-3	-1
2,7	6	1	-1	-24,67	-1	0	0	0	-2,1	-3	-1

2.4 Ulangi iterasi sampai perubahan bobot (Δwn = 0) tidak ada (Iterasi 3)

- Hitung Response Keluaran iterasi 3
- Perbaiki bobot pola yang mengandung kesalahan

MASUKAN			TARGET		y=	PERUBAHAN BOBOT			BOBOT BARU		
X1	X2	1	t	NET	f(NET)	ΔW1	ΔW2	Δb	W1	W2	b
INISIALISASI									-2,1	-3	-1
2,9	1	1	1	-10,09	-1	2,9	1	1	0,8	-2	0
2,8	3	1	-1	-3,76	-1	0	0	0	0,8	-2	0
2,3	5	1	-1	-8,16	-1	0	0	0	0,8	-2	0
2,7	6	1	-1	-9,84	-1	0	0	0	0,8	-2	0

- Untuk data IPK memiliki pola 0.8 x - 2 y = 0 dapat dihitung prediksinya menggunakan bobot yang terakhir didapat:
  *V = X1*W1 + X2*W2 = 0,8 * 2,85 - 2*1 = 2,28 -2 = 0,28*
  Y = sign(V) = sign(0,28) = 1 (Lulus)

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 11 (Neural Network), p 219-228

- Dataset:
  - TeamValue-Training.csv
  - TeamValue-Scoring.csv

- Pahami model neural network yang dihasilkan, perhatikan ketebalan garis neuron yang menyambungkan antar node

1. Business Understanding

Motivation:

- Juan is a performance analyst for a major professional athletic team
- His team has been steadily improving over recent seasons, and heading into the coming season management believes that by adding between two and four excellent players, the team will have an outstanding shot at achieving the league championship
- They have tasked Juan with identifying their best options from among a list of 59 players that may be available to them
- All of these players have experience; some have played professionally before and some have years of experience as amateurs
- None are to be ruled out without being assessed for their potential ability to add star power and productivity to the existing team
- The executives Juan works for are anxious to get going on contacting the most promising prospects, so Juan needs to quickly evaluate these athletes' past performance and make recommendations based on his analysis

Objectives:

- To evaluate each of the 59 prospects' past statistical performance in order to help him formulate recommendations based on his analysis

Latihan

- Lakukan training dengan neural network untuk dataset TeamValue-Training.csv
- Gunakan 10-fold cross validation
- Lakukan adjusment terhadap hidden layer dan neuron size, misal: hidden layer saya buat 3, neuron size masing-masing 5
- Apa yang terjadi, apakah ada peningkatan akurasi?

	NN	NN (HL 2, NS 3)	NN (HL 2, NS 5)	NN (HL 3, NS 3)	NN (HL 3, NS 5)	NN (HL 4, NS 3)	NN (HL 4, NS 5)
Accuracy							

Penentuan Hidden Layer

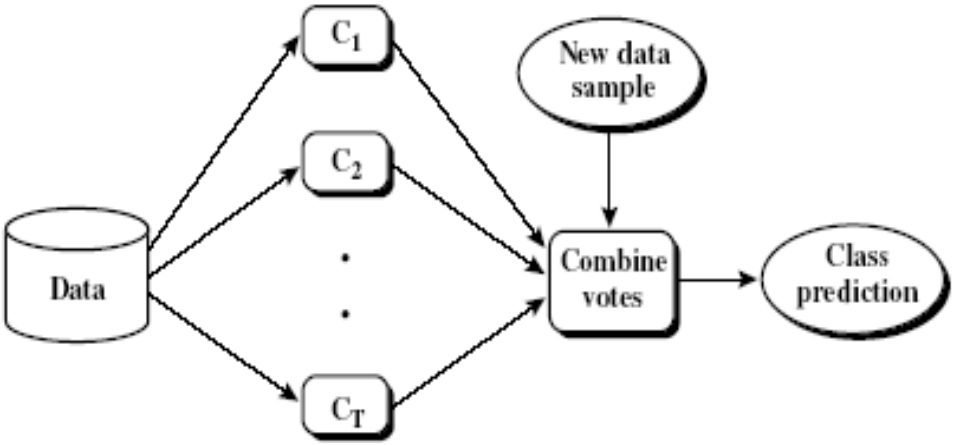
Hidden Layer	Capabilities
0	Only capable of representing linear separable functions or decisions
1	Can approximate any function that contains a continuous mapping from one finite space to another
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy

Penentuan Neuron Size

1. Trial and Error
2. Rule of Thumb:
   - Between the size of the input layer and the size of the output layer
   - 2/3 the size of the input layer, plus the size of the output layer
   - Less than twice the size of the input layer
3. Search Algorithm:
   - Greedy
   - Genetic Algorithm
   - Particle Swarm Optimization
   - etc

Techniques to Improve Classification Accuracy: Ensemble Methods

Ensemble Methods: Increasing the Accuracy



- Ensemble methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models, M1, M2, …, Mk, with the aim of creating an improved model M*
- Popular ensemble methods
  - Bagging: averaging the prediction over a collection of classifiers
  - Boosting: weighted vote with a collection of classifiers
  - Ensemble: combining a set of heterogeneous classifiers

Bagging: Boostrap Aggregation

- Analogy: Diagnosis based on multiple doctors' majority vote
- Training
  - Given a set D of *d* tuples, at each iteration *i*, a training set D_i of *d* tuples is sampled with replacement from D (i.e., bootstrap)
  - A classifier model M_i is learned for each training set D_i
- Classification: classify an unknown sample **X**
  - Each classifier M_i returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to **X**
- Prediction: can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple
- Accuracy
  - Often significantly better than a single classifier derived from D
  - For noise data: not considerably worse, more robust
  - Proved improved accuracy in prediction

Boosting

- Analogy: Consult several doctors, based on a combination of weighted diagnoses—weight assigned based on the previous diagnosis accuracy
- How boosting works?
  1. Weights are assigned to each training tuple
  2. A series of k classifiers is iteratively learned
  3. After a classifier M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1}, to pay more attention to the training tuples that were misclassified by M_i
  4. The final M* combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy
- Boosting algorithm can be extended for numeric prediction
- Comparing with bagging: Boosting tends to have greater accuracy, but it also risks overfitting the model to misclassified data

Adaboost (Freund and Schapire, 1997)

1. Given a set of *d* class-labeled tuples, ($\mathbf{X_1}$, y_1), …, ($\mathbf{X_d}$, y_d)
2. Initially, all the weights of tuples are set the same (1/d)
3. Generate k classifiers in k rounds. At round i,
   1. Tuples from D are sampled (with replacement) to form a training set D_i of the same size
   2. Each tuple's chance of being selected is based on its weight
   3. A classification model M_i is derived from D_i
   4. Its error rate is calculated using D_i as a test set
   5. If a tuple is misclassified, its weight is increased, o.w. it is decreased
4. Error rate: err($\mathbf{X_j}$) is the misclassification error of tuple $\mathbf{X_j}$. Classifier M_i error rate is the sum of the weights of the misclassified tuples:

$$error(M_i) = \sum_{j}^{d} w_j \times err(\mathbf{X_j})$$

5. The weight of classifier M_i's vote is $\log \frac{1-error(M_i)}{error(M_i)}$

Random Forest (Breiman 2001)

- Random Forest:
  - Each classifier in the ensemble is a *decision tree* classifier and is generated using a random selection of attributes at each node to determine the split
  - During classification, each tree votes and the most popular class is returned
- Two Methods to construct Random Forest:
  1. Forest-RI (*random input selection*): Randomly select, at each node, F attributes as candidates for the split at the node. The CART methodology is used to grow the trees to maximum size
  2. Forest-RC (*random linear combinations*): Creates new attributes (or features) that are a linear combination of the existing attributes (reduces the correlation between individual classifiers)
- Comparable in accuracy to Adaboost, but more robust to errors and outliers
- Insensitive to the number of attributes selected for consideration at each split, and faster than bagging or boosting

Classification of Class-Imbalanced Data Sets

- Class-imbalance problem: Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault, etc.
- Traditional methods assume a balanced distribution of classes and equal error costs: not suitable for class-imbalanced data
- Typical methods for imbalance data in 2-class classification:
  1. Oversampling: re-sampling of data from positive class
  2. Under-sampling: randomly eliminate tuples from negative class
  3. Threshold-moving: moves the decision threshold, t, so that the rare class tuples are easier to classify, and hence, less chance of costly false negative errors
  4. Ensemble techniques: Ensemble multiple classifiers introduced above
- Still difficult for class imbalance problem on multiclass tasks

4.2 Algoritma Klastering

4.2.1 Partitioning Methods

4.2.2 Hierarchical Methods

4.2.3 Density-Based Methods

4.2.4 Grid-Based Methods

What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Applications of Cluster Analysis

- Data reduction
  - Summarization: Preprocessing for regression, PCA, classification, and association analysis
  - Compression: Image processing: vector quantization
- Hypothesis generation and testing
- Prediction based on groups
  - Cluster & find characteristics/patterns for each group
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection: Outliers are often viewed as those “far away” from any cluster

Clustering: Application Examples

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market research

Basic Steps to Develop a Clustering Task

- Feature selection
  - Select info concerning the task of interest
  - Minimal information redundancy
- Proximity measure
  - Similarity of two feature vectors
- Clustering criterion
  - Expressed via a cost function or some rules
- Clustering algorithms
  - Choice of algorithms
- Validation of the results
  - Validation test (also, *clustering tendency* test)
- Interpretation of the results
  - Integration with applications

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters

- The quality of a clustering method depends on
  - the similarity measure used by the method
  - its implementation, and
  - Its ability to discover some or all of the hidden patterns

Measure the Quality of Clustering

- Dissimilarity/Similarity metric
  - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
  - The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
  - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
  - There is usually a separate "quality" function that measures the "goodness" of a cluster.
  - It is hard to define "similar enough" or "good enough"
    - The answer is typically highly subjective

Considerations for Cluster Analysis

- Partitioning criteria
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Clustering space
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Requirements and Challenges

- Scalability
  - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
  - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
  - User may give inputs on constraints
  - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
  - Discovery of clusters with arbitrary shape
  - Ability to deal with noisy data
  - Incremental clustering and insensitivity to input order
  - High dimensionality

Major Clustering Approaches 1

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue
- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches 2

- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM, COBWEB
- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

4.2.1 Partitioning Methods

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

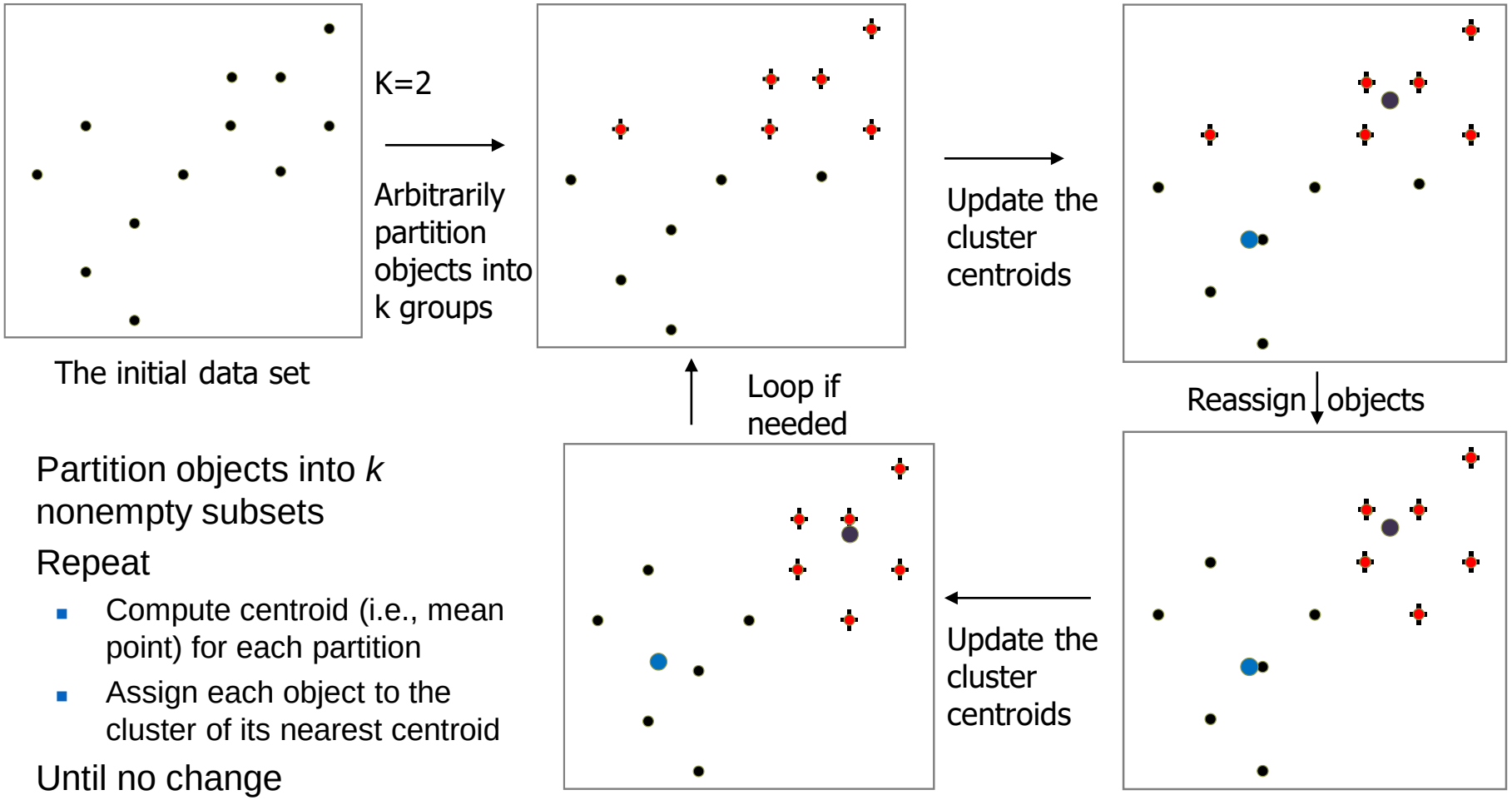
$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (d(p, c_i))^2$$

- Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
  3. Assign each object to the cluster with the nearest seed point
  4. Go back to Step 2, stop when the assignment does not change

An Example of K-Means Clustering



- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

Tahapan Algoritma k-Means

1. Pilih jumlah klaster k yang diinginkan
2. Inisialisasi k pusat klaster (centroid) secara random
3. Tempatkan setiap data atau objek ke klaster terdekat. Kedekatan dua objek ditentukan berdasar jarak. Jarak yang dipakai pada algoritma k-Means adalah *Euclidean distance* (*d*)

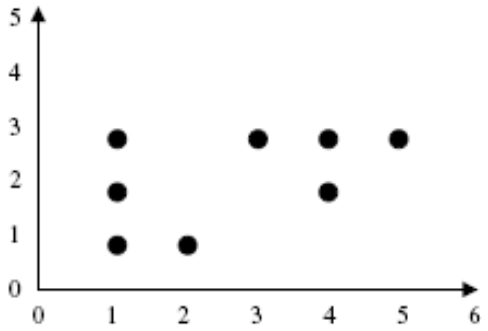
$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

   - x = x1, x2, . . . , xn, dan y = y1, y2, . . . , yn merupakan banyaknya *n* atribut(kolom) antara 2 record

4. Hitung kembali pusat klaster dengan keanggotaan klaster yang sekarang. Pusat klaster adalah rata-rata (mean) dari semua data atau objek dalam klaster tertentu
5. Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru. Jika pusat klaster sudah tidak berubah lagi, maka proses pengklasteran selesai. Atau, kembali lagi ke langkah nomor 3 sampai pusat klaster tidak berubah lagi (stabil) atau tidak ada penurunan yang signifikan dari nilai SSE (*Sum of Squared Errors*)

Contoh Kasus – Iterasi 1

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1



1. Tentukan jumlah klaster k=2
2. Tentukan centroid awal secara acak misal dari data disamping m1 =(1,1), m2=(2,1)
3. Tempatkan tiap objek ke klaster terdekat berdasarkan nilai centroid yang paling dekat selisihnya (jaraknya). Didapatkan hasil, anggota *cluster1 = {A,E,G}, cluster2={B,C,D,F,H}*

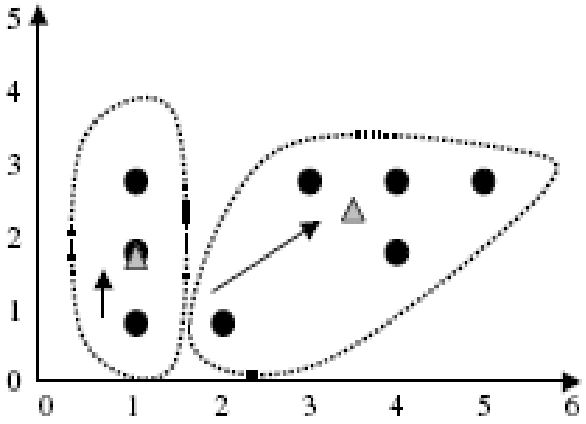
Nilai SSE yaitu:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

$$2^2 + 2{,}24^2 + 2{,}83^2 + 3{,}61^2 + 1^2 + 2{,}24^2 + 0^2 + 0^2 = 36$$

Interasi 2

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	2.00	2.24	C_1
b	2.83	2.24	C_2
c	3.61	2.83	C_2
d	4.47	3.61	C_2
e	1.00	1.41	C_1
f	3.16	2.24	C_2
g	0.00	1.00	C_1
h	1.00	0.00	C_2



4. Menghitung nilai centroid yang baru

$$m_1 = [(1+1+1)/3, (3+2+1)/3] = (1,2)$$

$$m_2 = [(3+4+5+4+2)/5, (3+3+3+2+1)/5] = (3{,}6;2{,}4)$$

5. Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru.

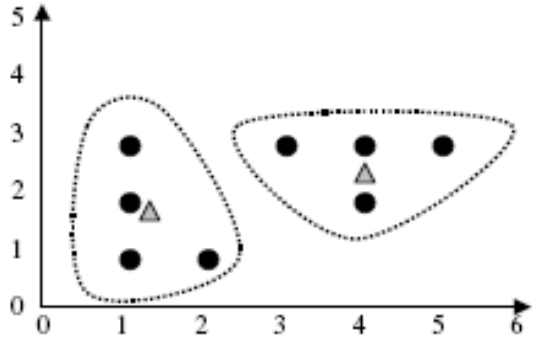
   Nilai SSE yang baru:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} d(p,m_i)^2 = 1^2 + 0.85^2 + 0.72^2 + 1.52^2 + 0^2 + 0.57^2 + 1^2$$

$$+1.41^2 = 7.88$$

Iterasi 3

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.00	2.67	C_1
b	2.24	0.85	C_2
c	3.16	0.72	C_2
d	4.12	1.52	C_2
e	0.00	2.63	C_1
f	3.00	0.57	C_2
g	1.00	2.95	C_1
h	1.41	2.13	C_2



4. Terdapat perubahan anggota cluster yaitu cluster1={A,E,G,H}, cluster2={B,C,D,F}, maka cari lagi nilai centroid yang baru yaitu: m1=(1,25;1,75) dan m2=(4;2,75)

5. Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru
Nilai SSE yang baru:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} d(p,m_i)^2 = 1.27^2 + 1.03^2 + 0.25^2 + 1.03^2 + 0.35^2 + 0.75^2$$

$$+0.79^2 + 1.06^2 = 6.25$$

Hasil Akhir

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.27	3.01	C_1
b	2.15	1.03	C_2
c	3.02	0.25	C_2
d	3.95	1.03	C_2
e	0.35	3.09	C_1
f	2.76	0.75	C_2
g	0.79	3.47	C_1
h	1.06	2.66	C_2

- Dapat dilihat pada tabel. Tidak ada perubahan anggota lagi pada masing-masing cluster
- Hasil akhir yaitu: cluster1={A,E,G,H}, dan cluster2={B,C,D,F}
  Dengan nilai SSE = 6,25 dan jumlah iterasi 3

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses, 2012, Chapter 6 k-Means Clustering, pp. 91-103 (CoronaryHeartDisease.csv)
- Gambarkan grafik (chart) dan pilih Scatter 3D Color untuk menggambarkan data hasil klastering yang telah dilakukan
- Analisis apa yang telah dilakukan oleh Sonia, dan apa manfaat k-Means clustering bagi pekerjaannya?

Latihan

- Lakukan pengukuran performance dengan menggunakan Cluster Distance Performance, untuk mendapatkan nilai Davies Bouldin Index (DBI)
- Nilai DBI semakin rendah berarti cluster yang kita bentuk semakin baik

Latihan

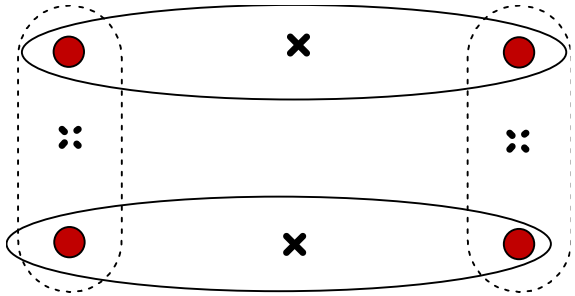
- Lakukan klastering terhadap data IMFdata.csv (http://romisatriawahono.net/lecture/dm/dataset)

Comments on the *K-Means* Method

- Strength:
  - *Efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t << n$.
  - Comparing: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- Comment: Often terminates at a *local optimal*
- Weakness
  - Applicable only to objects in a continuous n-dimensional space
    - Using the k-modes method for categorical data
    - In comparison, k-medoids can be applied to a wide range of data
  - Need to specify *k*, the *number* of clusters, in advance (there are ways to automatically determine the best *k* (see Hastie et al., 2009)
  - Sensitive to noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Variations of the *K-Means* Method

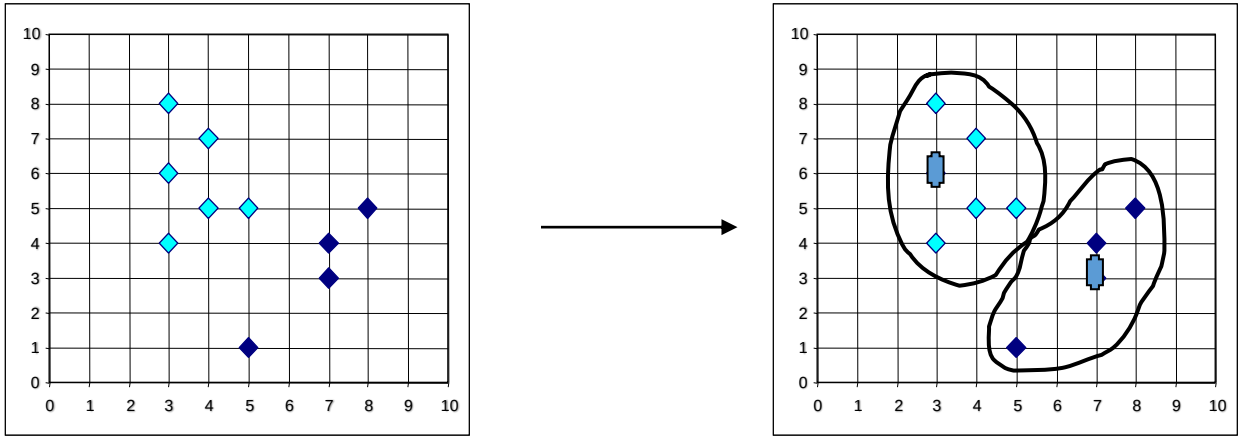
- Most of the variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means



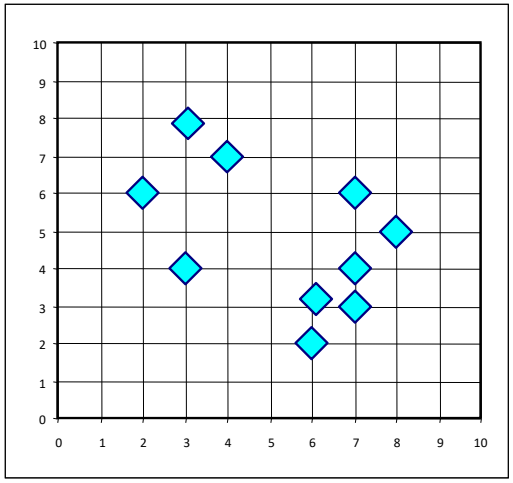
- Handling categorical data: *k-modes*
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

What Is the Problem of the K-Means Method?

- The k-means algorithm is sensitive to outliers!
  - Since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids:
  - Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

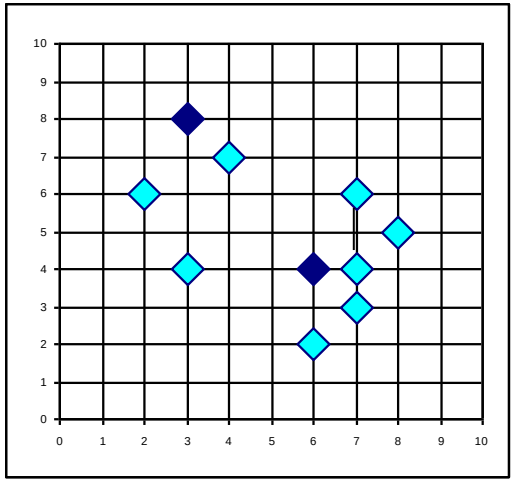


PAM: A Typical K-Medoids Algorithm



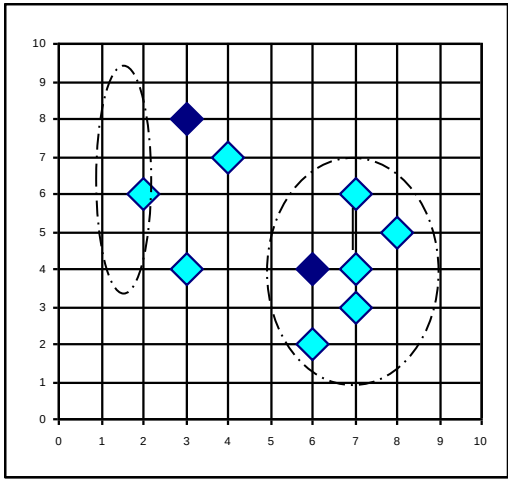
K=2

Arbitrary choose k object as initial medoids

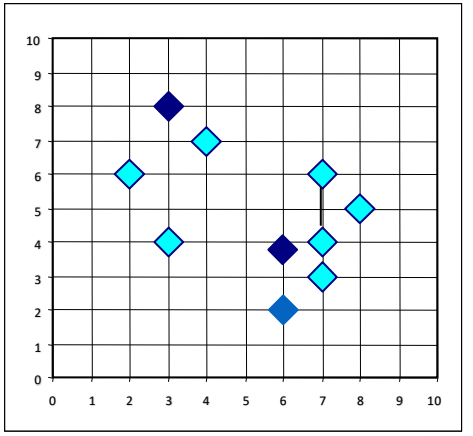


Assign each remaining object to nearest medoids

Total Cost = 20

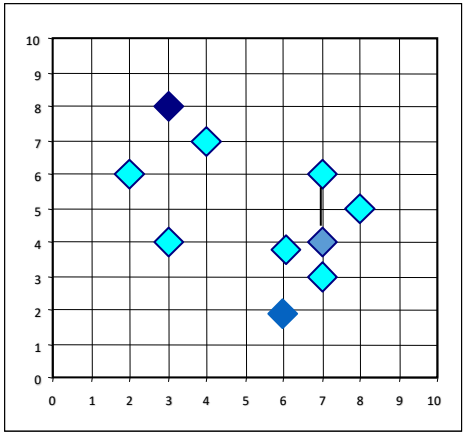


Randomly select a nonmedoid object, O_{ramdom}



Compute total cost of swapping

Total Cost = 26



Swapping O and O_{ramdom}

If quality is improved.

Do loop

Until no change

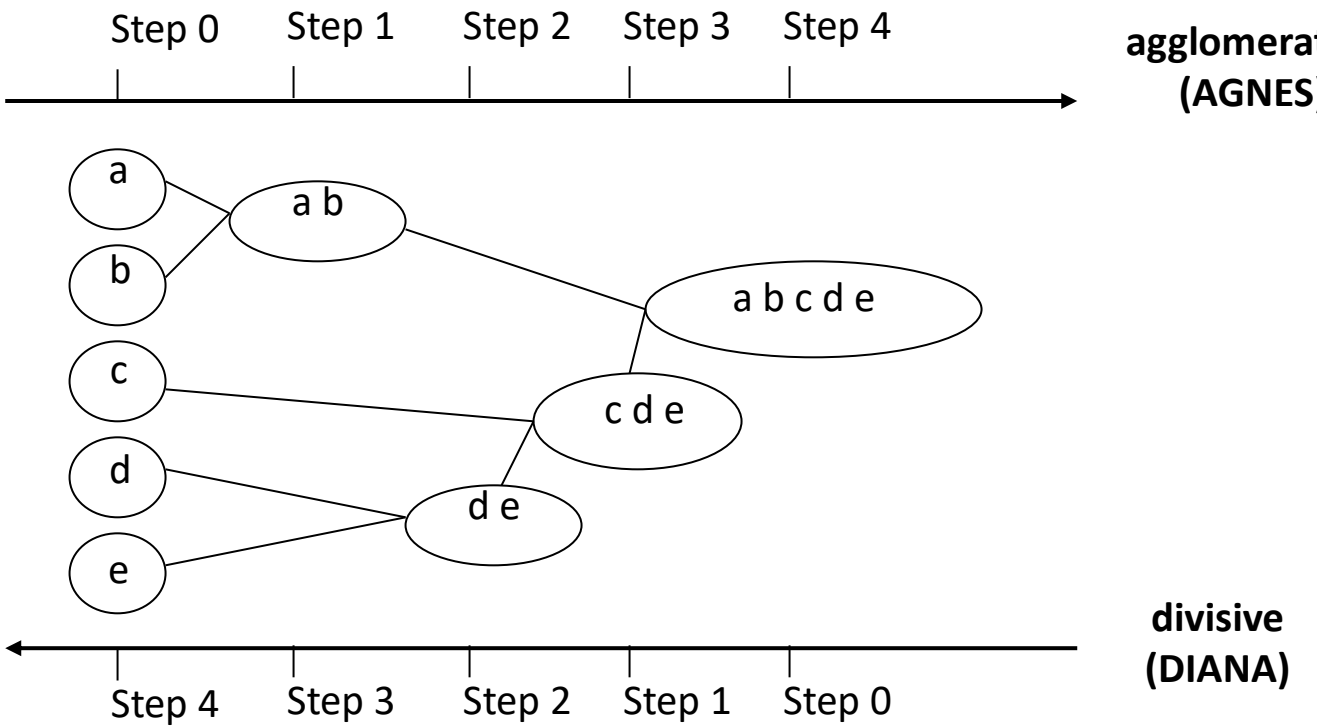
The K-Medoid Clustering Method

- K-Medoids Clustering: Find representative objects (medoids) in clusters
- PAM (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987)
  - Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - PAM works effectively for small data sets, but does not scale well for large data sets (due to the computational complexity)
- Efficiency improvement on PAM
  - CLARA (Kaufmann & Rousseeuw, 1990): PAM on samples
  - CLARANS (Ng & Han, 1994): Randomized re-sampling

4.2.2 Hierarchical Methods

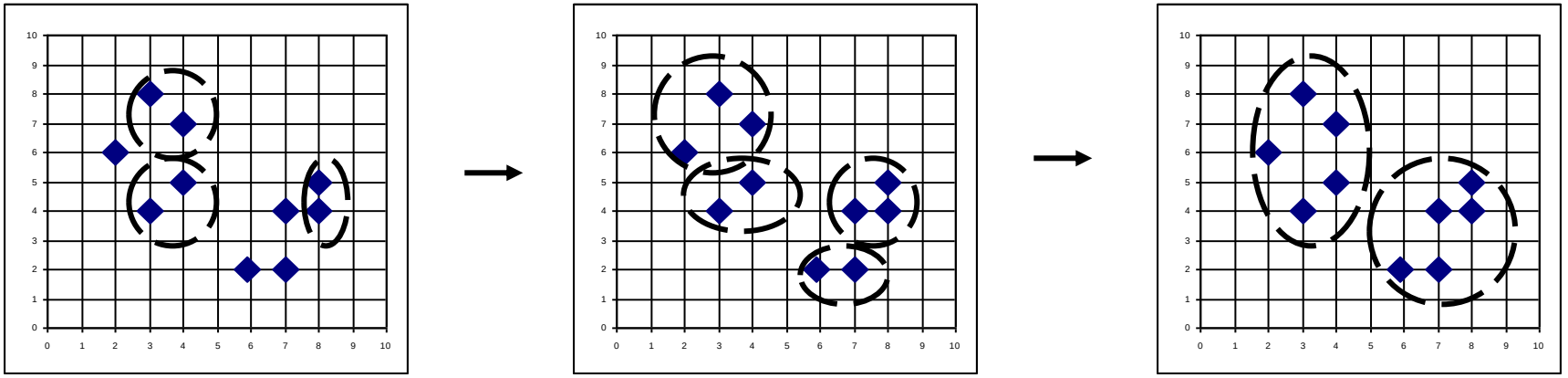
Hierarchical Clustering

- Use distance matrix as clustering criteria
- This method does not require the number of clusters k as an input, but needs a termination condition



AGNES (Agglomerative Nesting)

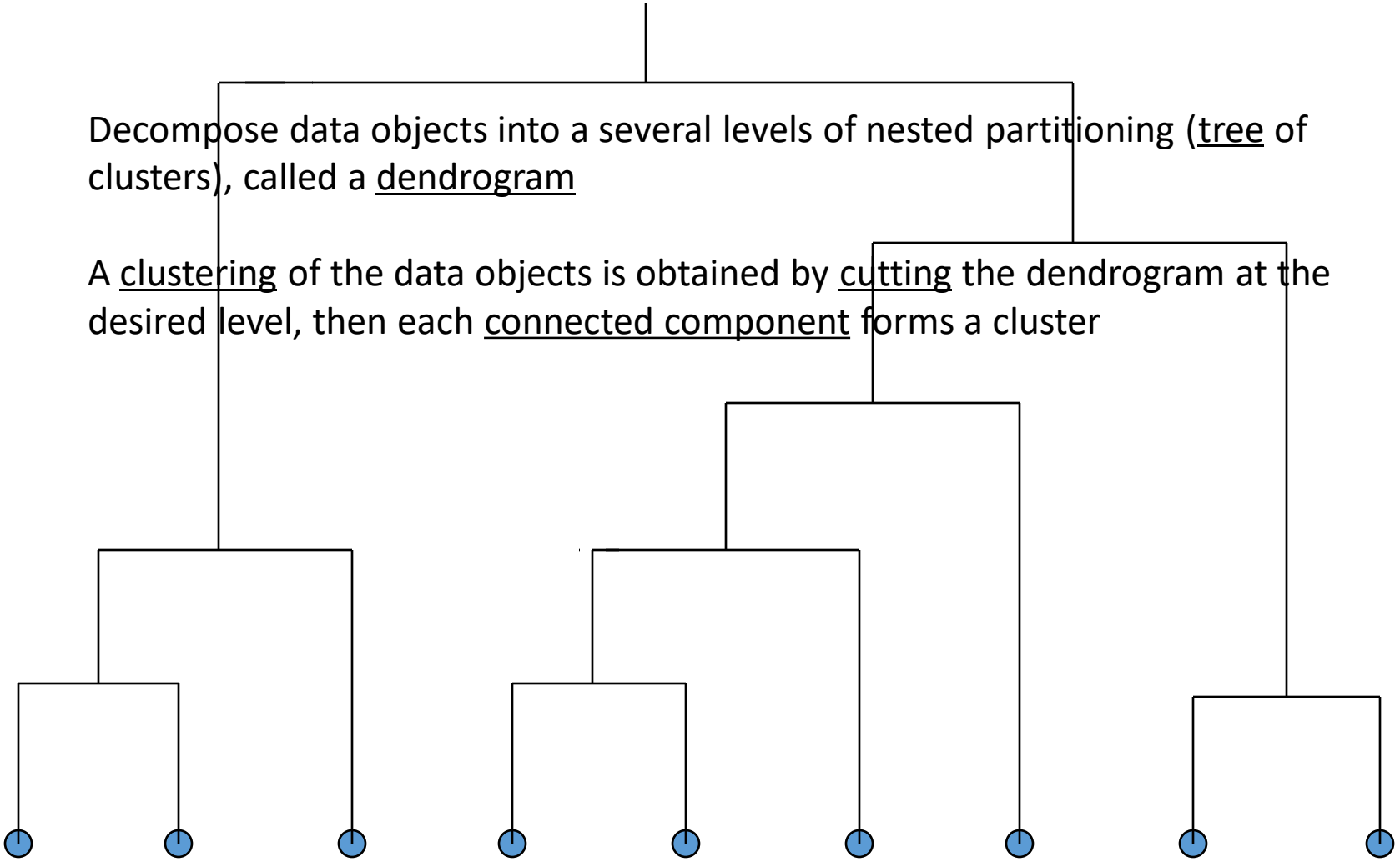
- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical packages, e.g., Splus
- Use the single-link method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



