

PERTEMUAN 9

Data warehouse and Business Intelligence



Pokok Bahasan

Pertemuan Ke-	Pokok Bahasan
1	Business Intelligence
2	Data Warehousing
3	Business Performance Management
4	Business Performance Management Methodologies
5	Pengantar Data Mining
6	Metode Learning Algoritma Data Mining
7	Review dan Quiz
8	UTS
9	Studi Kasus
10-15	Presentasi Tugas Kelompok
16	UAS



Rencana Pembelajaran

Tugas Kelompok

- ✓ Buat Kelompok maximal 4 orang/kelompok.
- ✓ Pengolahan data menggunakan tools rapidminer
- ✓ Menggunakan salahsatu metode data mining
- ✓ Dataset menggunakan data public atau private, setiap kelompok beda dataset
- ✓ Kumpulkan tugas kelompok tersebut berupa : makalah, dan powerpoint pada pertemuan 10 dan bisa dipresentasikan.
- ✓ Mengumpulkan draft artikel ilmiah.
- ✓ Nilai project & presentasi akan menjadi nilai kelompok, keaktifan dan nilai penguasaan materi.



Data warehouse dan Business Intelligence

Chapter 9:

Studi Kasus



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/)
 - Data.go.id
- 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)



Machine Learning Repository

Center for Machine Learning and Intelligent Systems



View ALL Data Sets

Browse Through: 360 Data Sets Table View <u>List View</u>

Default Task
Classification (262)
Regression (63)
Clustering (54)
Other (52)
Attribute Type
Categorical (37)

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)

lable view List v						
<u>Name</u>	<u>Data Types</u>	<u>Default Task</u>	Attribute Types	# Instances	<u>#</u> <u>Attributes</u>	<u>Year</u>
<u>Abalone</u>	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
UCI 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
Aa Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
N. a.						

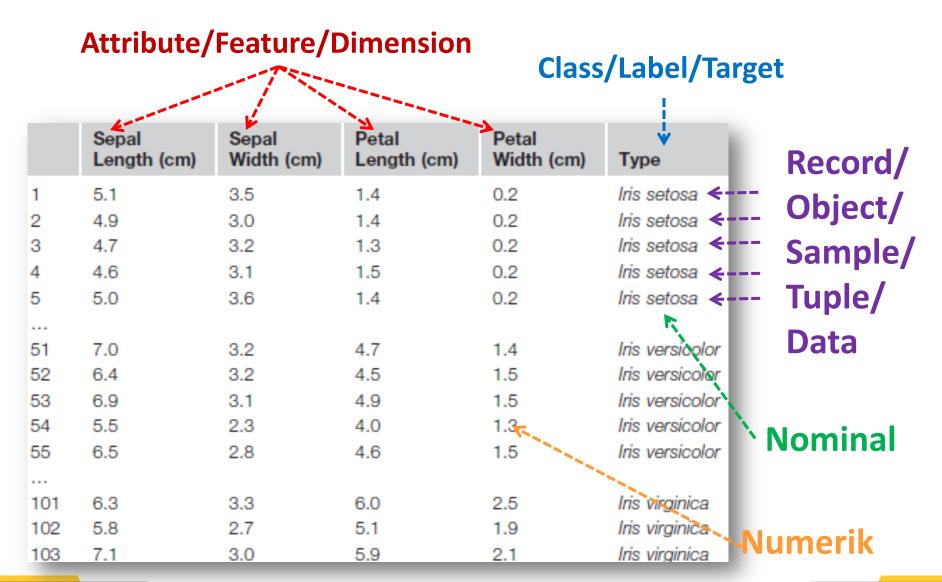


Data dan Format Data

- Data menyebutkan obyek-obyek dari sebuah konsep
 - Ditunjukkan sebagai baris dari tabel
- Metadata menggambarkan karakteristik dari konsep tersebut
 - Ditunjukkan sebagai kolom dari tabel
- Dukungan Format data
 - Oracle, IBM DB2, Microsoft SQL Server, MySQL,
 PostgreSQL, Ingres, Excel, Access, SPSS, CSV files
 dan berbagai format lain



Dataset (Himpunan Data)





2. Metode Data Mining (DM)

1. Estimation (Estimasi):

 Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc

2. Prediction/Forecasting (Prediksi/Peramalan):

 Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc

3. Classification (Klasifikasi):

Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, DynamicCC4.5),
 Naive Bayes, K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression, etc

4. Clustering (Klastering):

K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc.

