DATA MINING

PERTEMUAN KE-3

Proses Data Mining

5. Proses Data Mining

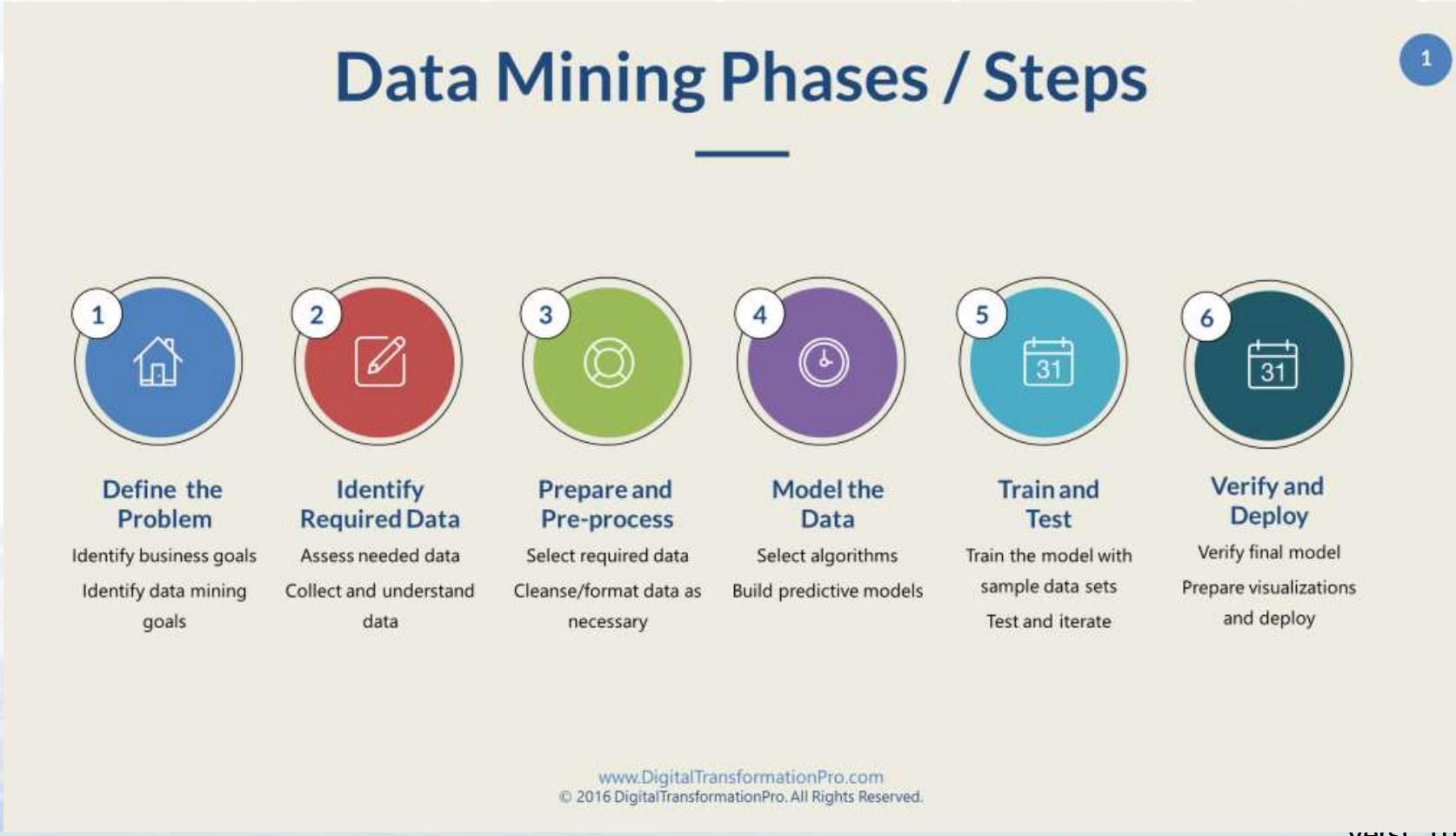
5.1 • Proses dan Tools Data Mining

5.2 • Penerapan Proses Data Mining

5.3 • Evaluasi Model Data Mining

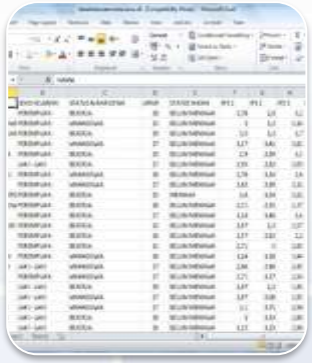
5.4 • Proses Data Mining berbasis CRISP-DM

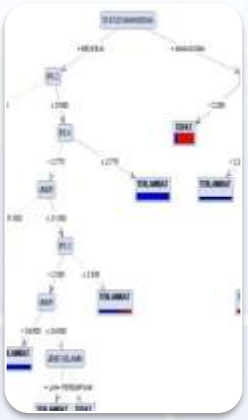
5.1 Proses dan Tools Data Mining

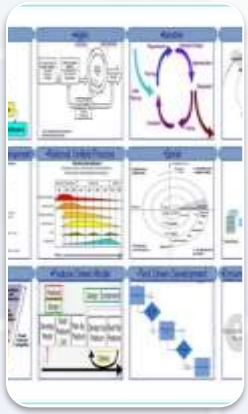


I.4

Proses Data Mining







1. Himpunan Data

(Pemahaman dan Pengolahan Data)

2. Metode Data Mining

(Pilih Metode Sesuai Karakter Data)

3. Pengetahuan

(Pola/Model/Rumus/ Tree/Rule/Cluster)

4. Evaluation

(Akurasi, AUC, RMSE, Lift Ratio,…)

DATA PRE-PROCESSING
Data Cleaning
Data Integration
Data Reduction
Data Transformation

Estimation
Prediction
Classification
Clustering
Association

Versi : 01

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

I.6

Public Data Set (UCI Repository)

UCI

About Citation Policy Donate a Data Set Contact

Search

Repository Web

Google

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

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

Versi : 01

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

Versi : 01

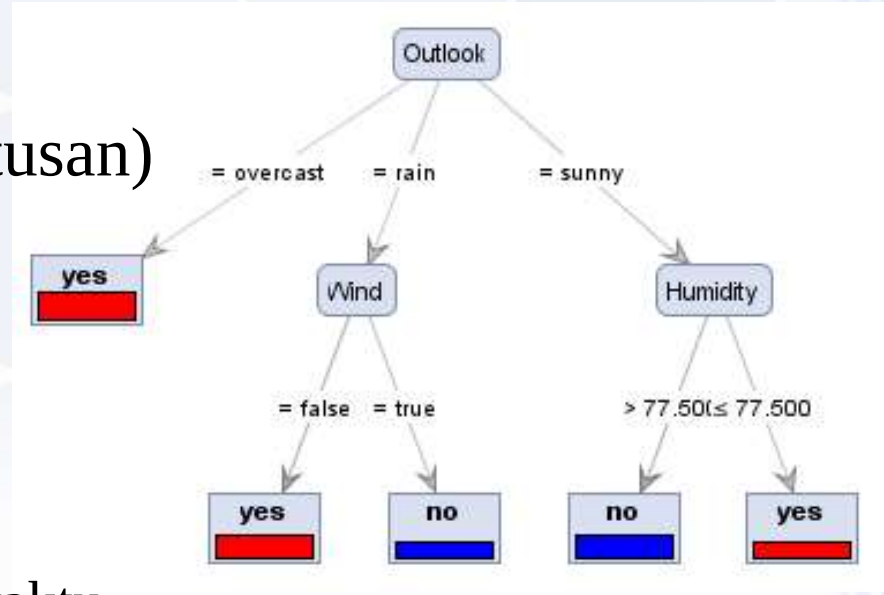
2. Metode Data Mining (DM)

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

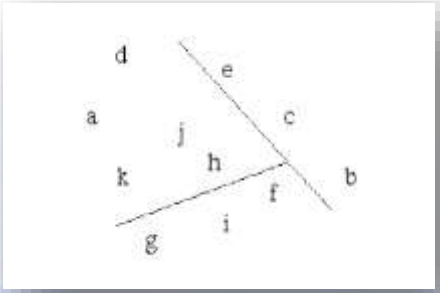
2. Decision Tree (Pohon Keputusan)

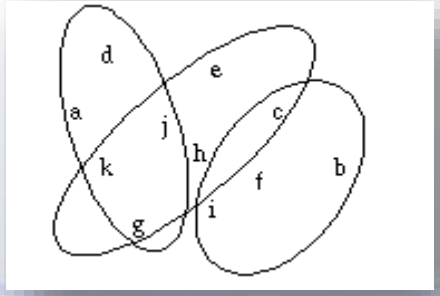


3. Tingkat Korelasi

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)

I.11 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

5.2 Tools Data Mining

Sejarah 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

- Menyediakan prosedur data mining dan machine learning termasuk: 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 proyek data mining Weka dan statistika R