Dendrogram: Shows How Clusters are Merged

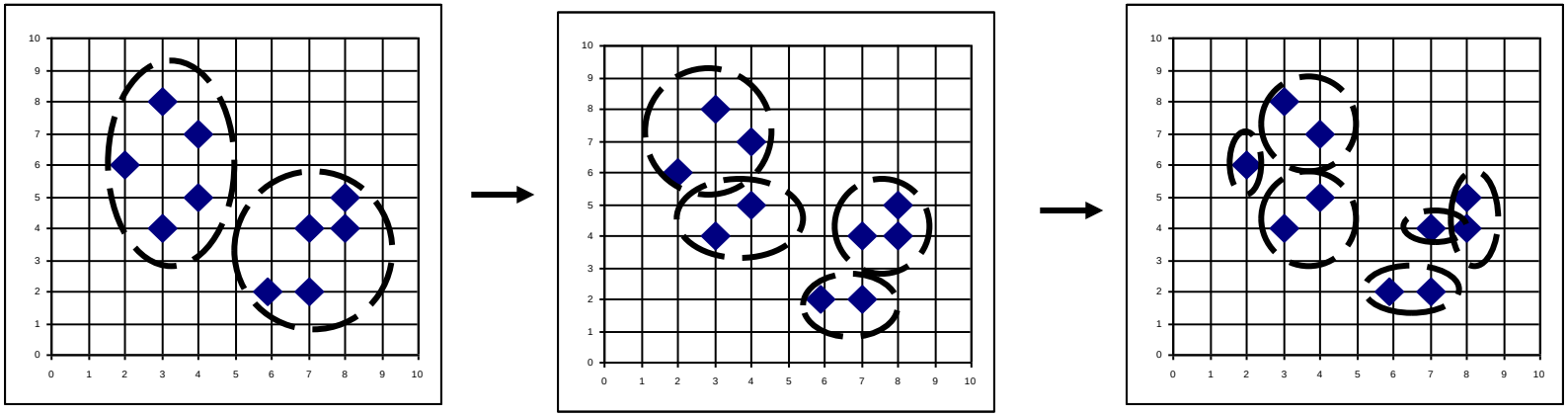
Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster

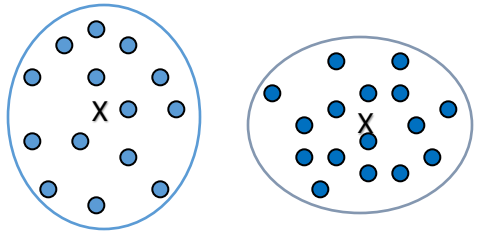


DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



Distance between Clusters



- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = avg(t_{ip}, t_{jq})$
- Centroid: distance between the centroids of two clusters, i.e., $dist(K_i, K_j) = dist(C_i, C_j)$
- Medoid: distance between the medoids of two clusters, i.e., $dist(K_i, K_j) = dist(M_i, M_j)$
  - Medoid: a chosen, centrally located object in the cluster

Centroid, Radius and Diameter of a Cluster (for numerical data sets)

- Centroid: the "middle" of a cluster

$$C_m = \frac{\sum_{i=1}^{N} (t_{ip})}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

$$R_m = \sqrt{\frac{\sum_{i=1}^{N} (t_{ip} - c_m)^2}{N}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D_m = \sqrt{\frac{\sum_{i=1}^{N} \sum_{i=1}^{N} (t_{ip} - t_{iq})^2}{N(N-1)}}$$

4.2.3 Density-Based Methods

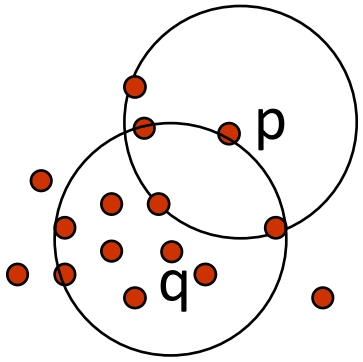
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-Based Clustering: Basic Concepts

- Two parameters*:*
  - *Eps*: Maximum radius of the neighbourhood
  - *MinPts*: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(q)$: {p belongs to D | dist(p,q) ≤ Eps}
- Directly density-reachable: A point *p* is directly density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if
  - *p* belongs to $N_{Eps}(q)$
  - core point condition:

$$|N_{Eps}(q)| \geq MinPts$$

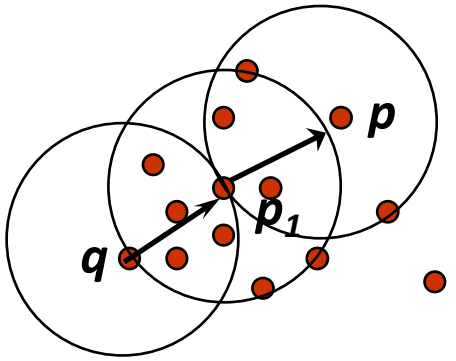


MinPts = 5

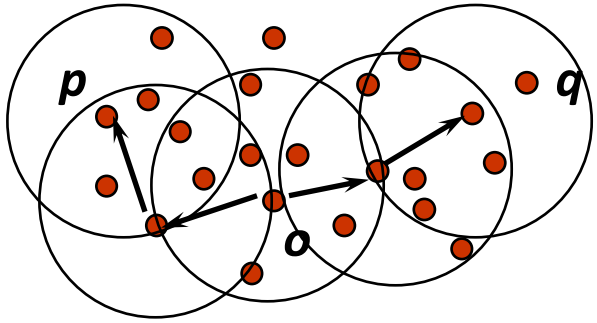
Eps = 1 cm

Density-Reachable and Density-Connected

- Density-reachable:
  - A point *p* is density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

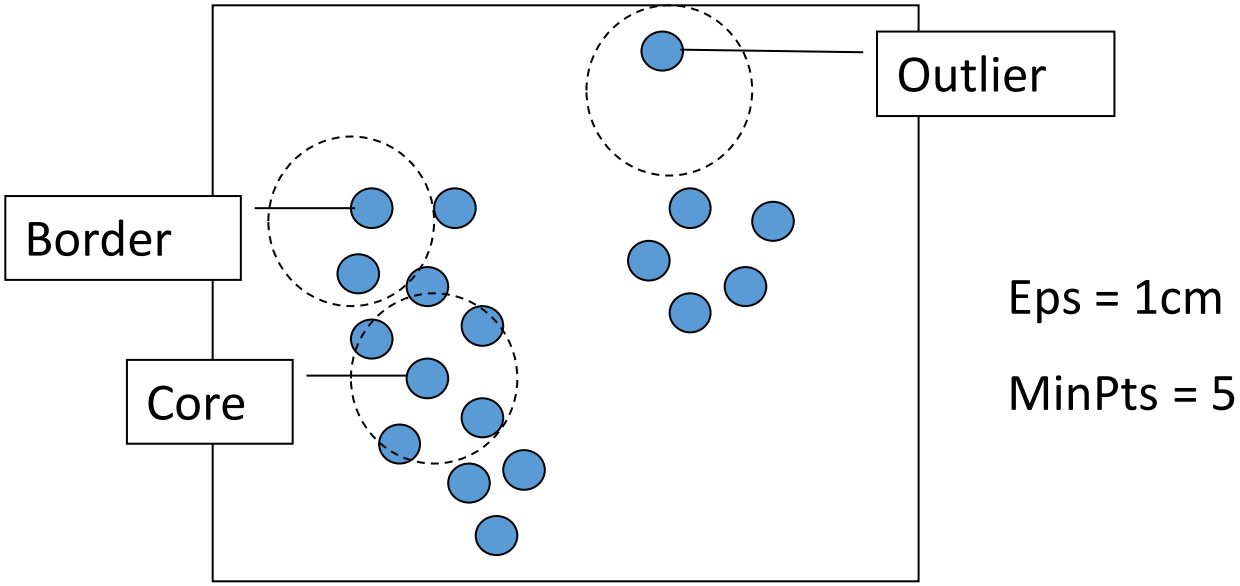


- Density-connected
  - A point *p* is density-connected to a point *q* w.r.t. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*



DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



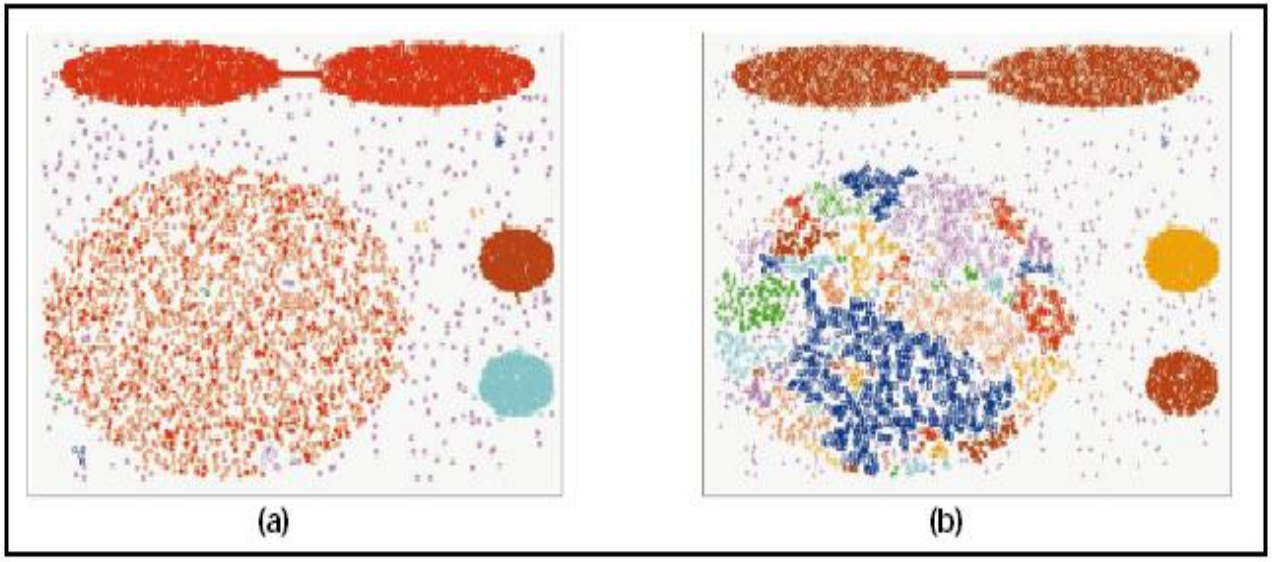
DBSCAN: The Algorithm

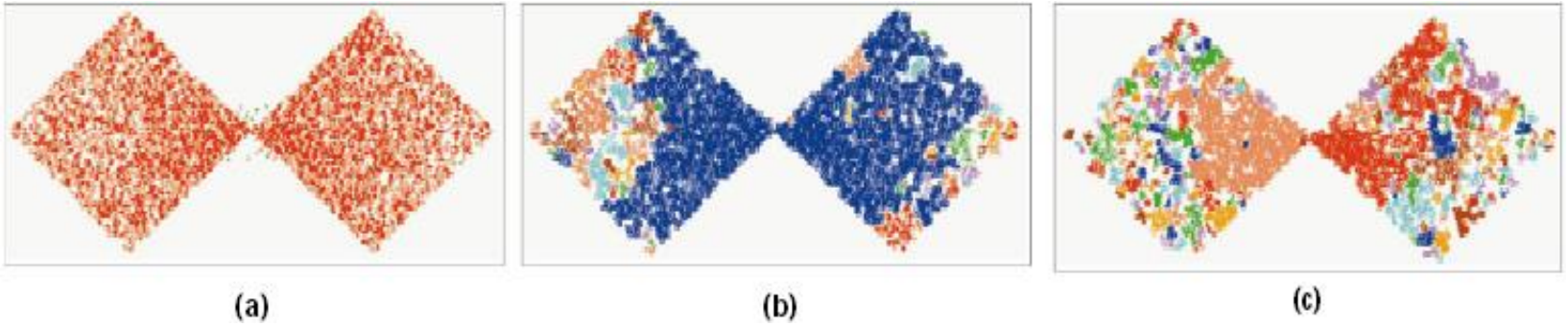
1. Arbitrary select a point *p*
2. Retrieve all points density-reachable from *p* w.r.t. *Eps* and *MinPts*
3. If *p* is a core point, a cluster is formed
4. If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database
5. Continue the process until all of the points have been processed

If a spatial index is used, the computational complexity of DBSCAN is O(nlogn), where n is the number of database objects. Otherwise, the complexity is $O(n^2)$

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.





http://webdocs.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Produces a special order of the database wrt its density-based clustering structure
  - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

OPTICS: Some Extension from DBSCAN

- Index-based: k = # of dimensions, N: # of points
  - Complexity: O(N*logN)
- Core Distance of an object p: the smallest value ε such that the ε-neighborhood of p has at least MinPts objects

  Let $N_\varepsilon(p)$: ε-neighborhood of p, ε is a distance value

  $\text{Core-distance}_{\varepsilon, \text{MinPts}}(p)$ = Undefined if $\text{card}(N_\varepsilon(p)) < \text{MinPts}$

  MinPts-distance(p), otherwise
- Reachability Distance of object p from core object q is the min radius value that makes p density-reachable from q

  $\text{Reachability-distance}_{\varepsilon, \text{MinPts}}(p, q)$ =

  Undefined if q is not a core object

  max(core-distance(q), distance (q, p)), otherwise

Core Distance & Reachability Distance

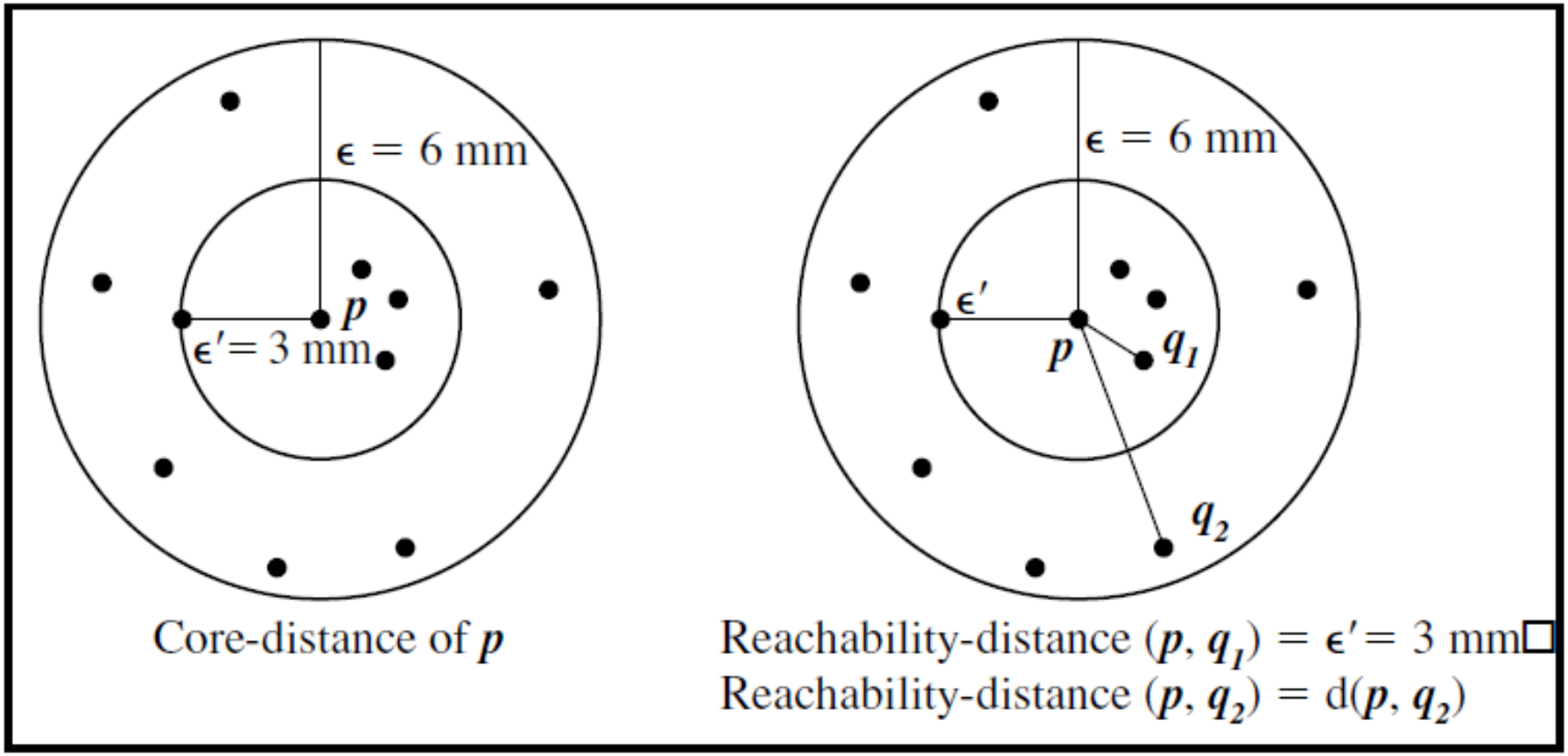
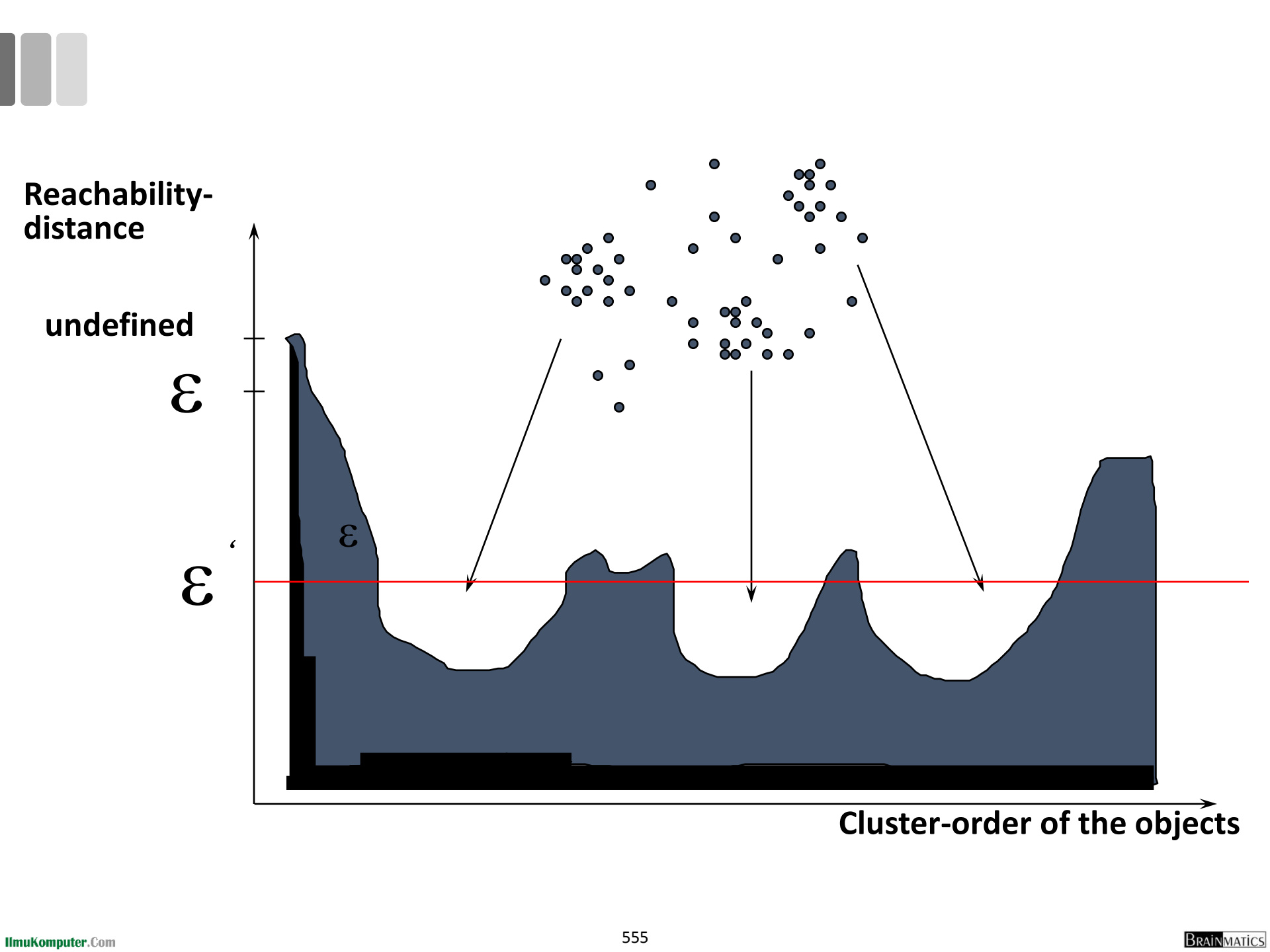
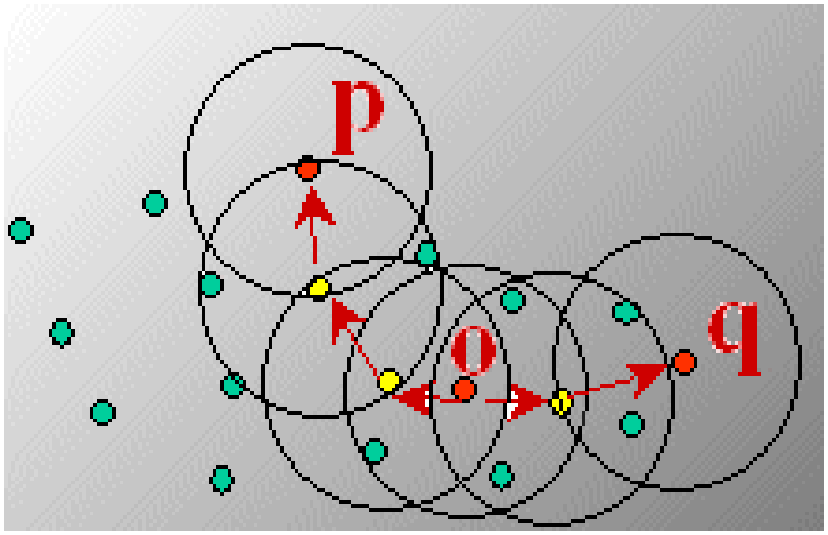


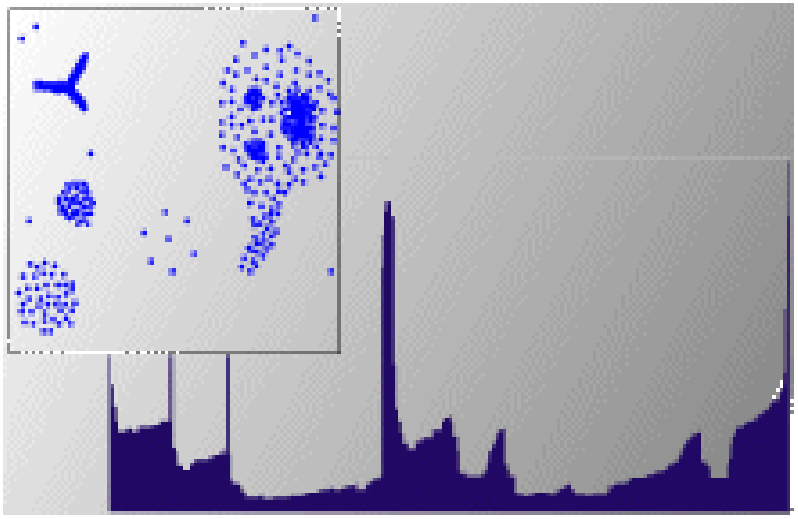
Figure 10.16: OPTICS terminology. Based on [ABKS99].

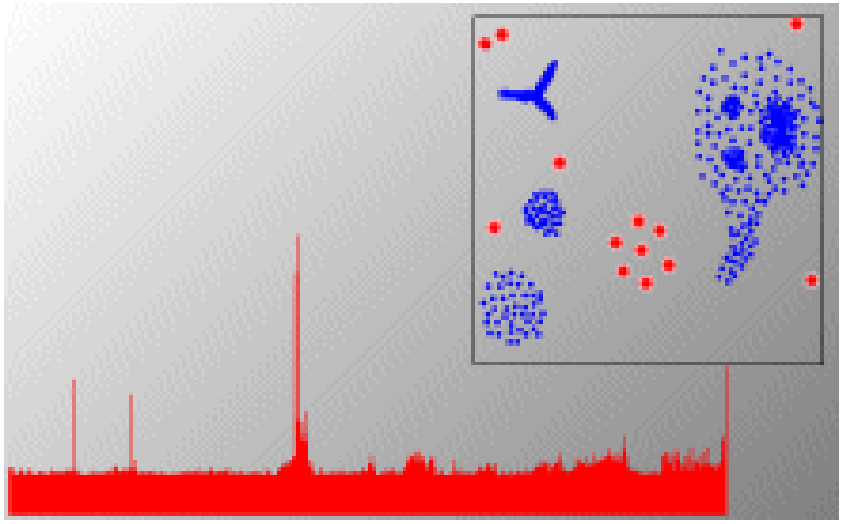


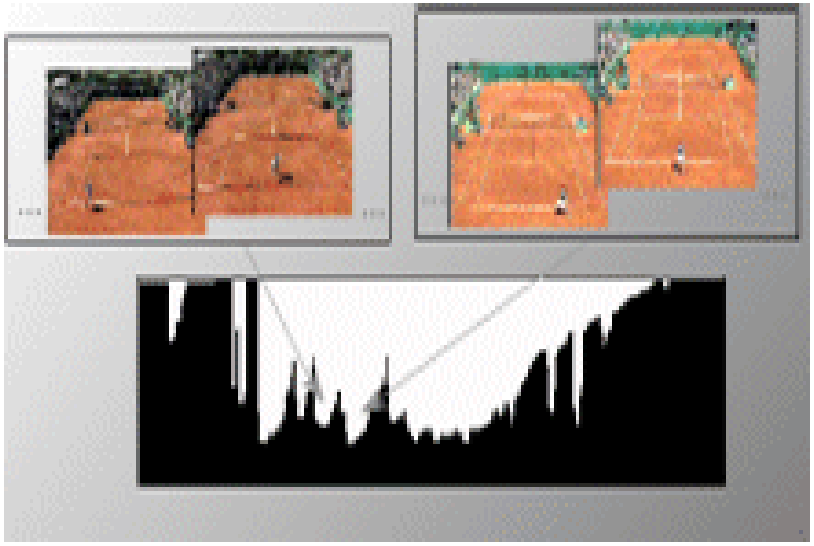
Density-Based Clustering: OPTICS & Applications:

http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/OPTICS/Demo









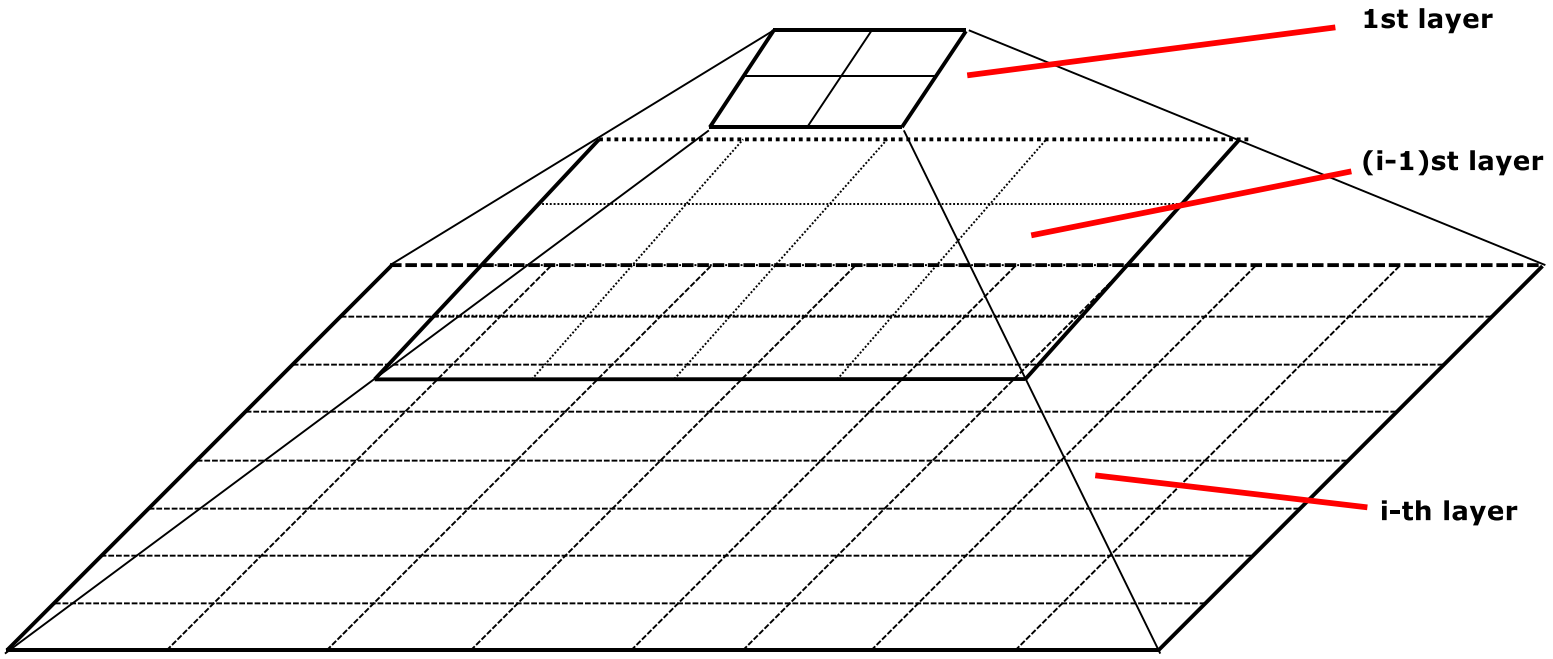
4.2.4 Grid-Based Methods

Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - CLIQUE: Agrawal, et al. (SIGMOD'98)
    - Both grid-based and subspace clustering

STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



The STING Clustering Method

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
  - *count*, *mean*, *s*, *min*, *max*
  - type of distribution—*normal*, *uniform*, etc.
- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

STING Algorithm and Its Analysis

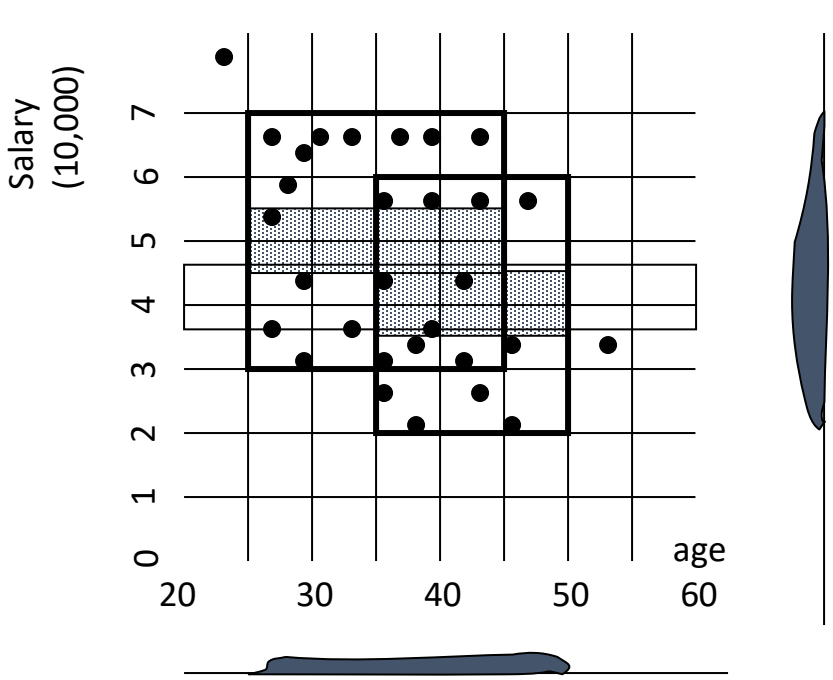
- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
  - Query-independent, easy to parallelize, incremental update
  - *O(K),* where *K* is the number of grid cells at the lowest level
- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

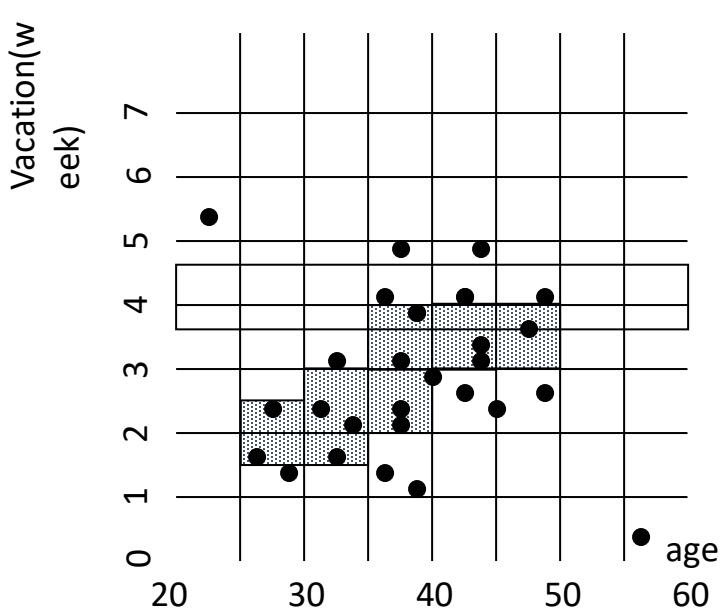
CLIQUE (Clustering In QUEst)

- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
  - It partitions each dimension into the same number of equal length interval
  - It partitions an m-dimensional data space into non-overlapping rectangular units
  - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - A cluster is a maximal set of connected dense units within a subspace

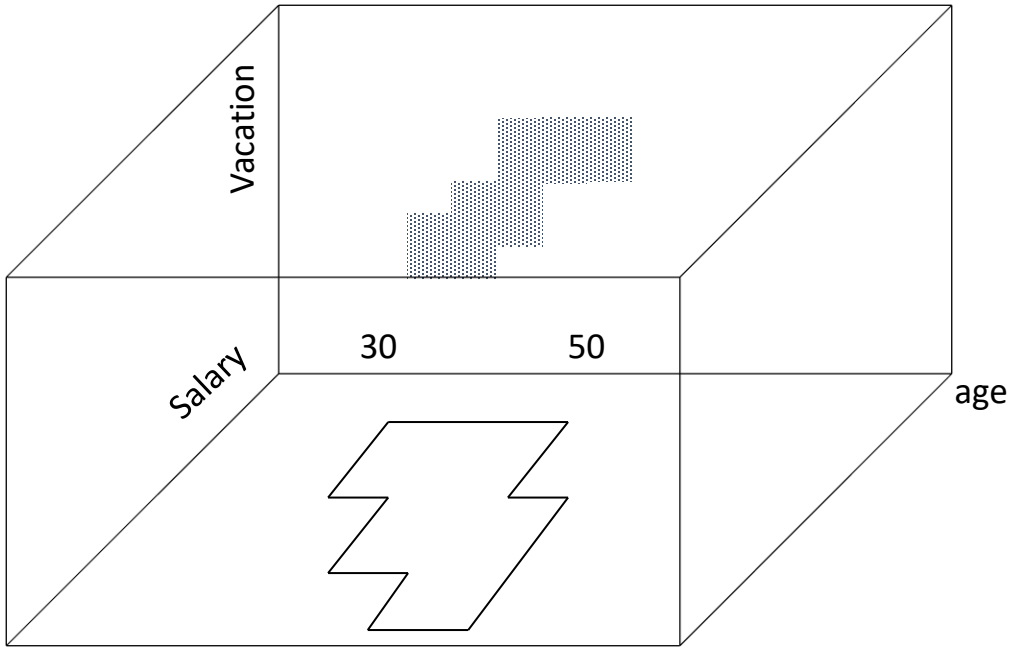
CLIQUE: The Major Steps

1. Partition the data space and find the number of points that lie inside each cell of the partition.
2. Identify the subspaces that contain clusters using the Apriori principle
3. Identify clusters
    1. Determine dense units in all subspaces of interests
    2. Determine connected dense units in all subspaces of interests.
4. Generate minimal description for the clusters
    1. Determine maximal regions that cover a cluster of connected dense units for each cluster
    2. Determination of minimal cover for each cluster





τ = 3



Strength and Weakness of *CLIQUE*

- Strength
  - automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
  - insensitive to the order of records in input and does not presume some canonical data distribution
  - scales linearly with the size of input and has good scalability as the number of dimensions in the data increases
- Weakness
  - The accuracy of the clustering result may be degraded at the expense of simplicity of the method

4.3 Algoritma Asosiasi

4.3.1 Frequent Itemset Mining Methods

4.3.2 Pattern Evaluation Methods

What Is Frequent Pattern Analysis?

