



DATA MINING

PERTEMUAN KE-12

Algoritma Estimasi dan Forecasting

Algoritma Estimasi dan Forecasting

27. Linear Regression

28. Time Series Forecasting



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 |

I.5 2. Identifikasikan Atribut dan Label

$$Y = a + bX$$

Dimana:

Y = Variabel terikat (Dependen)

X = Variabel tidak terikat (Independen)

a = konstanta

b = koefisien regresi (kemiringan); besaran Response yang ditimbulkan oleh variabel

$$a = \frac{(\sum y) (\sum x^2) - (\sum x) (\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x) (\sum y)}{n(\sum x^2) - (\sum x)^2}$$

I.6 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)

$$a = \frac{(\Sigma y) (\Sigma x^2) - (\Sigma x) (\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{(72) (4876) - (220) (1640)}{10 (4876) - (220)^2}$$

$$\mathbf{a = -27,02}$$

- Menghitung Koefisien Regresi (b)

$$b = \frac{n(\Sigma xy) - (\Sigma x) (\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{10 (1640) - (220) (72)}{10 (4876) - (220)^2}$$

$$\mathbf{b = 1,56}$$

5. Buatlah 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?

$$5 = -27,02 + 1,56X$$

$$1,56X = 5 + 27,02$$

$$X = 32,02 / 1,56$$

$$X = \mathbf{20,52}$$

Jadi **Prediksi Suhu Ruangan** yang paling sesuai untuk mencapai target Cacat Produksi adalah sekitar **20,52°C**

STUDI KASUS CRISP-DM

Heating Oil Consumption – Estimation

(*Matthew North, Data Mining for the Masses, 2012,*

Chapter 8 Estimation, pp. 127-140)

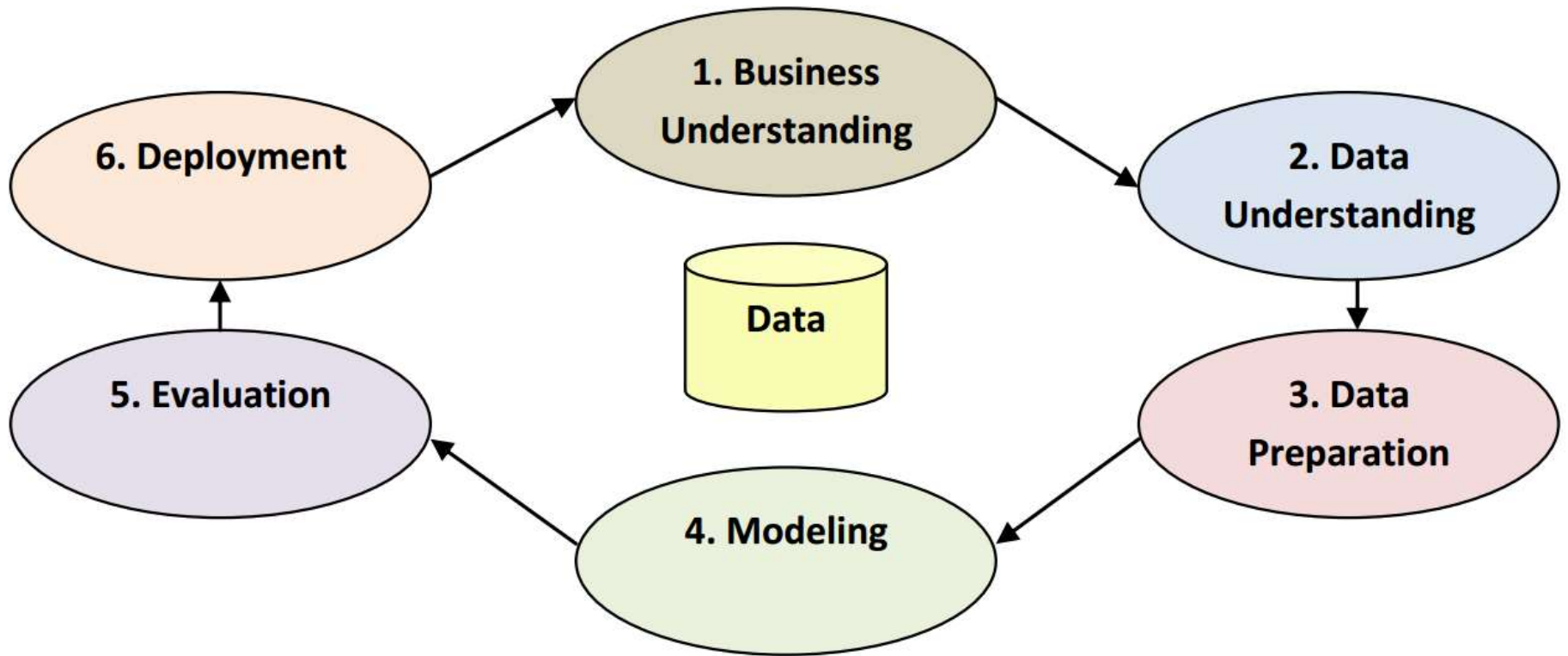
Dataset: HeatingOil-Training.csv dan HeatingOil-Scoring.csv