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

I.16 Tipe Nilai Atribut pada Rapidminer

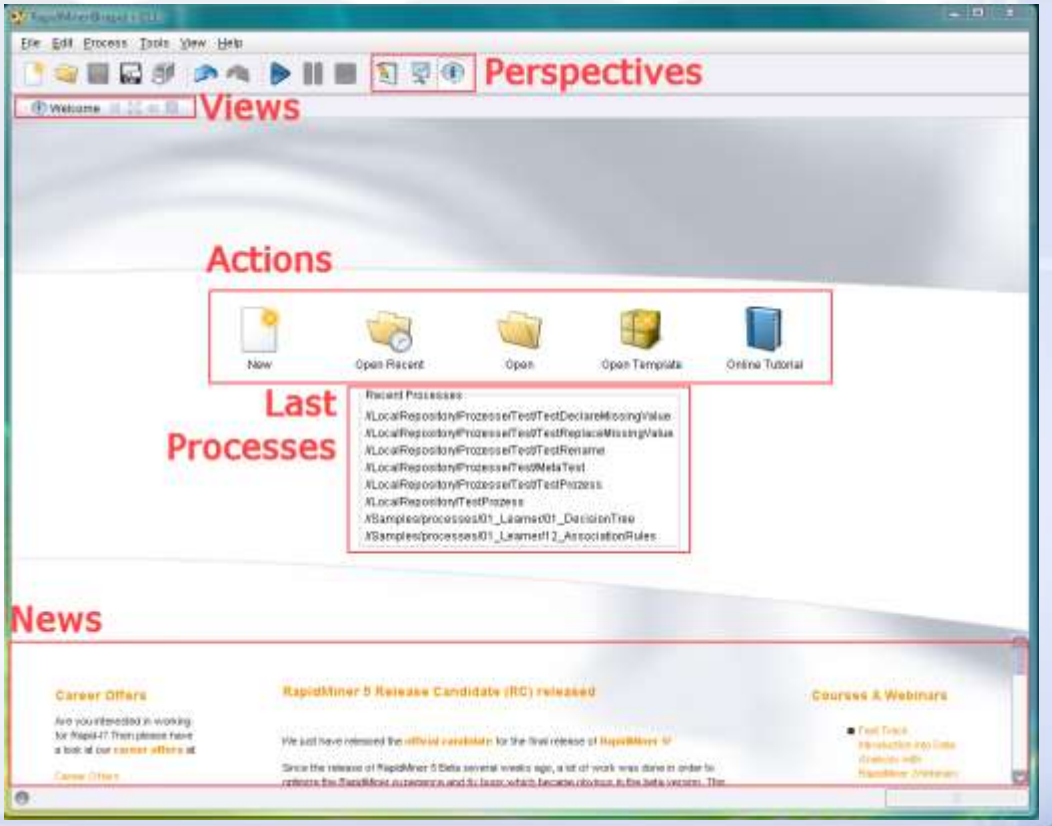
1. nominal: nilai secara kategori
2. binominal: nominal dua nilai
3. polynominal: nominal lebih dari dua nilai
4. numeric: nilai numerik secara umum
5. integer: bilangan bulat
6. real: bilangan nyata
7. text: teks bebas tanpa struktur
8. date_time: tanggal dan waktu
9. date: hanya tanggal
10. time: hanya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

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

Perspektif dan View

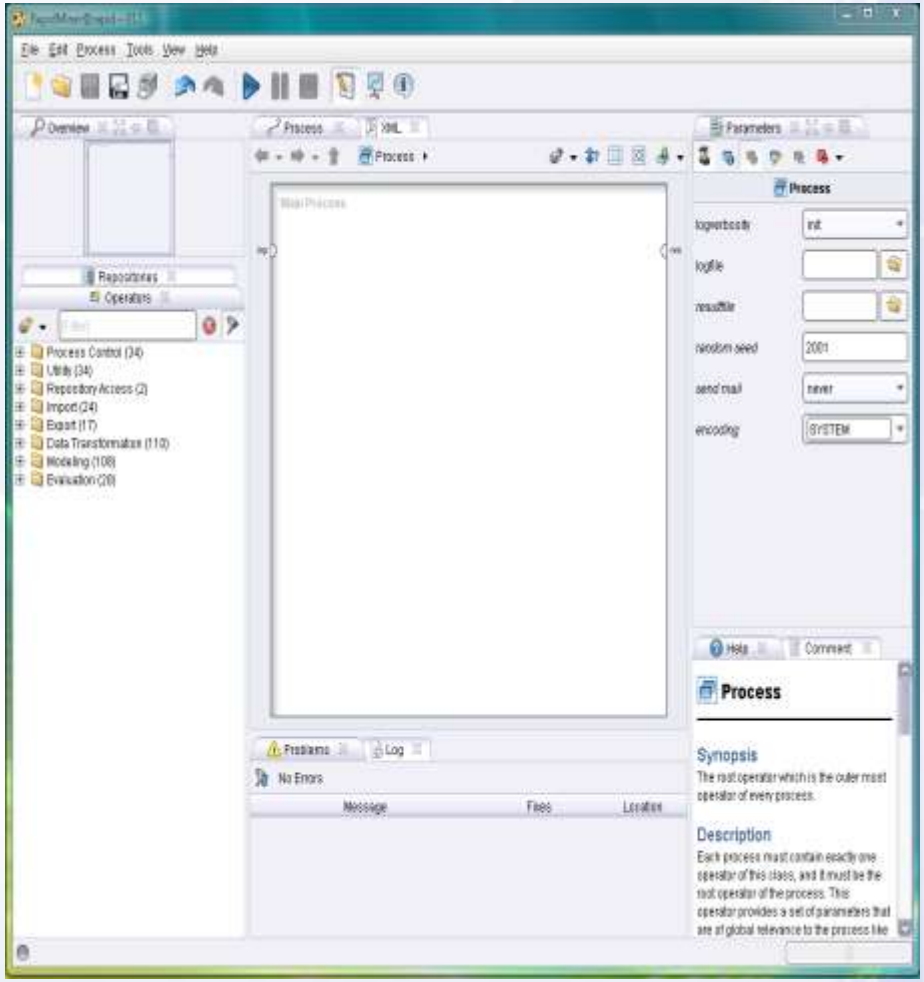
1. Perspektif Selamat Datang (Welcome perspective)
2. Perspektif Desain (Design perspective)
3. Perspektif Hasil (Result perspective)



Versi : 01

Perspektif Desain

- Perspektif pusat di mana semua proses analisa dibuat dan dimanage



- Pindah ke Perspektif Desain dengan:

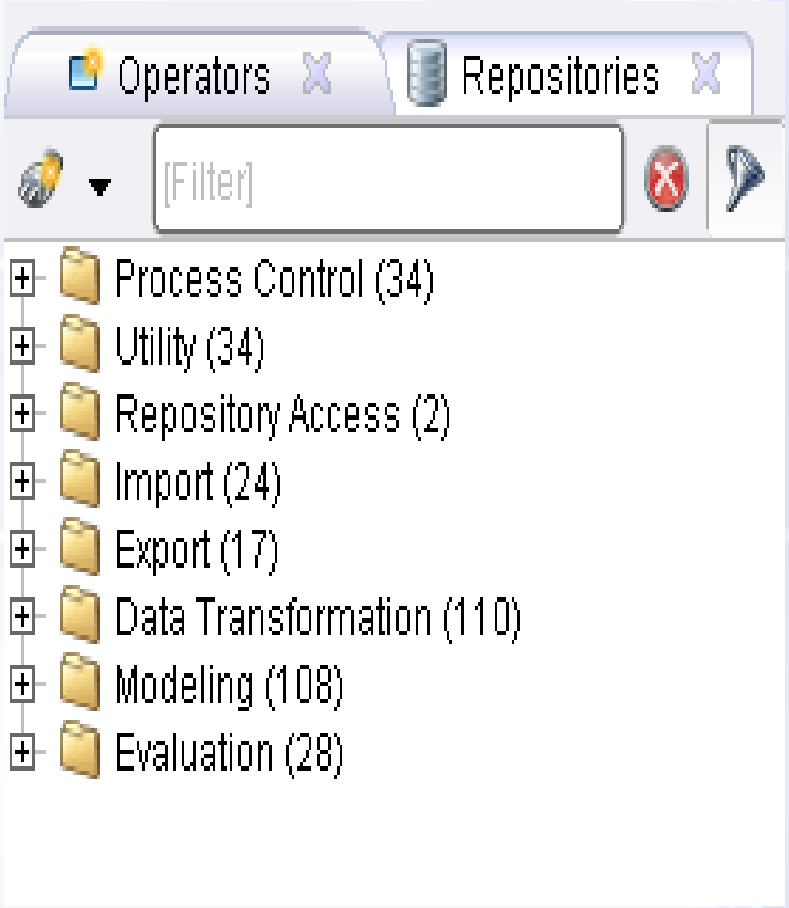
  - Klik tombol paling kiri
  - Atau gunakan menu
    View → Perspectives → Design

Versi : 01

View Operator

I.20

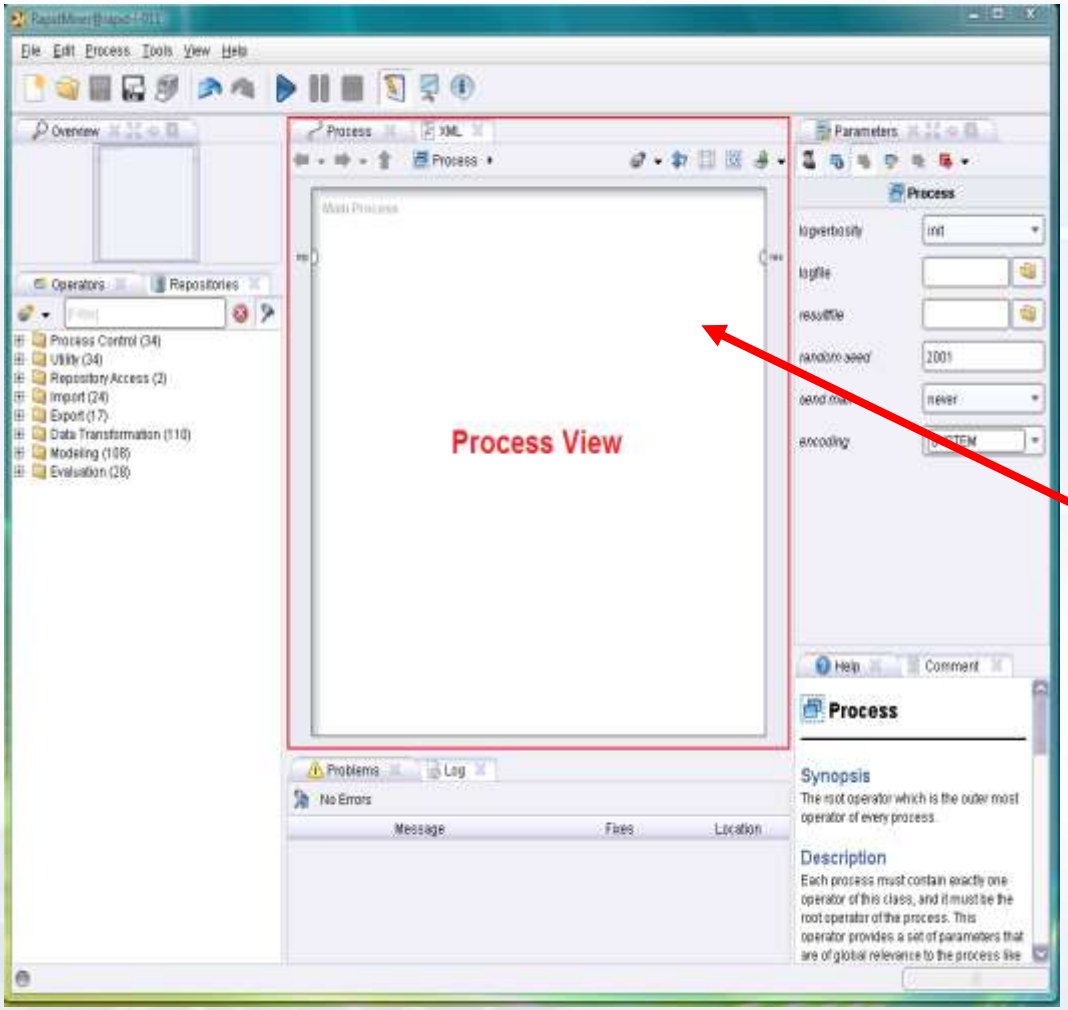
- Process Control
  Untuk mengontrol aliran proses, seperti *loop* atau *conditional branch*
- Utility
  Untuk mengelompokkan *subprocess*, juga *macro* dan *logger*
- Repository Access
  Untuk membaca dan menulis repositori
- Import
  Untuk membaca data dari berbagai format eksternal
- Export
  Untuk 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