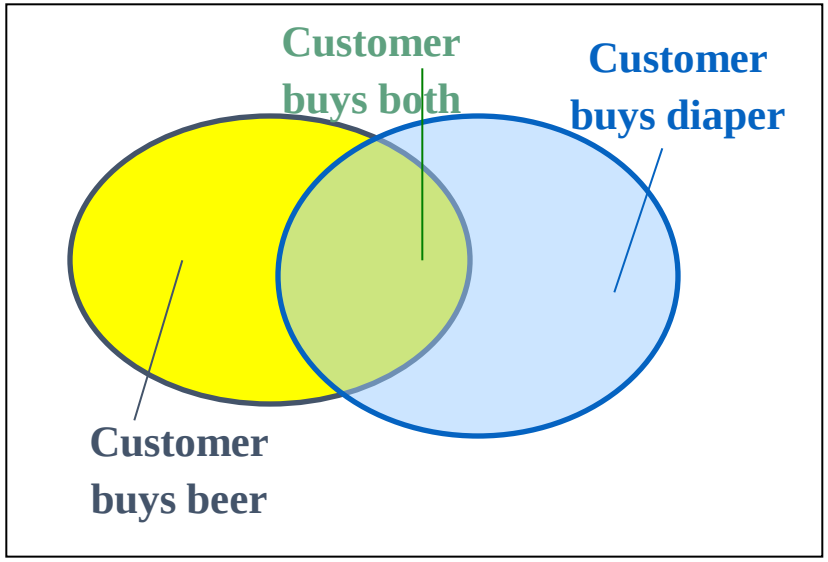
- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Why Is Freq. Pattern Mining Important?

- Freq. pattern: An intrinsic and important property of datasets
- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: discriminative, frequent pattern analysis
  - Cluster analysis: frequent pattern-based clustering
  - Data warehousing: iceberg cube and cube-gradient
  - Semantic data compression: fascicles
  - Broad applications

Basic Concepts: Frequent Patterns

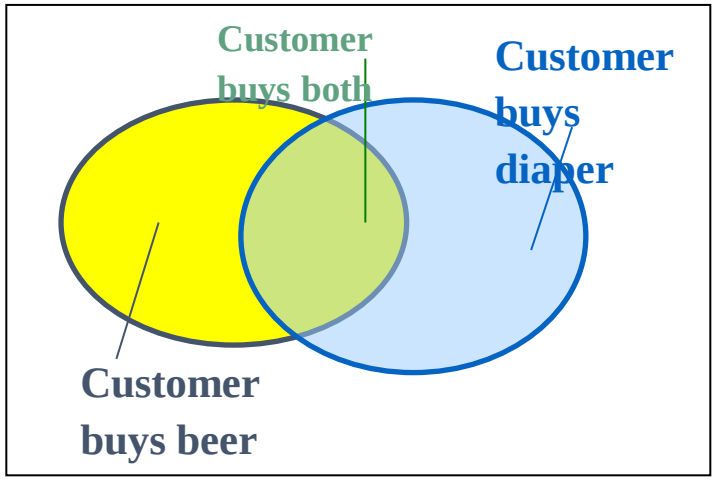
Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold

Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, *s*, probability that a transaction contains $X \cup Y$
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

Let minsup = 50%, minconf = 50%

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

Closed Patterns and Max-Patterns

- A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{1}{1}\binom{0}{0}\binom{0}{0}$ $= 2^{100} - 1 = 1.27*10^{30}$ sub-patterns!
- Solution: *Mine closed patterns and max-patterns instead*
- An itemset X is closed if X is *frequent* and there exists *no super-pattern* $Y \supset X$, *with the same support* as X (proposed by Pasquier, et al. @ ICDT'99)
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern $Y \supset X$ (proposed by Bayardo @ SIGMOD'98)
- Closed pattern is a lossless compression of freq. patterns
  - Reducing the # of patterns and rules

Closed Patterns and Max-Patterns

- Exercise. DB = {<a_1, …, a_{100}>, < a_1, …, a_{50}>}
  - Min_sup = 1.
- What is the set of closed itemset?
  - <a_1, …, a_{100}>: 1
  - < a_1, …, a_{50}>: 2
- What is the set of max-pattern?
  - <a_1, …, a_{100}>: 1
- What is the set of all patterns?
  - !!

Computational Complexity of Frequent Itemset Mining

- How many itemsets are potentially to be generated in the worst case?
  - The number of frequent itemsets to be generated is senstive to the minsup threshold
  - When minsup is low, there exist potentially an exponential number of frequent itemsets
  - The worst case: MN where M: # distinct items, and N: max length of transactions
- The worst case complexty vs. the expected probability

  Ex. Suppose Walmart has 104 kinds of products
  - The chance to pick up one product 10-4
  - The chance to pick up a particular set of 10 products: ~10-40
  - What is the chance this particular set of 10 products to be frequent 103 times in 109 transactions?

4.3.1 Frequent Itemset Mining Methods

Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FPGrowth: A Frequent Pattern-Growth Approach
- ECLAT: Frequent Pattern Mining with Vertical Data Format

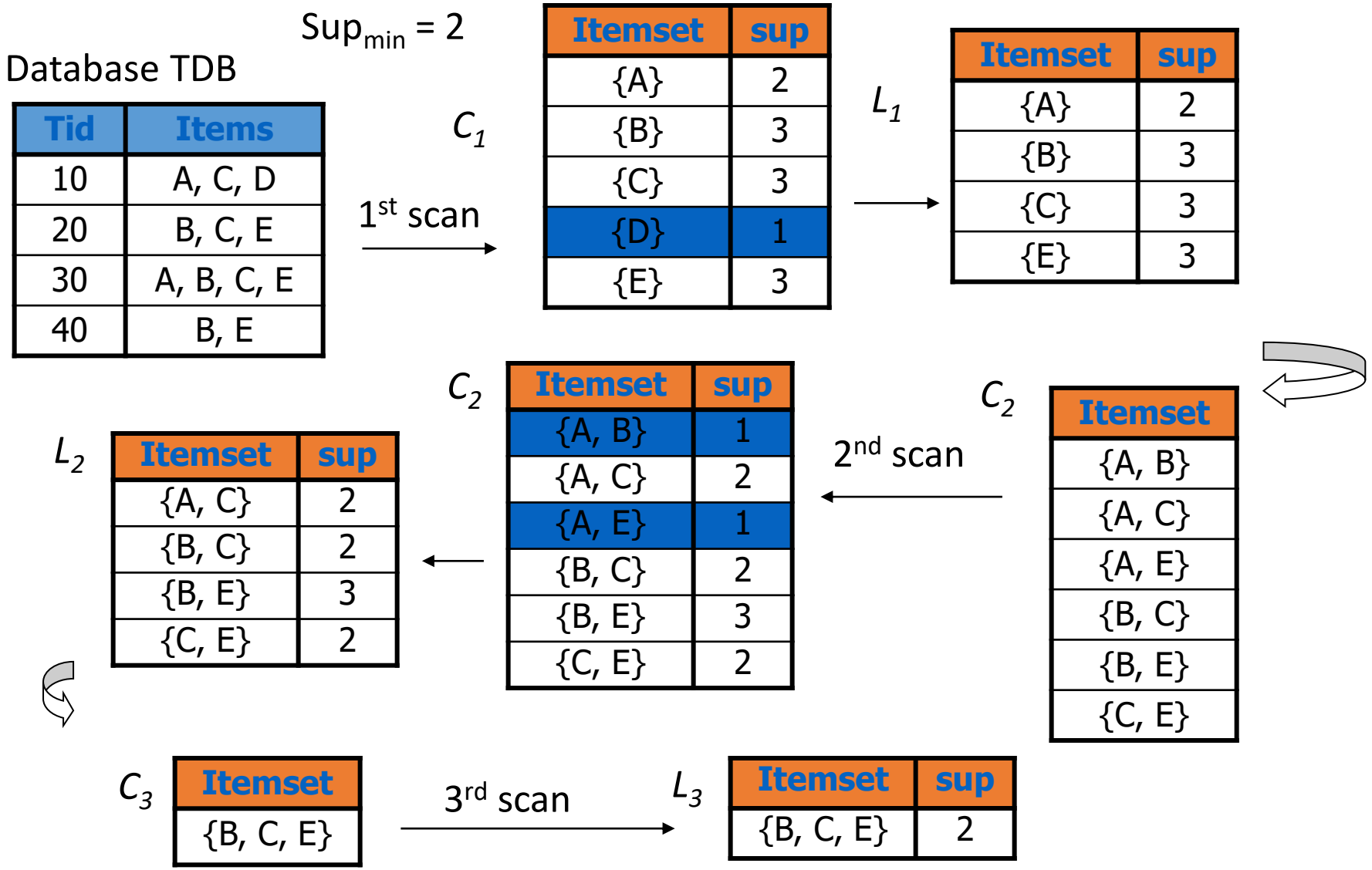
The Downward Closure Property and Scalable Mining Methods

- The downward closure property of frequent patterns
  - Any subset of a frequent itemset must be frequent
  - If {beer, diaper, nuts} is frequent, so is {beer, diaper}
  - i.e., every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Scalable mining methods: Three major approaches
  - Apriori (Agrawal & Srikant@VLDB'94)
  - Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD'00)
  - Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)

Apriori: A Candidate Generation & Test Approach

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested! (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- Method:
  1. Initially, scan DB once to get frequent 1-itemset
  2. Generate length (k+1) candidate itemsets from length k frequent itemsets
  3. Test the candidates against DB
  4. Terminate when no frequent or candidate set can be generated

The Apriori Algorithm—An Example



The Apriori Algorithm (Pseudo-Code)

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};

for (k = 1; L_k !=∅; k++) **do begin**

  C_{k+1} = candidates generated from L_k;

  for each transaction t in database do

    increment the count of all candidates in C_{k+1} that are contained in t

  L_{k+1} = candidates in C_{k+1} with min_support

  end

return $\cup_k L_k$;

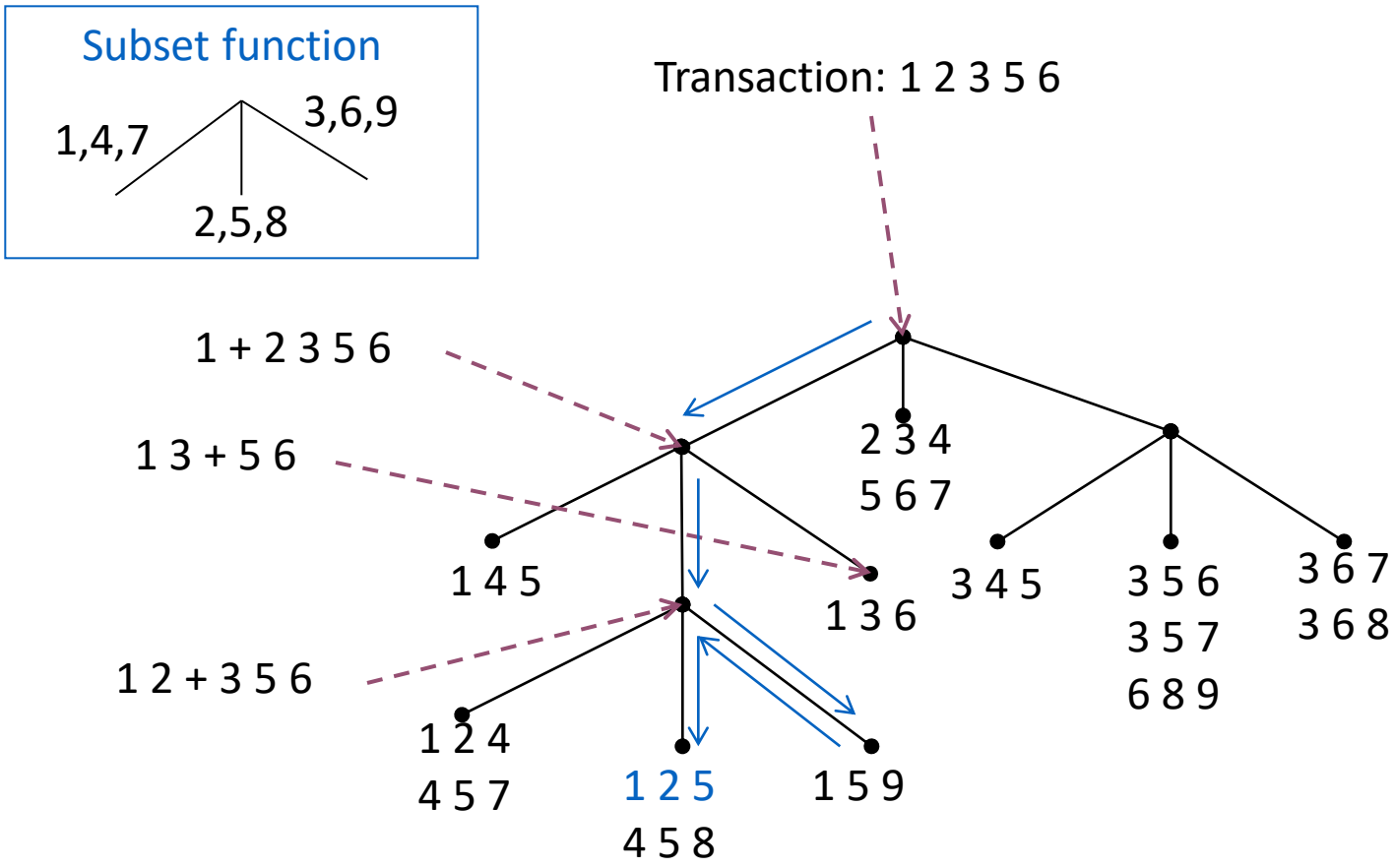
Implementation of Apriori

- How to generate candidates?
  - Step 1: self-joining L_k
  - Step 2: pruning
- Example of Candidate-generation
  - L_3={*abc, abd, acd, ace, bcd*}
  - Self-joining: $L_3 * L_3$
    - *abcd* from *abc* and *abd*
    - *acde* from *acd* and *ace*
  - Pruning:
    - *acde* is removed because *ade* is not in L_3
  - C_4 = {*abcd*}

How to Count Supports of Candidates?

- Why counting supports of candidates a problem?
  - The total number of candidates can be very huge
  - One transaction may contain many candidates
- Method:
  - Candidate itemsets are stored in a *hash-tree*
  - *Leaf* node of hash-tree contains a list of itemsets and counts
  - *Interior* node contains a hash table
  - *Subset function*: finds all the candidates contained in a transaction

Counting Supports of Candidates Using Hash Tree



Candidate Generation: An SQL Implementation

- SQL Implementation of candidate generation
  - Suppose the items in L_{k-1} are listed in an order
  - Step 1: self-joining L_{k-1}

    insert into $\boldsymbol{C_k}$

    select $\boldsymbol{p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}}$

    from $\boldsymbol{L_{k-1}\ p, L_{k-1}\ q}$

    where $\boldsymbol{p.item_1=q.item_1, \ldots, p.item_{k-2}=q.item_{k-2}, p.item_{k-1} < q.item_{k-1}}$
  - Step 2: pruning

    forall ***itemsets c in*** $\boldsymbol{C_k}$ do

    forall ***(k-1)-subsets s of c*** do

    if *(s is not in* L_{k-1}*)* **then delete** *c* **from** C_k
- Use object-relational extensions like UDFs, BLOBs, and Table functions for efficient implementation

  (S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. SIGMOD'98)

Pattern-Growth Approach: Mining Frequent Patterns Without Candidate Generation

- Bottlenecks of the Apriori approach
  - Breadth-first (i.e., level-wise) search
  - Candidate generation and test
    - Often generates a huge number of candidates
- The FPGrowth Approach *(J. Han, J. Pei, and Y. Yin, SIGMOD' 00)*
  - Depth-first search
  - Avoid explicit candidate generation
- Major philosophy: Grow long patterns from short ones using local frequent items only
  - "abc" is a frequent pattern
  - Get all transactions having "abc", i.e., project DB on abc: DB|abc
  - "d" is a local frequent item in DB|abc → abcd is a frequent pattern

Construct FP-tree from a Transaction Database

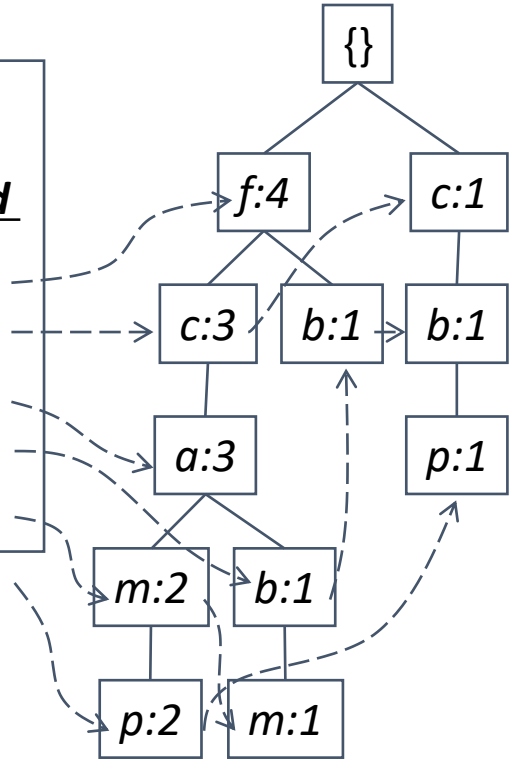
TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support = 3

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

Header Table

Item	frequency	head
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



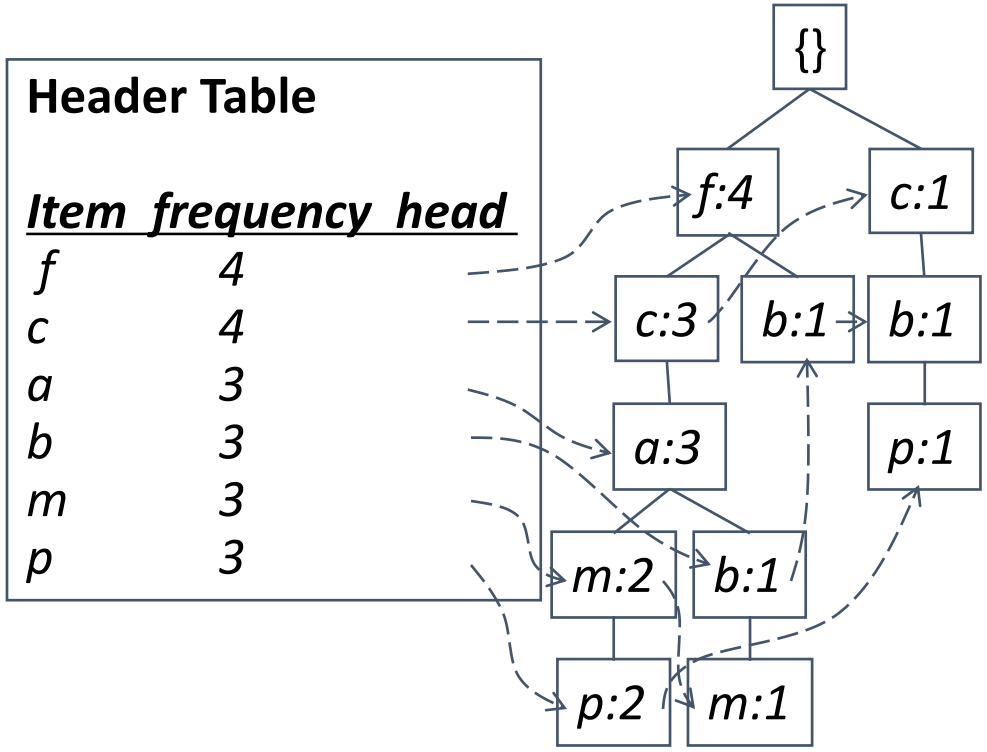
F-list = f-c-a-b-m-p

Partition Patterns and Databases

- Frequent patterns can be partitioned into subsets according to f-list
  - F-list = f-c-a-b-m-p
  - Patterns containing p
  - Patterns having m but no p
  - …
  - Patterns having c but no a nor b, m, p
  - Pattern f
- Completeness and non-redundency

Find Patterns Having P From P-conditional Database

- Starting at the frequent item header table in the FP-tree
- Traverse the FP-tree by following the link of each frequent item *p*
- Accumulate all of *transformed prefix paths* of item *p* to form *p*'s conditional pattern base

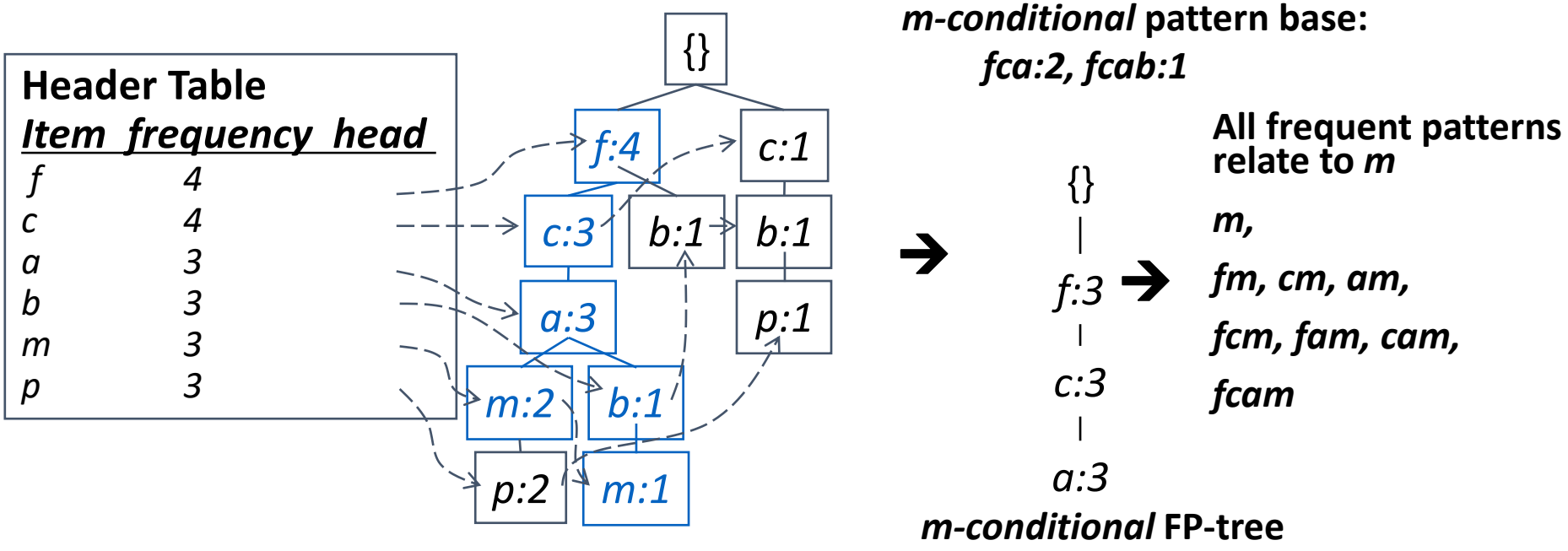


***Conditional* pattern bases**

item	*cond. pattern base*
c	***f:3***
a	***fc:3***
b	***fca:1, f:1, c:1***
m	***fca:2, fcab:1***
p	***fcam:2, cb:1***

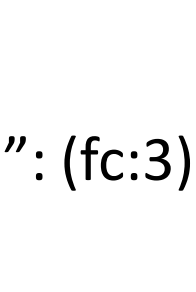
From Conditional Pattern-bases to Conditional FP-trees

- For each pattern-base
  - Accumulate the count for each item in the base
  - Construct the FP-tree for the frequent items of the pattern base



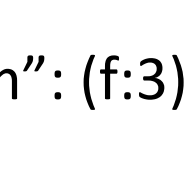
Recursion: Mining Each Conditional FP-tree

Cond. pattern base of “am”: (fc:3)

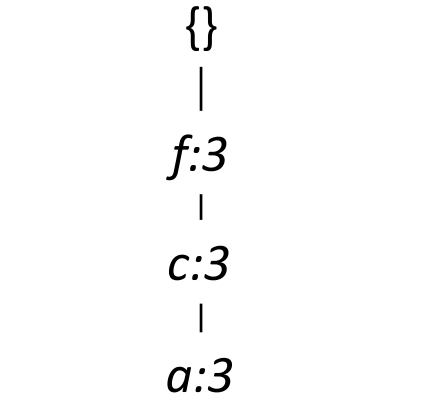


***am-conditional* FP-tree**

Cond. pattern base of “cm”: (f:3)



***cm-conditional* FP-tree**



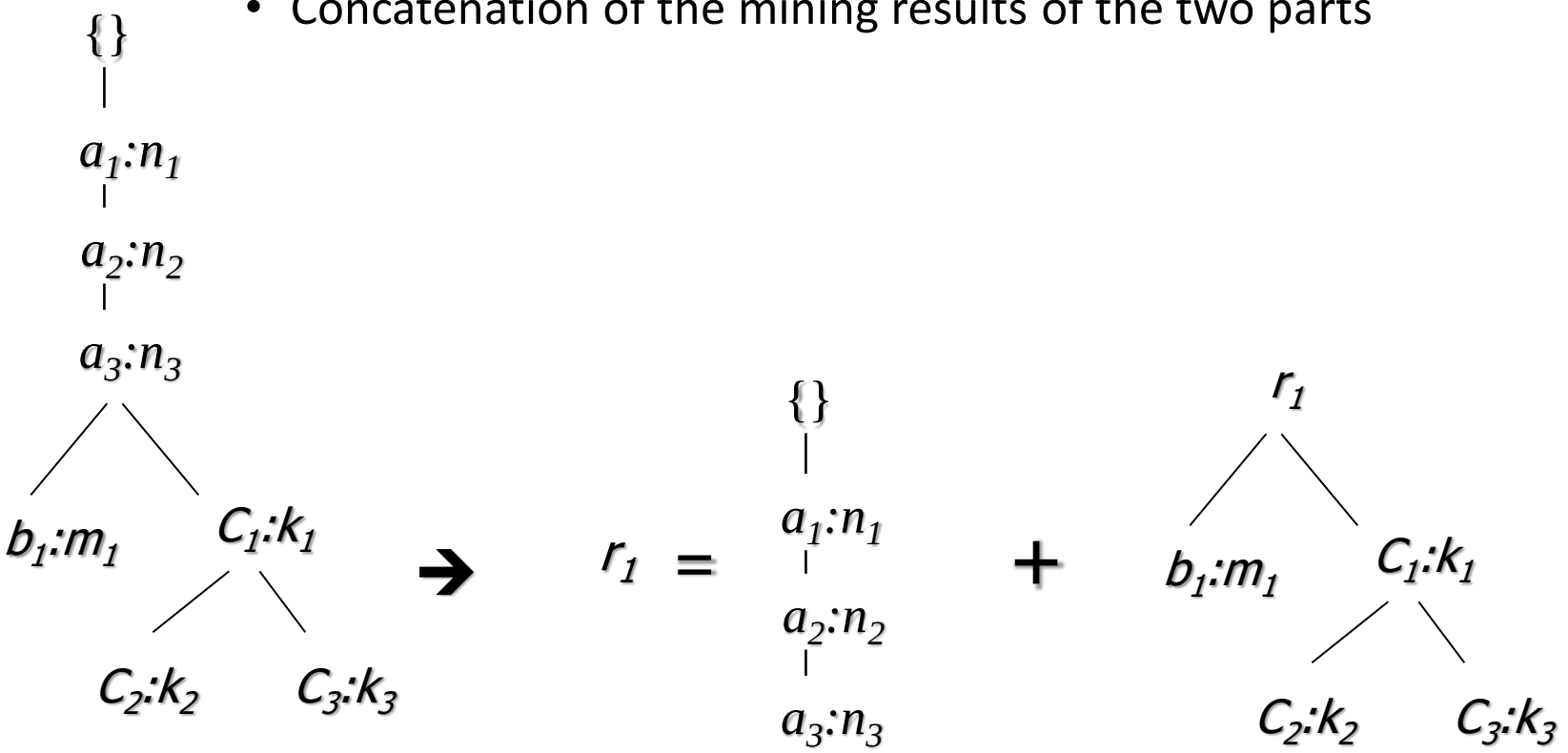
***m-conditional* FP-tree**

Cond. pattern base of “cam”: (f:3)

***cam-conditional* FP-tree**

A Special Case: Single Prefix Path in FP-tree

- Suppose a (conditional) FP-tree T has a shared single prefix-path P
- Mining can be decomposed into two parts
  - Reduction of the single prefix path into one node
  - Concatenation of the mining results of the two parts



Benefits of the FP-tree Structure

- Completeness
  - Preserve complete information for frequent pattern mining
  - Never break a long pattern of any transaction
- Compactness
  - Reduce irrelevant info—infrequent items are gone
  - Items in frequency descending order: the more frequently occurring, the more likely to be shared
  - Never be larger than the original database (not count node-links and the *count* field)

The Frequent Pattern Growth Mining Method

- Idea: Frequent pattern growth
  - Recursively grow frequent patterns by pattern and database partition

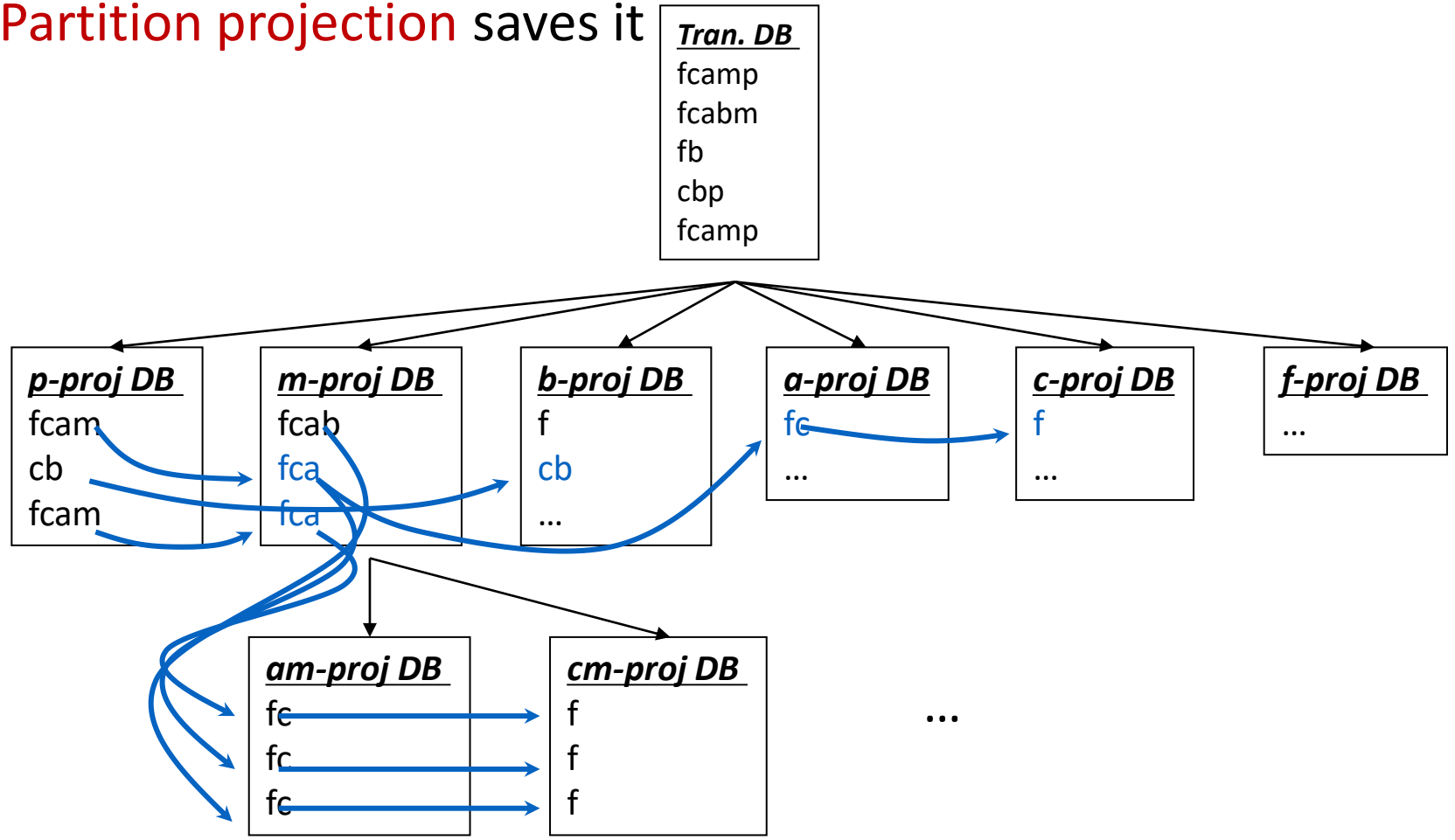
- Method
  1. For each frequent item, construct its conditional pattern-base, and then its conditional FP-tree
  2. Repeat the process on each newly created conditional FP-tree
  3. Until the resulting FP-tree is empty, or it contains only one path—single path will generate all the combinations of its sub-paths, each of which is a frequent pattern

Scaling FP-growth by Database Projection

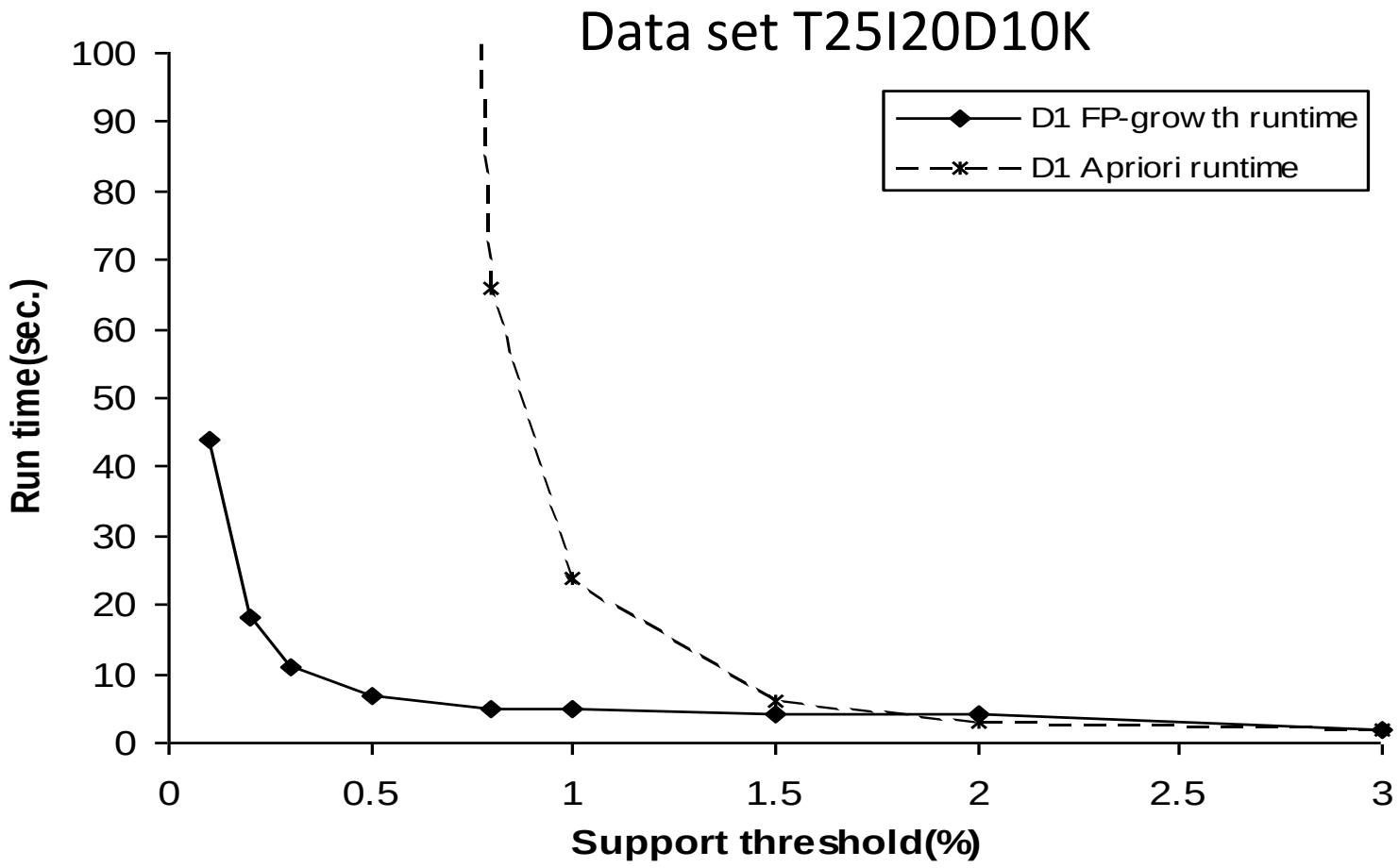
- What about if FP-tree cannot fit in memory? DB projection
- First partition a database into a set of projected DBs
- Then construct and mine FP-tree for each projected DB
- Parallel projection vs. partition projection techniques
  - Parallel projection
    - Project the DB in parallel for each frequent item
    - Parallel projection is space costly
    - All the partitions can be processed in parallel
  - Partition projection
    - Partition the DB based on the ordered frequent items
    - Passing the unprocessed parts to the subsequent partitions