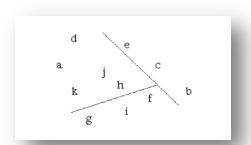
5. Association (Asosiasi):

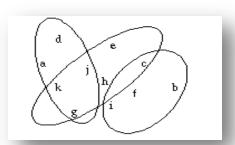
FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc.

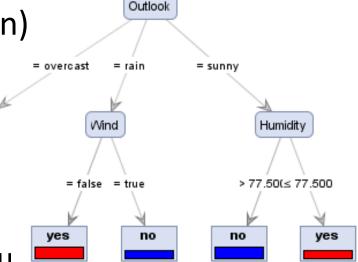


3. Pengetahuan (Pola/Model)

- 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

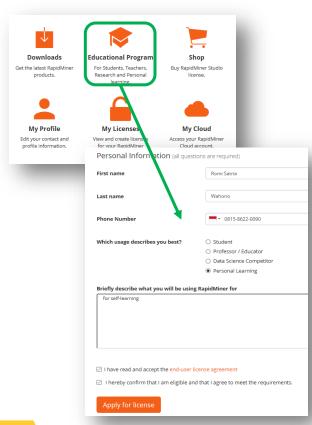
 Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna

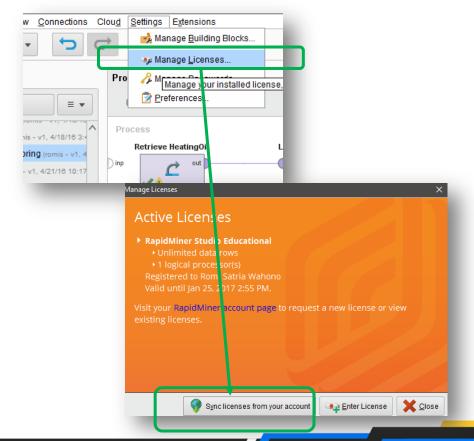
Keseimbangan diantaranya ketiganya diperlukan karena belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna



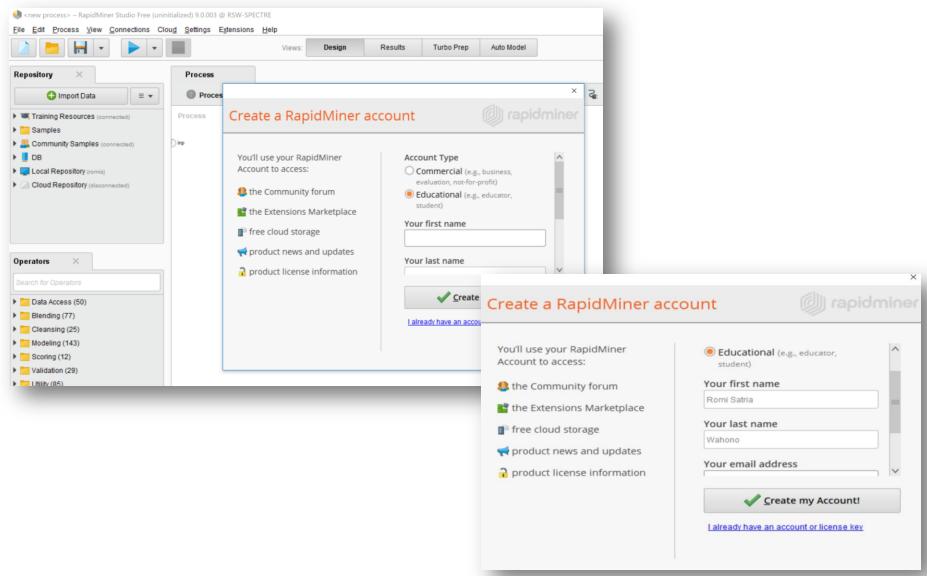
Instalasi dan Registrasi Lisensi Rapidminer