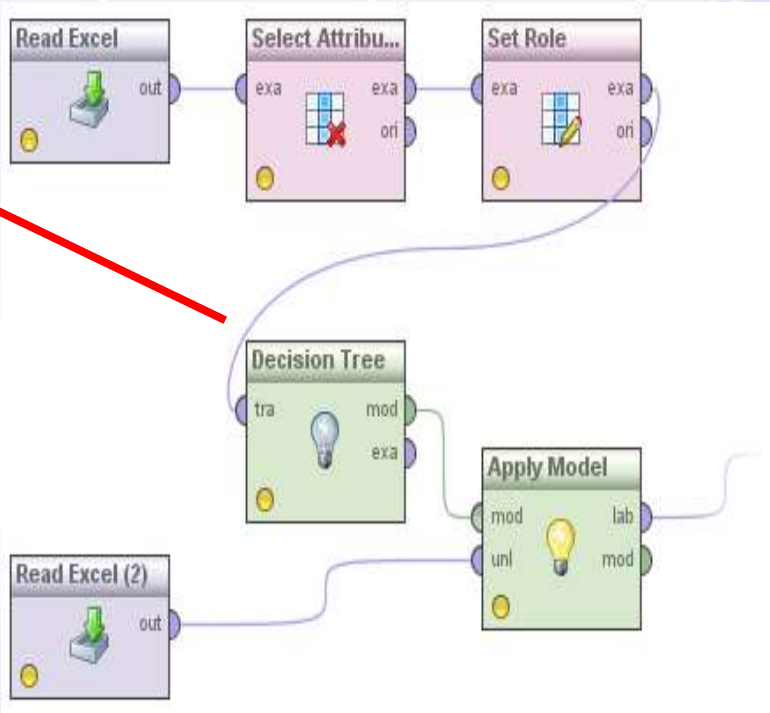


Versi : 01

I.21

View Proses

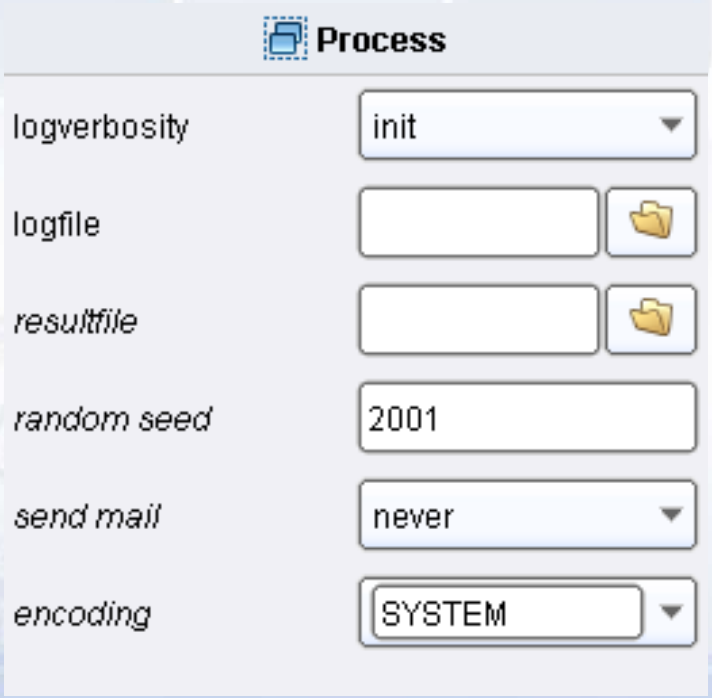




Versi : 01

View Parameter

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

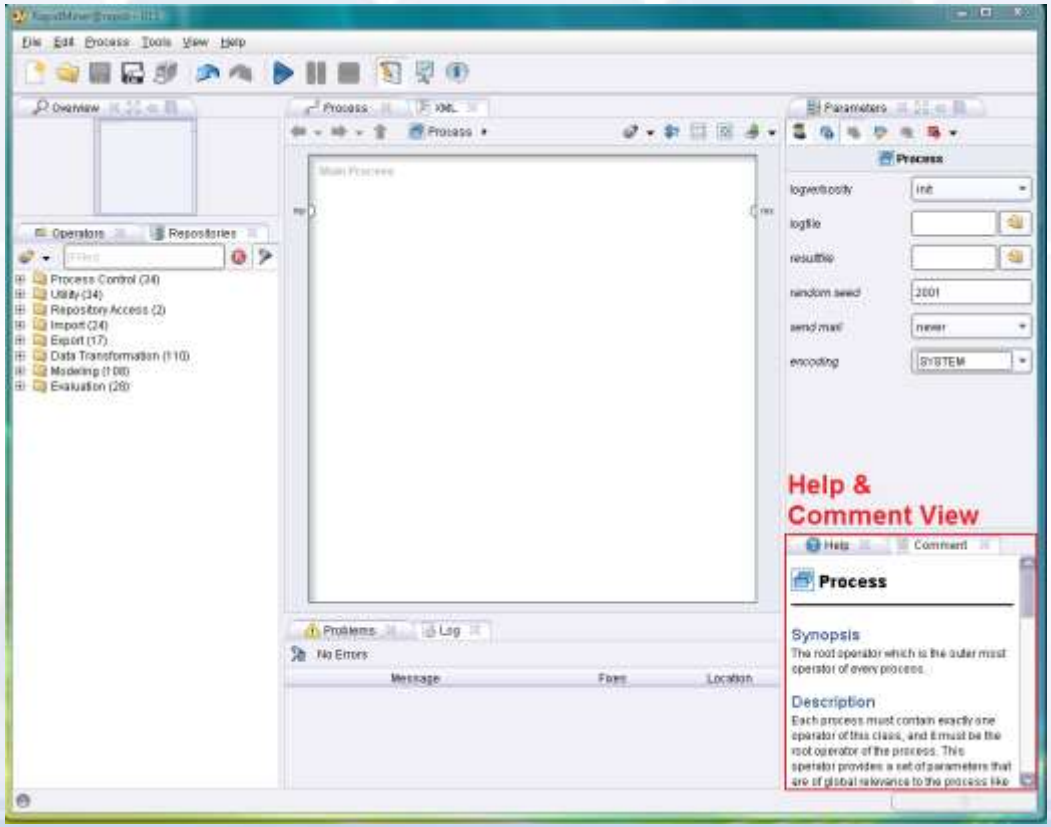


Versi : 01

I.23

View Help dan View Comment

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



Help | Comment

Process

Synopsis

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

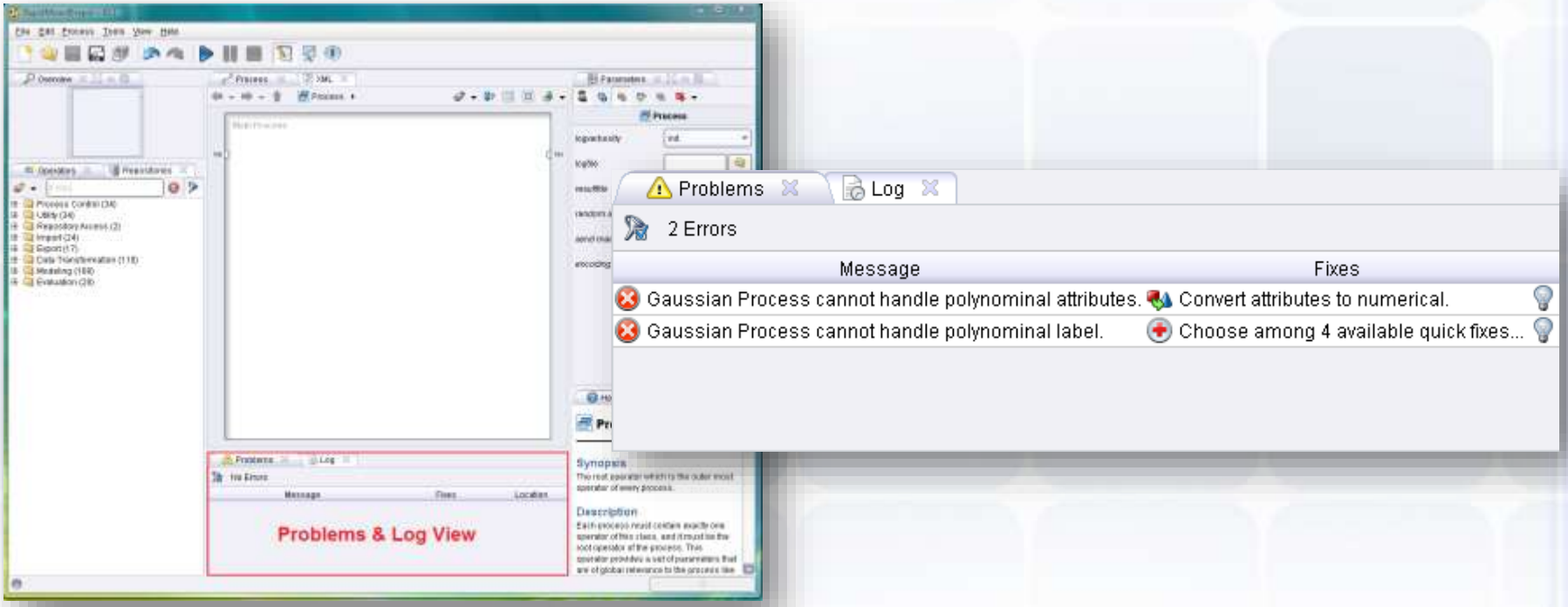
Description