Partition-Based Projection

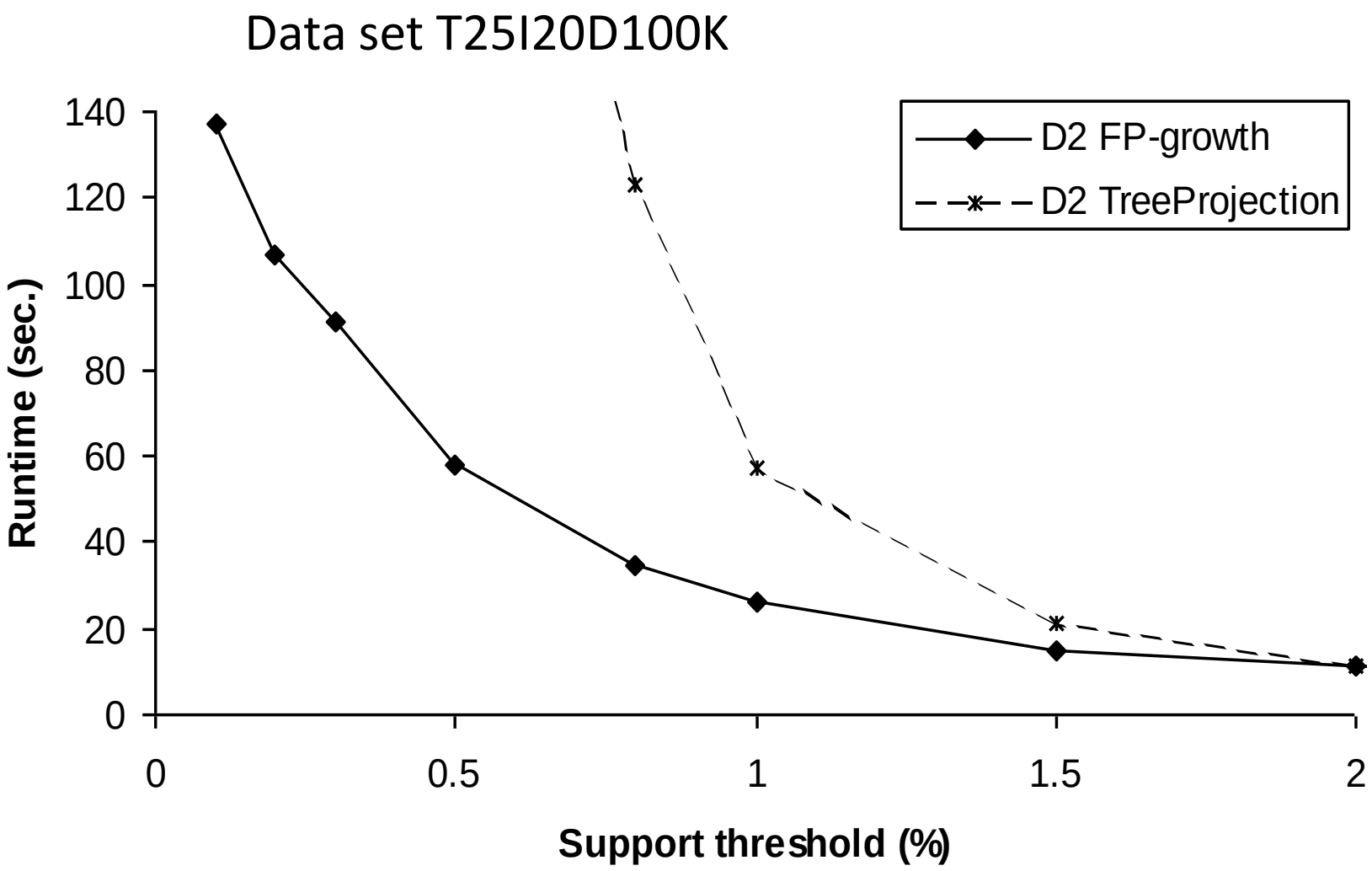
- Parallel projection needs a lot of disk space
- Partition projection saves it



FP-Growth vs. Apriori: Scalability With the Support Threshold



FP-Growth vs. Tree-Projection: Scalability with the Support Threshold



Advantages of the Pattern Growth Approach

- Divide-and-conquer:
  - Decompose both the mining task and DB according to the frequent patterns obtained so far
  - Lead to focused search of smaller databases
- Other factors
  - No candidate generation, no candidate test
  - Compressed database: FP-tree structure
  - No repeated scan of entire database
  - Basic ops: counting local freq items and building sub FP-tree, no pattern search and matching
- A good open-source implementation and refinement of FPGrowth
  - FPGrowth+ (Grahne and J. Zhu, FIMI'03)

Further Improvements of Mining Methods

- AFOPT *(Liu, et al. @ KDD'03)*
  - A "push-right" method for mining condensed frequent pattern (CFP) tree
- Carpenter *(Pan, et al. @ KDD'03)*
  - Mine data sets with small rows but numerous columns
  - Construct a row-enumeration tree for efficient mining
- FPgrowth+ *(Grahne and Zhu, FIMI'03)*
  - Efficiently Using Prefix-Trees in Mining Frequent Itemsets, Proc. ICDM'03 Int. Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, FL, Nov. 2003
- TD-Close *(Liu, et al, SDM'06)*

Extension of Pattern Growth Mining Methodology

- Mining closed frequent itemsets and max-patterns
  - CLOSET (DMKD'00), FPclose, and FPMax (Grahne & Zhu, Fimi'03)
- Mining sequential patterns
  - PrefixSpan (ICDE'01), CloSpan (SDM'03), BIDE (ICDE'04)
- Mining graph patterns
  - gSpan (ICDM'02), CloseGraph (KDD'03)
- Constraint-based mining of frequent patterns
  - Convertible constraints (ICDE'01), gPrune (PAKDD'03)
- Computing iceberg data cubes with complex measures
  - H-tree, H-cubing, and Star-cubing (SIGMOD'01, VLDB'03)
- Pattern-growth-based Clustering
  - MaPle (Pei, et al., ICDM'03)
- Pattern-Growth-Based Classification
  - Mining frequent and discriminative patterns (Cheng, et al, ICDE'07)

Tahapan Algoritma FP Growth

1. Penyiapan Dataset
2. Pencarian Frequent Itemset (Item yang sering muncul)
3. Dataset diurutkan Berdasarkan Priority
4. Pembuatan FP-Tree Berdasarkan Item yang sudah diurutkan
5. Pembangkitan Conditional Pattern Base
6. Pembangkitan Conditional FP-tree
7. Pembangkitan Frequent Pattern
8. Mencari Support
9. Mencari Confidence

1. Penyiapan Dataset

Customer	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1	1	0	0	0	1	1	0	0	0
2	1	1	0	1	1	0	0	1	1	1
3	0	0	0	1	1	0	0	0	0	1
4	1	0	1	0	0	0	0	0	0	0
5	0	0	1	1	0	0	1	0	0	0
6	1	1	0	0	0	1	0	0	0	0
7	0	0	0	0	1	0	0	0	1	1
8	0	0	1	1	1	1	1	1	0	0
9	1	1	0	0	0	0	0	0	1	0
10	0	0	1	0	0	0	1	0	0	0
11	1	1	1	0	0	0	0	0	0	0
12	0	0	0	0	1	1	1	0	0	0
Frequent	6	5	5	4	5	4	5	2	3	3
Priority	1	2	3	6	4	7	5	10	8	9

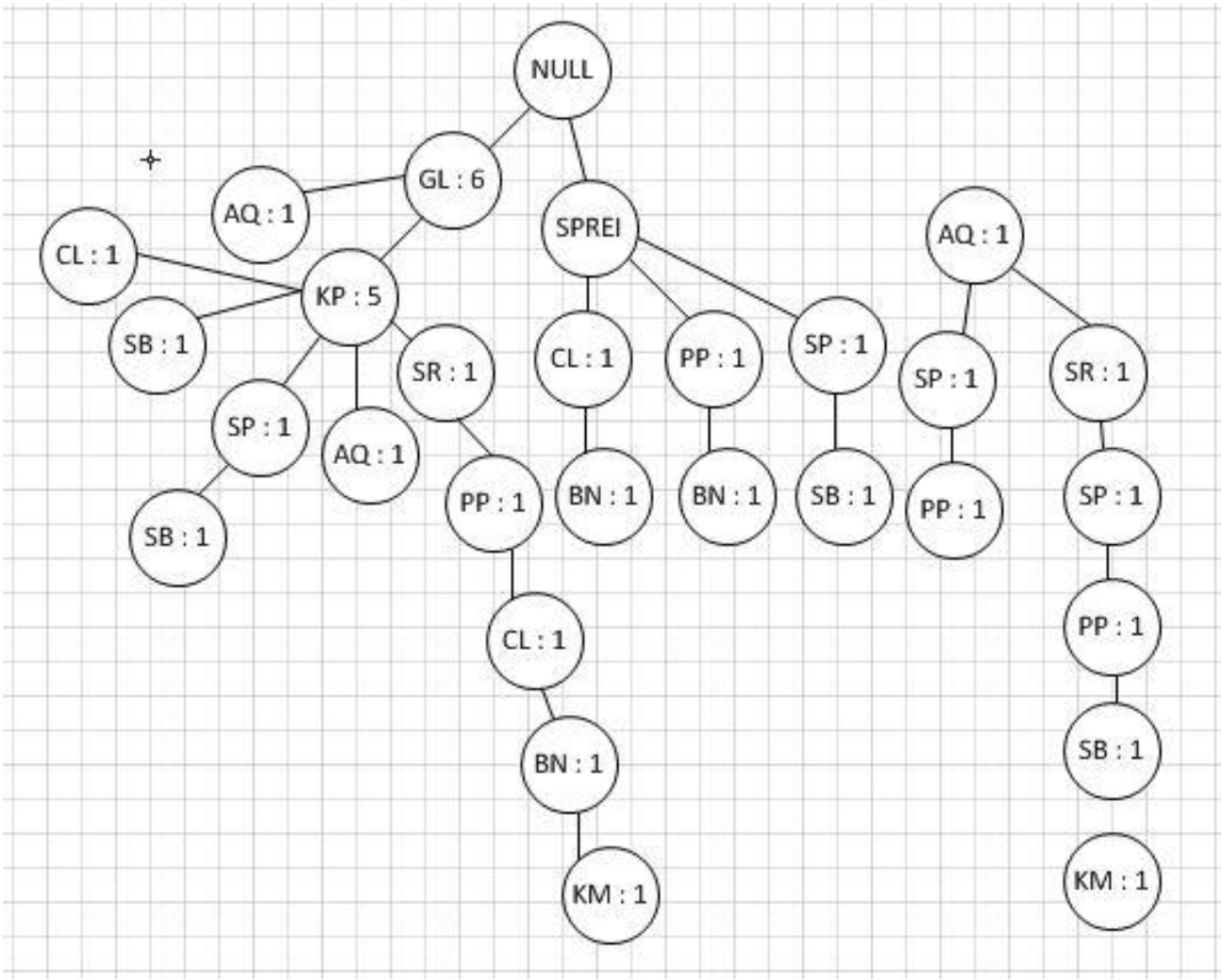
2. Pencarian Frequent Itemset

Frequent	
Gula	6
Kopi	5
Aqua	5
Popok	4
Sprei	5
Sabun	4
Sampo	5
Kemeja	2
Celana	3
Boneka	3

3. Dataset diurutkan Berdasarkan Priority

Customer	Item	Urutan
1	GL, KP, SB, SP	GL, KP, SP, SB
2	GL, KP,PP, SR, KM, CL, BN	GL, KP, SR, PP, CL, BN, KM
3	PP, SR, BN	SR, PP, BN
4	GL, AQ	GL, AQ
5	AQ, PP, SP	AQ, SP, PP
6	GL, KP, SB	GL, KP, SB
7	SR, CL, BN	SR, CL, BN
8	AQ, PP, SR, SB, SP, KM	AQ, SR, SP, PP, SB, KM
9	GL, KP, CL	GL, KP, CL
10	AQ, SP	AQ, SP
11	GL, KP, AQ	GL, KP, AQ
13	SR, SB, SP	SR, SP, SB

4. Pembuatan FP-Tree



5. Pembangkitan Conditional Pattern Base

Conditional Pattern Base
Kemeja : {{GL, KP, SR, PP, CL, BN :1}, {AQ, SR, SP, PP, SB :1}}
Boneka : {{GL, KP, SR, PP, CL :1}, {SR, CL:1}, {SR,PP}}
Celana : {{GL, KP, SR, PP:1}, {GL, KP :1}, {SR:1}}
Sabun : {{GL, KP, SP :1}, {GL, KP :1}, {AQ, SR, SP, PP :1}, {SR, SP :1}}
Popok : {{GL, KP, SR :1}, {SR :1}, {AQ, SP :1}, {AQ, SR, SP :1}}
Sampo : {{GL, KP :1}, {AQ :2}, {AQ, SR :1}, {SR :1}}
Sprei : {{GL, KP :1}, {AQ :1}}
Aqua : {{GL :1}, {GL, KP :1}}
Kopi : {GL : 5}

6. Pembangkitan Conditional FP-tree

Conditional FP Tree
Kemeja : (GL :1, KP : 1), (AQ : 1, SR :1)
Boneka : (GL :1, KP : 1), (SR :2)
Celana : (GL : 2, KP :2), (SR :1)
Sabun : (GL :2, KP :2), (SR : 1), (AQ :1)
Popok : (SR :1, AQ :2)
Sampo : (GL :1, KP :1),
Sprei : (GL : 1, KP :2), (SR :1)
Aqua : (GL :2, KP :2)
Kopi : (GL :1)

7. Pembangkitan Frequent Pattern

Pattern Generated
Kemeja : {GL, KM : 1}, {KP, KM : 1}, {AQ, KM : 1}
Boneka : {GL, BN : 1}, {SR, BN : 1}
Celana : {GL, CL : 2}, {KP, CL : 2}
Sabun : {GL, SB : 2}, {KP, SB : 2}, {SR, SB : 1}, {AQ, SB : 1}
Popok : {GL, PP : 1}, {KP, PP : 1}, {SR, PP : 1}, {AQ, PP :1}
Sampo : {GL, SP : 1}, {KP, SP : 1}, {AQ, SP : 2}, {SR, SP : 1}
Sprei : {GL, SR : 1}, {KP, SR : 1}, {AQ, SR : 1}
Aqua : {GL, AQ : 2}, {KP, AQ : 2}
Kopi : {GL, KP : 5}

Frequent 2 Itemset

Frequent 2 Item	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
Gula	6	5	2	1	1	2	1	1	2	1
Kopi	5	5	1	1	1	2	1	1	2	1
Aqua	2	1	5	2	1	1	3	1	0	0
Popok	1	1	2	4	3	1	2	2	1	2
Sprei	1	1	1	3	5	2	2	2	2	3
Sabun	2	2	1	1	2	4	3	1	0	0
Sampo	1	1	3	2	2	3	5	1	0	0
Kemeja	1	1	1	2	2	1	1	2	1	1
Celana	2	2	0	1	2	0	0	1	3	2
Boneka	1	1	0	2	3	0	0	1	2	3

8. Mencari Support 2 Itemset

Support (2 Items)	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
Gula	0,5	0,4167	0,1667	0,0833	0,0833	0,16667	0,0833	0,08333	0,16667	0,08333
Kopi	0,4167	0,4167	0,0833	0,0833	0,0833	0,16667	0,0833	0,08333	0,16667	0,08333
Aqua	0,1667	0,0833	0,4167	0,1667	0,0833	0,08333	0,25	0,08333	0	0
Popok	0,0833	0,0833	0,1667	0,3333	0,25	0,08333	0,1667	0,16667	0,08333	0,16667
Sprei	0,0833	0,0833	0,0833	0,25	0,4167	0,16667	0,1667	0,16667	0,16667	0,25
Sabun	0,1667	0,1667	0,0833	0,0833	0,1667	0,33333	0,25	0,08333	0	0
Sampo	0,0833	0,0833	0,25	0,1667	0,1667	0,25	0,4167	0,08333	0	0
Kemeja	0,0833	0,0833	0,0833	0,1667	0,1667	0,08333	0,0833	0,16667	0,08333	0,08333
Celana	0,1667	0,1667	0	0,0833	0,1667	0	0	0,08333	0,25	0,16667
Boneka	0,0833	0,0833	0	0,1667	0,25	0	0	0,08333	0,16667	0,25

9. Mencari Confidence 2 Itemset

Confidence	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
Gula	1	0,833333	0,333333	0,166667	0,16666667	0,333333	0,166667	0,166667	0,333333	0,166667
Kopi	1	1	0,2	0,2	0,2	0,4	0,2	0,2	0,4	0,2
Aqua	0,4	0,2	1	0,4	0,2	0,2	0,6	0,2	0	0
Popok	0,25	0,25	0,5	1	0,75	0,25	0,5	0,5	0,25	0,5
Sprei	0,2	0,2	0,2	0,6	1	0,4	0,4	0,4	0,4	0,6
Sabun	0,5	0,5	0,25	0,25	0,5	1	0,75	0,25	0	0
Sampo	0,2	0,2	0,6	0,4	0,4	0,6	1	0,2	0	0
Kemeja	0,5	0,5	0,5	1	1	0,5	0,5	1	0,5	0,5
Celana	0,666667	0,666667	0	0,333333	0,66666667	0	0	0,333333	1	0,666667
Boneka	0,333333	0,333333	0	0,666667	1	0	0	0,333333	0,666667	1

4.3.2 Pattern Evaluation Methods

Interestingness Measure: Correlations (Lift)

- *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%] is misleading
  - The overall % of students eating cereal is 75% > 66.7%.
- *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%] is more accurate, although with lower support and confidence
- Measure of dependent/correlated events: lift

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

$$lift(B, C) = \frac{2000/5000}{3000/5000 * 3750/5000} = 0.89$$

$$lift(B, \neg C) = \frac{1000/5000}{3000/5000 * 1250/5000} = 1.33$$

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

Are *lift* and χ^2 Good Measures of Correlation?

- *"Buy walnuts $\Rightarrow$ buy milk* [1%, 80%]" is misleading if 85% of customers buy milk
- Support and confidence are not good to indicate correlations
- Over 20 interestingness measures have been proposed (see Tan, Kumar, Sritastava @KDD'02)
- Which are good ones?

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1 . . . 1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 . . . 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 . . . 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 . . . 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 . . . 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 . . . 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 . . . 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 . . . 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 . . . 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 . . . 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 . . . 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$
			$P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 . . . 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$
			$P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 . . . 1	$P(A,B)$
c	confidence	0 . . . 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 . . . 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 . . . 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 . . . 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 . . . 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 . . . ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 . . . ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 . . . ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 . . . ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 . . . ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Null-Invariant Measures

Table 6: Properties of interestingness measures. Note that none of the measures satisfies all the properties.

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
ϕ	ϕ-coefficient	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0 \cdots 1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	$0 \cdots 1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0 \cdots 1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0 \cdots 1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0 \cdots 1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0 \cdots 1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0 \cdots 1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5 \cdots 1 \cdots \infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0 \cdots \sqrt{P(A,B)} \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25 \cdots 0 \cdots 0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0 \cdots 1 \cdots \infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0 \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}} - 1)^{1/2}[2 - \sqrt{3} - \frac{1}{\sqrt{3}}] \cdots 0 \cdots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No**	No	No	No	No

where: P1: $O(\mathbf{M}) = 0$ if $det(\mathbf{M}) = 0$, *i.e.*, whenever A and B are statistically independent.
P2: $O(\mathbf{M_2}) > O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [k \ -k; \ -k \ k]$.
P3: $O(\mathbf{M_2}) < O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [0 \ k; \ 0 \ -k]$ or $\mathbf{M_2} = \mathbf{M_1} + [0 \ 0; \ k \ -k]$.
O1: Property 1: Symmetry under variable permutation.
O2: Property 2: Row and Column scaling invariance.
O3: Property 3: Antisymmetry under row or column permutation.
O3': Property 4: Inversion invariance.
O4: Property 5: Null invariance.
Yes*: Yes if measure is normalized.
No*: Symmetry under row or column permutation.
No**: No unless the measure is symmetrized by taking $\max(M(A,B), M(B,A))$.

Comparison of Interestingness Measures

- Null-(transaction) invariance is crucial for correlation analysis
- Lift and χ^2 are not null-invariant
- 5 null-invariant measures

	Milk	No Milk	Sum (row)
Coffee	m, c	~m, c	c
No Coffee	m, ~c	~m, ~c	~c
Sum(col.)	m	~m	Σ

Measure	Definition	Range	Null-Invariant
$\chi^2(a,b)$	$\sum_{i,j=0,1} \frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(a,b)$	$\frac{P(ab)}{P(a)P(b)}$	$[0,\infty]$	No
$AllConf(a,b)$	$\frac{sup(ab)}{max\{sup(a),sup(b)\}}$	$[0,1]$	Yes
$Coherence(a,b)$	$\frac{sup(ab)}{sup(a)+sup(b)-sup(ab)}$	$[0,1]$	Yes
$Cosine(a,b)$	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	$[0,1]$	Yes
$Kulc(a,b)$	$\frac{sup(ab)}{2}(\frac{1}{sup(a)}+\frac{1}{sup(b)})$	$[0,1]$	Yes
$MaxConf$(a,b)	$max\{\frac{sup(ab)}{sup(a)}, \frac{sup(ab)}{sup(b)}\}$	$[0,1]$	Yes

Table 3. Interestingness measure definitions.

Null-transactions w.r.t. m and c

Kulczynski measure (1927)

Null-invariant

Data set	mc	$\overline{mc}$	$m\overline{c}$	$\overline{mc}$	χ^2	$Lift$	$AllConf$	$Coherence$	$Cosine$	$Kulc$	$MaxConf$
D_1	10,000	1,000	1,000	100,000	90557	9.26	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0	1	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	670	8.44	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	24740	25.75	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	8173	9.18	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	965	1.97	0.01	0.01	0.10	0.5	0.99

Table 2. Example data sets.

Subtle: They disagree

Analysis of DBLP Coauthor Relationships

Recent DB conferences, removing balanced associations, low sup, etc.

ID	Author a	Author b	$sup(ab)$	$sup(a)$	$sup(b)$	$Coherence$	$Cosine$	$Kulc$
1	Hans-Peter Kriegel	Martin Ester	28	146	54	0.163 (2)	0.315 (7)	0.355 (9)
2	Michael Carey	Miron Livny	26	104	58	0.191 (1)	0.335 (4)	0.349 (10)
3	Hans-Peter Kriegel	Joerg Sander	24	146	36	0.152 (3)	0.331 (5)	0.416 (8)
4	Christos Faloutsos	Spiros Papadimitriou	20	162	26	0.119 (7)	0.308 (10)	0.446 (7)
5	Hans-Peter Kriegel	Martin Pfeifle	18	146	18	0.123 (6)	0.351 (2)	0.562 (2)
6	Hector Garcia-Molina	Wilburt Labio	16	144	18	0.110 (9)	0.314 (8)	0.500 (4)
7	Divyakant Agrawal	Wang Hsiung	16	120	16	0.133 (5)	0.365 (1)	0.567 (1)
8	Elke Rundensteiner	Murali Mani	16	104	20	0.148 (4)	0.351 (3)	0.477 (6)
9	Divyakant Agrawal	Oliver Po	12	120	12	0.100 (10)	0.316 (6)	0.550 (3)
10	Gerhard Weikum	Martin Theobald	12	106	14	0.111 (8)	0.312 (9)	0.485 (5)

Table 5. Experiment on DBLP data set.

Advisor-advisee relation: Kulc: high, coherence: low, cosine: middle

Tianyi Wu, Yuguo Chen and Jiawei Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", Proc. 2007 Int. Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Sept. 2007

Which Null-Invariant Measure Is Better?

- IR (Imbalance Ratio): measure the imbalance of two itemsets A and B in rule implications

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

- Kulczynski and Imbalance Ratio (IR) together present a clear picture for all the three datasets D_4 through D_6
  - D_4 is balanced & neutral
  - D_5 is imbalanced & neutral
  - D_6 is very imbalanced & neutral

Data	*mc*	$\overline{m}c$	$m\overline{c}$	$\overline{mc}$	*all_conf.*	*max_conf.*	*Kulc.*	*cosine*	IR
D_1	10,000	1,000	1,000	100,000	0.91	0.91	0.91	0.91	0.0
D_2	10,000	1,000	1,000	100	0.91	0.91	0.91	0.91	0.0
D_3	100	1,000	1,000	100,000	0.09	0.09	0.09	0.09	0.0
D_4	1,000	1,000	1,000	100,000	0.5	0.5	0.5	0.5	0.0
D_5	1,000	100	10,000	100,000	0.09	0.91	0.5	0.29	0.89
D_6	1,000	10	100,000	100,000	0.01	0.99	0.5	0.10	0.99

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 5 (Association Rules), p 85-97

- Analisis, bagaimana data mining bisa bermanfaat untuk membantu Roger, seorang City Manager

1. Business Understanding

- Motivation:
  - Roger is a city manager for a medium-sized, but steadily growing city
  - The city has limited resources, and like most municipalities, there are more needs than there are resources
  - He feels like the citizens in the community are fairly active in various community organizations, and believes that he may be able to get a number of groups to work together to meet some of the needs in the community
  - He knows there are churches, social clubs, hobby enthusiasts and other types of groups in the community
  - What he doesn't know is if there are connections between the groups that might enable natural collaborations between two or more groups that could work together on projects around town

- Objectives:
  - To find out if there are any existing associations between the different types of groups in the area

4.4 Algoritma Estimasi dan Forecasting

4.4.1 Linear Regression

4.4.2 Time Series Forecasting

4.4.1 Linear Regression

Tahapan Algoritma Linear Regression

1. Siapkan data
2. Identifikasi Atribut dan Label
3. Hitung X^2, Y^2, XY dan total dari masing-masingnya
4. Hitung a dan b berdasarkan persamaan yang sudah ditentukan
5. Buat Model Persamaan Regresi Linear Sederhana

1. Persiapan Data

Tanggal	Rata-rata Suhu Ruangan (X)	Jumlah Cacat (Y)
1	24	10
2	22	5
3	21	6
4	20	3
5	22	6
6	19	4
7	20	5
8	23	9
9	24	11
10	25	13

2. Identifikasikan Atribut dan Label

$$\mathbf{Y = a + bX}$$

Dimana:

Y = Variabel terikat (Dependen)
X = Variabel tidak terikat (Independen)
a = konstanta
b = koefisien regresi (kemiringan); besaran Response yang ditimbulkan oleh variabel

$$a = \frac{(\Sigma y)\,(\Sigma x^2) - (\Sigma x)\,(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)\,(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

3. Hitung X^2, Y^2, XY dan total dari masing-masingnya

Tanggal	Rata-rata Suhu Ruangan (X)	Jumlah Cacat (Y)	X^2	Y^2	XY
1	24	10	576	100	240
2	22	5	484	25	110
3	21	6	441	36	126
4	20	3	400	9	60
5	22	6	484	36	132
6	19	4	361	16	76
7	20	5	400	25	100
8	23	9	529	81	207
9	24	11	576	121	264
10	25	13	625	169	325
	220	72	4876	618	1640

4. Hitung a dan b berdasarkan persamaan yang sudah ditentukan

- Menghitung Koefisien Regresi (a)

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{(72)(4876) - (220)(1640)}{10(4876) - (220)^2}$$

a = -27,02

- Menghitung Koefisien Regresi (b)

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{10(1640) - (220)(72)}{10(4876) - (220)^2}$$

b = 1,56

5. Buatkan Model Persamaan Regresi Linear Sederhana

$$Y = a + bX$$

$$Y = -27{,}02 + 1{,}56X$$

Pengujian

1. Prediksikan Jumlah Cacat Produksi jika suhu dalam keadaan tinggi (Variabel X), contohnya: 30°C

   Y = -27,02 + 1,56X

   Y = -27,02 + 1,56(30)

   =19,78

2. Jika Cacat Produksi (Variabel Y) yang ditargetkan hanya boleh 5 unit, maka berapakah suhu ruangan yang diperlukan untuk mencapai target tersebut?

   5= -27,02 + 1,56X

   1,56X = 5+27,02

   X= 32,02/1,56

   X =20,52

   Jadi Prediksi Suhu Ruangan yang paling sesuai untuk mencapai target Cacat Produksi adalah sekitar **$20,52^0$C**

7.1.2 Studi Kasus CRISP-DM

Heating Oil Consumption – Estimation

(Matthew North, Data Mining for the Masses, 2012,
Chapter 8 Estimation, pp. 127-140*)*

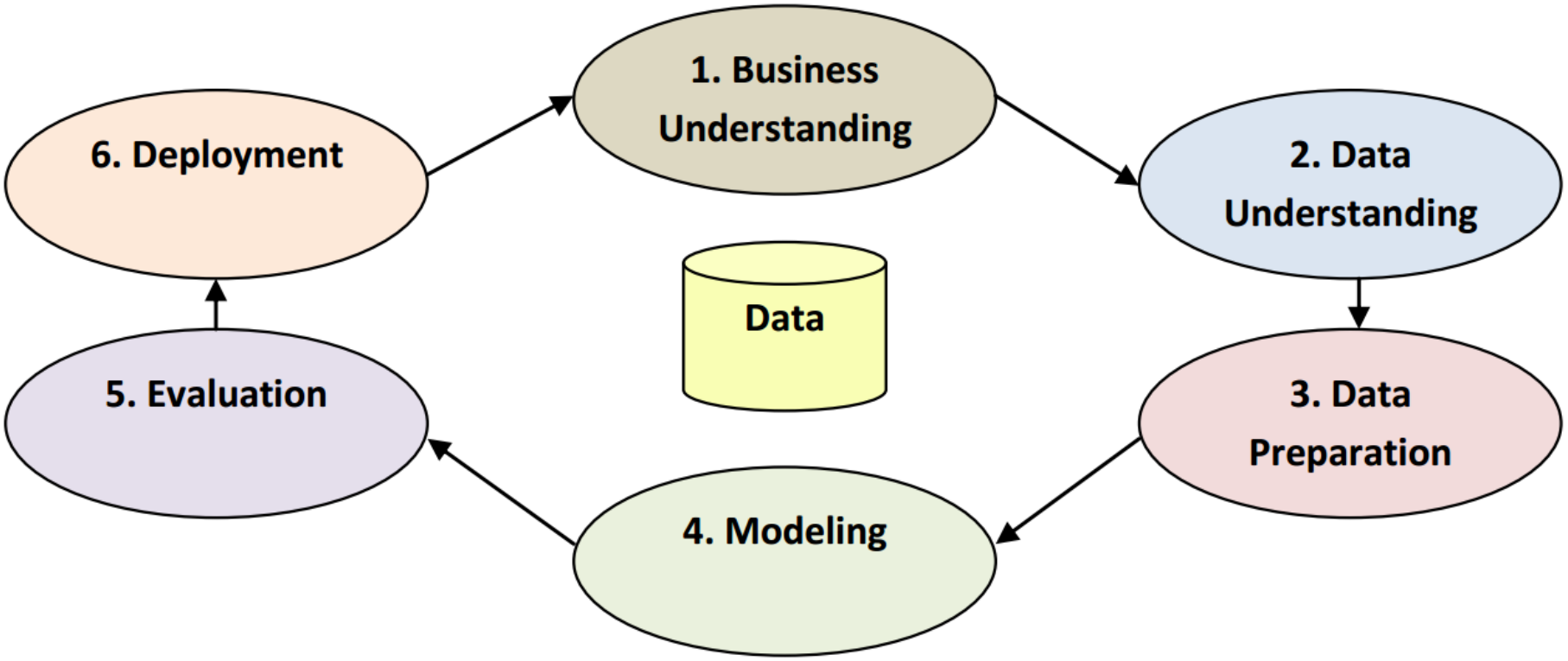
Dataset: HeatingOil-Training.csv dan HeatingOil-Scoring.csv

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses, 2012, Chapter 8 Estimation, pp. 127-140 tentang Heating Oil Consumption

- Dataset: HeatingOil-Training.csv dan HeatingOil-Scoring.csv

CRISP-DM


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

Context and Perspective

- Sarah, the regional sales manager is back for more help
- Business is booming, her sales team is signing up thousands of new clients, and she wants to be sure the company will be able to meet this new level of demand, she now is hoping we can help her do some prediction as well
- She knows that there is some correlation between the attributes in her data set (things like temperature, insulation, and occupant ages), and she's now wondering if she can use the previous data set to predict heating oil usage for new customers
- You see, these new customers haven't begun consuming heating oil yet, there are a lot of them (42,650 to be exact), and she wants to know how much oil she needs to expect to keep in stock in order to meet these new customers' demand
- Can she use data mining to examine household attributes and known past consumption quantities to anticipate and meet her new customers' needs?

1. Business Understanding

- Sarah's new data mining objective is pretty clear: she wants to anticipate demand for a consumable product
- We will use a linear regression model to help her with her desired predictions
- She has data, 1,218 observations that give an attribute profile for each home, along with those homes' annual heating oil consumption
- She wants to use this data set as training data to predict the usage that 42,650 new clients will bring to her company
- She knows that these new clients' homes are similar in nature to her existing client base, so the existing customers' usage behavior should serve as a solid gauge for predicting future usage by new customers.