- Instal Rapidminer versi 7
- Registrasi account di rapidminer.com, dan lakukan dapatkan lisensi Educational Program untuk mengolah data tanpa batasan record











Fitur Rapidminer

- Menyediakan prosedur data mining dan machine learning termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi
- Proses data mining tersusun atas operatoroperator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI
- Mengintegrasikan proyek data mining Weka dan statistika R



Atribut Pada Rapidminer

- 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

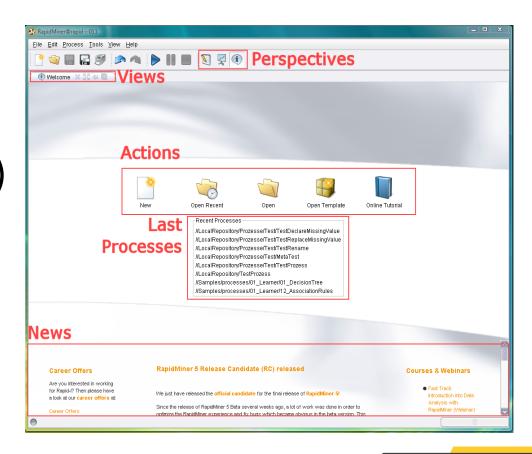
- 1. nominal: nilai secara kategori
- 2. binominal: nominal dua nilai
- polynominal: nominal lebih dari dua nilai
- 4. numeric: nilai numerik secara umum
- 5. integer: bilangan bulat
- 6. real: bilangan nyata
- 7. text: teks bebas tanpa struktur
- 8. date time: tanggal dan waktu
- 9. date: hanya tanggal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



Perspektif dan View

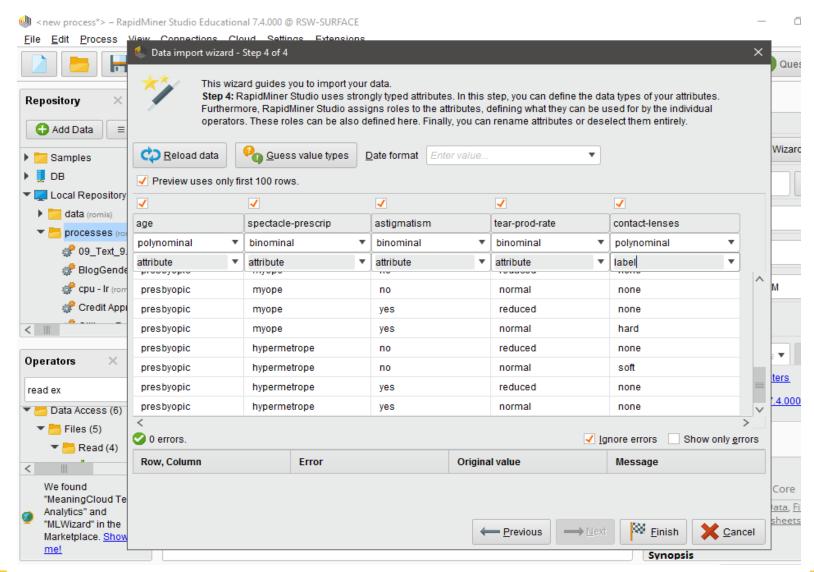
1. Perspektif Selamat Datang (Welcome perspective)

- Perspektif Desain (Design perspective)
- 3. Perspektif Hasil (Result perspective)