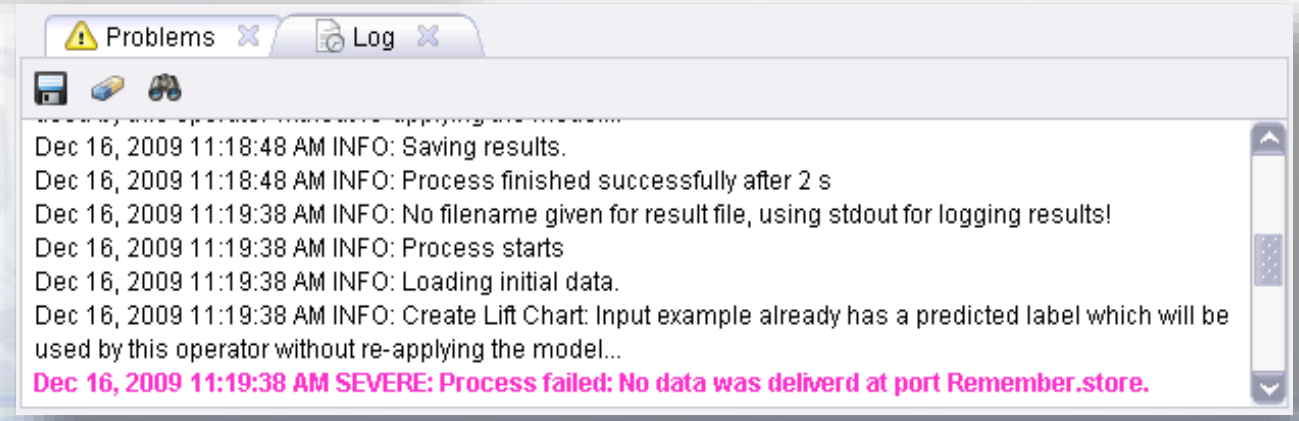
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like

Versi : 01

I.24

View Problems dan View Log





Versi : 01

Operator dan Proses

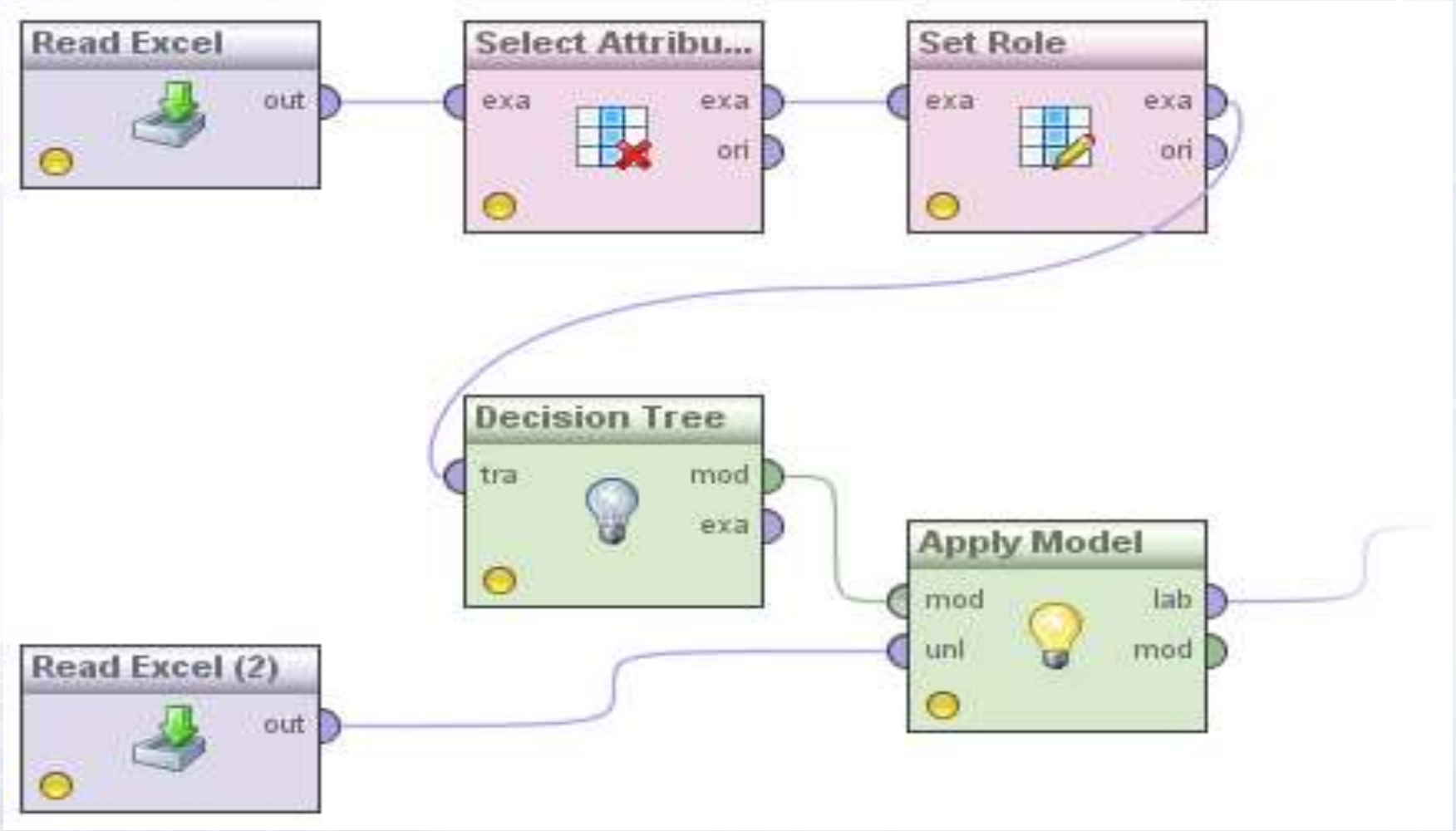
- Proses data mining pada dasarnya adalah proses analisa yang berisi alur kerja dari komponen data mining
- Komponen dari proses ini disebut operator, yang didefinisikan dengan:
  1. Deskripsi input
  2. Deskripsi output
  3. Aksi yang dilakukan
  4. Parameter yang diperlukan