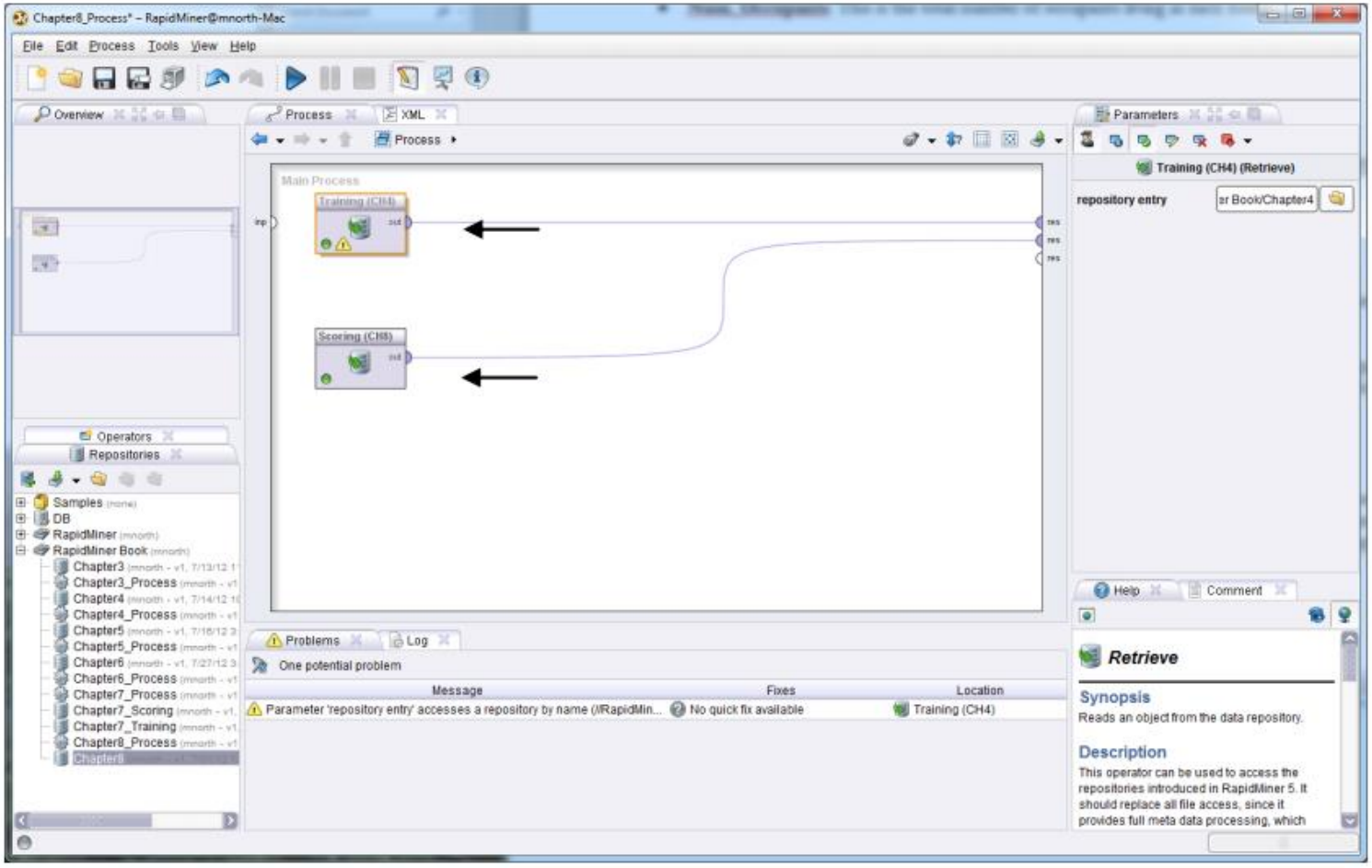
2. Data Understanding

We create a data set comprised of the following attributes:

- Insulation: This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation
- Temperature: This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit
- Heating_Oil: This is the total number of units of heating oil purchased by the owner of each home in the most recent year
- Num_Occupants: This is the total number of occupants living in each home
- Avg_Age: This is the average age of those occupants
- Home_Size: This is a rating, on a scale of one to eight, of the home's overall size. The higher the number, the larger the home

3. Data Preparation

- A CSV data set for this chapter's example is available for download at the book's companion web site (https://sites.google.com/site/dataminingforthemasses/)



3. Data Preparation

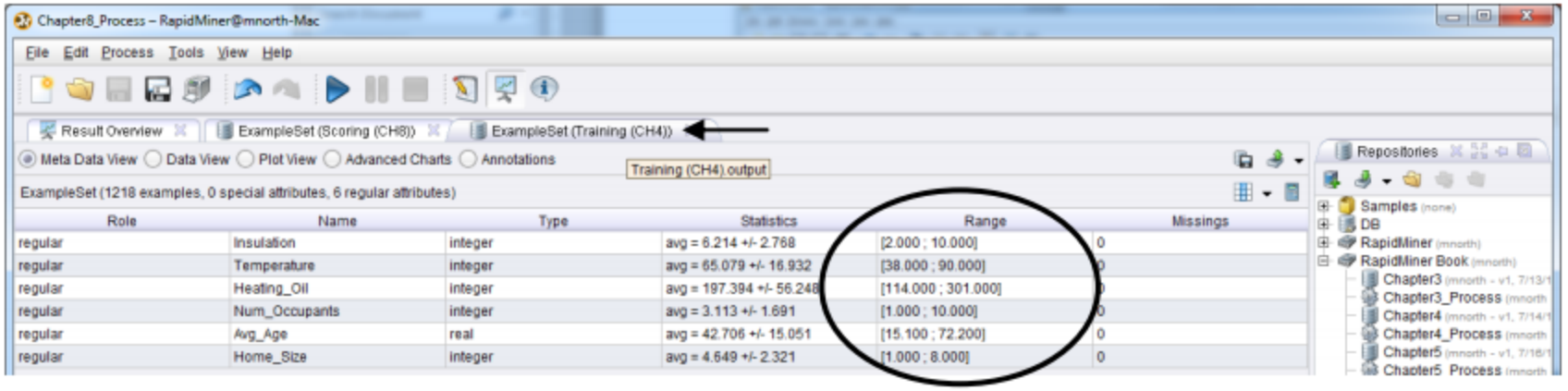


Figure 8-2. Value ranges for the training data set's attributes.

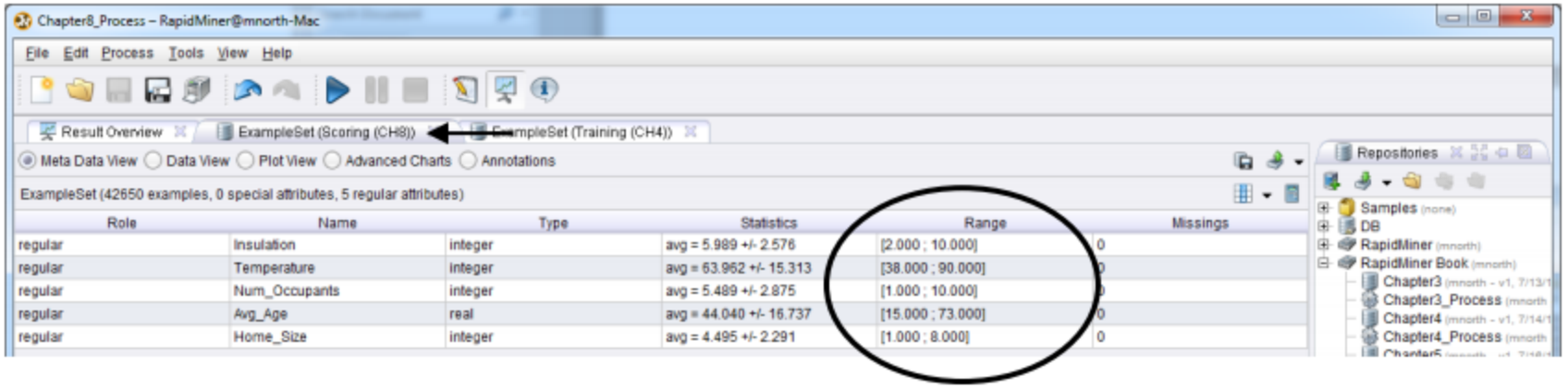
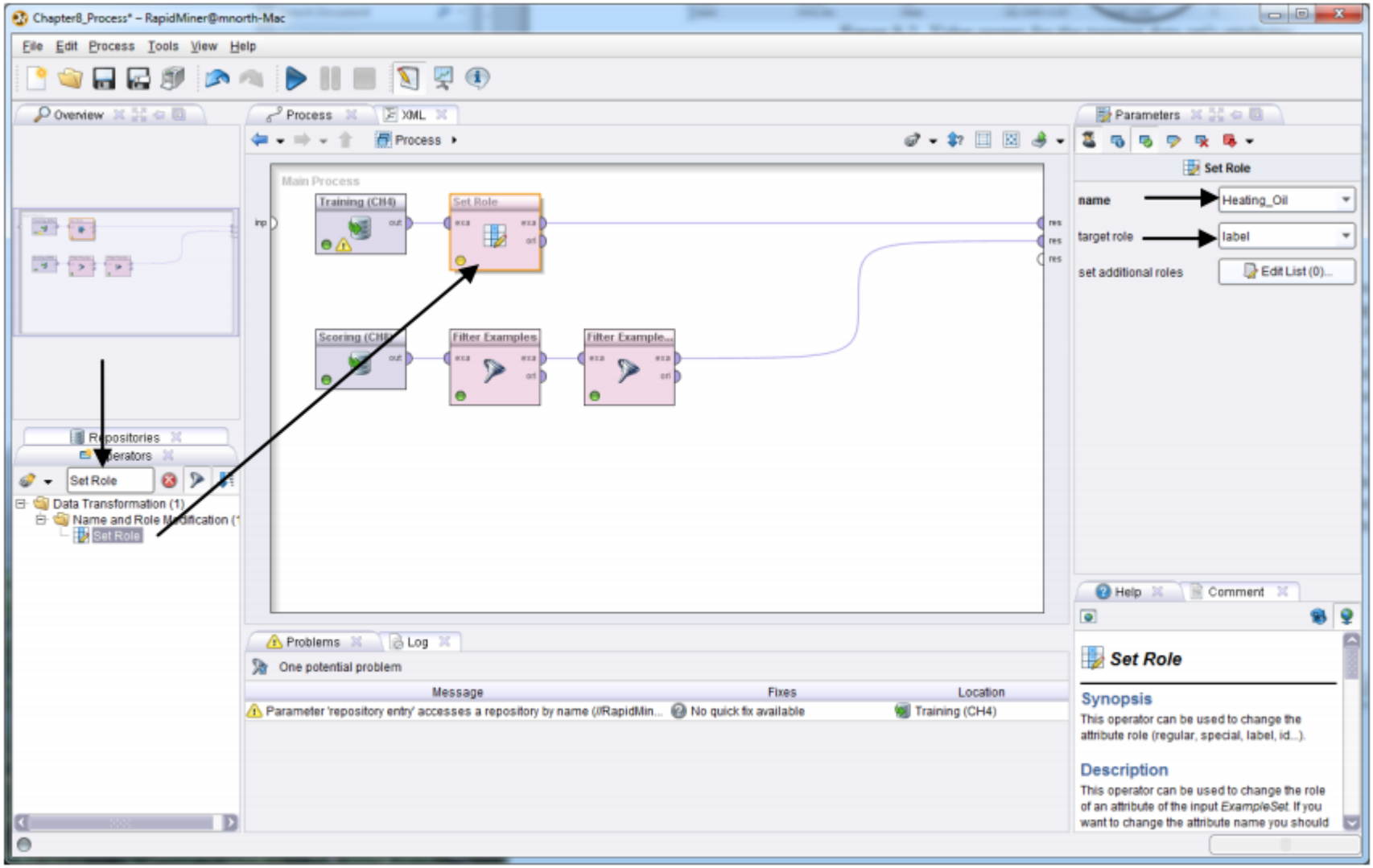
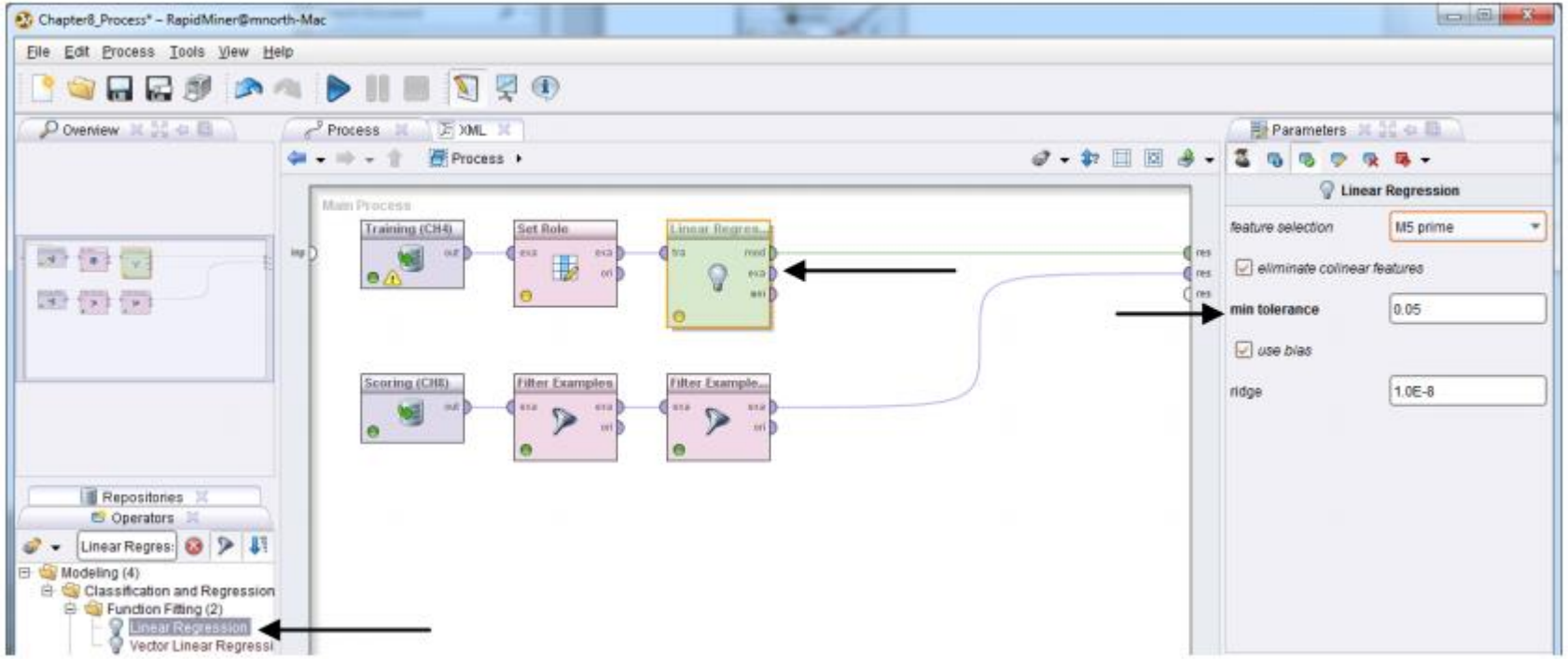


Figure 8-3. Value ranges for the scoring data set's attributes.

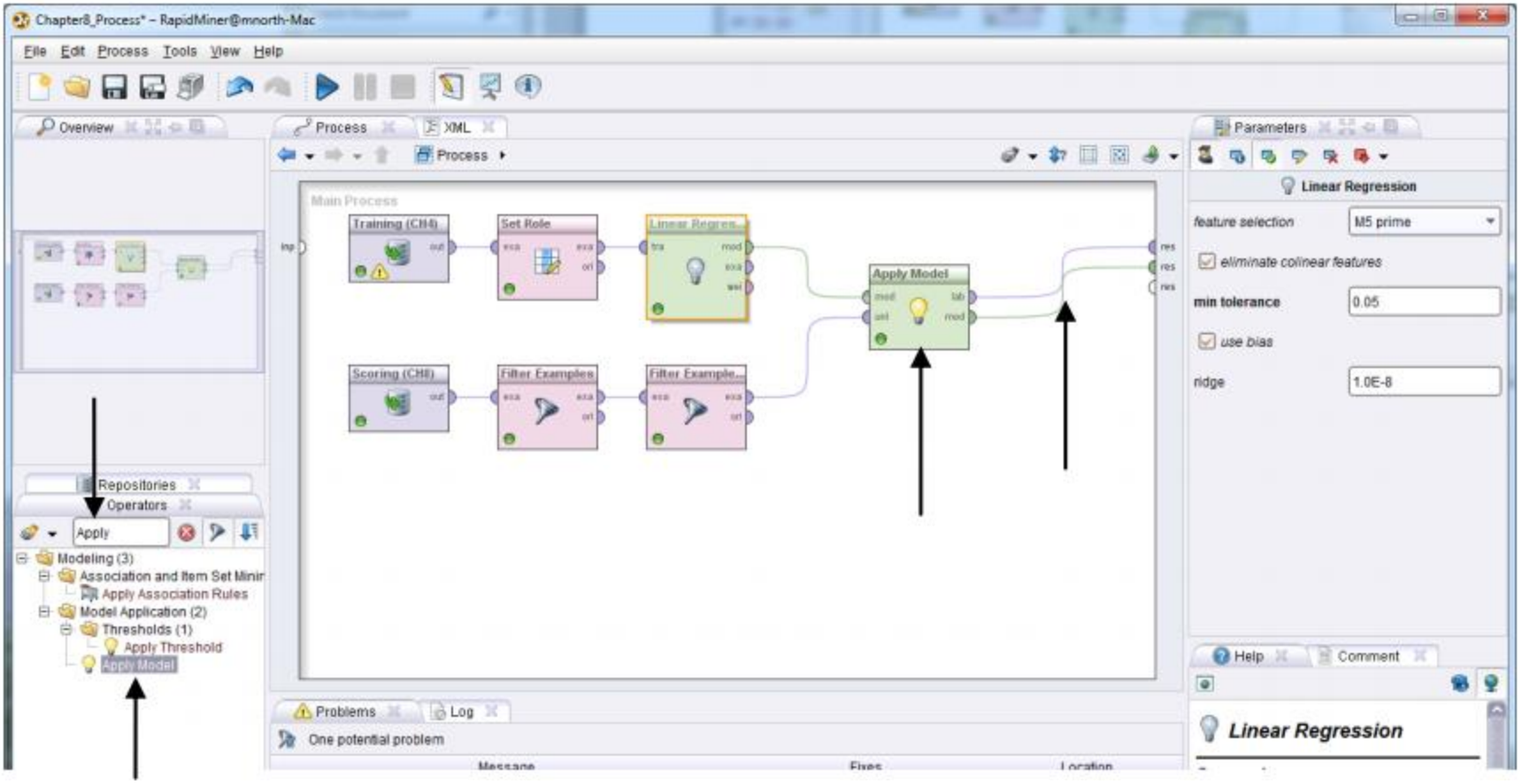
3. Data Preparation



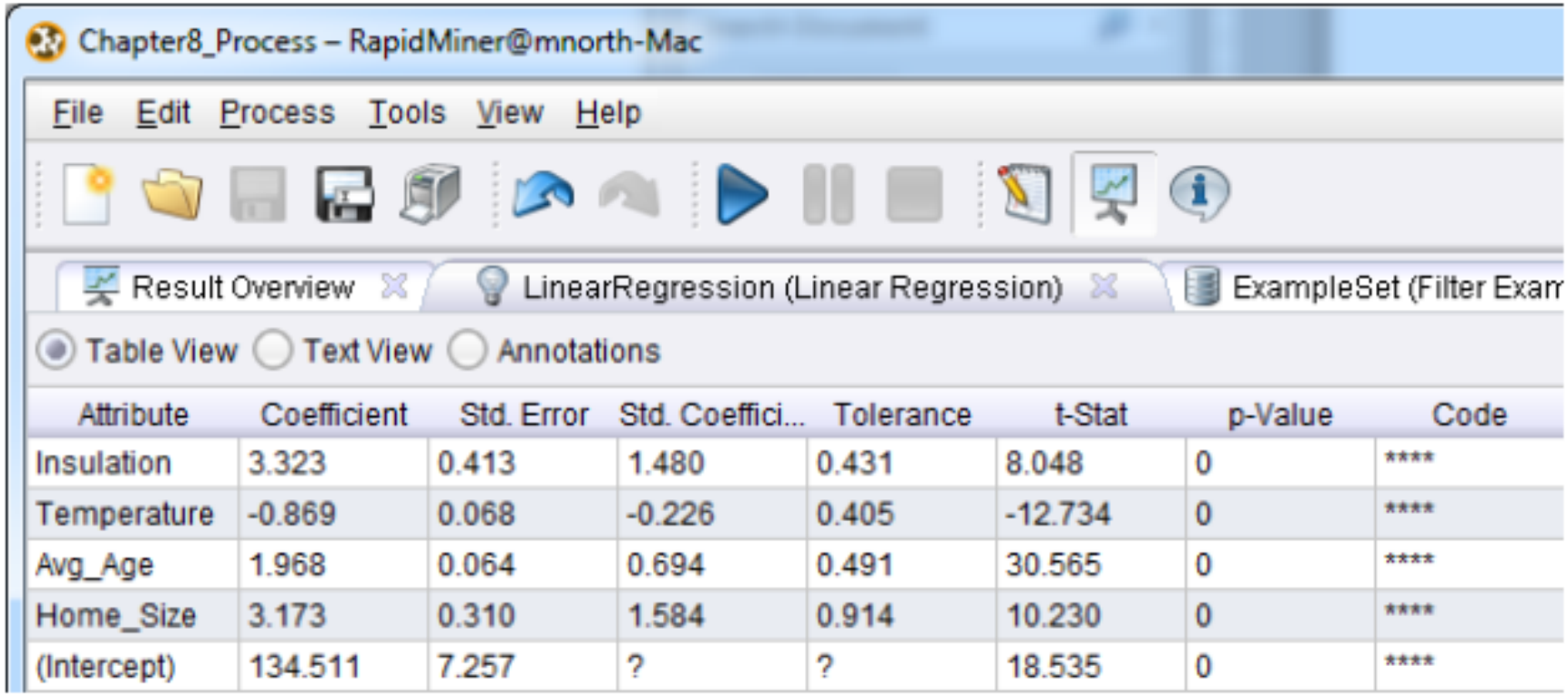
4. Modeling


Chapter8_Process* – RapidMiner@mnorth-Mac
File Edit Process Tools View Help
Overview
Process
XML
Process
Main Process
Training (CH4)
Set Role
Linear Regres...
Scoring (CH8)
Filter Examples
Filter Example...
Parameters
Linear Regression
feature selection
M5 prime
eliminate colinear features
min tolerance
0.05
use bias
ridge
1.0E-8
Repositories
Operators
Linear Regres
Modeling (4)
Classification and Regression
Function Fitting (2)
Linear Regression
Vector Linear Regressi

4. Modeling

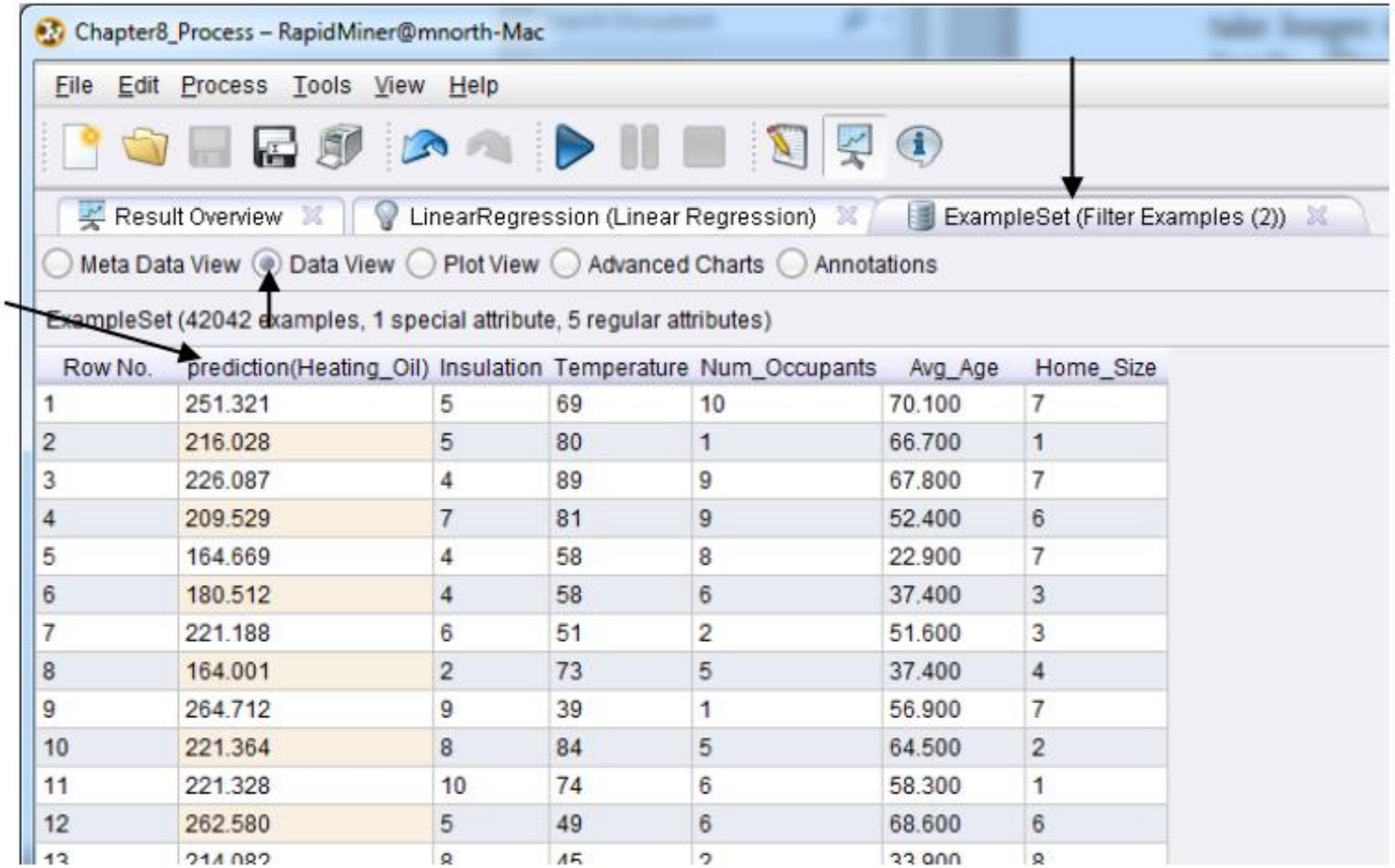


5. Evaluation



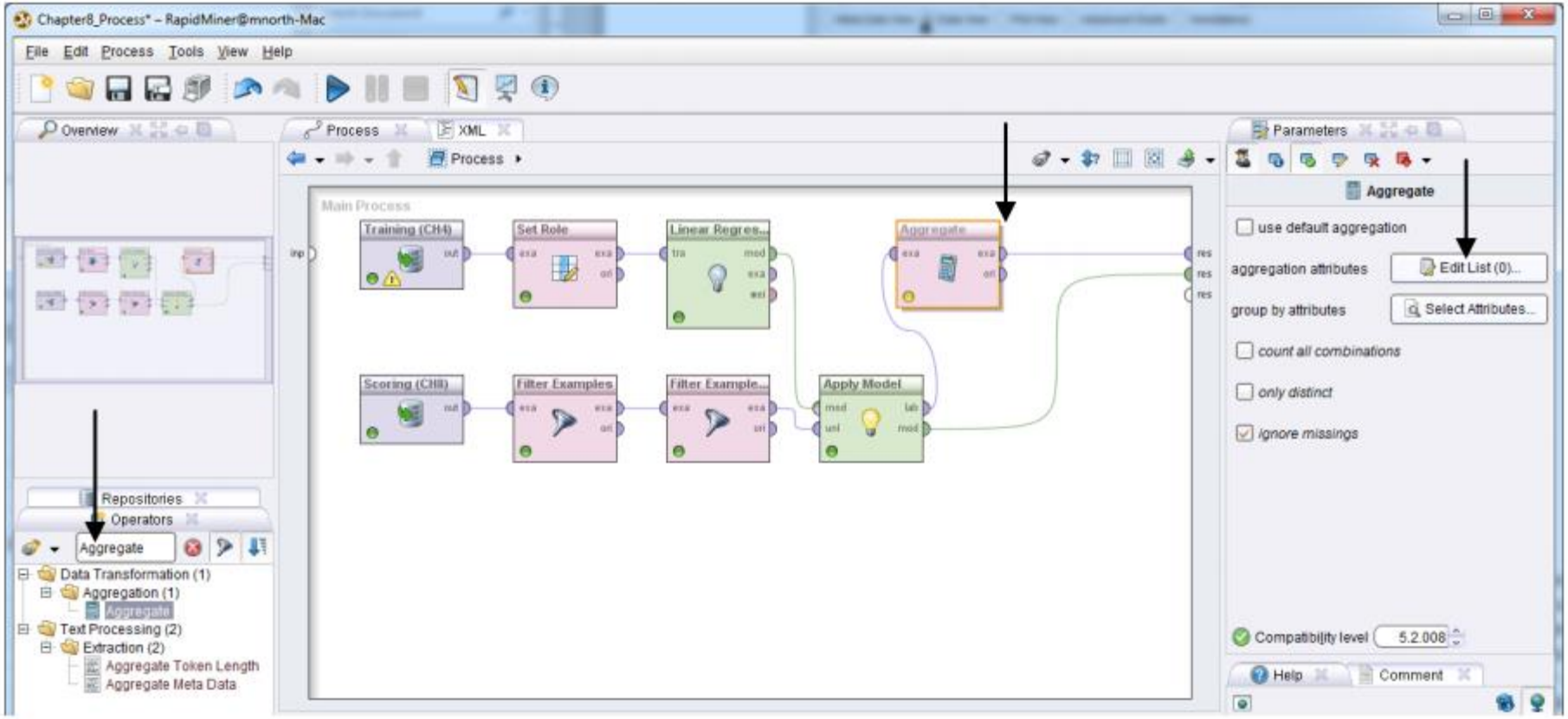
Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
Insulation	3.323	0.413	1.480	0.431	8.048	0	****
Temperature	-0.869	0.068	-0.226	0.405	-12.734	0	****
Avg_Age	1.968	0.064	0.694	0.491	30.565	0	****
Home_Size	3.173	0.310	1.584	0.914	10.230	0	****
(Intercept)	134.511	7.257	?	?	18.535	0	****

5. Evaluation

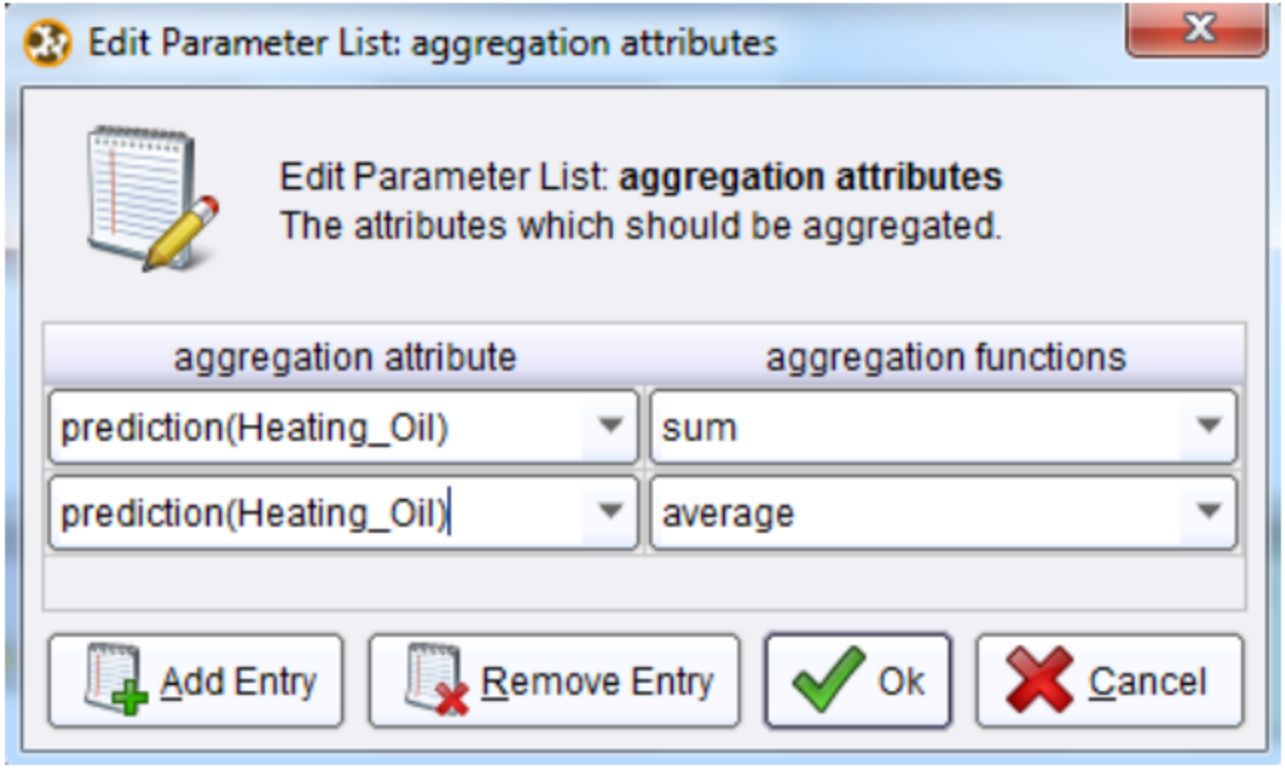


Row No.	prediction(Heating_Oil)	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	89	9	67.800	7
4	209.529	7	81	9	52.400	6
5	164.669	4	58	8	22.900	7
6	180.512	4	58	6	37.400	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4
9	264.712	9	39	1	56.900	7
10	221.364	8	84	5	64.500	2
11	221.328	10	74	6	58.300	1
12	262.580	5	49	6	68.600	6
13	214.082	8	45	2	33.900	8

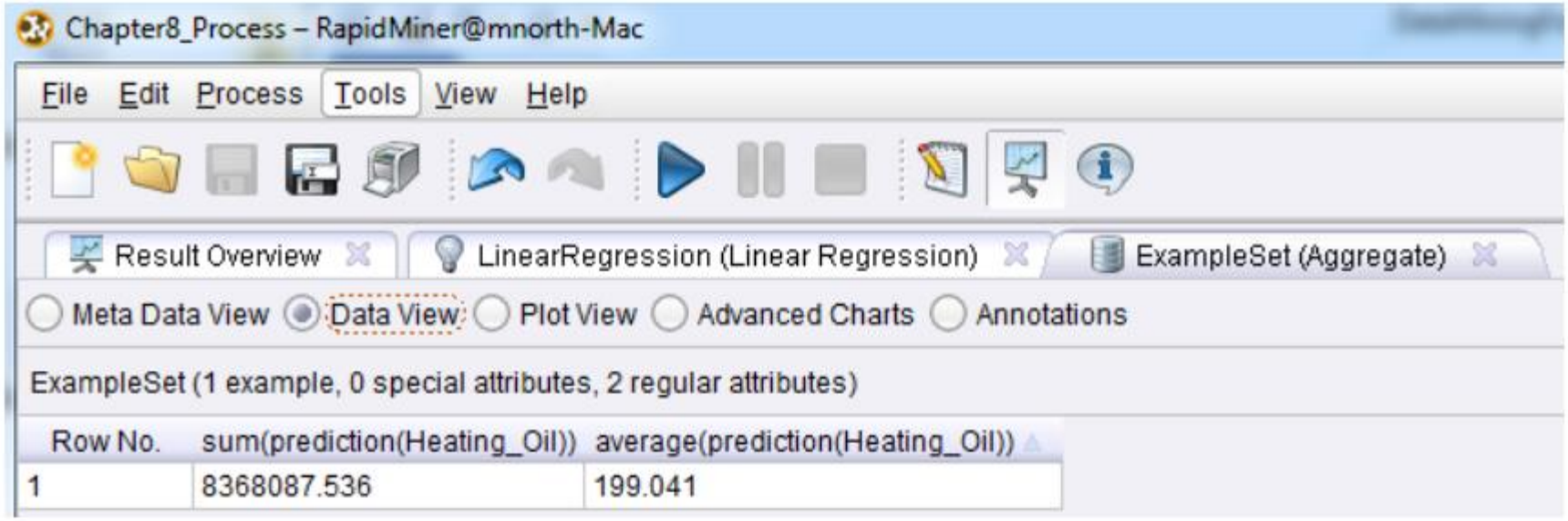
6. Deployment



6. Deployment



6. Deployment



4.4.2 Time Series Forecasting

Time Series Forecasting

- Time series forecasting is one of the oldest known predictive analytics techniques
  - It has existed and been in widespread use even before the term "predictive analytics" was ever coined
- Independent or predictor variables are not strictly necessary for univariate time series forecasting, but are strongly recommended for multivariate time series
- Time series forecasting methods:
  1. Data Driven Method: There is no difference between a predictor and a target. Techniques such as time series averaging or smoothing are considered data-driven approaches to time series forecasting
  2. Model Driven Method: Similar to "conventional" predictive models, which have independent and dependent variables, but with a twist: the independent variable is now time

Data Driven Methods

- There is no difference between a predictor and a target
- The predictor is also the target variable
- Data Driven Methods:
  - Naïve Forecast
  - Simple Average
  - Moving Average
  - Weighted Moving Average
  - Exponential Smoothing
  - Holt's Two-Parameter Exponential Smoothing

Model Driven Methods

- In model-driven methods, time is the predictor or independent variable and the time series value is the dependent variable
- Model-based methods are generally preferable when the time series appears to have a "global" pattern
- The idea is that the model parameters will be able to capture these patterns
  - Thus enable us to make predictions for any step ahead in the future under the assumption that this pattern is going to repeat
- For a time series with local patterns instead of a global pattern, using the model-driven approach requires specifying how and when the patterns change, which is difficult

Model Driven Methods

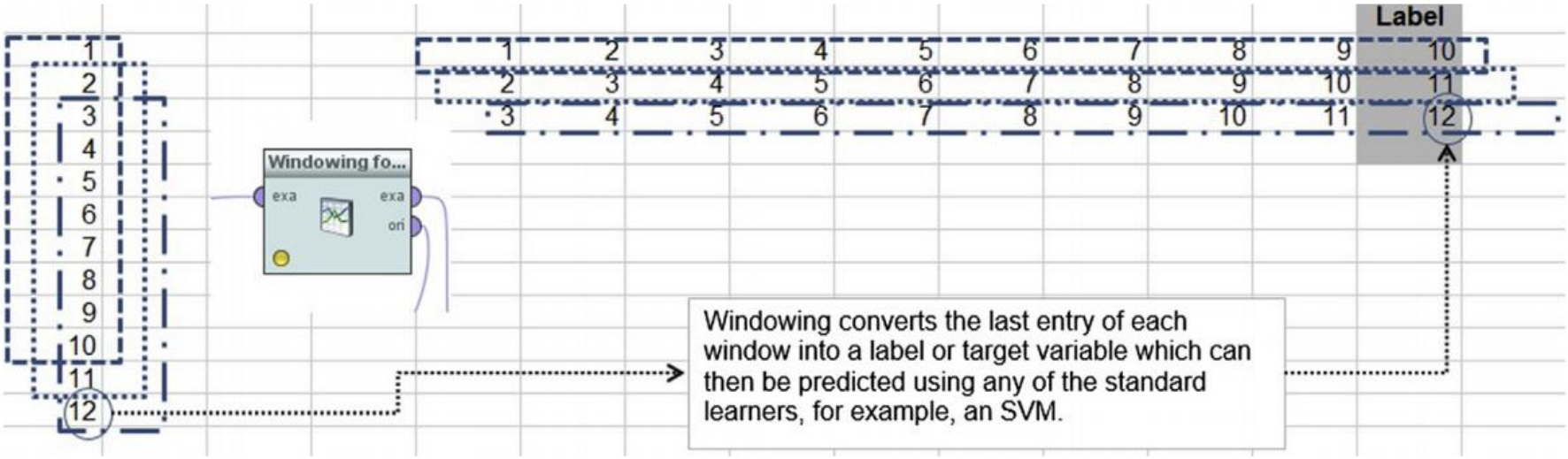
- Linear Regression
- Polynomial Regression
- Linear Regression with Seasonality
- Autoregression Models and ARIMA