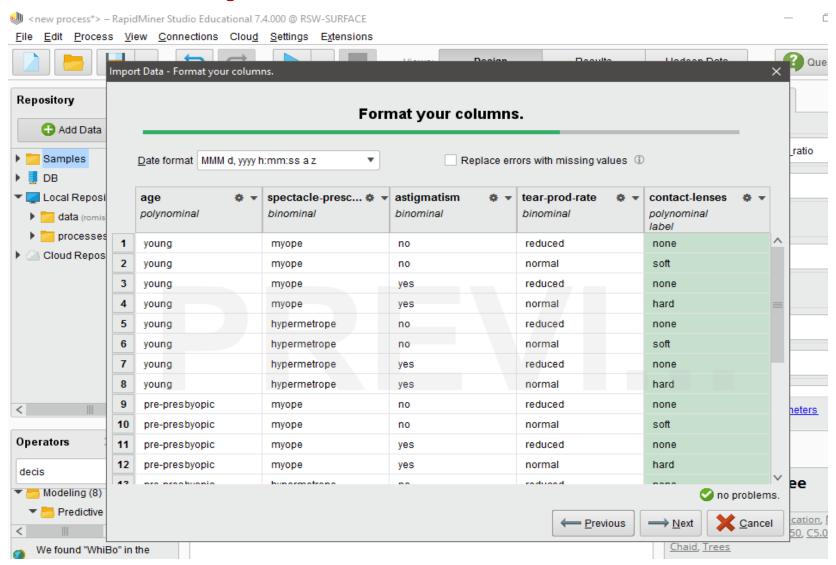


Read Excel Operator





Import Data Function





Perspektif Desain

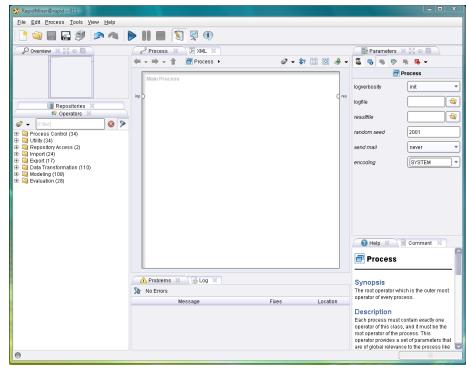
Perspektif pusat di mana

semua proses analisa dibuat dan dimanage

 Pindah ke Perspektif Desain dengan:



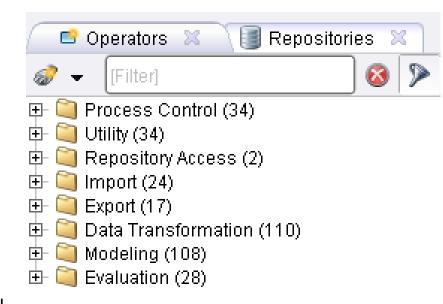
- Klik tombol paling kiri
- Atau gunakan menu
 View → Perspectives →
 Design





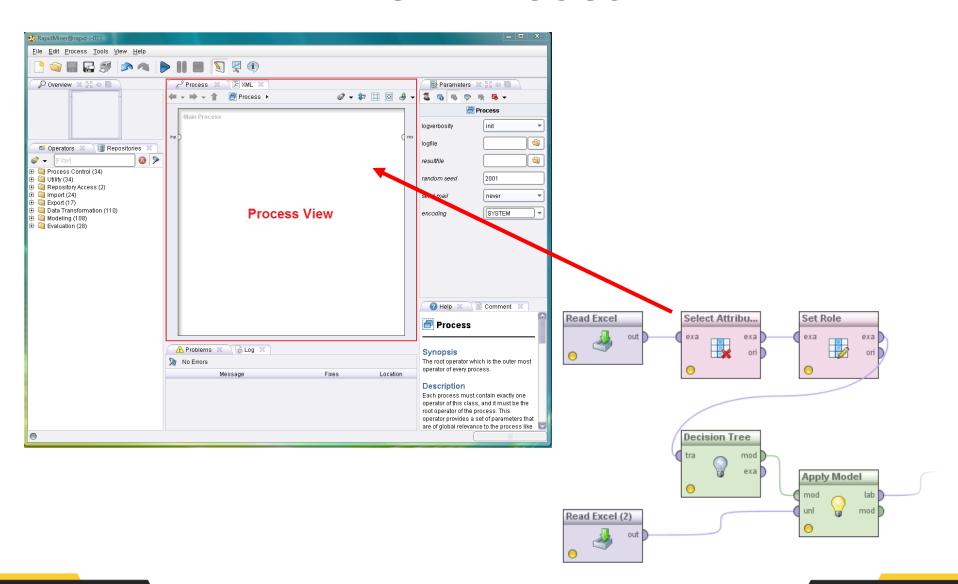
View Operator

- 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 yang sesungguhnya seperti klasifikasi, regresi, clustering, aturan asosiasi dll
- Evaluation
 Untuk menghitung kualitas dan perfomansi dari model