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



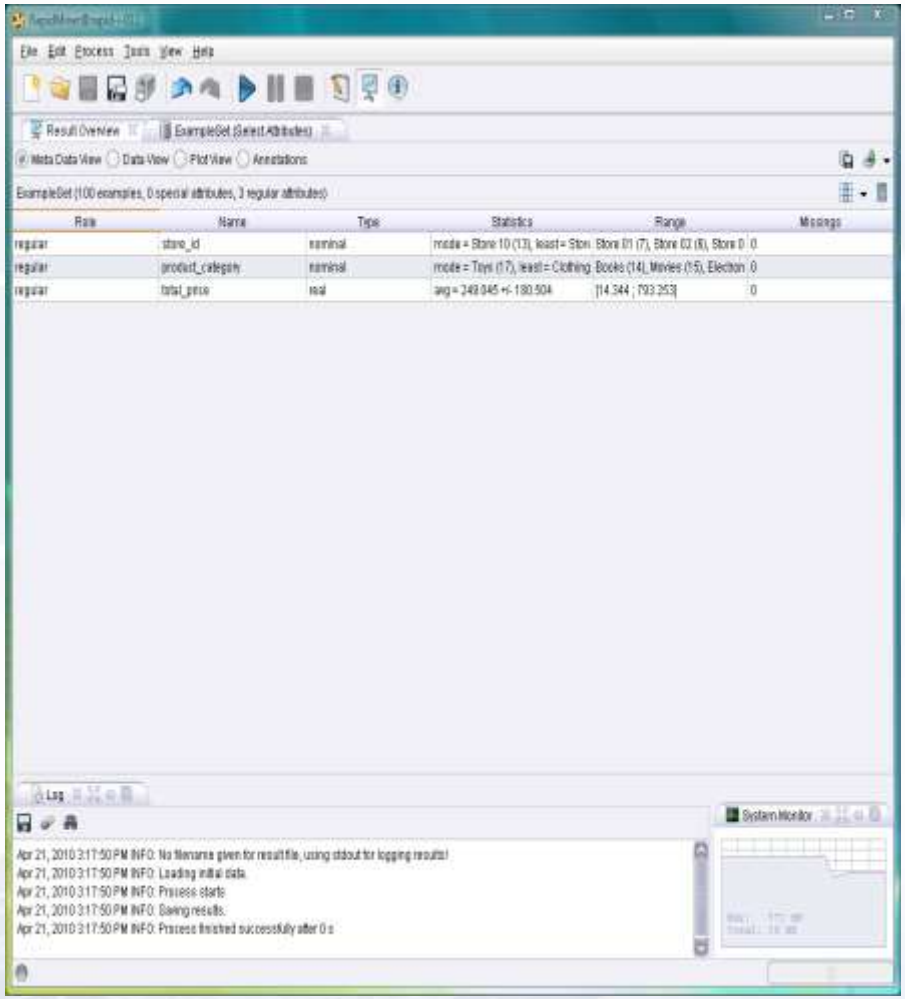
Versi : 01

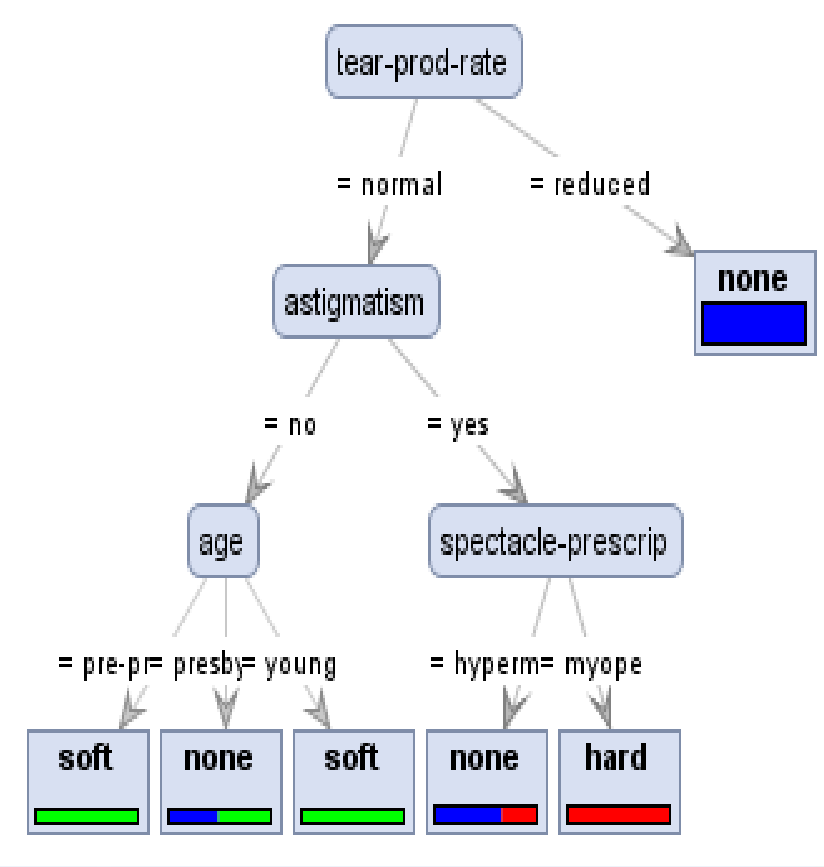
Menjalankan Proses

Proses dapat dijalankan dengan:

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

Melihat Hasil



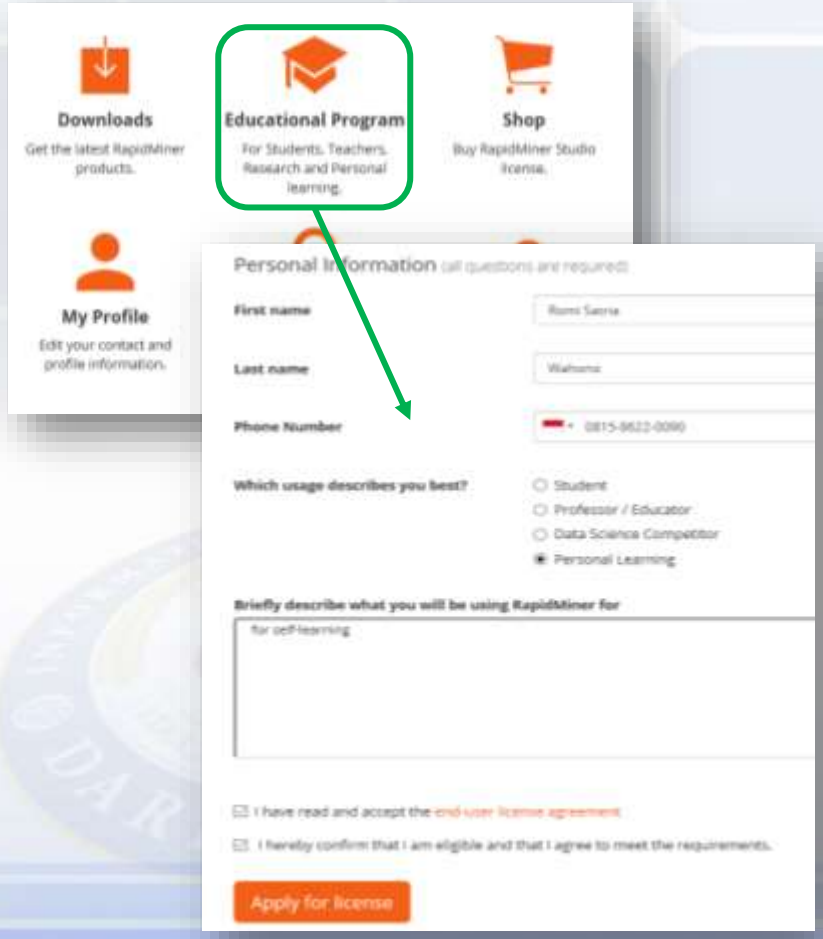


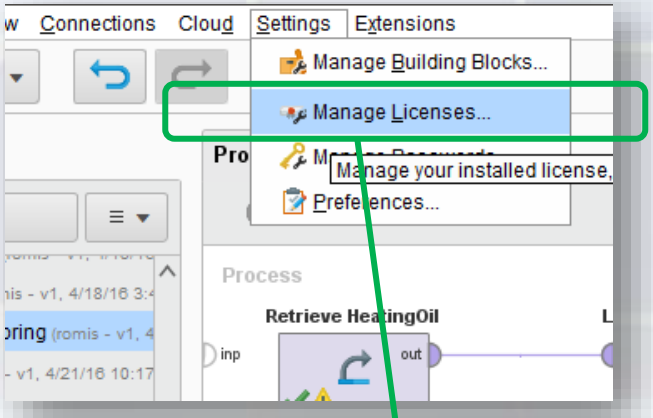
Versi : 01

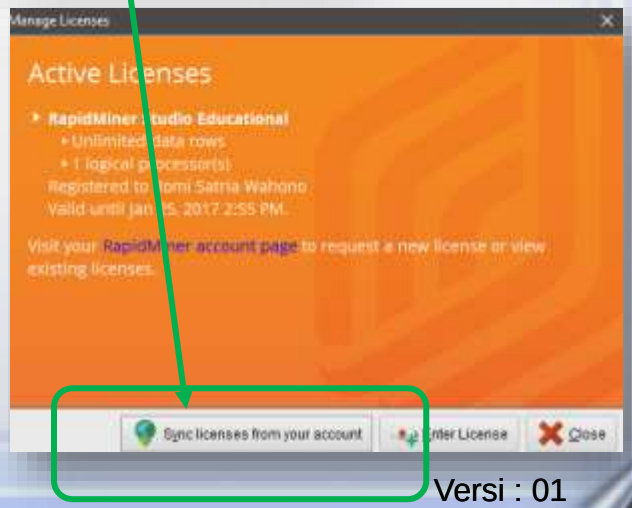
Instalasi dan Registrasi Lisensi Rapidminer

I.30

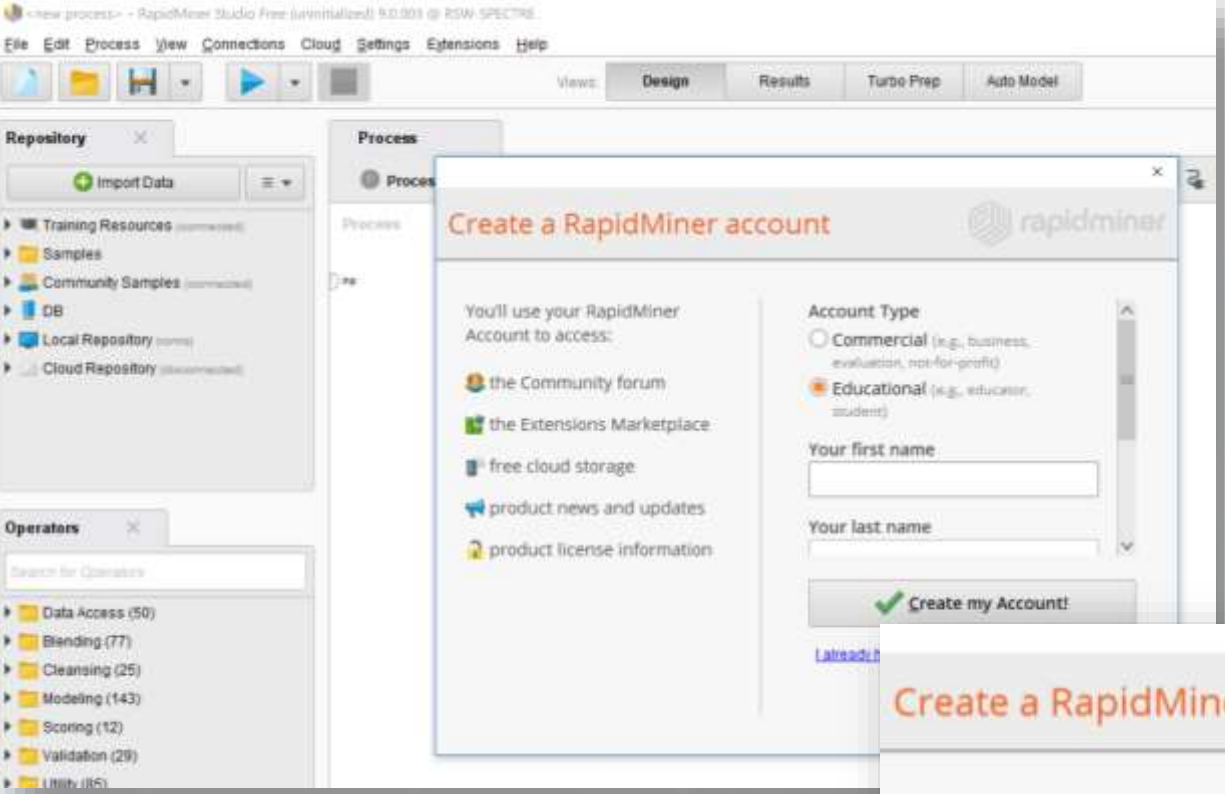
- Instal Rapidminer versi 7
- Registrasi account di rapidminer.com, dan lakukan dapatkan lisensi Educational Program untuk mengolah data tanpa batasan record