How to Implement

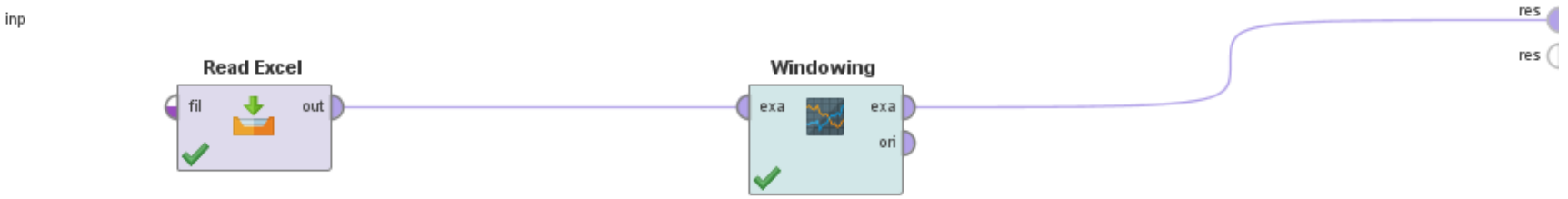
- RapidMiner's approach to time series is based on two main data transformation processes
- The first is windowing to transform the time series data into a generic data set:
  - This step will convert the last row of a window within the time series into a label or target variable
- We apply any of the "learners" or algorithms to predict the target variable and thus predict the next time step in the series

Windowing Concept

- The parameters of the Windowing operator allow changing the size of the windows, the overlap between consecutive windows (step size), and the prediction horizon, which is used for forecasting
- The prediction horizon controls which row in the raw data series ends up as the label variable in the transformed series



Rapidminer Windowing Operator



Date	inputYt
Jan 1, 2009	0.709
Feb 1, 2009	1.886
Mar 1, 2009	1.293
Apr 1, 2009	0.822
May 1, 2009	-0.173
Jun 1, 2009	0.552
Jul 1, 2009	1.169
Aug 1, 2009	1.604
Sep 1, 2009	0.949
Oct 1, 2009	0.080
Nov 1, 2009	-0.040
Dec 1, 2009	1.381
Jan 1, 2010	0.761

Date	label	inputYt-5	inputYt-4	inputYt-3	inputYt-2	inputYt-1	inputYt-0
Jun 1, 2009	1.169	0.709	1.886	1.293	0.822	-0.173	0.552
Jul 1, 2009	1.604	1.886	1.293	0.822	-0.173	0.552	1.169
Aug 1, 2009	0.949	1.293	0.822	-0.173	0.552	1.169	1.604
Sep 1, 2009	0.080	0.822	-0.173	0.552	1.169	1.604	0.949
Oct 1, 2009	-0.040	-0.173	0.552	1.169	1.604	0.949	0.080
Nov 1, 2009	1.381	0.552	1.169	1.604	0.949	0.080	-0.040
Dec 1, 2009	0.761	1.169	1.604	0.949	0.080	-0.040	1.381
Jan 1, 2010	2.312	1.604	0.949	0.080	-0.040	1.381	0.761
Feb 1, 2010	1.795	0.949	0.080	-0.040	1.381	0.761	2.312
Mar 1, 2010	0.586	0.080	-0.040	1.381	0.761	2.312	1.795
Apr 1, 2010	-0.077	-0.040	1.381	0.761	2.312	1.795	0.586
May 1, 2010	0.613	1.381	0.761	2.312	1.795	0.586	-0.077

Window size = 6
Step size = 1
Horizon = 1

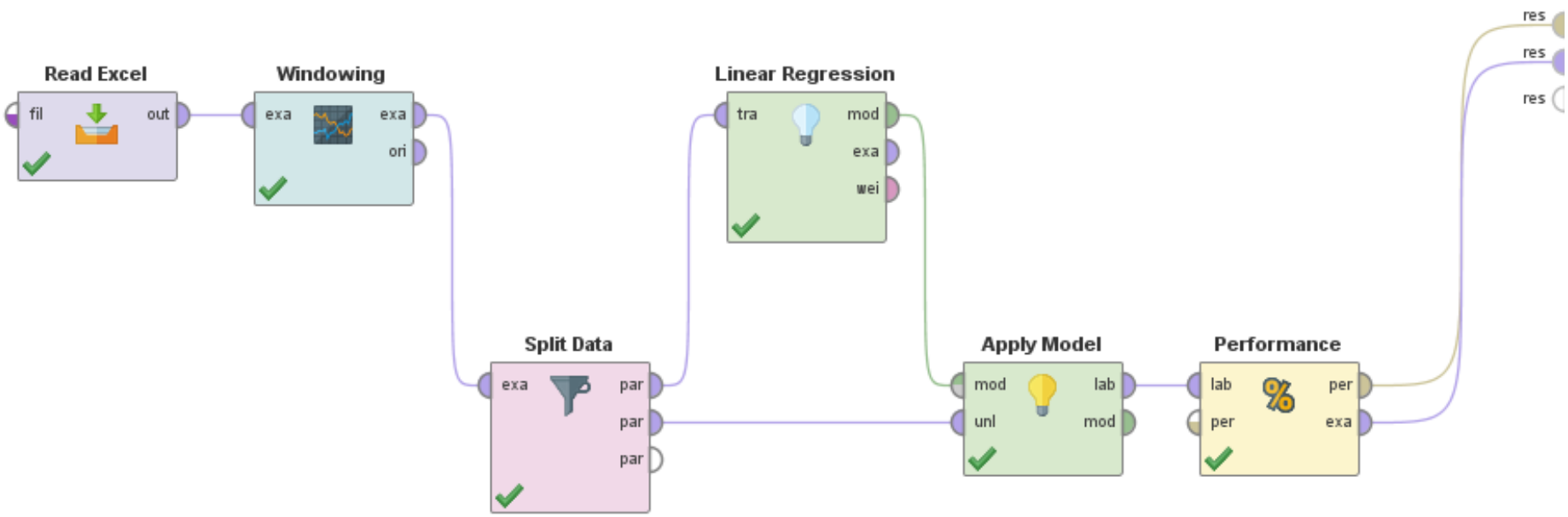
Using data from 6 rows (Jan 2009 – Jun 2009) of the window, a learner can be trained to predict the label which is the value of the time series in the next time step (Jul 2009) and so on.

Windowing Operator Parameters

- Window size: Determines how many "attributes" are created for the cross-sectional data
  - Each row of the original time series within the window width will become a new attribute
  - We choose w = 6
- Step size: Determines how to advance the window
  - Let us use s = 1
- Horizon: Determines how far out to make the forecast
  - If the window size is 6 and the horizon is 1, then the seventh row of the original time series becomes the first sample for the "label" variable
  - Let us use h = 1

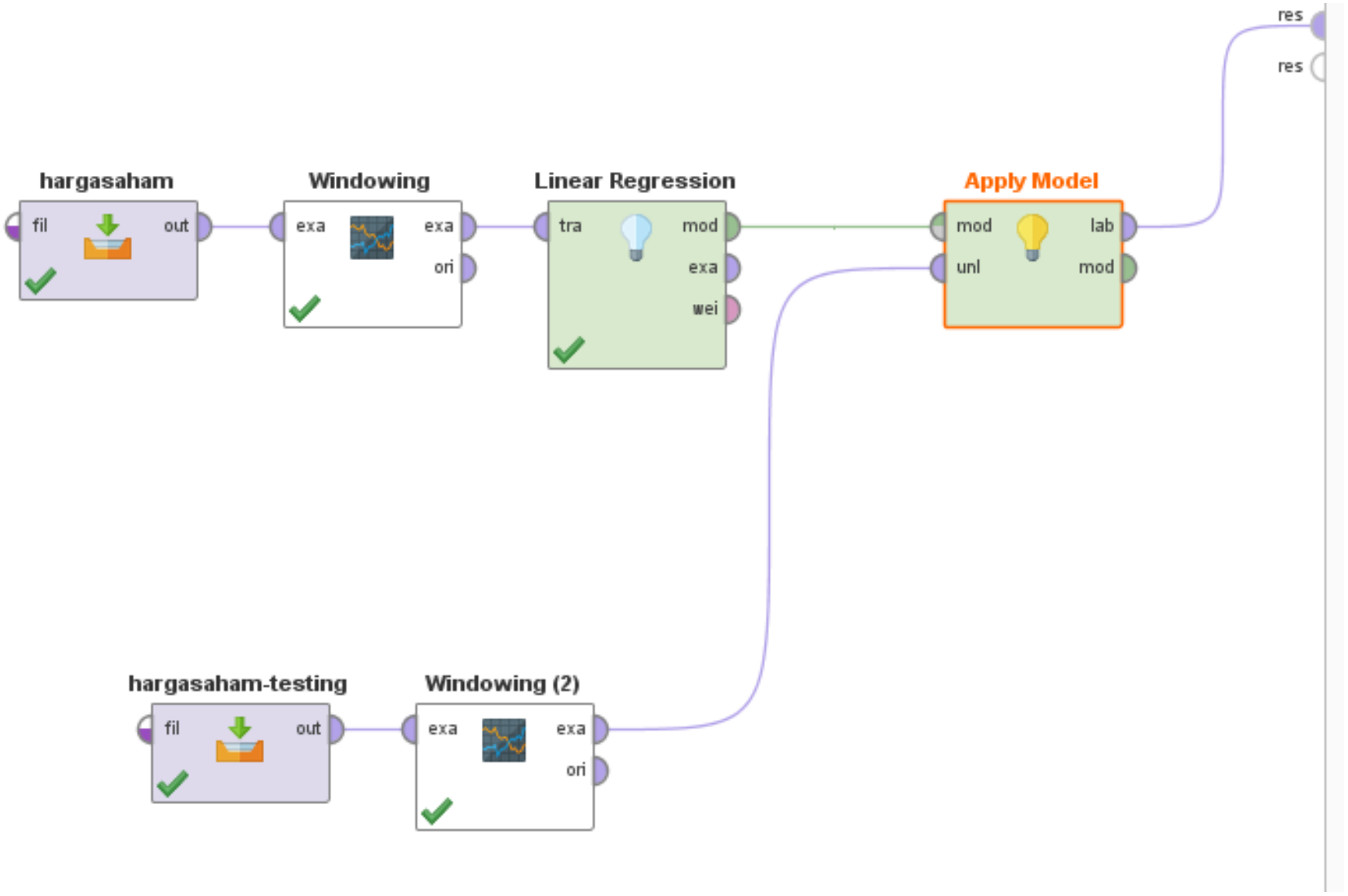
Latihan

- Lakukan training dengan menggunakan linear regression pada dataset hargasaham-training-uni.xls
- Gunakan Split Data untuk memisahkan dataset di atas, 90% training dan 10% untuk testing
- Harus dilakukan proses Windowing pada dataset
- Plot grafik antara label dan hasil prediksi dengan menggunakan chart


Read Excel
fil
out
Windowing
exa
exa
ori
Linear Regression
tra
mod
exa
wei
Split Data
exa
par
par
par
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa
res
res
res

Latihan

- Lakukan training dengan menggunakan linear regression pada dataset hargasaham-training.xls
- Terapkan model yang dihasilkan untuk data hargasaham-testing-kosong.xls
- Harus dilakukan proses Windowing pada dataset
- Plot grafik antara label dan hasil prediksi dengan menggunakan chart


hargasaham
fil
out
Windowing
exa
exa
ori
Linear Regression
tra
mod
exa
wei
Apply Model
mod
unl
lab
mod
res
res
hargasaham-testing
fil
out
Windowing (2)
exa
exa
ori

5. Text Mining

5.1 Text Mining Concepts

5.2 Text Clustering

5.3 Text Classification

5.4 Data Mining Law

5.1 Text Mining Concepts

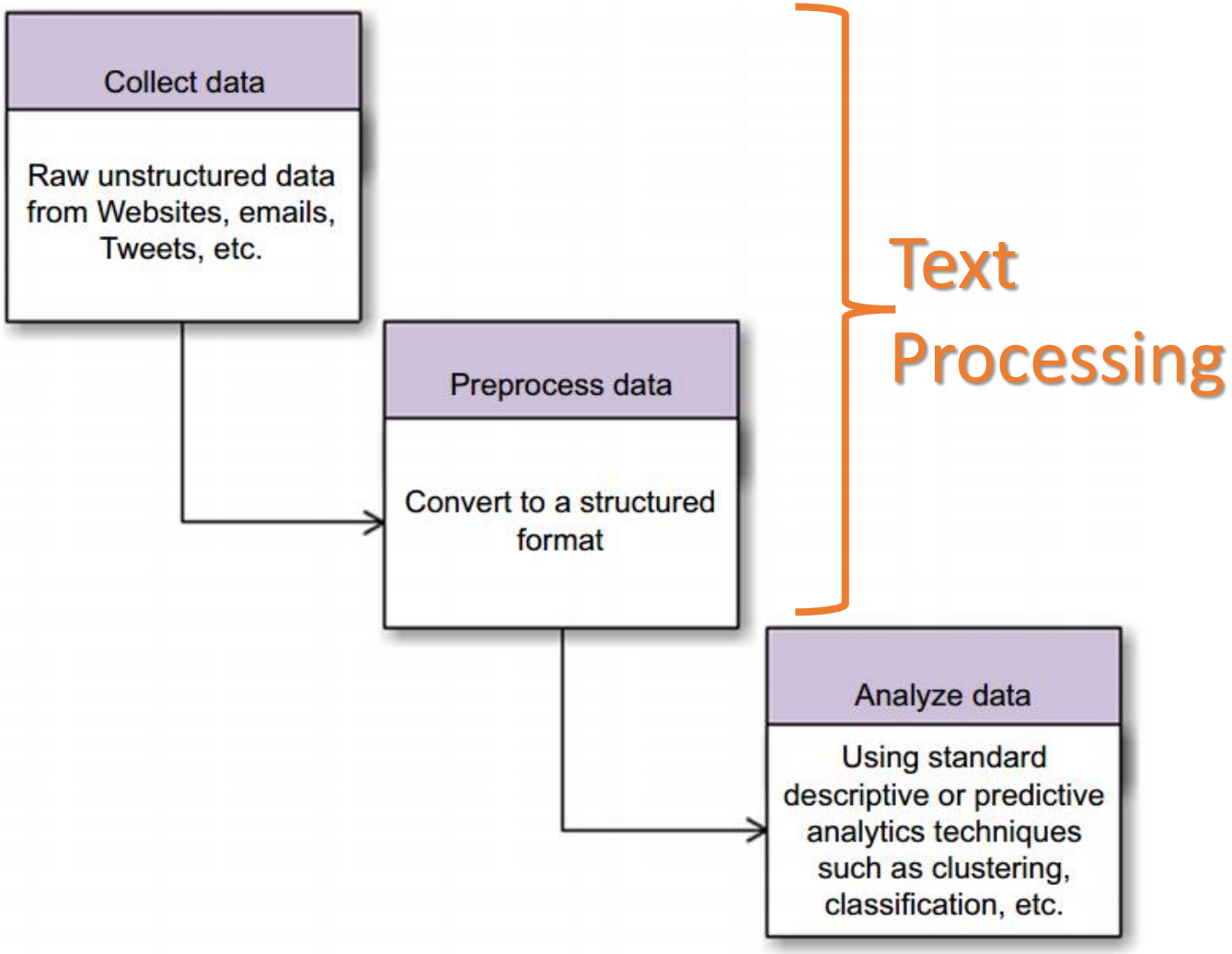
Data Mining vs Text Mining

1. Text Mining:
   - Mengolah data tidak terstruktur dalam bentuk text, web, social media, dsb
   - Menggunakan metode text processing untuk mengkonversi data tidak terstruktur menjadi terstruktur
     - Kemudian diolah dengan data mining

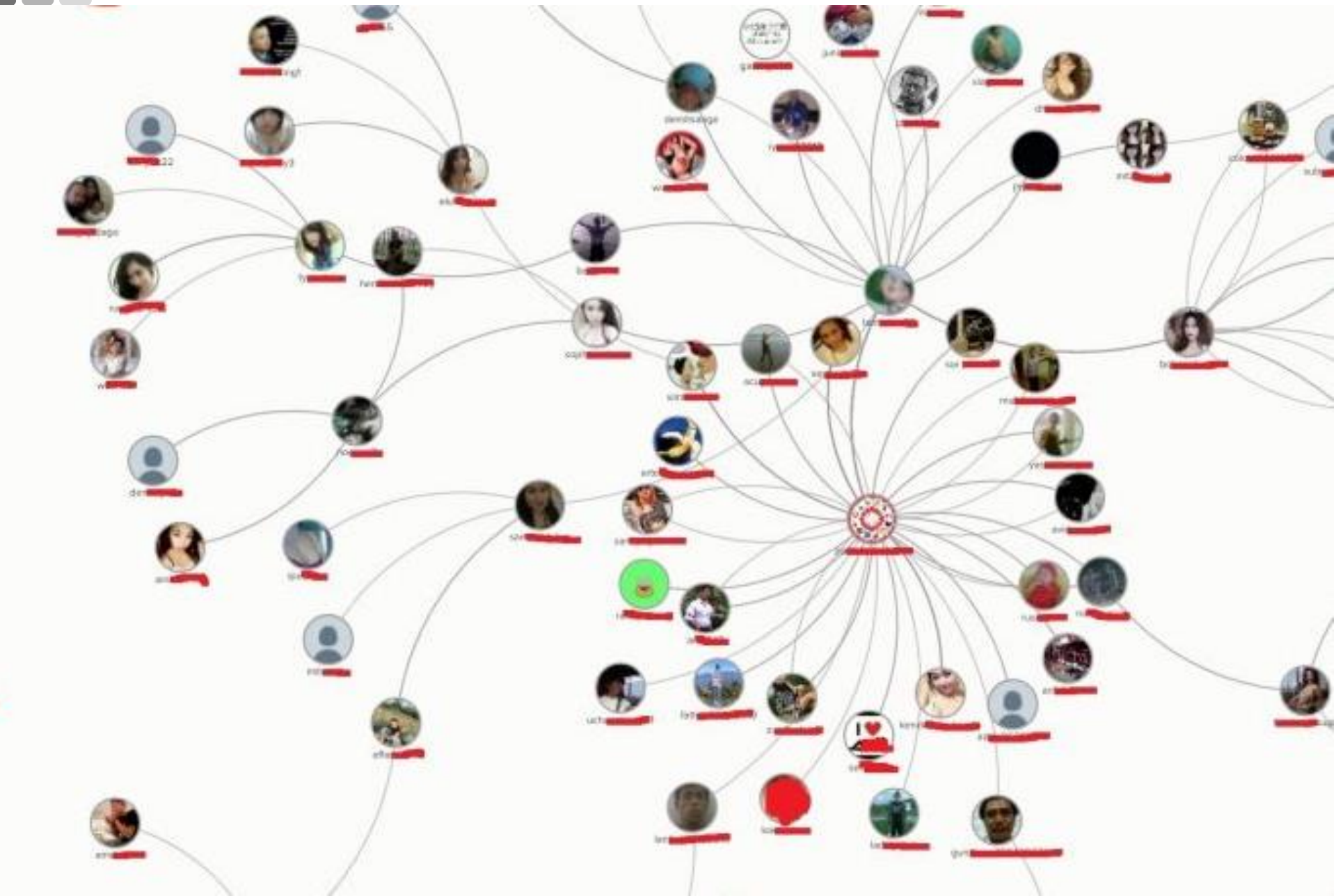
2. Data Mining:
   - Mengolah data terstruktur dalam bentuk tabel yang memiliki atribut dan kelas
   - Menggunakan metode data mining, yang terbagi menjadi metode estimasi, forecasting, klasifikasi, klastering atau asosiasi
     - Yang dasar berpikirnya menggunakan konsep statistika atau heuristik ala machine learning

How Text Mining Works

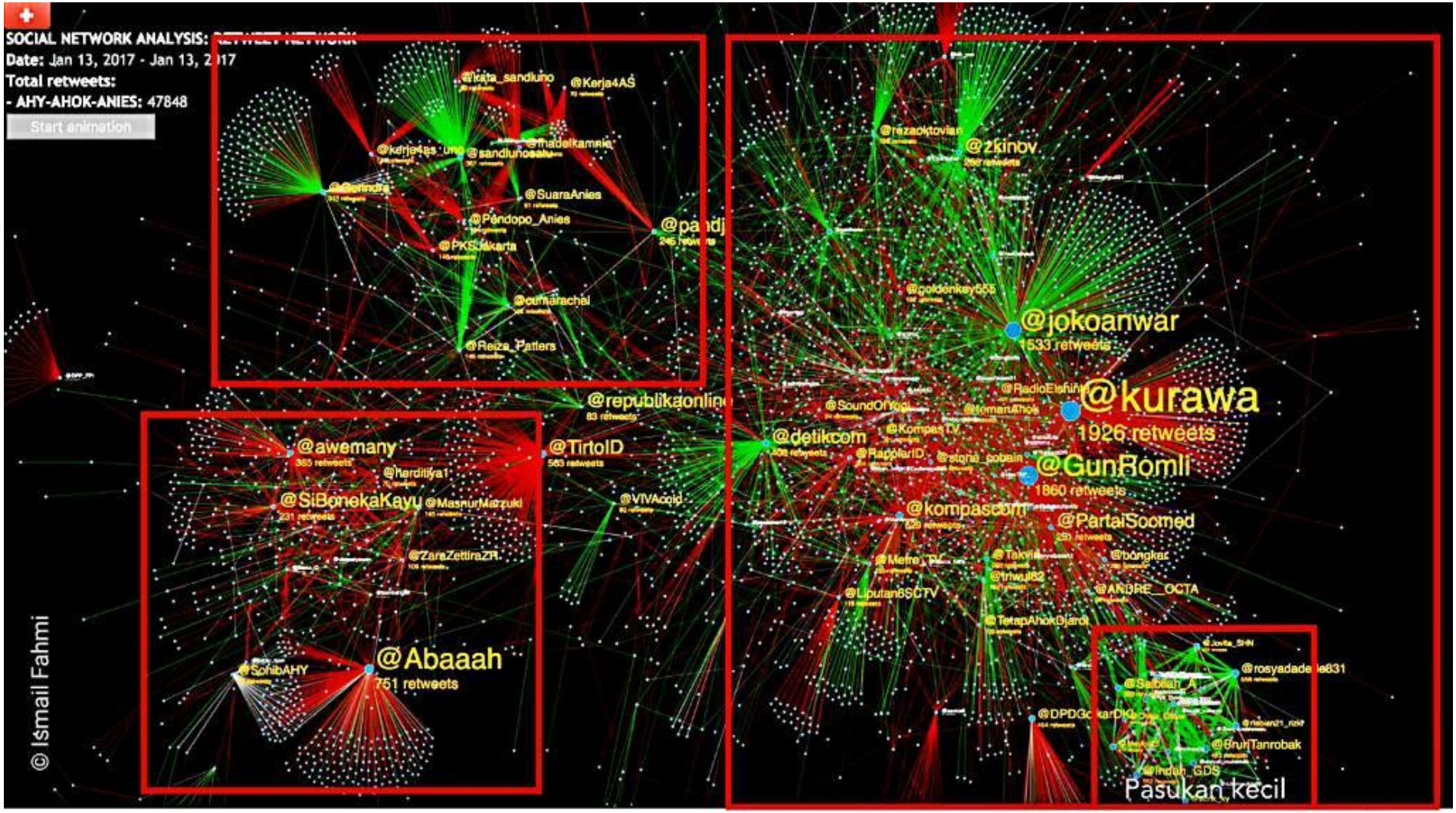
- The fundamental step is to convert text into semi-structured data
- Then apply the data mining methods to classify, cluster, and predict



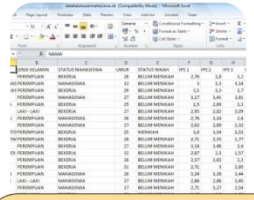
Text Mining: Jejak Pornografi di Indonesia



Text Mining: AHY-AHOK-ANIES

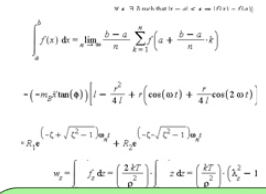


Proses Data Mining



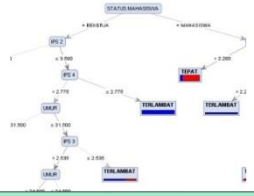
1. Himpunan Data

(**Pahami** dan Persiapkan Data)



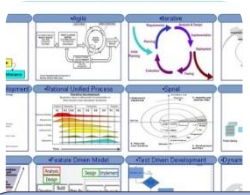
2. Metode Data Mining

(**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
Text Processing

MODELING

Estimation
Prediction
Classification
Clustering
Association

MODEL

Formula
Tree
Cluster
Rule
Correlation

KINERJA

Akurasi
Tingkat Error
Jumlah Cluster

MODEL

Atribute/Faktor
Korelasi
Bobot

Word, Token and Tokenization

Document 1	This is a book on data mining.
Document 2	This book describes data mining and text mining using RapidMiner.

- Words are separated by a special character: a blank space
- Each word is called a token
- The process of discretizing words within a document is called tokenization
- For our purpose here, each sentence can be considered a separate document, although what is considered an individual document may depend upon the context
- For now, a document here is simply a sequential collection of tokens

Matrix of Terms

- We can impose some form of structure on this raw data by creating a matrix, where:
  - the columns consist of all the tokens found in the two documents
  - the cells of the matrix are the counts of the number of times a token appears
- Each token is now an attribute in standard data mining parlance and each document is an example

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Term Document Matrix (TDM)

- Basically, unstructured raw data is now transformed into a format that is recognized, not only by the human users as a data table, but more importantly by all the machine learning algorithms which require such tables for training
- This table is called a document vector or term document matrix (TDM) and is the cornerstone of the preprocessing required for text mining

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	Using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

TF–IDF

- We could have also chosen to use the TF–IDF scores for each term to create the document vector
- N is the number of documents that we are trying to mine
- N_k is the number of documents that contain the keyword, k

$$TF = n_k/n \qquad IDF = \log_2 (N/N_k)$$

$$TF - IDF = n_k/n * \log_2 (N/N_k)$$

ExampleSet (2 examples, 0 special attributes, 12 regular attributes)

Row No.	RapidMiner	This	a	and	book	data	describes	is	mining	on	text	using
1	0	0	0.577	0	0	0	0	0.577	0	0.577	0	0
2	0.447	0	0	0.447	0	0	0.447	0	0	0	0.447	0.447

Stopwords

- In the two sample text documents was the occurrence of common words such as “a,” “this,” “and,” and other similar terms
- Clearly in larger documents we would expect a larger number of such terms that do not really convey specific meaning
- Most grammatical necessities such as articles, conjunctions, prepositions, and pronouns may need to be filtered before we perform additional analysis
  - Such terms are called stopwords and usually include most articles, conjunctions, pronouns, and prepositions
  - Stopword filtering is usually the second step that follows immediately after tokenization
- Notice that our document vector has a significantly reduced size after applying standard English stopword filtering

Row No.	RapidMiner	book	data	describes	mining	text	using
1	0	1	1	0	1	0	0
2	1	1	1	1	2	1	1

Stopwords Bahasa Indonesia

- Lakukan googling dengan keyword: stopwords bahasa Indonesia
- Download stopword bahasa Indonesia dan gunakan di Rapidminer

Stemming

- Words such as “recognized,” “recognizable,” or “recognition” in different usages, but contextually they may all imply the same meaning, for example:
  - “Einstein is a well-recognized name in physics”
  - “The physicist went by the easily recognizable name of Einstein”
  - “Few other physicists have the kind of name recognition that Einstein has”
  - The so-called root of all these highlighted words is “recognize”
- By reducing terms in a document to their basic stems, we can simplify the conversion of unstructured text to structured data because we now only take into account the occurrence of the root terms
- This process is called stemming. The most common stemming technique for text mining in English is the Porter method (Porter, 1980)

A Typical Sequence of Preprocessing Steps to Use in Text Mining

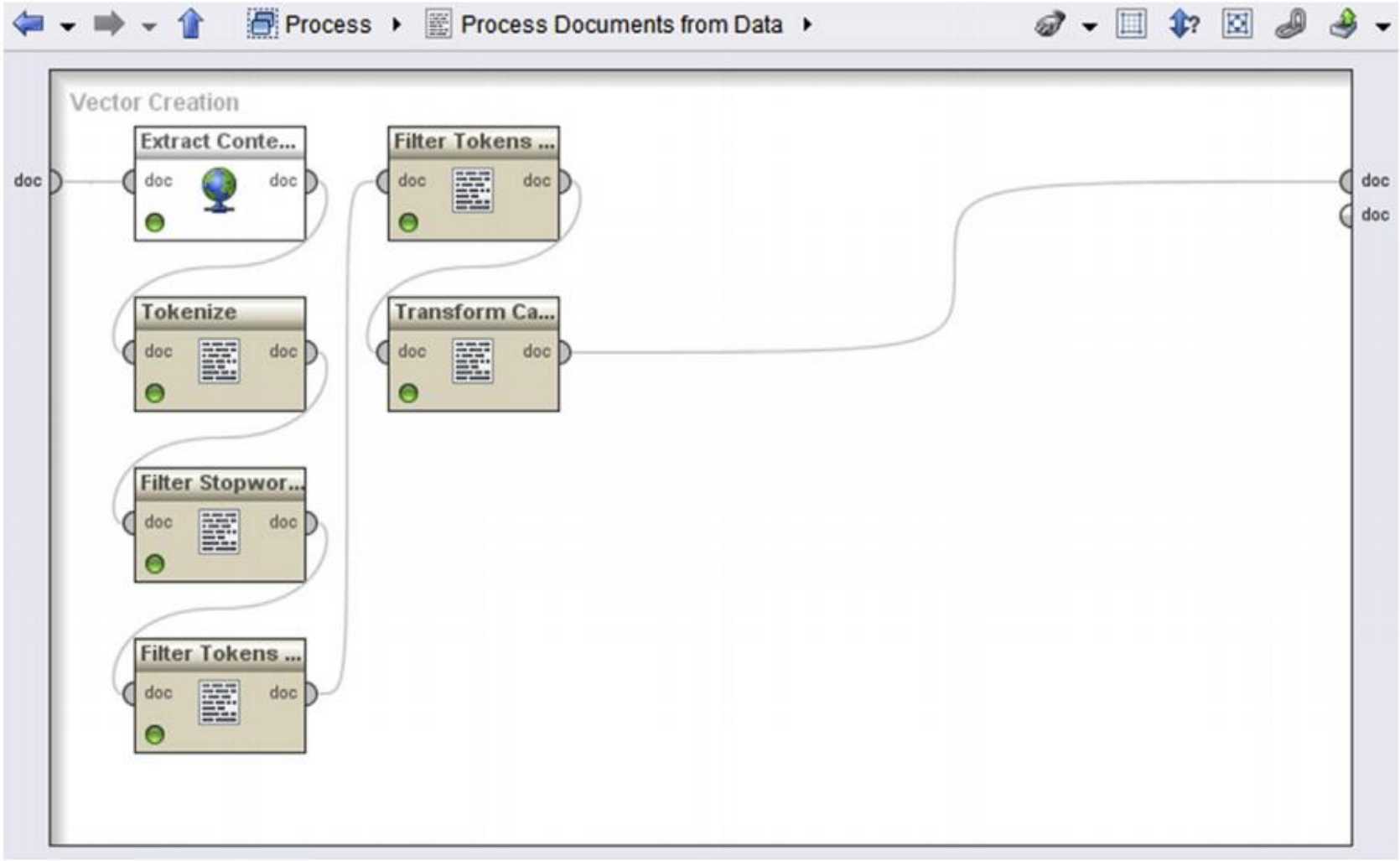
Step	Action	Result
1	Tokenize	Convert each word or term in a document into a distinct attribute
2	Stopword removal	Remove highly common grammatical tokens/words
3	Filtering	Remove other very common tokens
4	Stemming	Trim each token to its most essential minimum
5	n-grams	Combine commonly occurring token pairs or tuples (more than 2)

N-Grams

- There are families of words in the spoken and written language that typically go together
  - The word “Good” is usually followed by either “Morning,” “Afternoon,” “Evening,” “Night,” or in Australia, “Day”
  - Grouping such terms, called **n-grams**, and analyzing them statistically can present new insights
- Search engines use word n-gram models for a variety of applications, such as:
  - Automatic translation, identifying speech patterns, checking misspelling, entity detection, information extraction, among many different use cases

Row...	label	RapidMiner	book	book_data	book_descr...	data	data_mining	describes	describes_data	mining	mining_text	mining_usi...	text_0	text_mining	using	using_RapidMiner
1	text1	0	0.447	0.447	0	0.447	0.447	0	0	0.447	0	0	0	0	0	0
2	text2	0.243	0.243	0	0.243	0.243	0.243	0.243	0.243	0.485	0.243	0.243	0.243	0.243	0.243	0.243

Rapidminer Process of Text Mining



5.2 Text Clustering

Latihan

- Lakukan eksperimen mengikuti buku Matthew North (Data Mining for the Masses) Chapter 12 (Text Mining), 2012, p 189-215

- Datasets: Federalist Papers

- Pahami alur text mining yang dilakukan dan sesuaikan dengan konsep yang sudah dipelajari

1. Business Understanding

- Motivation:
  - Gillian is a historian, and she has recently curated an exhibit on the Federalist Papers, the essays that were written and published in the late 1700's
  - The essays were published anonymously under the author name 'Publius', and no one really knew at the time if 'Publius' was one individual or many
  - After Alexander Hamilton died in 1804, some notes were discovered that revealed that he (Hamilton), James Madison and John Jay had been the authors of the papers
  - The notes indicated specific authors for some papers, but not for others:
    - John Jay was revealed to be the author for papers 3, 4 and 5
    - James Madison for paper 14
    - Hamilton for paper 17
    - Paper 18 had no author named, but there was evidence that Hamilton and Madison worked on that one together
- Objective:
  - Gillian would like to analyze paper 18's content in the context of the other papers with known authors, to see if she can generate some evidence that the suspected collaboration between Hamilton and Madison is in fact

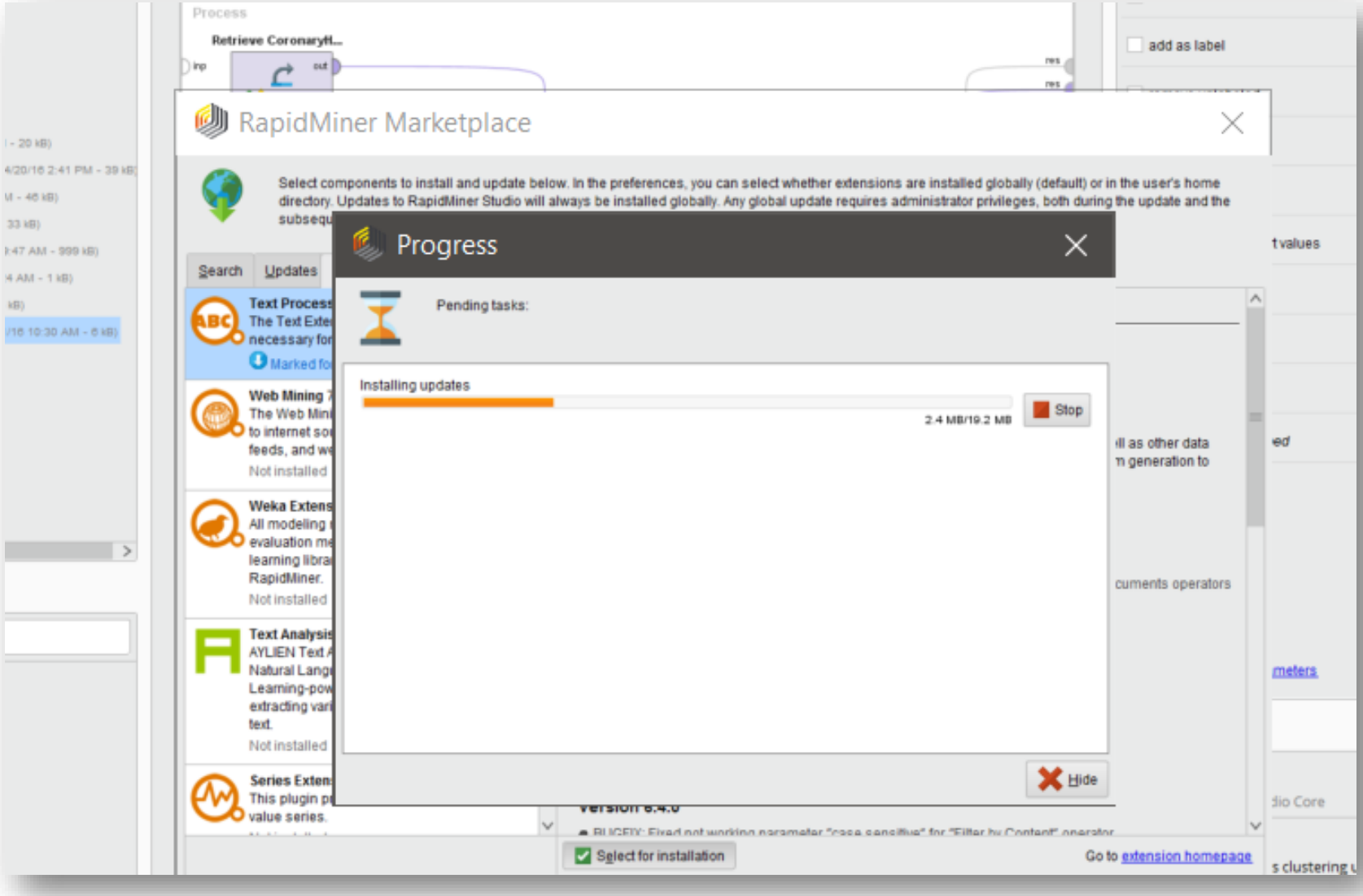
2. Data Understanding

- The Federalist Papers are available through a number of sources:
  - They have been re-published in book form, they are available on a number of different web sites
  - Their text is archived in many libraries throughout the world