View Proses





View Parameter

 Operator kadang memerlukan parameter untuk bisa berfungsi

 Setelah operator dipilih di view Proses, parameternya ditampilkan di view ini

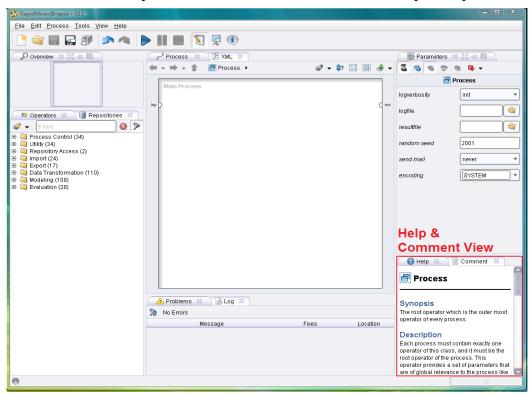






View Help dan View Comment

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





Synopsis

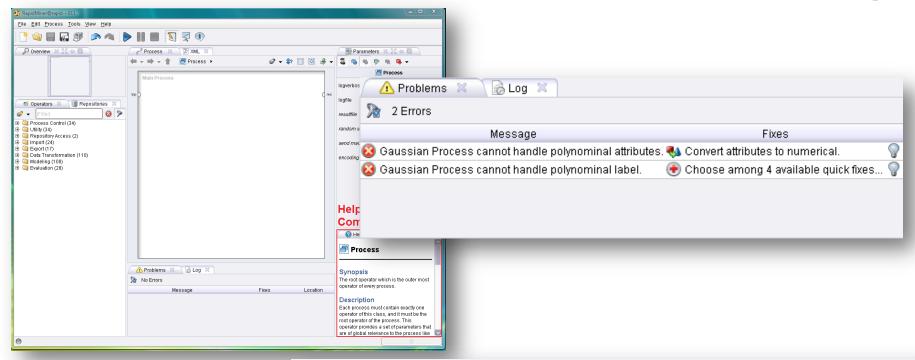
The root operator which is the outer most operator of every process.

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



View Problems dan View Log







Operator dan Proses

- Proses data mining pada dasarnya adalah proses analisa yang berisi alur kerja dari komponen data mining
- Komponen dari proses ini disebut operator, yang didefinisikan dengan:
 - 1. Deskripsi input
 - 2. Deskripsi output
 - 3. Aksi yang dilakukan
 - 4. Parameter yang diperlukan



Operator dan Proses

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

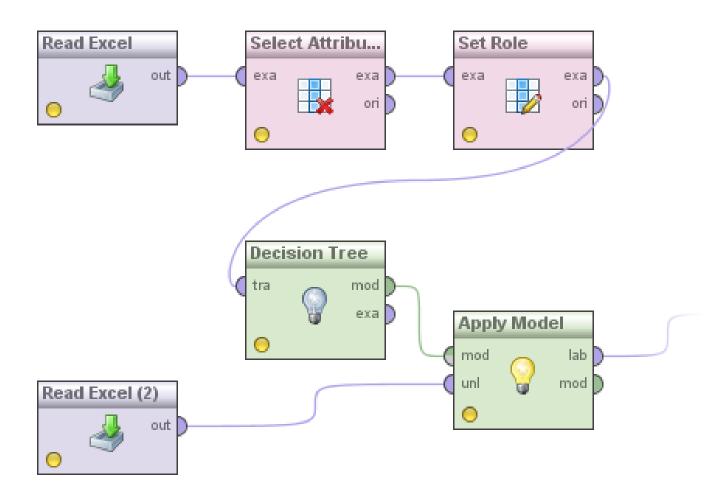




- Indikator status dari operator:
 - Lampu status: merah (tak tersambung), kuning (lengkap tetapi belum dijalankan), hijau (sudah behasil dijalankan)
 - Segitiga warning: bila ada pesan status
 - Breakpoint: bila ada breakpoint sebelum/sesudahnya
 - Comment: bila ada komentar
 - Subprocess: bila mempunyai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