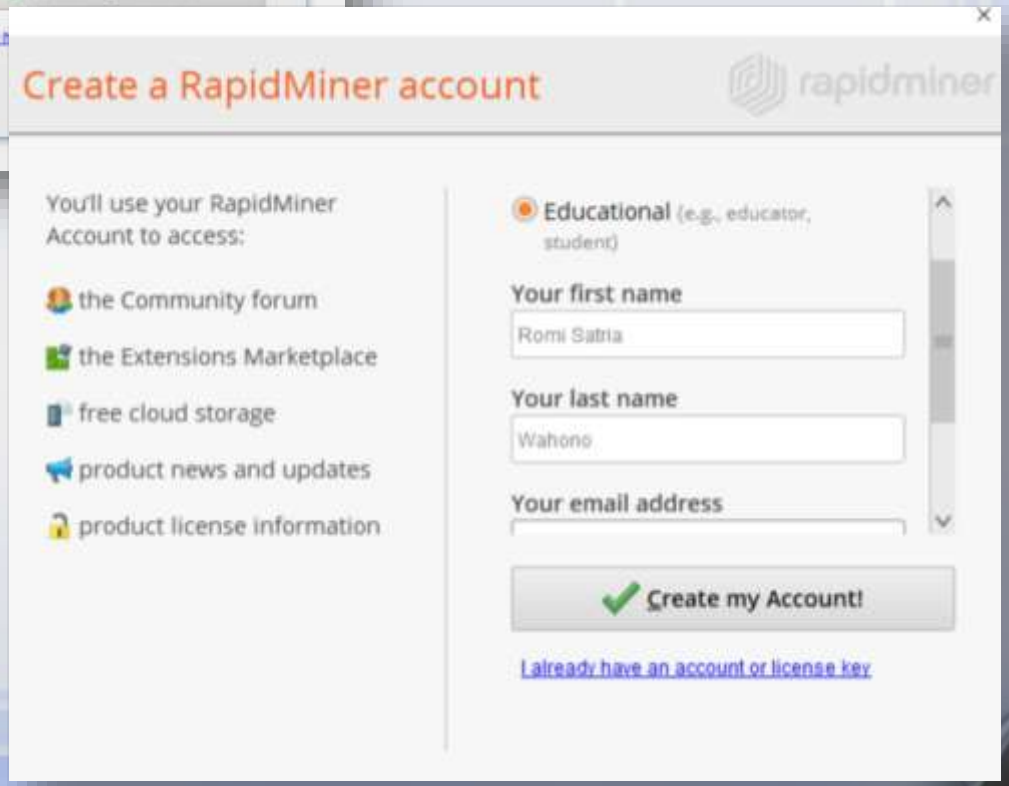




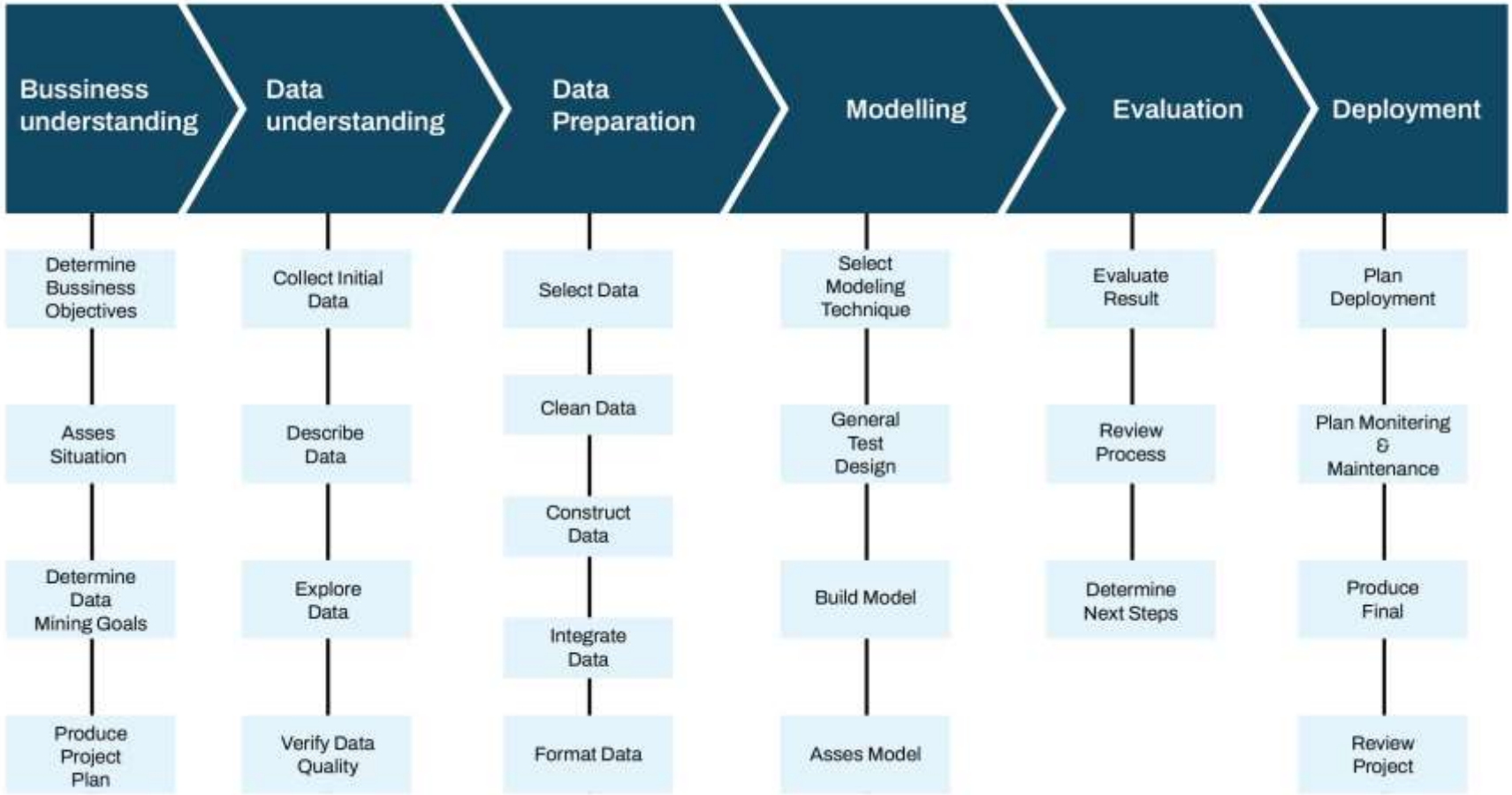


Versi : 01

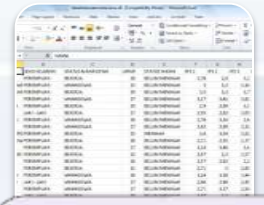

File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Repository
Import Data
Training Resources
Samples
Community Samples
DB
Local Repository
Cloud Repository
Process
Operators
Data Access (50)
Blending (77)
Cleansing (25)
Modeling (143)
Scoring (12)
Validation (29)
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

6.1 Penerapan Proses Data Mining

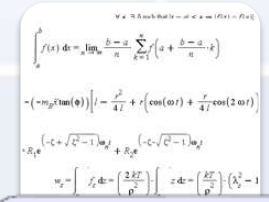


Proses Data Mining



1. Himpunan Data

(Pemahaman dan Pengolahan Data)

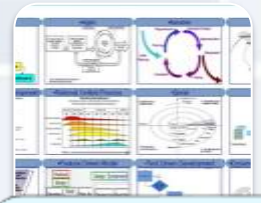


2. Metode Data Mining

(Pilih Metode Sesuai Karakter Data)

3. Pengetahuan

(Pola/Model/Rumus/ Tree/Rule/Cluster)



4. Evaluation

(Akurasi, AUC, RMSE, Lift Ratio,…)

DATA PRE-PROCESSING
Data Cleaning
Data Integration
Data Reduction
Data Transformation