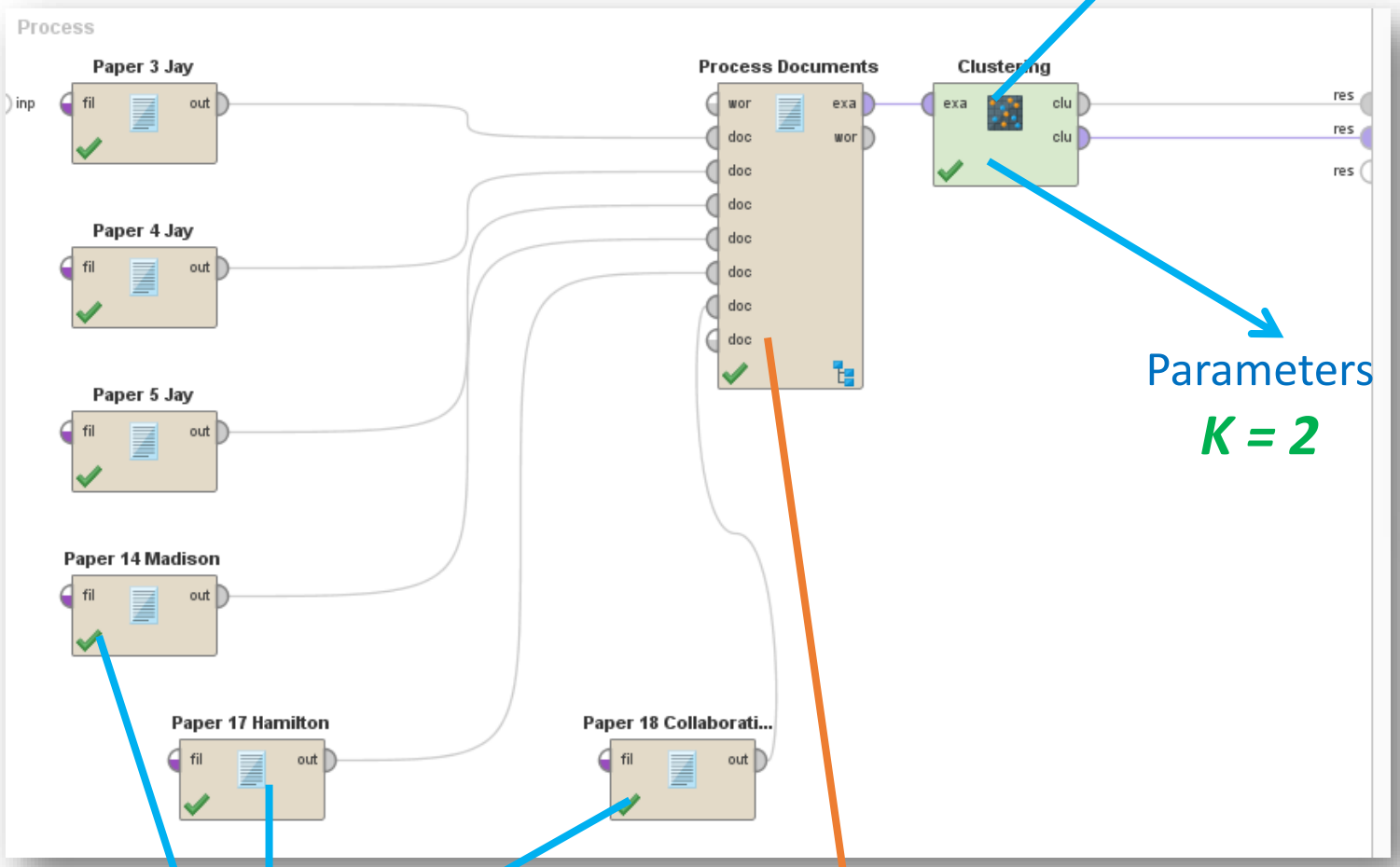
- Gillian's data set is simple (6 dataset):
  - Federalist03_Jay.txt
  - Federalist04_Jay.txt
  - Federalist05_Jay.txt
  - Federalist14_Madison.txt
  - Federalist17_Hamilton.txt
  - Federalist18_Collaboration.txt (*suspected*)

Modeling

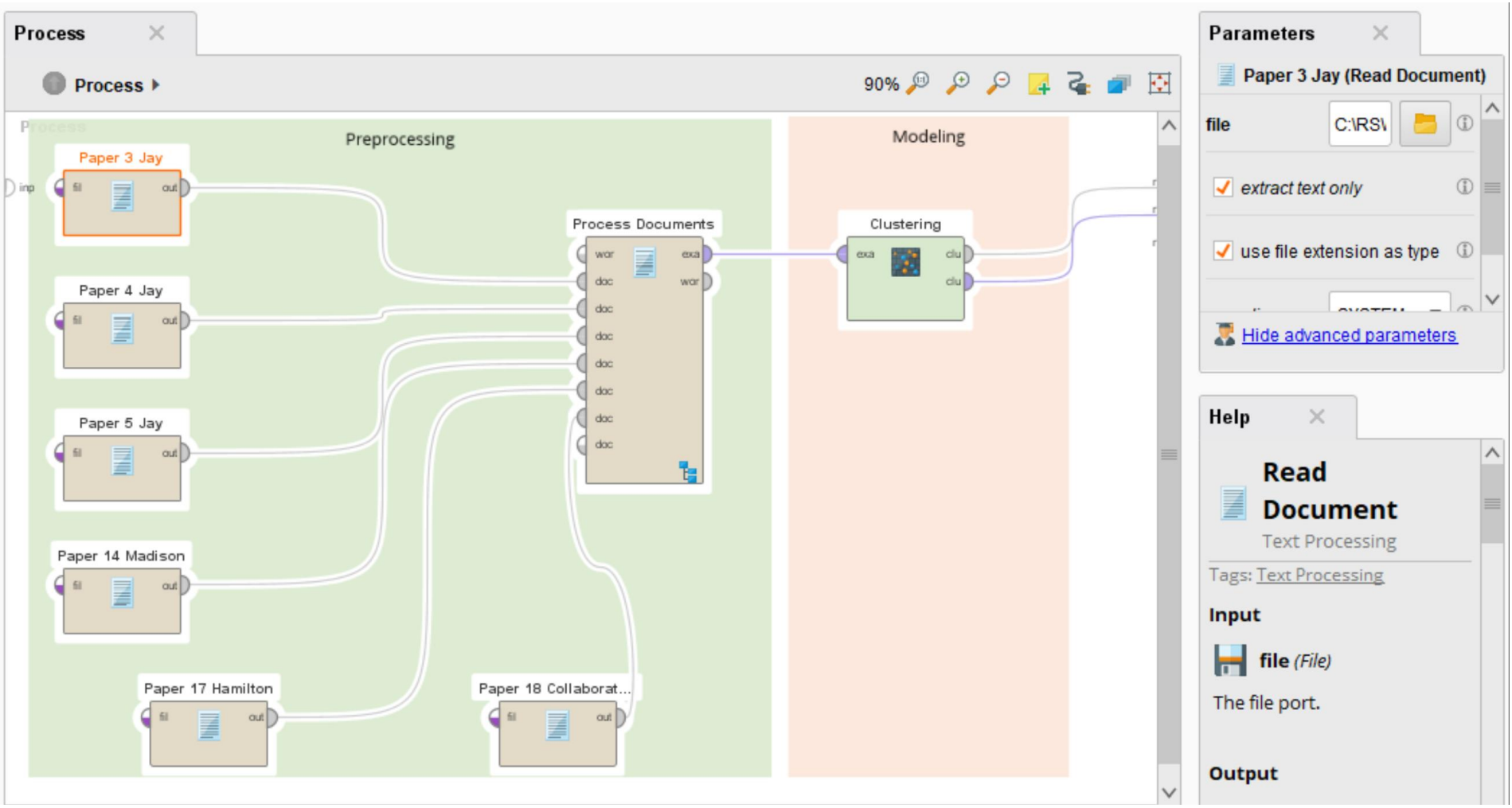
Text Processing Extension Installation



Modeling


Process
Paper 3 Jay
Paper 4 Jay
Paper 5 Jay
Paper 14 Madison
Paper 17 Hamilton
Paper 18 Collaborati...
Process Documents
Clustering
Operator
K-Means
Parameters
K = 2
Parameters
Clustering (k-Means)
add cluster attribute
add as label
remove unlabeled
k
2
max runs
10
Operator
Read Document
Text Processing
Tokenize
Filter Stopwords (En...
Transform Cases
Stem (Porter)

Modelling with Annotation



Evaluation

- Gillian feels confident that paper 18 is a collaboration that John Jay did not contribute to
- His vocabulary and grammatical structure was quite different from those of Hamilton and Madison

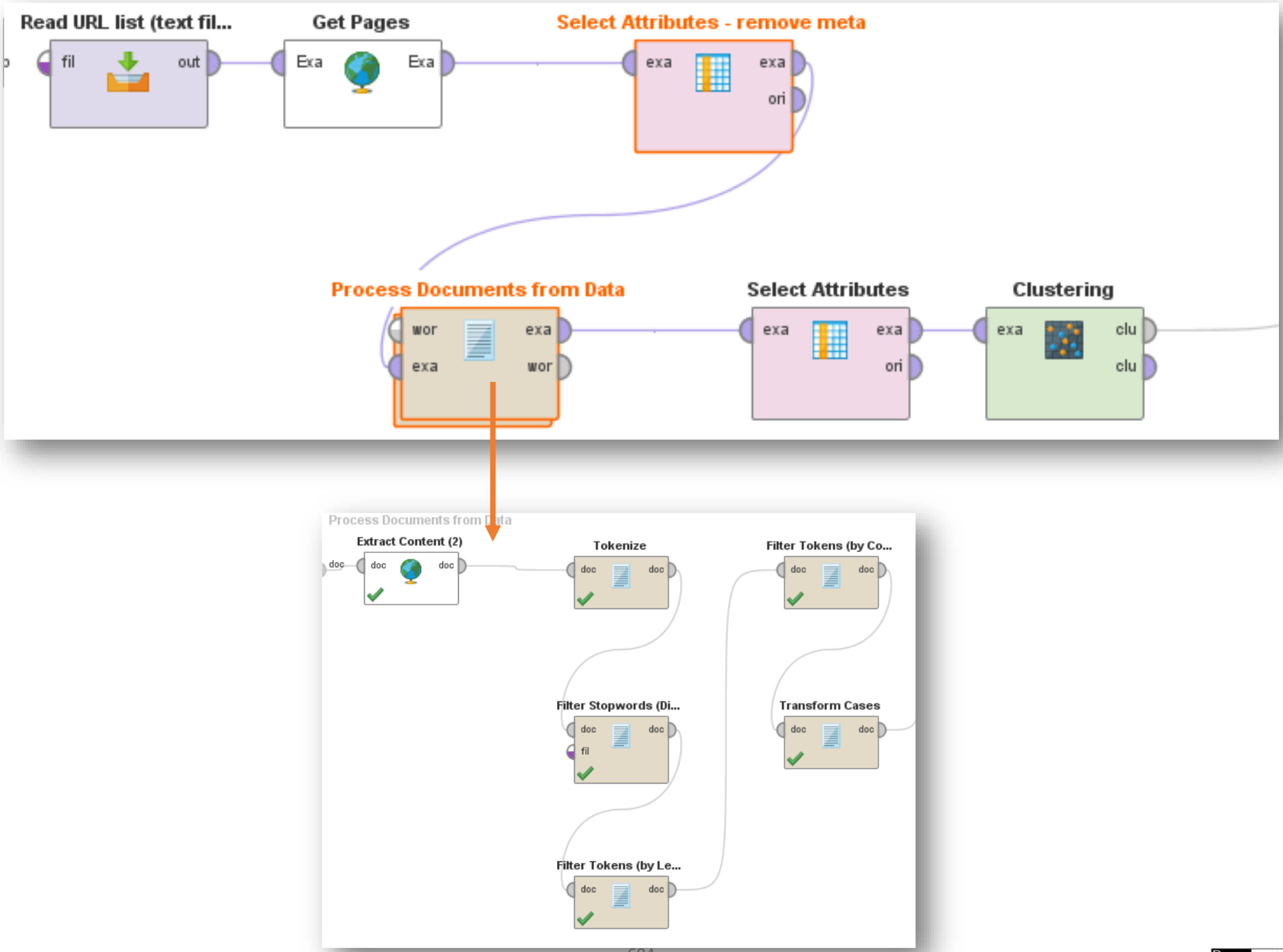
ExampleSet (Process Documents) | Cluster Model (Clustering)

ExampleSet (6 examples, 7 special attributes, 1523 regular attributes) Filter (6 / 6 examples): all

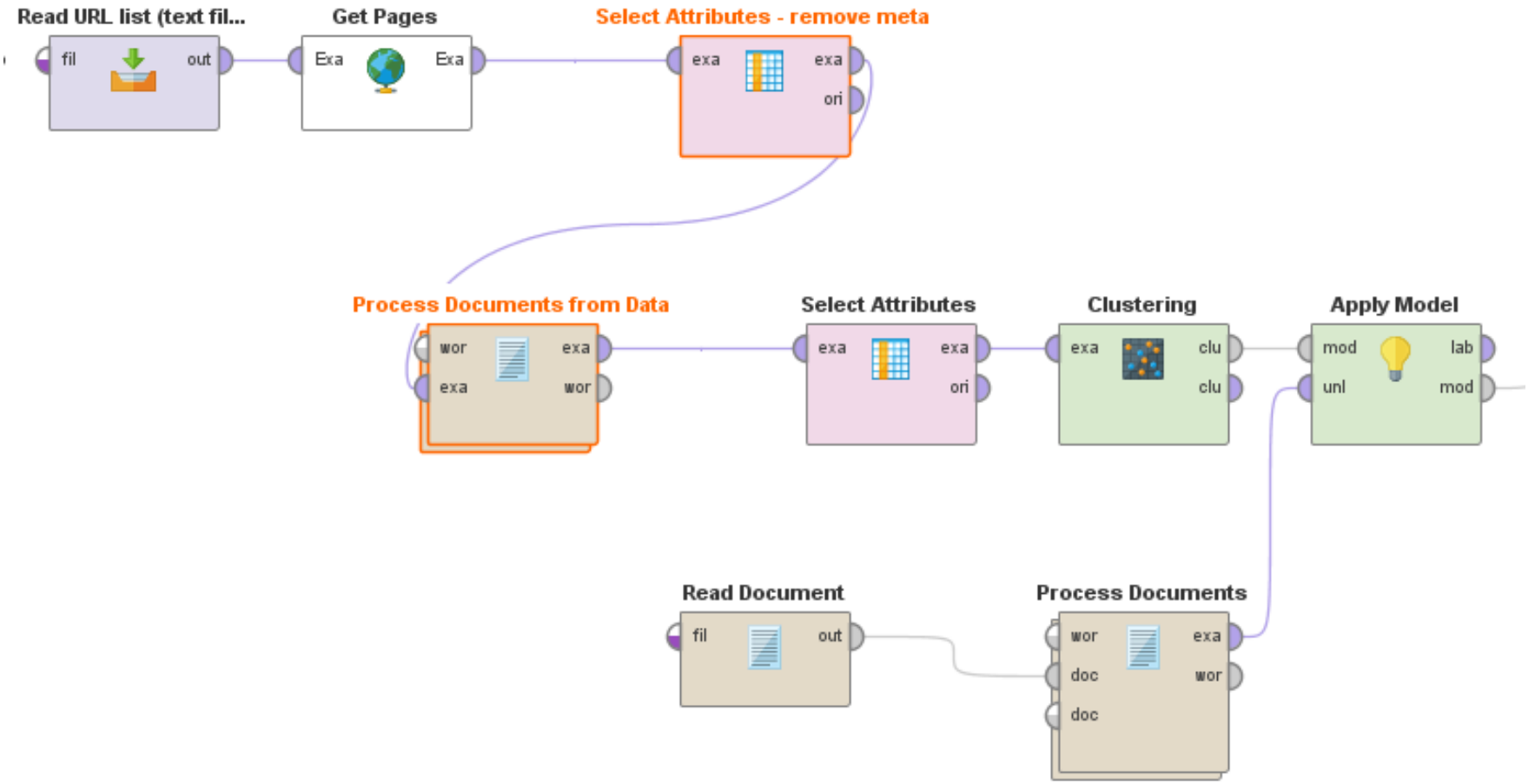
Row No.	id	cluster	file_type	metadata_file	metadata_d...	metadata_p...	metadata_si...
1	1	cluster_0	txt	\Federalist03_Jay.txt	Jun 11, 2015 ...	C:\RSWLectu...	8760
2	2	cluster_0	txt	\Federalist04_Jay.txt	Jun 11, 2015 ...	C:\RSWLectu...	9729
3	3	cluster_0	txt	\Federalist05_Jay.txt	Jun 8, 2015 1...	C:\RSWLectu...	8276
4	4	cluster_1	txt	\Federalist14_Madison.txt	Jun 8, 2015 1...	C:\RSWLectu...	12741
5	5	cluster_1	txt	\Federalist17_Hamilton.txt	Jun 8, 2015 1...	C:\RSWLectu...	9720
6	6	cluster_1	txt	\Federalist18_Collaboration.txt	Jun 8, 2015 1...	C:\RSWLectu...	12962

Latihan

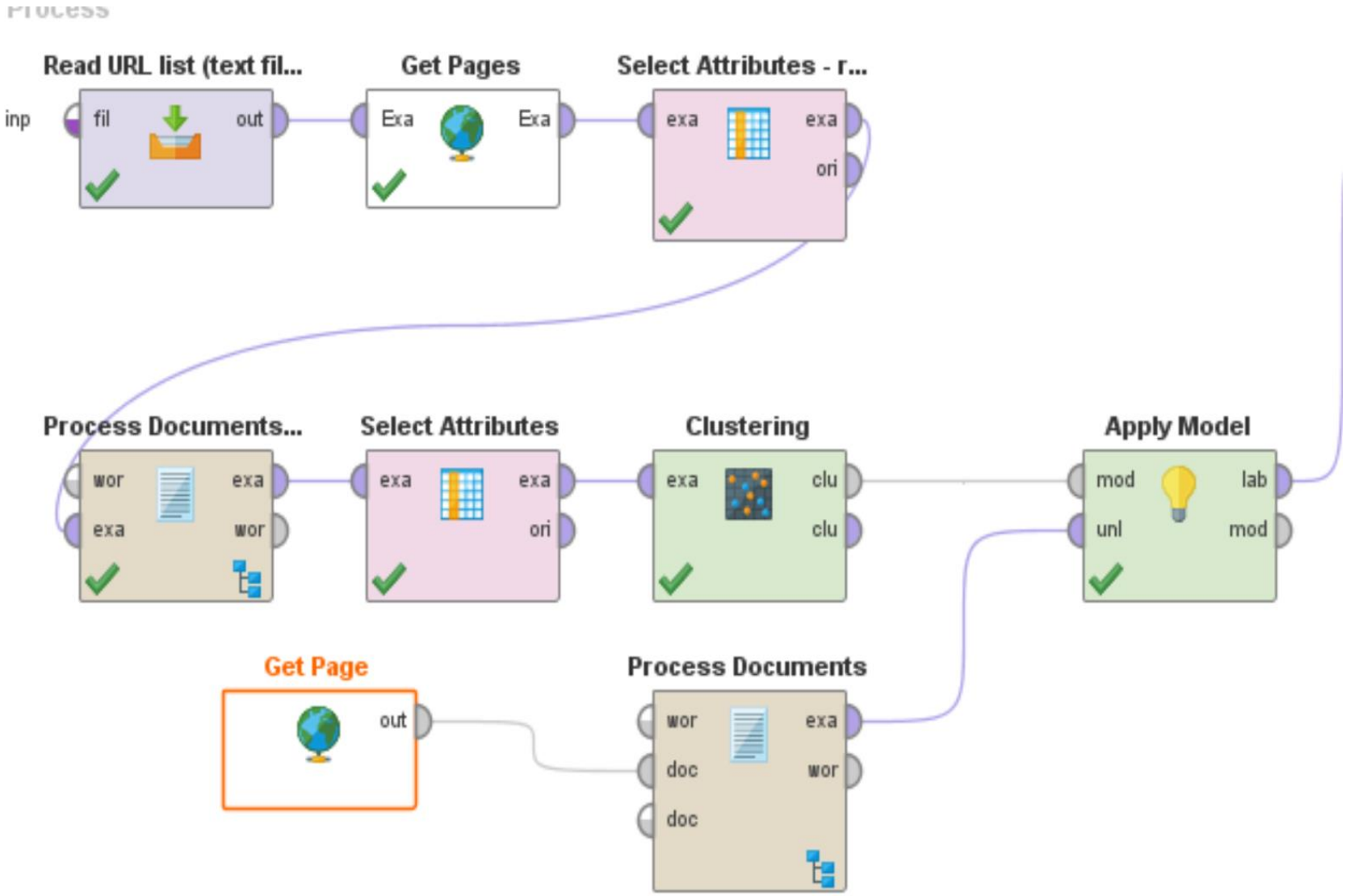
- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 1: Keyword Clustering, p 284-287
- Datasets (file pages.txt):
  1. https://www.cnnindonesia.com/olahraga
  2. https://www.cnnindonesia.com/ekonomi
- Gunakan stopword Bahasa Indonesia (ada di folder dataset), dengan operator Stopword (Dictionary) dan pilih file stopword-indonesia.txt
- Untuk mempermudah, copy/paste file 09_Text_9.3.1_keyword_clustering_webmining.rmp ke Repository dan kemudian buka di Rapidminer
  - Pilih file pages.txt yang berisis URL pada Read URL


Read URL list (text fil...
fil
out
Get Pages
Exa
Exa
Select Attributes - remove meta
exa
exa
ori
Process Documents from Data
wor
exa
exa
wor
Select Attributes
exa
exa
ori
Clustering
exa
clu
clu
Process Documents from Data
Extract Content (2)
doc
doc
doc
Tokenize
doc
doc
Filter Tokens (by Co...
doc
doc
Filter Stopwords (Di...
doc
doc
fil
Transform Cases
doc
doc
Filter Tokens (by Le...
doc
doc

Testing Model (Read Document)



Testing Model (Get Page)



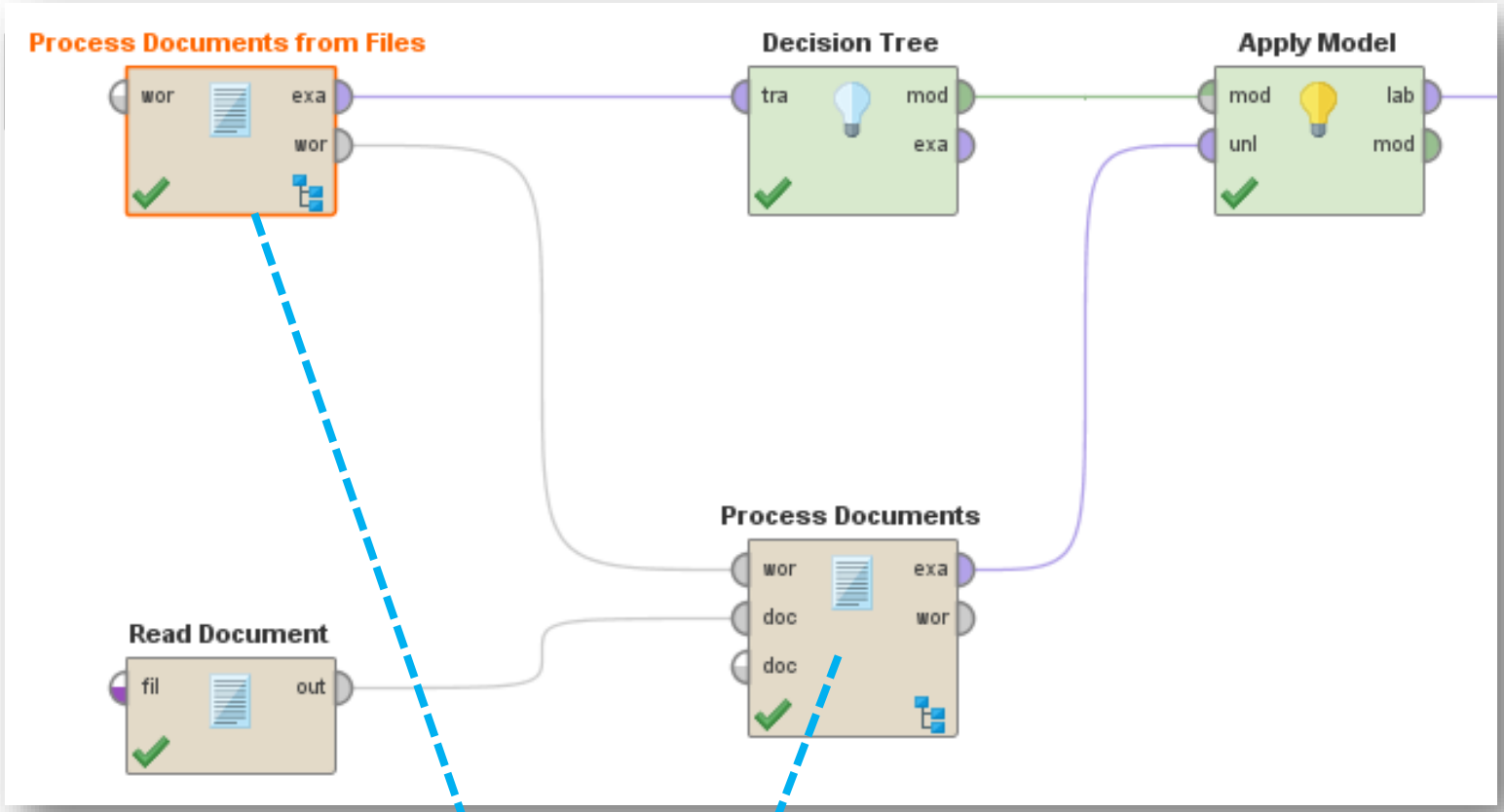
5.3 Text Classification

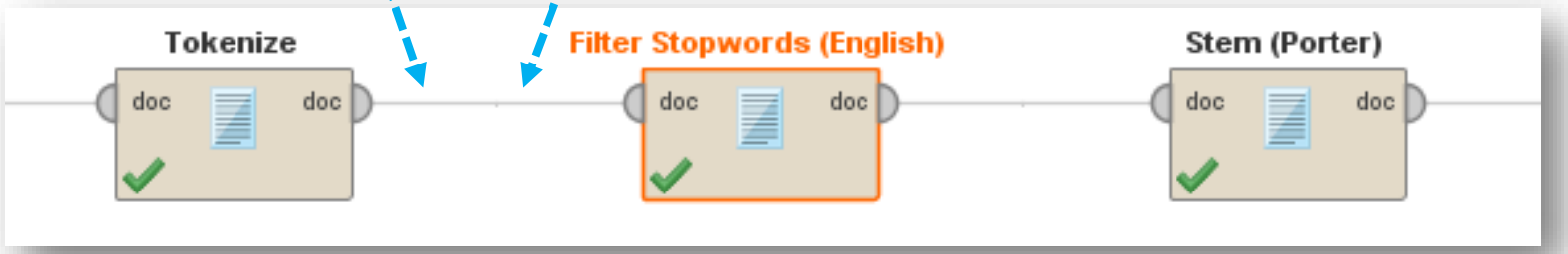
Latihan

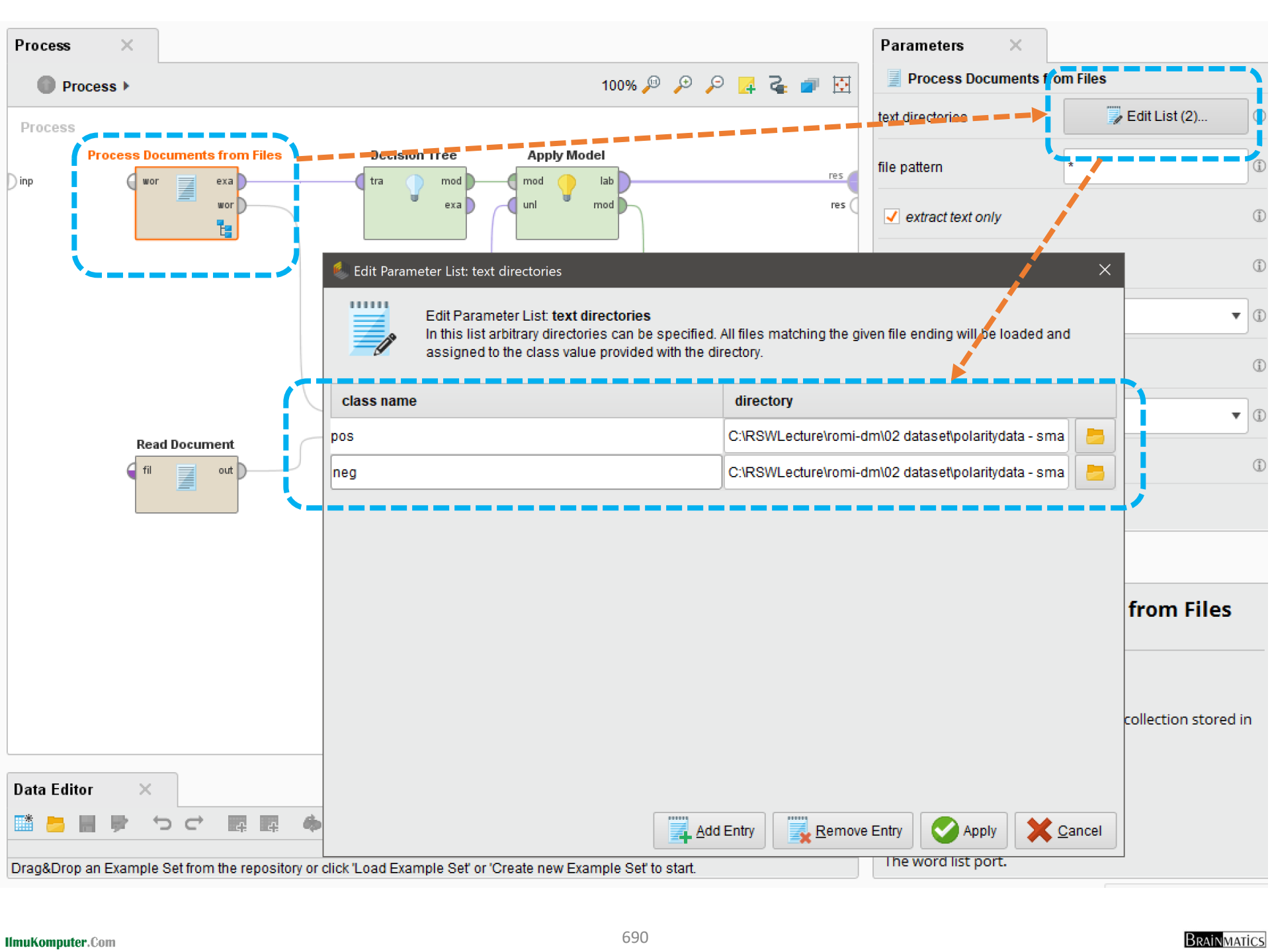
- Dengan berbagai konsep dan teknik yang anda kuasai, lakukan text classification pada dataset polarity data - small

- Gunakan algoritma Decision Tree untuk membentuk model

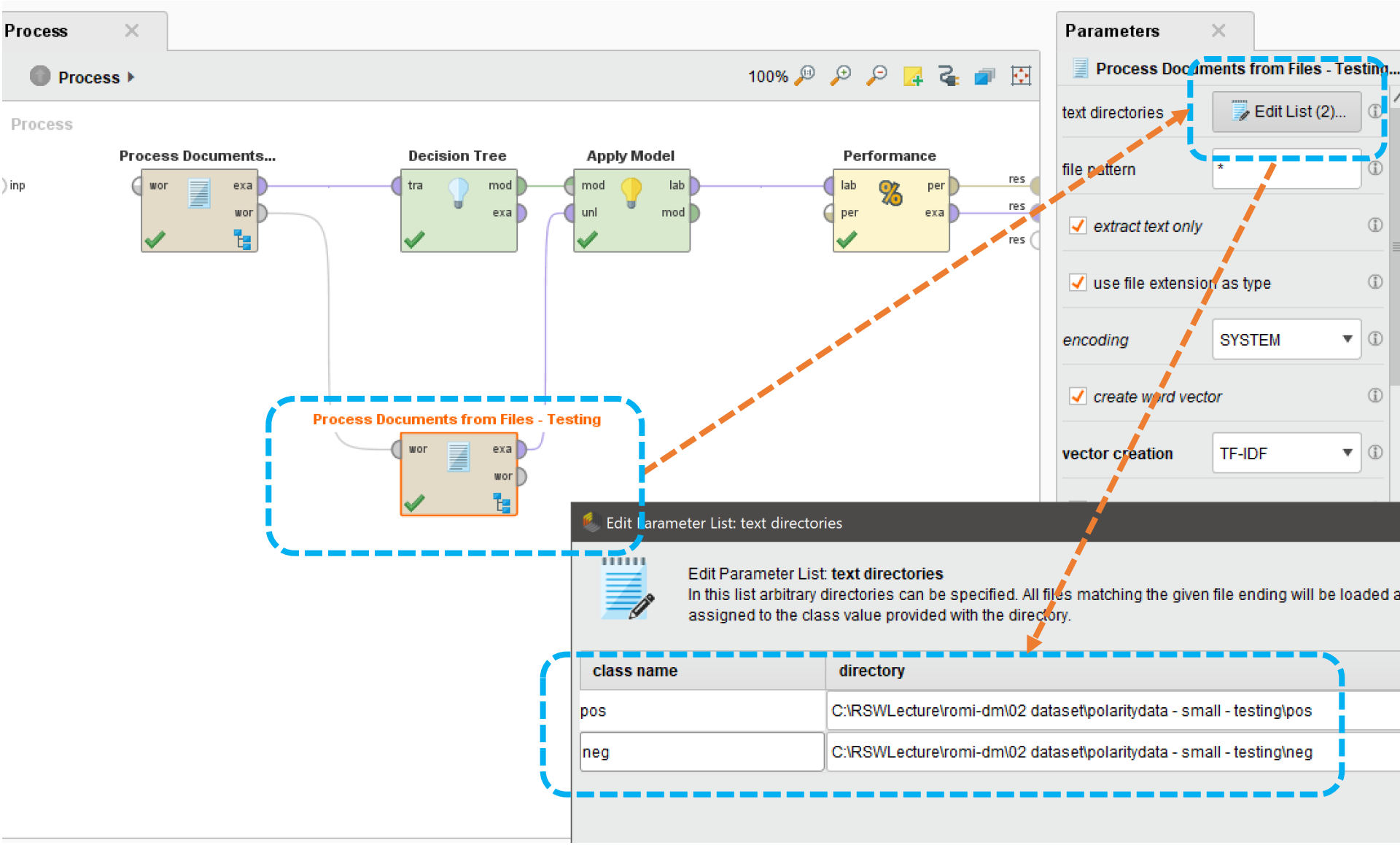
- Ambil 1 artikel di dalam folder polaritydata - small – testing , misalnya dalam folder pos, uji apakah artikel tersebut diprediksi termasuk sentiment negative atau positive


Process Documents from Files
wor
exa
wor
Decision Tree
tra
mod
exa
Apply Model
mod
lab
unl
mod
Process Documents
wor
exa
doc
wor
doc
Read Document
fil
out


Tokenize
doc
doc
Filter Stopwords (English)
doc
doc
Stem (Porter)
doc
doc


Process
Process
100%
Process
Process Documents from Files
Decision Tree
Apply Model
inp
wor
exa
wor
tra
mod
exa
mod
lab
unl
mod
res
res
Read Document
fil
out
Parameters
Process Documents from Files
text directories
Edit List (2)...
file pattern
*
extract text only
Edit Parameter List: text directories
Edit Parameter List: text directories
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.
class name
directory
pos
C:\RSWLecture\romi-dm\02 dataset\polaritydata - sma
neg
C:\RSWLecture\romi-dm\02 dataset\polaritydata - sma
Add Entry
Remove Entry
Apply
Cancel
from Files
collection stored in
Data Editor
Drag&Drop an Example Set from the repository or click 'Load Example Set' or 'Create new Example Set' to start.
The word list port.