Mendesain Proses

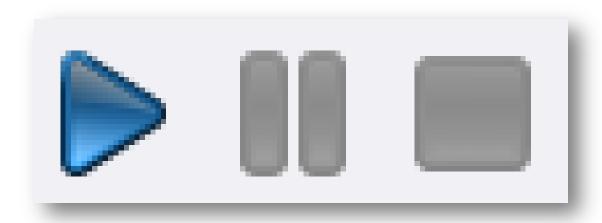




Menjalankan Proses

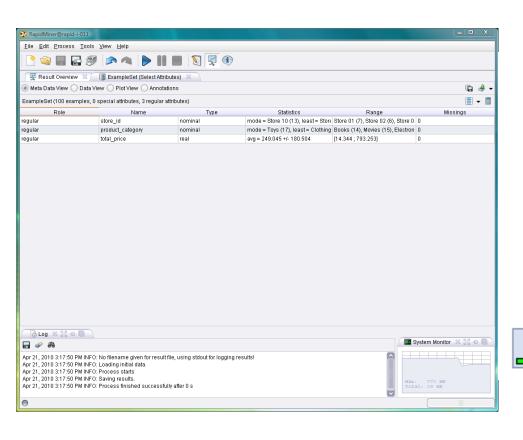
Proses dapat dijalankan dengan:

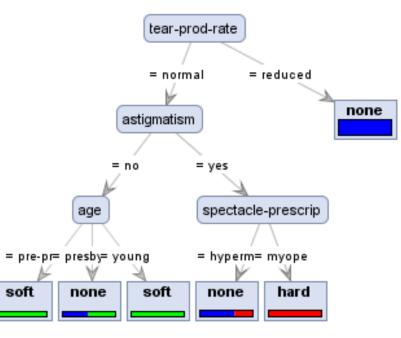
- Menekan tombol Play
- Memilih menu Process → Run
- Menekan kunci F11





Melihat Hasil





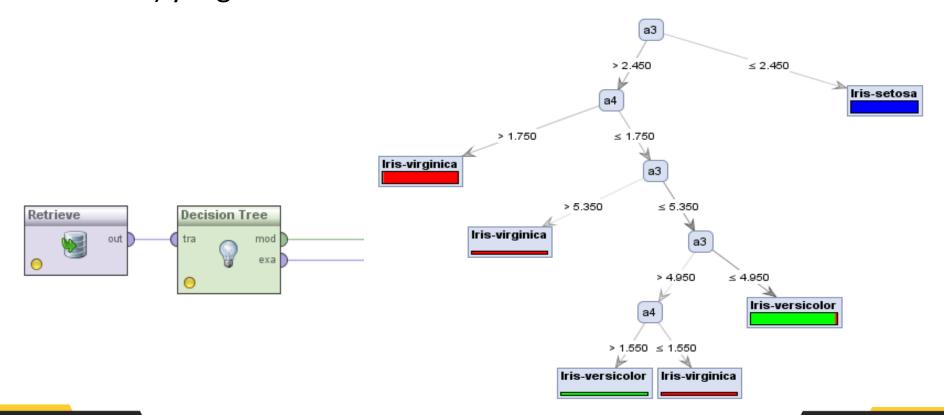


Studi Kasus 1



Latihan: Penentuan Jenis Bunga Iris

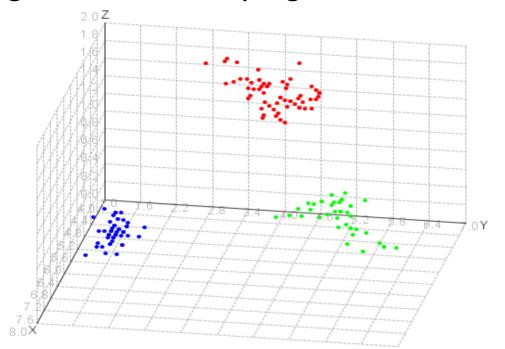
- 1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree
- Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk





Latihan: Klastering Jenis Bunga Iris

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



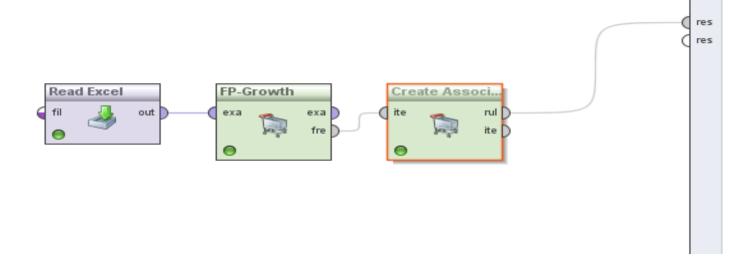


Studi Kasus 2

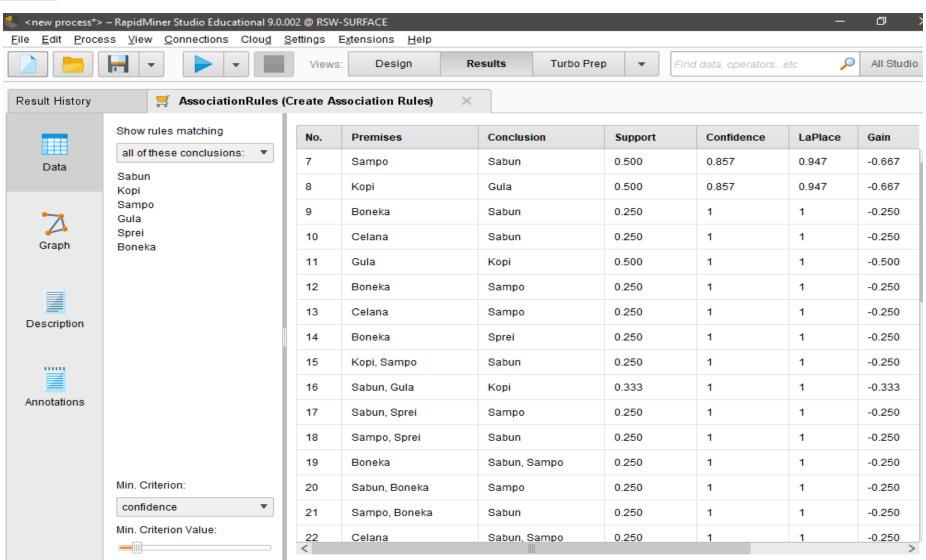


Latihan: Aturan Asosiasi Data Transaksi

- Lakukan training pada data transaksi (transaksi.xlsx)
- 2. Pilih metode yang tepat supaya menghasilkan pola









Ketentuan Pengerjaan Studi Kasus

- 1. Pengerjaan studi kasus menggunakan tools rapidminer
- 2. Hasil nilai tersebut dianalisa hipotesanya



Studi Kasus 3



Latihan: Klasifikasi Breast Cancer

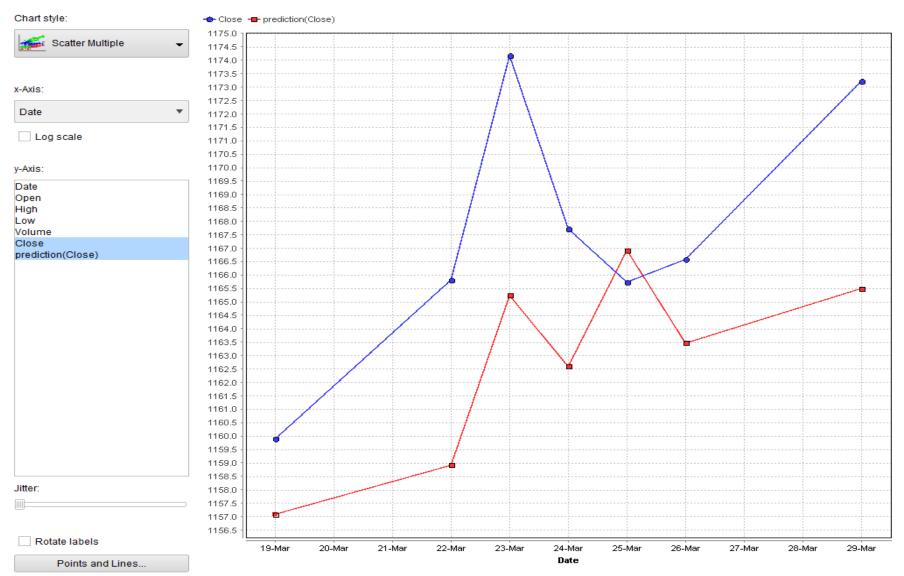
- Lakukan training pada data breast cancer (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra)
- 2. Gunakan operator Split Data untuk memecah data secara otomatis menjadi dua dengan perbandingan 0.9:0.1, di mana 0.9 untuk training dan 0.1 untuk testing
- 3. Pilih metode yang tepat supaya menghasilkan pola yang bisa menguji data testing 10%