Estimation
Prediction
Classification
Clustering
Association

Versi : 01

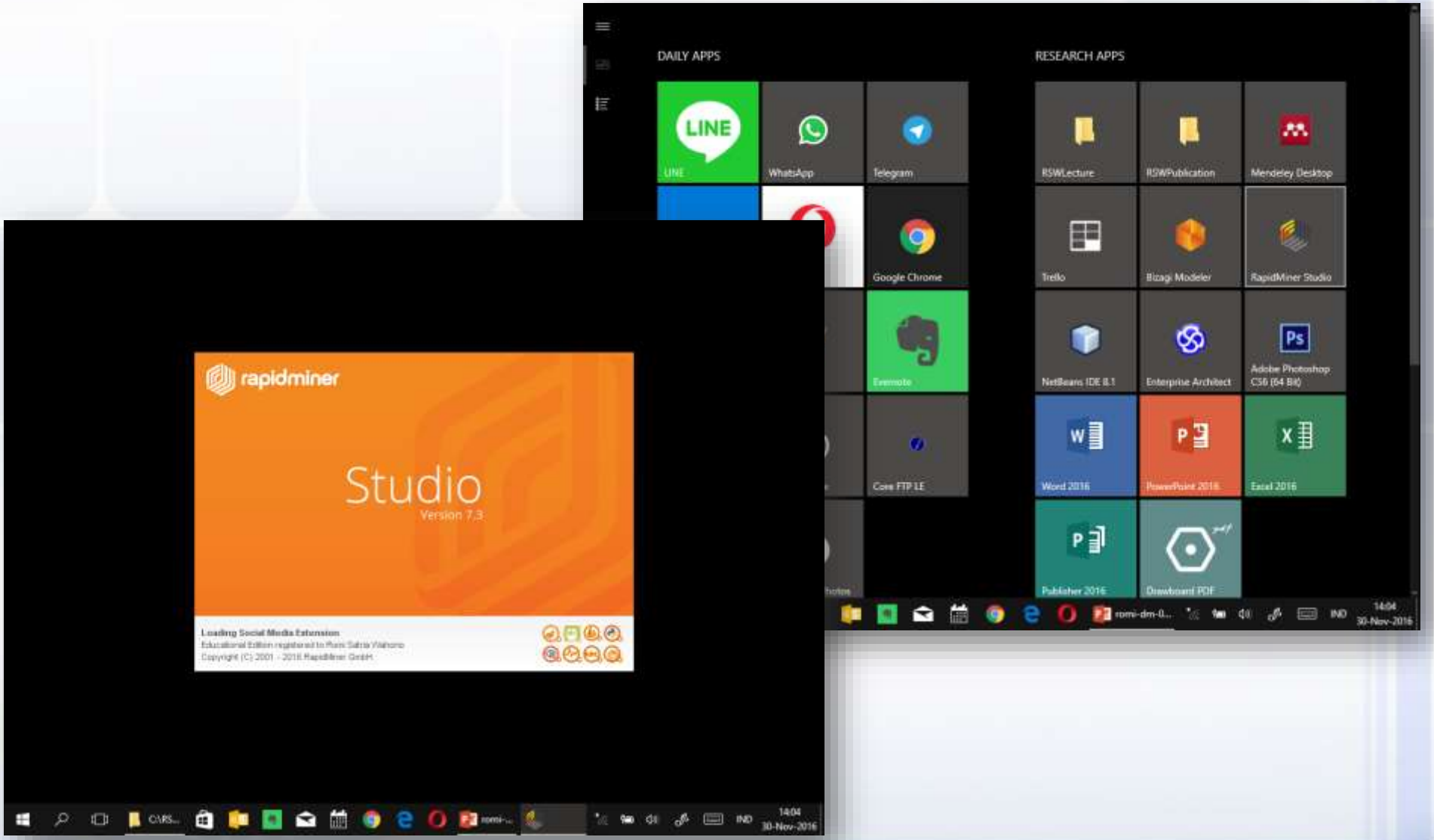
Latihan: Rekomendasi Main Golf

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

I.35

Buka Rapidminer



Versi : 01

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.

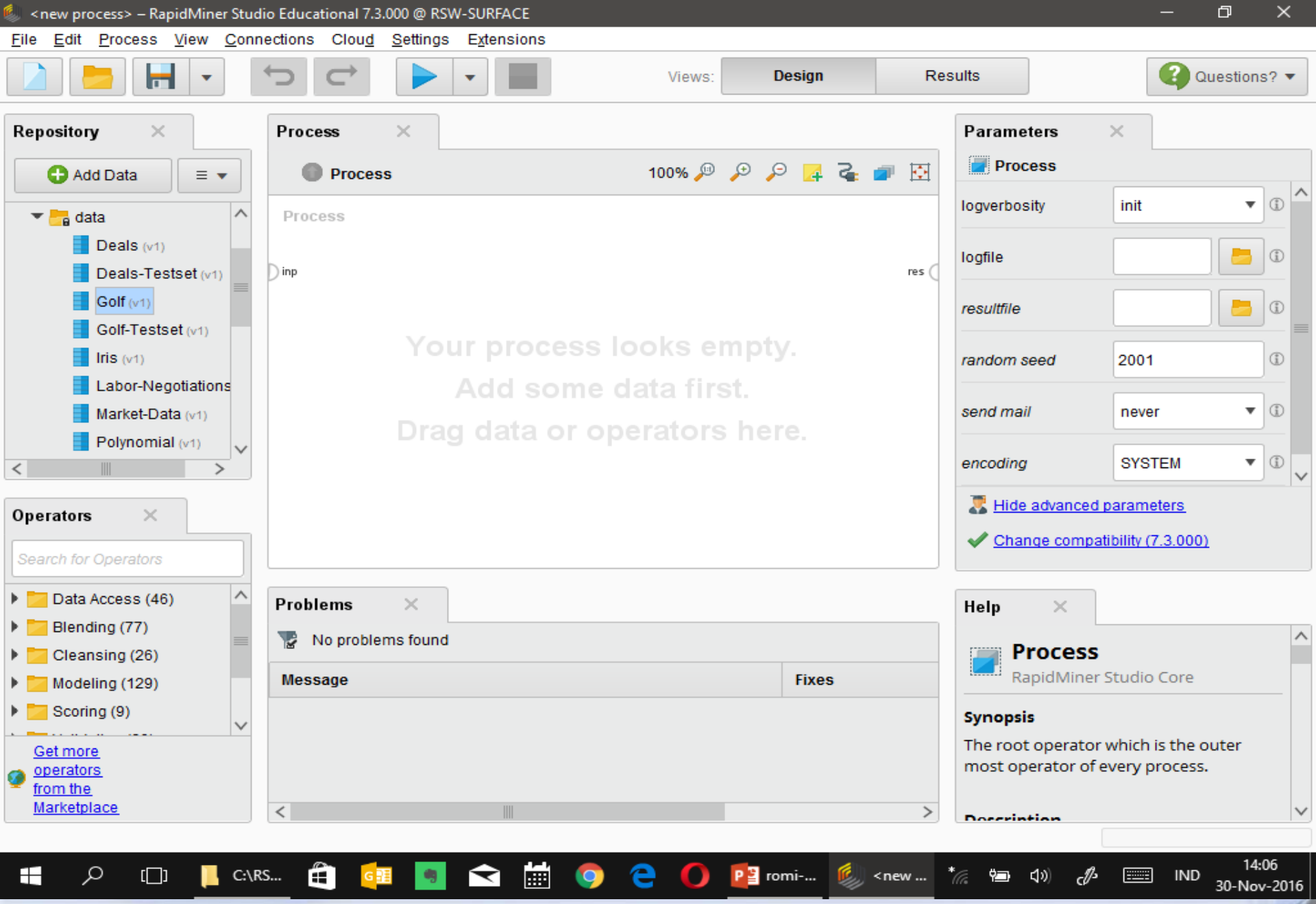
Credit Risk Modeling
Model credit default risk by training an optimized Support Vector Machine (SVM) model.

Market Basket Analysis
Find products frequently purchased together and turn them into rules for recommendations.

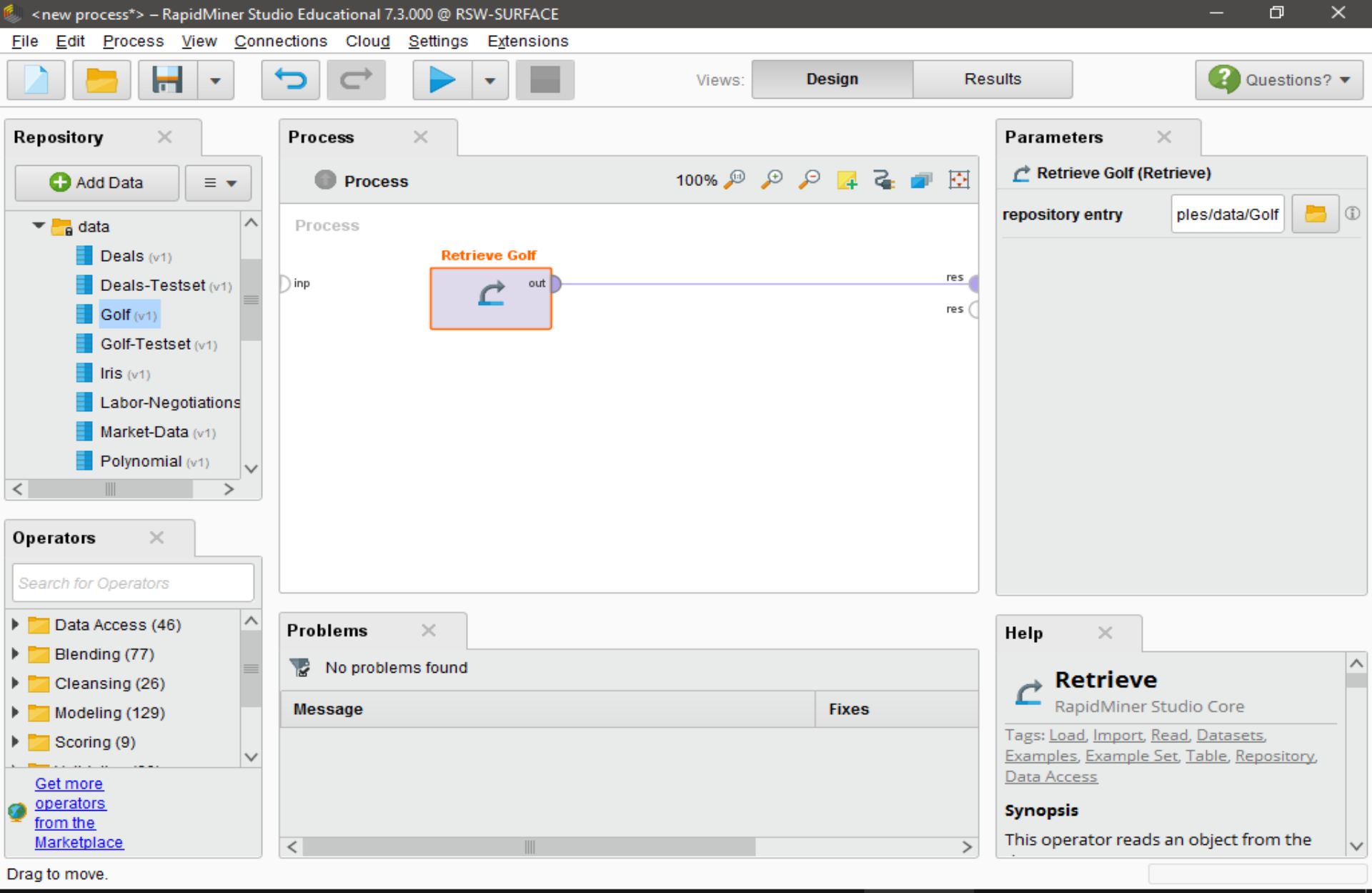
Predictive Maintenance
Model equipment failures to schedule maintenance pre-emptively.