Ukur Akurasi dari polaritydata-small-testing

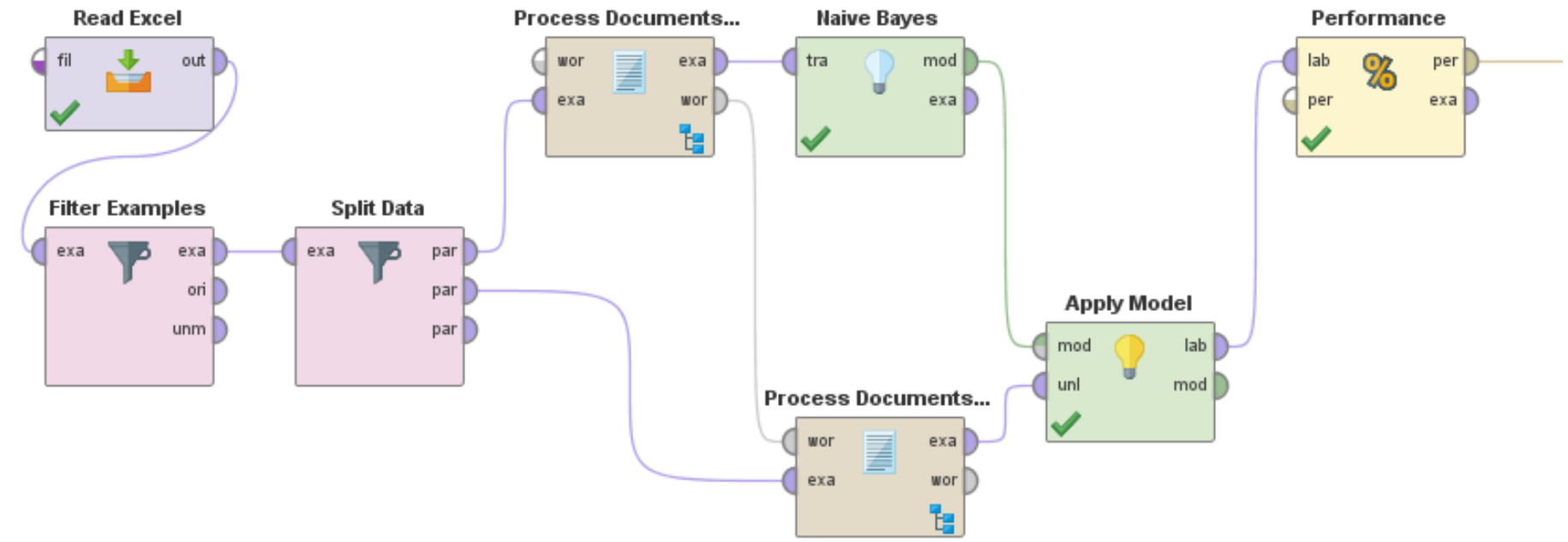


Latihan

- Dengan berbagai konsep dan teknik yang anda kuasai, lakukan text classification pada dataset polarity data
- Terapkan beberapa metode feature selection, baik filter maupun wrapper
- Lakukan komparasi terhadap berbagai algoritma klasifikasi, dan pilih yang terbaik

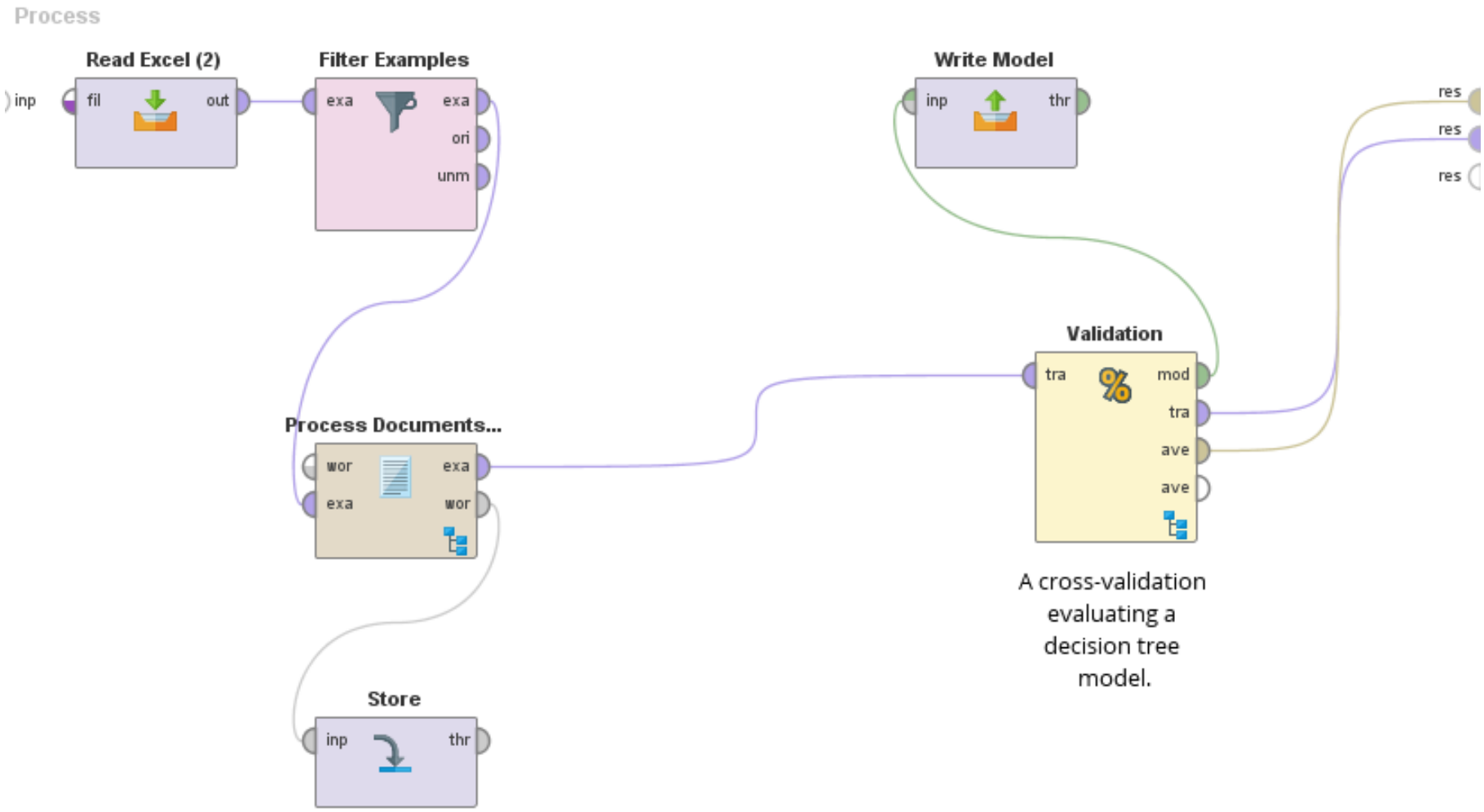
Latihan

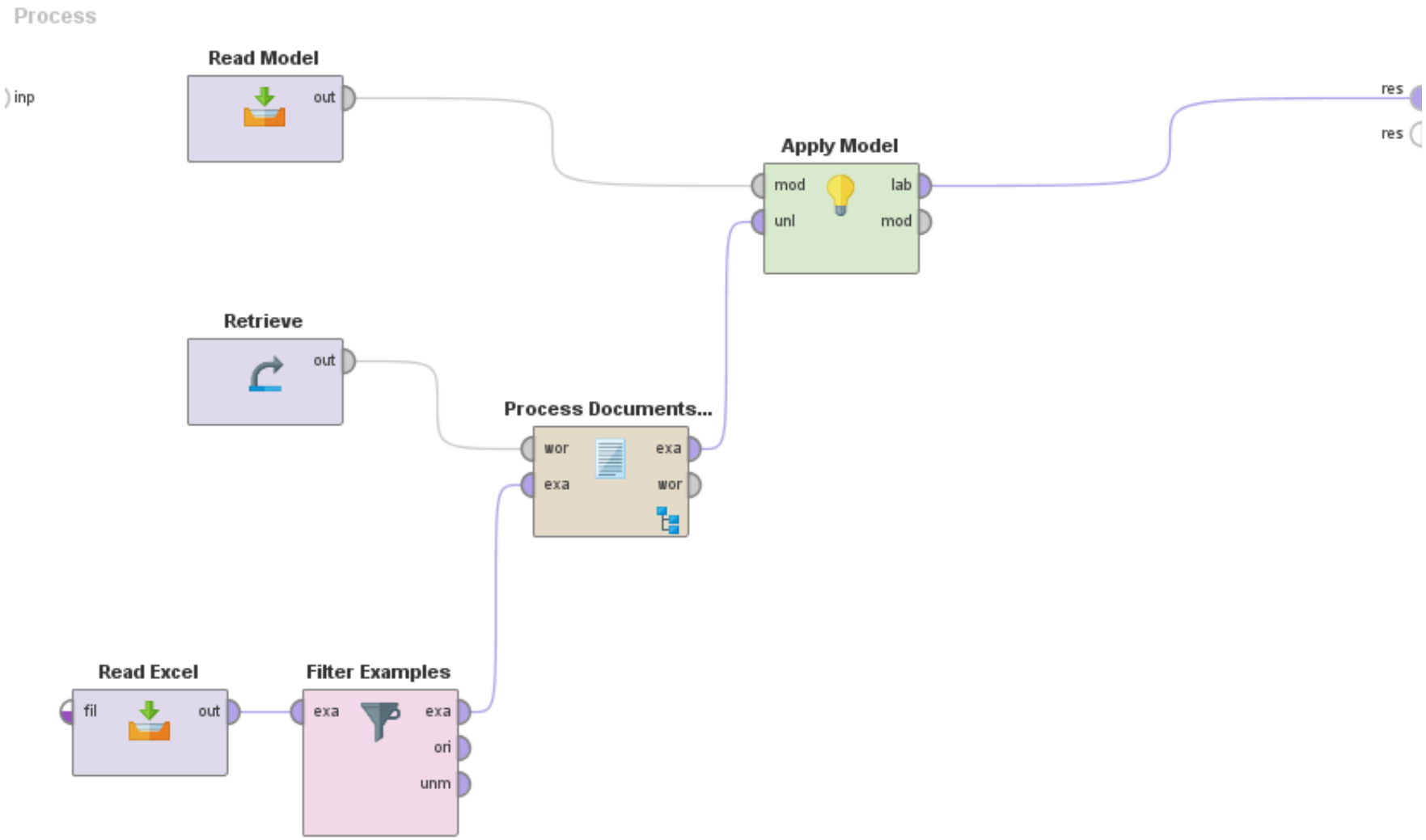
- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 2: Predicting the Gender of Blog Authors, p 287-301
- Datasets: blog-gender-dataset.xslx
- Split Data: 50% data training dan 50% data testing
- Gunakan algoritma Naïve Bayes
- Apply model yang dihasilkan untuk data testing
- Ukur performance nya

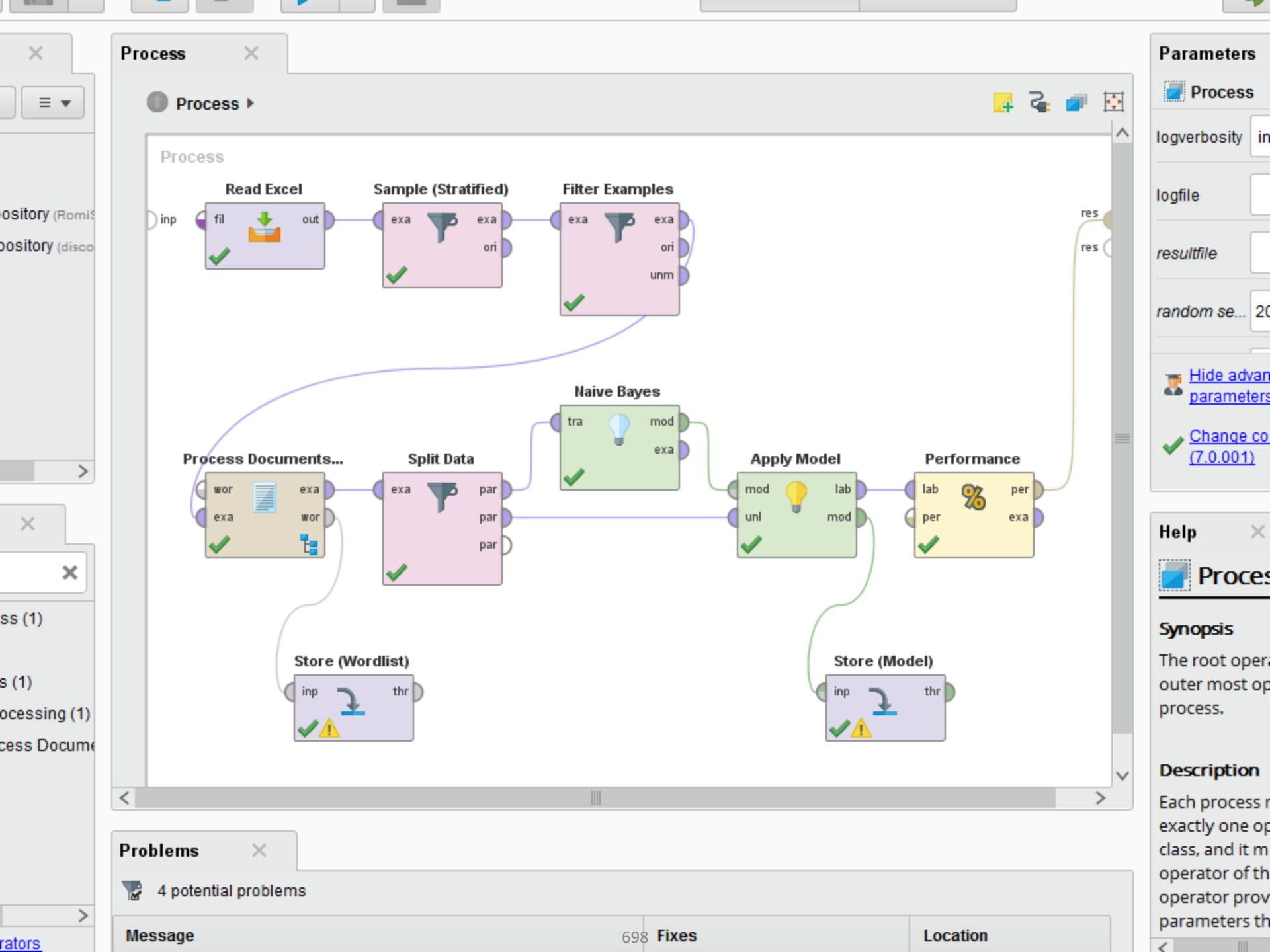


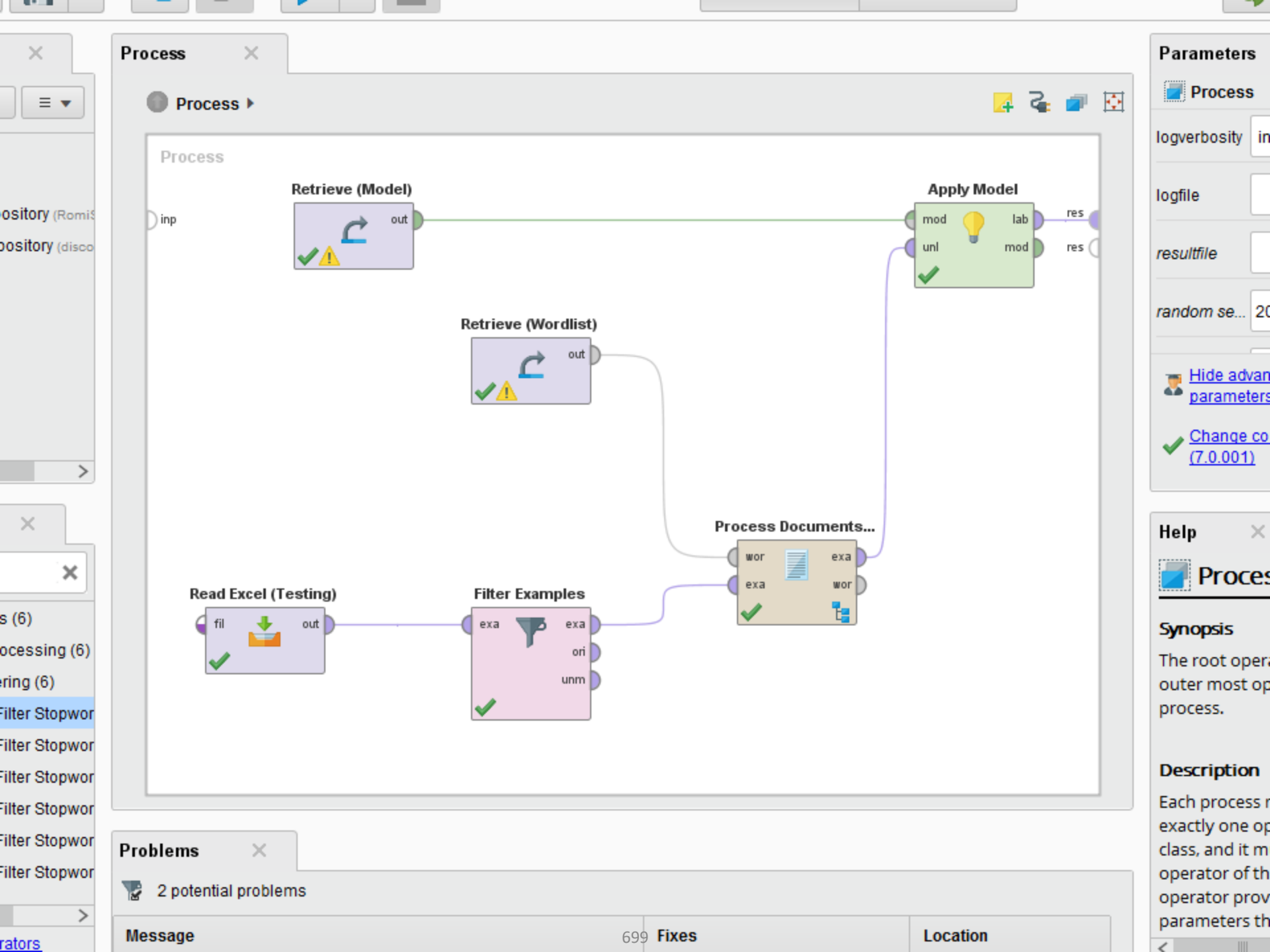
Latihan

- Lakukan eksperimen mengikuti buku Vijay Kotu (Predictive Analytics and Data Mining) Chapter 9 (Text Mining), Case Study 2: Predicting the Gender of Blog Authors, p 287-301
- Datasets:
  - blog-gender-dataset.xslx
  - blog-gender-dataset-testing.xlsx
- Gunakan 10-fold X validation dan operator write model (read model), store (retrieve)


Process
Read Excel (2)
inp
fil
out
Filter Examples
exa
exa
ori
unm
Write Model
inp
thr
res
res
res
Validation
tra
mod
tra
ave
ave
Process Documents...
wor
exa
exa
wor
Store
inp
thr
A cross-validation evaluating a decision tree model.


Process
inp
Read Model
out
Apply Model
mod
unl
lab
mod
res
res
Retrieve
out
Process Documents...
wor
exa
exa
wor
Read Excel
fil
out
Filter Examples
exa
exa
ori
unm


Process
Process
Process
Read Excel
Sample (Stratified)
Filter Examples
Naive Bayes
Process Documents...
Split Data
Apply Model
Performance
Store (Wordlist)
Store (Model)
Parameters
Process
logverbosity
logfile
resultfile
random se...
Hide advan
parameters
Change co
(7.0.001)
Help
Proces
Synopsis
The root opera
outer most op
process.
Description
Each process r
exactly one op
class, and it m
operator of th
operator prov
parameters th
Problems
4 potential problems
Message
Fixes
Location


Process
Process
Process
Retrieve (Model)
inp
out
Apply Model
mod
unl
lab
mod
res
res
Retrieve (Wordlist)
out
Process Documents...
wor
exa
exa
wor
Read Excel (Testing)
fil
out
Filter Examples
exa
exa
ori
unm
Parameters
Process
logverbosity
logfile
resultfile
random se...
Hide advan
parameters
Change co
(7.0.001)
Help
Proces
Synopsis
The root opera
outer most op
process.
Description
Each process r
exactly one op
class, and it m
operator of th
operator prov
parameters th
Problems
2 potential problems
Message
Fixes
Location

Post-Test

1. Jelaskan perbedaan antara data, informasi dan pengetahuan!
2. Jelaskan apa yang anda ketahui tentang data mining!
3. Sebutkan peran utama data mining!
4. Sebutkan pemanfaatan dari data mining di berbagai bidang!
5. Pengetahuan atau pola apa yang bisa kita dapatkan dari data di bawah?

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	SMAN 7	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

5.4 Data Mining Laws

Tom Khabaza, Nine Laws of Data Mining, 2010
(*http://khabaza.codimension.net/index_files/9laws.htm*)

Data Mining Laws

1. Business objectives are the origin of every data mining solution
2. Business knowledge is central to every step of the data mining process
3. Data preparation is more than half of every data mining process
4. There is no free lunch for the data miner
5. There are always patterns
6. Data mining amplifies perception in the business domain
7. Prediction increases information locally by generalisation
8. The value of data mining results is not determined by the accuracy or stability of predictive models
9. All patterns are subject to change

Tom Khabaza, Nine Laws of Data Mining, 2010
(*http://khabaza.codimension.net/index_files/9laws.htm*)

1 Business Goals Law

Business objectives are the origin of every data mining solution

- This defines the field of data mining: data mining is concerned with solving business problems and achieving business goals
- Data mining is not primarily a technology; it is a process, which has one or more business objectives at its heart
- Without a business objective, there is no data mining
- The maxim: "Data Mining is a Business Process"

2 Business Knowledge Law

Business knowledge is central to every step of the data mining process

- A naive reading of CRISP-DM would see business knowledge used at the start of the process in defining goals, and at the end of the process in guiding deployment of results
- This would be to miss a key property of the data mining process, that business knowledge has a central role in every step

2 Business Knowledge Law

1. Business understanding must be based on business knowledge, and so must the mapping of business objectives to data mining goals
2. Data understanding uses business knowledge to understand which data is related to the business problem, and how it is related
3. Data preparation means using business knowledge to shape the data so that the required business questions can be asked and answered
4. Modelling means using data mining algorithms to create predictive models and interpreting both the models and their behaviour in business terms – that is, understanding their business relevance
5. Evaluation means understanding the business impact of using the models
6. Deployment means putting the data mining results to work in a business process

3 Data Preparation Law

Data preparation is more than half of every data mining process

- Maxim of data mining: most of the effort in a data mining project is spent in data acquisition and preparation, and informal estimates vary from 50 to 80 percent
- The purpose of data preparation is:
  1. To put the data into a form in which the data mining question can be asked
  2. To make it easier for the analytical techniques (such as data mining algorithms) to answer it

4 No Free Lunch Theory

There is No Free Lunch for the Data Miner (NFL-DM)

The right model for a given application can only be discovered by experiment

- Axiom of machine learning: if we knew enough about a problem space, we could choose or design an algorithm to find optimal solutions in that problem space with maximal efficiency
- Arguments for the superiority of one algorithm over others in data mining rest on the idea that data mining problem spaces have one particular set of properties, or that these properties can be discovered by analysis and built into the algorithm
- However, these views arise from the erroneous idea that, in data mining, the data miner formulates the problem and the algorithm finds the solution
- In fact, the data miner both formulates the problem and finds the solution – the algorithm is merely a tool which the data miner uses to assist with certain steps in this process

4 No Free Lunch Theory

- If the problem space were well-understood, the data mining process would not be needed
  - Data mining is the process of searching for as yet unknown connections
- For a given application, there is not only one problem space
  - Different models may be used to solve different parts of the problem
  - The way in which the problem is decomposed is itself often the result of data mining and not known before the process begins
- The data miner manipulates, or "shapes", the problem space by data preparation, so that the grounds for evaluating a model are constantly shifting
- There is no technical measure of value for a predictive model
- The business objective itself undergoes revision and development during the data mining process
  - so that the appropriate data mining goals may change completely

5 Watkins' Law

There are always patterns

- This law was first stated by David Watkins
- There is always something interesting to be found in a business-relevant dataset, so that even if the expected patterns were not found, something else useful would be found
- A data mining project would not be undertaken unless business experts expected that patterns would be present, and it should not be surprising that the experts are usually right

6 Insight Law

Data mining amplifies perception in the business domain

- How does data mining produce insight? This law approaches the heart of data mining – why it must be a business process and not a technical one
  - Business problems are solved by people, not by algorithms
- The data miner and the business expert "see" the solution to a problem, that is the patterns in the domain that allow the business objective to be achieved
  - Thus data mining is, or assists as part of, a perceptual process
  - Data mining algorithms reveal patterns that are not normally visible to human perception
- The data mining process integrates these algorithms with the normal human perceptual process, which is active in nature
- Within the data mining process, the human problem solver interprets the results of data mining algorithms and integrates them into their business understanding

7 Prediction Law

Prediction increases information locally by generalisation

- "Predictive models" and "predictive analytics" means "predict the most likely outcome"
- Other kinds of data mining models, such as clustering and association, are also characterised as "predictive"; this is a much looser sense of the term:
  - A clustering model might be described as "predicting" the group into which an individual falls
  - An association model might be described as "predicting" one or more attributes on the basis of those that are known
- What is "prediction" in this sense? What do classification, regression, clustering and association algorithms and their resultant models have in common?
  - The answer lies in "scoring", that is the application of a predictive model to a new example
  - The available information about the example in question has been increased, locally, on the basis of the patterns found by the algorithm and embodied in the model, that is on the basis of generalisation or induction

8 Value Law

The value of data mining results is not determined by the accuracy or stability of predictive models

- Accuracy and stability are useful measures of how well a predictive model makes its predictions
  - Accuracy means how often the predictions are correct
  - Stability means how much the predictions would change if the data used to create the model were a different sample from the same population
- The value of a predictive model arises in two ways:
  - The model's predictions drive improved (more effective) action
  - The model delivers insight (new knowledge) which leads to improved strategy

9 Law of Change

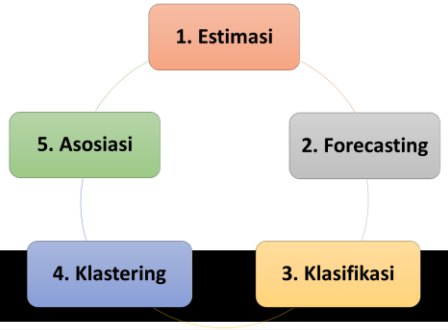
All patterns are subject to change

- The patterns discovered by data mining do not last forever
- In marketing and CRM applications of data mining, it is well-understood that patterns of customer behaviour are subject to change over time
  - Fashions change, markets and competition change, and the economy changes as a whole; for all these reasons, predictive models become out-of-date and should be refreshed regularly or when they cease to predict accurately
  - The same is true in risk and fraud-related applications of data mining. Patterns of fraud change with a changing environment and because criminals change their behaviour in order to stay ahead of crime prevention efforts

Tugas Menyelesaikan Masalah Organisasi

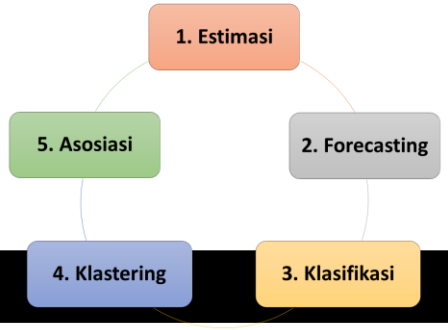
- Analisis masalah dan kebutuhan yang ada di organisasi lingkungan sekitar anda
- Kumpulkan dan review dataset yang tersedia, dan hubungkan masalah dan kebutuhan tadi dengan data yang tersedia (analisis dari 5 peran data mining)
  - Bila memungkinkan pilih beberapa peran sekaligus untuk mengolah data tersebut, misalnya: lakukan association (analisis faktor), sekaligus estimation atau clustering
- Lakukan proses CRISP-DM untuk menyelesaikan masalah yang ada di organisasi sesuai dengan data yang didapatkan
  - Pada proses data preparation, lakukan data cleaning (replace missing value, replace, filter attribute) sehingga data siap dimodelkan
  - Lakukan juga komparasi algoritma dan *feature selection* untuk memilih pola dan model terbaik
  - Rangkumkan evaluasi dari pola/model/knowledge yang dihasilkan dan relasikan hasil evaluasi dengan deployment yang dilakukan
- Rangkumkan dalam bentuk slide dengan contoh studi kasus Sarah untuk membantu bidang marketing

Studi Kasus Organisasi



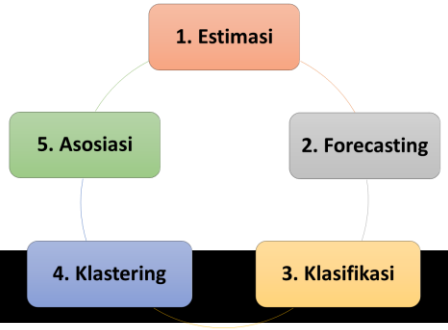
Organisasi	Masalah	Tujuan	Dataset
KPK	• Sulitnya mengidentifikasi profil koruptor • Tidak patuhnya WL dalam LHKPN	• Klasifikasi Profil Pelaku Korupsi • Asosiasi Atribut Pelaku Korupsi • Klasifikasi Kepatuhan LHKPN • Estimasi Penentuan Angka Tuntutan	• LHKPN • Penuntutan
BSM	Sulit mengidentifikasi faktor apa yang mempengaruhi kualitas pembiayaan	Klasifikasi kualitas profil nasabah	Data pembiayaan nasabah
LKPP	Banyaknya konsultasi dan pertanyaan dari berbagai instansi yg harus dijawab	• Asosiasi pola pertanyaan instansi • Klasifikasi jenis pertanyaan	Data konsultasi
BPPK	Sulitnya penanganan tweet dari masyarakat, apakah terkait pertanyaan, keluhan atau saran	Klasifikasi dan Klastering text mining dari keluhan atau pertanyaan atau saran di media sosial	Data twitter masyarakat
Universitas Siliwangi	Tingkat kelulusan tepat waktu belum maksimal (apakah dikarenakan faktor jurusan atau faktor lain?)	Klasifikasi data kelulusan mahasiswa	Data mahasiswa

Studi Kasus Organisasi



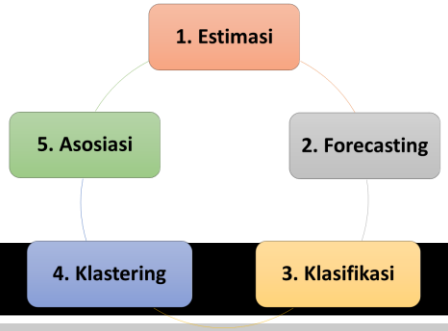
Organisasi	Masalah	Tujuan	Dataset
Kemenkeu (DJPB)	Sulit menentukan faktor refinement indicator kinerja	1. Seberapa erat hubungan antar komponen terhadap potensi penyempurnaan 2. Klastering data kinerja organsiasi	Data kinerja organisasi
Kemenkeu (DJPB)	Sulit menentukan arah opini hasil audit kementerian	1. Melihat hubungan beberapa data terhadap opini 2. Klasifikasi profil kementerian	Data profil kementerian
Kemenkeu (DJPB)	Banyaknya pelaporan kanwil yang harus dianalisis dengan beragam atribut	1. Melihat hubungan beberapa indikator laporan kanwil terhadap akurasi 2. Klastering data pelaporan kanwil 3. Klasifikasi akurasi pelaporan kanwil	Data pelaporan kanwil
Kemenkeu (DJPB)	Sulit menentukan prioritas monitoring kanwil	1. Klastering data profil kanwil 2. Melihat hubungan beberapa atribut terhadap klaster profil kanwil	Data transaksi dan profil kanwil

Studi Kasus Organisasi



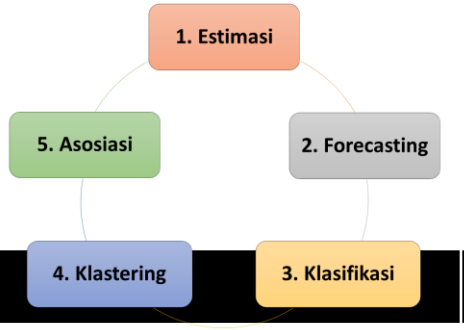
Organisasi	Masalah	Tujuan	Dataset
Kemenkeu (SDM)	Kebijakan masalah reward dan punishment untuk pegawai sering tidak efektif	Klasifikasi profil pegawai yang sering telat dan disiplin, sehingga terdeteksi lebih dini	Pegawai
Kemenkeu (SDM)	Rasio perempuan yang menjabat eselon 4/3/2/1 hanya 15%, padahal masuk PNS rasionya hampir imbang	• Klasifikasi dan klastering profile pejabat eselon 4/3/2/1 • Asosiasi jabatan dan atribut profile pegawai	Pegawai
Bank Indonesia	Peredaran uang palsu yang semakin banyak di Indonesia	• Asosiasi jumlah peredaran uang palsu dengan profil wilayah Indonesia • Klastering wilayah peredaran uang palsu	Peredaran Uang Palsu
Adira Finance	Rasio kredit macet yang semakin meninggi	• Klasifikasi kualitas kreditur yang lancar dan macet • Forecasting jumlah kredit macet • Tingkat hubungan kredit macet dengan berbagai atribut	Kreditur

Studi Kasus Organisasi



Organisasi	Masalah	Tujuan	Dataset
Kemsos Trima	Kompleksnya parameter penentuan tingkat kemiskinan rumah tangga di Indonesia	Klasifikasi profil rumah tangga miskin di kabupaten Brebes	Rumah tangga miskin di kabupaten Brebes
Kemsos Rahmat	Sulitnya menentukan rumah tangga yang diprioritaskan menerima bantuan sosial	Klastering profile rumah tangga miskin yang belum menerima bantuan	Rumah tangga miskin di kabupaten Belitung
Kemsos Septian	Sulitnya menentukan jenis penyakit kronis yang diprioritaskan menerima program Penerima Bantuan Iuran jaminan kesehatan (PBIJK)	Klasifikasi penyakit kronis yang diderita anggota rumah tangga miskin	Anggota rumah tangga di kabupaten Belitung
Kemsos Bayani	Sulitnya menentukan rumah tangga miskin di Indonesia		

Studi Kasus Organisasi



Organisasi	Masalah	Tujuan	Dataset
Kemsos Oki	Kompleksnya parameter penentuan tingkat kemiskinan rumah tangga di Indonesia	Klastering profil rumah tangga miskin di kabupaten Belitung	Data Terpadu Kesejahteraan Sosial (DTKS) kabupaten Belitung
Kemsos Dharma	Penerima Program Keluarga Harapan (PKH) yang tidak tepat sasaran	Klasifikasi faktor utama atribut yang berpengaruh pada penerima PKH	Data Terpadu Kesejahteraan Sosial (DTKS) kabupaten Belitung
Kemsos Dewi	Penentuan kebijakan penerima bantuan program rehabilitasi sosial rumah tidak layak huni	Klastering spesifikasi rumah pada data rumah tangga Data Terpadu Kesejahteraan Sosial (DTKS)	Data Terpadu Kesejahteraan Sosial (DTKS) kabupaten Seram bagian barat
Kemsos Dicksan	Banyaknya penerima bantuan sosial yang tidak tepat	Klastering profil rumah tangga miskin dari Data Terpadu	Data Terpadu Kesejahteraan

Terima Kasih

Romi Satria Wahono
romi@romisatriawahono.net
http://romisatriawahono.net
08118228331