Latihan: Forecasting Harga Saham

- 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
- Lakukan pengujian terhadap data baru (hargasaham-testing.xls), untuk model yang dihasilkan dari tahapan 1
- 4. Lakukan plot berupa grafik dari data yang terbentuk dengan menggunakan Scatter Multiple







Latihan: Forecasting Harga Saham

(Univariat)

ori



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		

Window size = 6
Step size = 1
Horizon = 1

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	

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.

res (



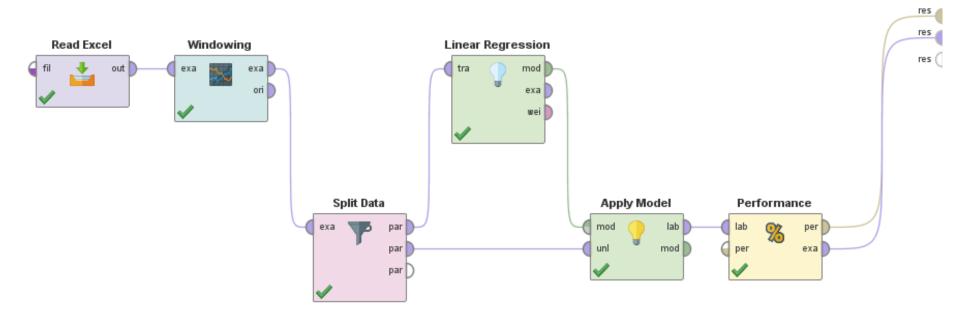
Latihan

- Lakukan training dengan menggunakan linear regression pada dataset hargasaham-traininguni.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



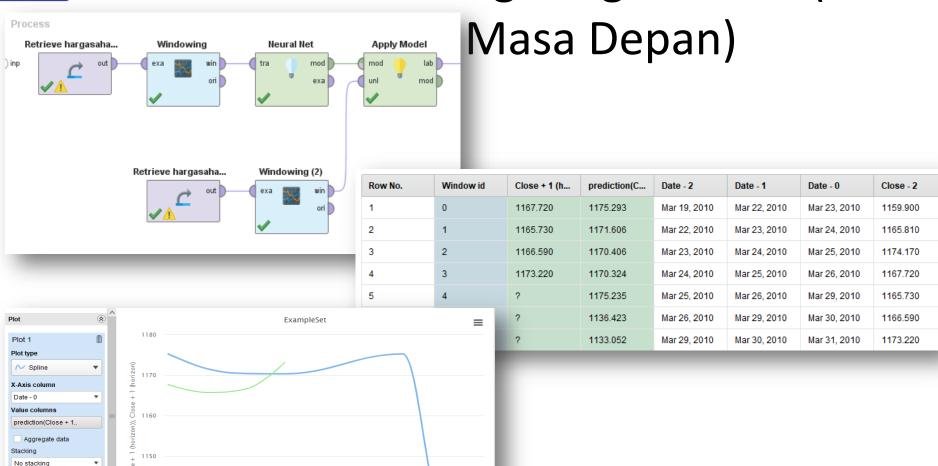
Forecasting Harga Saham (Data Lampau)

Mileter Class





Forecasting Harga Saham (Data



Plot style >>>

Add new plot

General

X-Axis

1140

Monday, Mar 29, 17:00

Date - 0

— Close + 1 (horizon)

25. Mar

prediction(Close + 1 (horizon))

prediction(Close + 1 (horizon)): 1,136.423190449824



Ketentuan Pengerjaan Studi Kasus

- 1. Pengerjaan studi kasus menggunakan tools rapidminer
- 2. Hasil nilai tersebut dianalisa hipotesanya