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.
IND
14:06
30-Nov-2016

Versi : 01


<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

Versi : 01

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

<new process*> – RapidMiner Studio Educational 7.3.000 @ RSW-SURFACE

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results Questions?

Result History | ExampleSet (Retrieve Golf)

Data
Statistics
Charts
Advanced Charts
Annotations

Filter (5 / 5 attributes): Search for Attributes

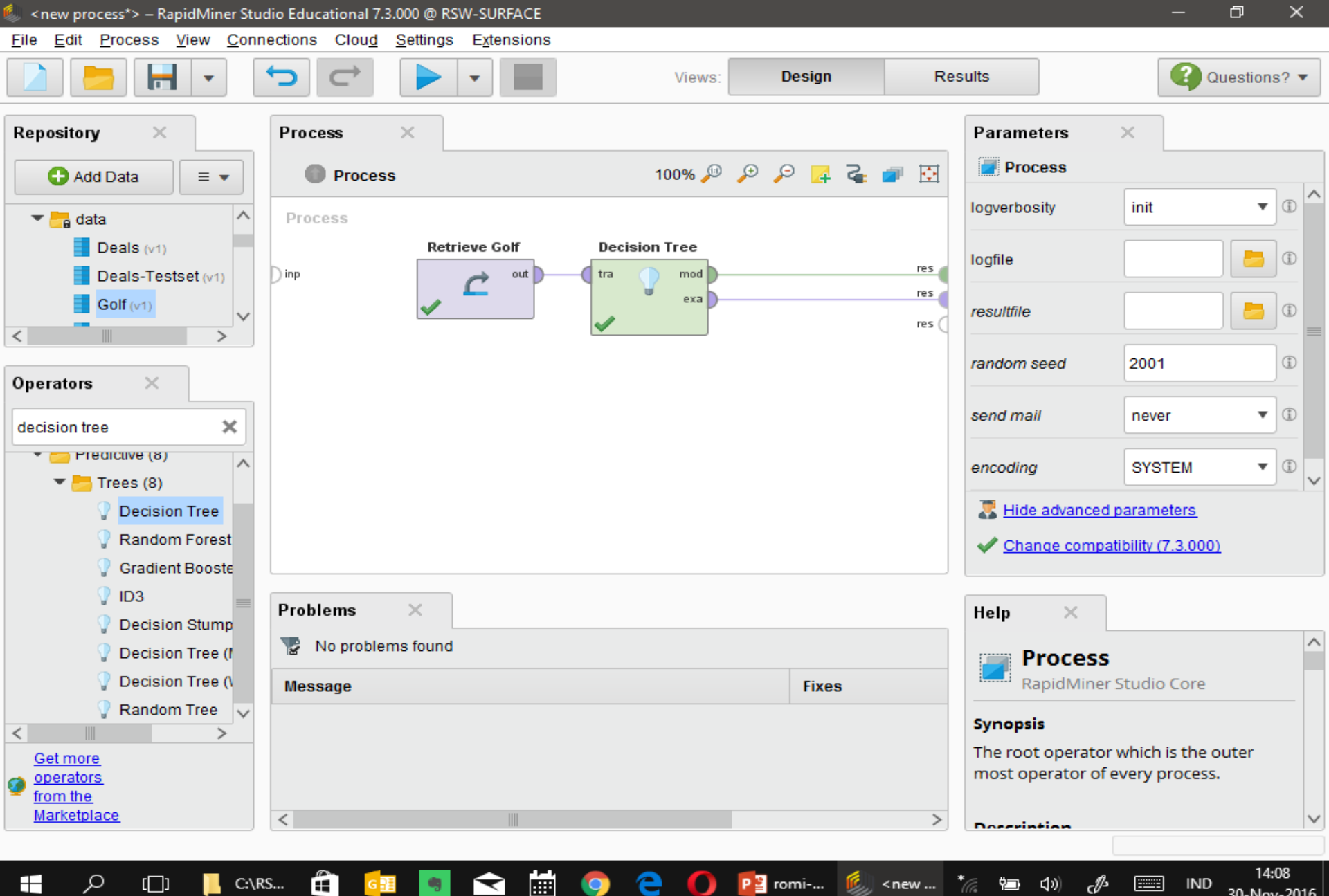
Name	Type	Missing	Statistics		
Play (Label)	Nominal	0	 Open chart	Least: no (5)	Most: yes (9)
Outlook	Nominal	0	Least: overcast (4)	Most: rain (5)	Values: rain (5), sunny
Temperature	Integer	0	Min: 64	Max: 85	Average: 73.571
Humidity	Integer	0	Min: 65	Max: 96	Average: 80.286
Wind	Nominal	0	Least: true (6)	Most: false (8)	Values: false (8), true (

Showing attributes 1 - 5

Examples: 14 Special Attributes: 1 Regular Attributes: 4

IND 14:07 30-Nov-2016

Versi : 01


<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.
IND
14:08
30-Nov-2016

Versi : 01

Graph

Description

Annotations

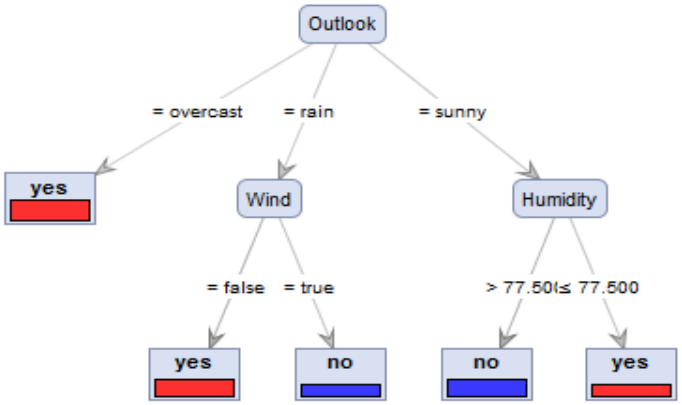
Zoom

Mode

Tree

Node Labels

Edge Labels



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

Graph

Description

Annotations

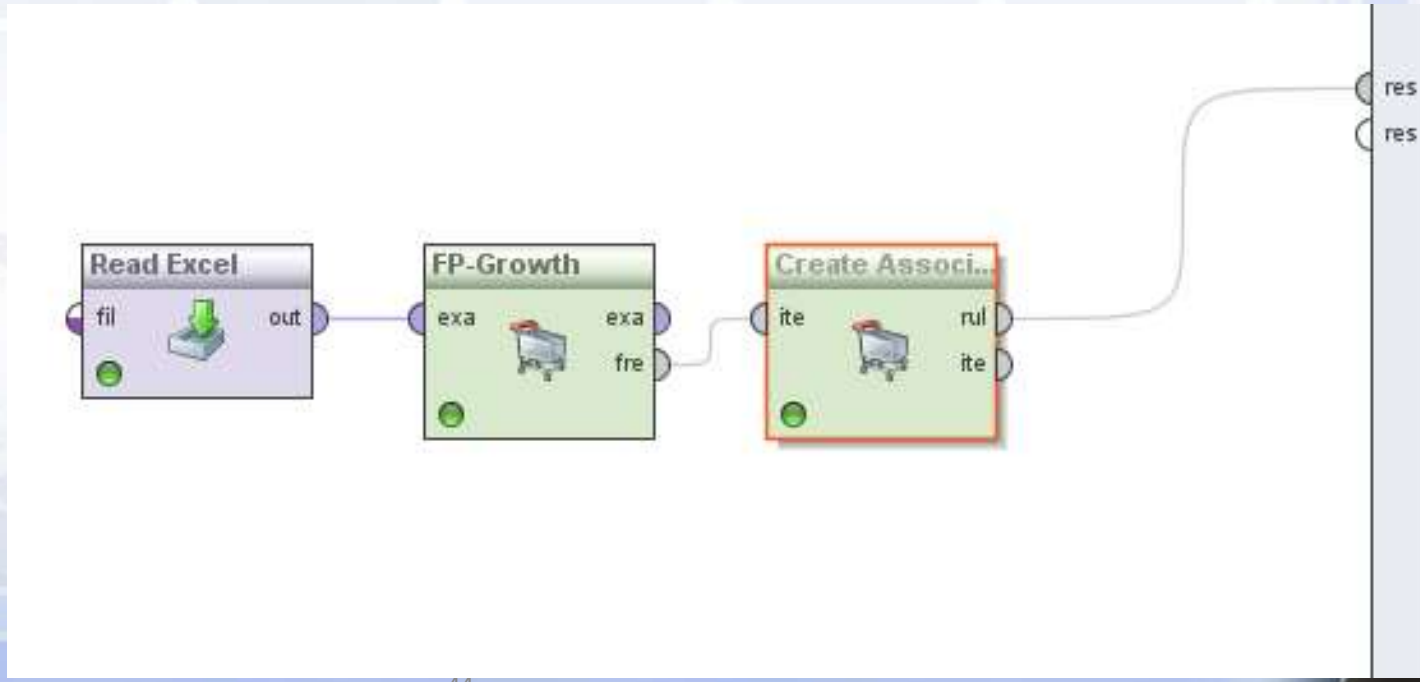
Tree

```
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: Aturan Asosiasi Data Transaksi

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

2. Pilih metode yang tepat supaya menghasilkan pola



<new process*> – RapidMiner Studio Educational 9.0.002 @ RSW-SURFACE

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design | Results | Turbo Prep

Find data, operators...etc | All Studio

Result History | AssociationRules (Create Association Rules)

Data

Graph

Description

Annotations

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

Versi : 01

Review dan Latihan