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



Konteks dan Prespektif Kasus (1)

- Sarah, seorang manajer penjualan regional berkeinginan untuk mendapatkan bantuan lebih lanjut
- Bisnis sedang berkembang pesat, tim penjualannya merekrut ribuan klien baru, dan dia ingin memastikan perusahaannya mampu memenuhi tingkat permintaan baru ini, dia berharap mendapatkan bantuan untuk melakukan beberapa prediksi
- Dia mengetahui bahwa ada beberapa korelasi antara atribut dalam kumpulan datanya (seperti suhu, isolasi, dan usia penghuni), dan dia sekarang bertanya-tanya apakah dia dapat menggunakan kumpulan data sebelumnya untuk memprediksi penggunaan minyak pemanas untuk pelanggan baru.

Konteks dan Prespektif Kasus(2)

- Kita tahu bahwa, para pelanggan baru ini belum mulai mengonsumsi minyak pemanas, jumlahnya banyak (**tepatnya 42.650**), dan dia ingin tahu berapa banyak minyak yang harus dia simpan untuk memenuhi kebutuhan baru ini. permintaan pelanggan
- Bisakah dia menggunakan data mining untuk memeriksa atribut rumah tangga dan mengetahui jumlah konsumsi di masa lalu untuk mengantisipasi dan memenuhi kebutuhan pelanggan barunya?

1. Pemahaman Bisnis

- **Tujuan** penambangan data (Data Mining) Sarah cukup jelas: dia ingin mengantisipasi permintaan akan produk konsumsi
- **Gunakan model regresi linier** untuk membantunya mendapatkan prediksi yang diinginkan
- Dia memiliki data, **1.218 observasi** yang memberikan profil atribut untuk setiap rumah, beserta konsumsi minyak pemanas tahunan rumah tersebut.
- Dia ingin menggunakan kumpulan data ini sebagai data pelatihan untuk memprediksi penggunaan yang akan diberikan oleh **42.650 klien baru** ke perusahaannya
- Dia tahu bahwa rumah klien baru ini memiliki sifat yang serupa dengan basis kliennya yang sudah ada, sehingga perilaku penggunaan pelanggan lama harus berfungsi sebagai ukuran yang kuat untuk memprediksi penggunaan di masa depan oleh pelanggan baru.

2. Pemahaman Bisnis

Buat kumpulan data yang terdiri dari atribut berikut:

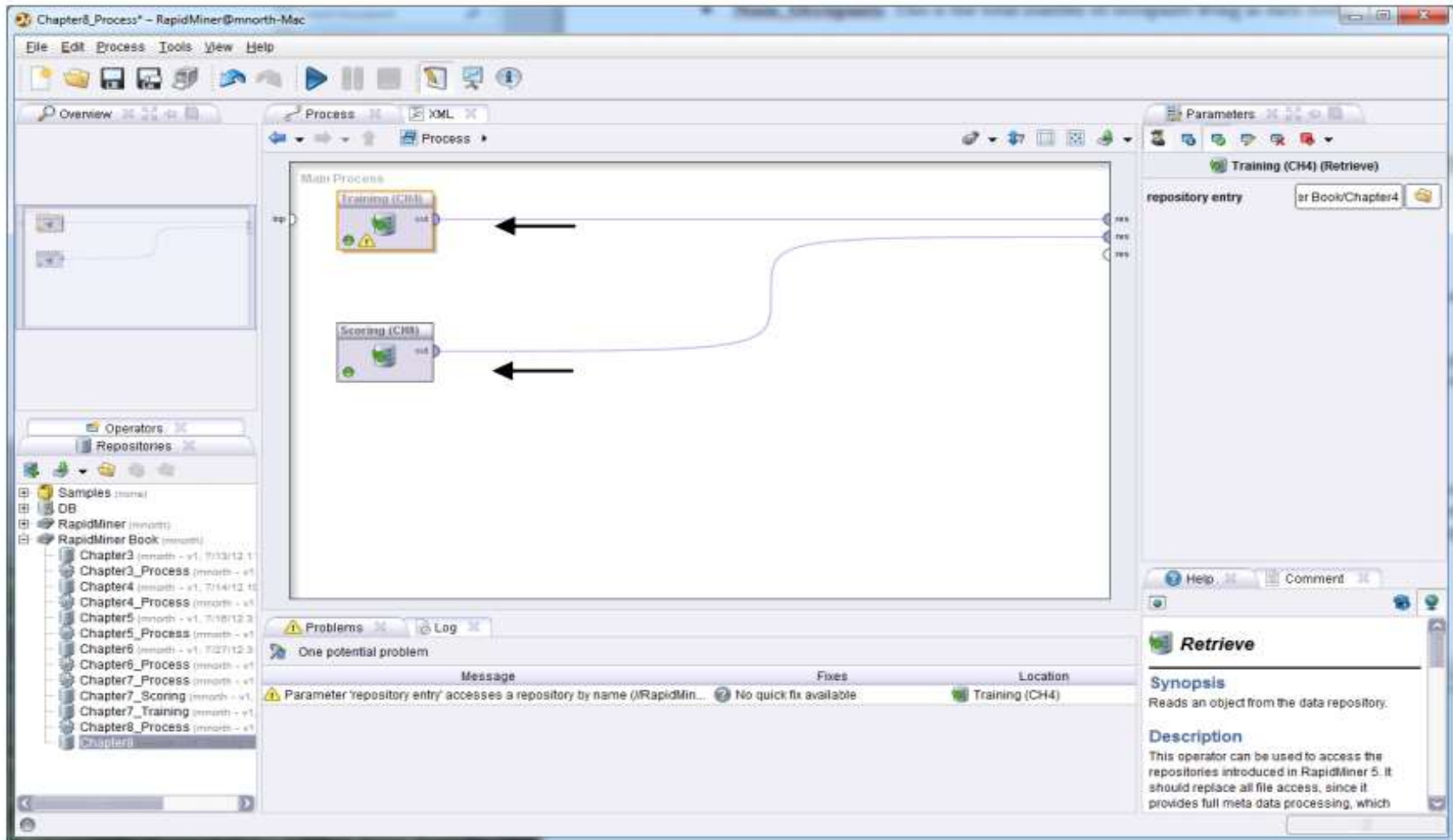
- **Isolasi:** Ini adalah tingkat kepadatan, berkisar antara satu hingga sepuluh, yang menunjukkan **ketebalan insulasi setiap rumah**. Rumah dengan kepadatan satu memiliki insulasi yang buruk, sedangkan rumah dengan kepadatan sepuluh memiliki insulasi yang sangat baik
- **Suhu:** Ini adalah suhu **lingkungan luar ruangan rata-rata** di setiap rumah selama setahun terakhir, diukur dalam derajat Fahrenheit

(Lanjutan) Pemahaman Bisnis

- **Heating_Oil**: Ini adalah jumlah total unit minyak pemanas yang dibeli oleh pemilik setiap rumah dalam satu tahun terakhir
- **Num_Occupants**: Ini adalah jumlah total penghuni yang tinggal di setiap rumah
- **Avg_Age**: Ini adalah usia rata-rata penghuni tersebut
- **Home_Size**: Ini adalah penilaian, dalam skala satu sampai delapan, dari ukuran rumah secara keseluruhan. Semakin tinggi angkanya, semakin besar rumahnya

3. Data Preparation

- Dataset CSV untuk contoh bab ini tersedia untuk diunduh di situs web berikut (<https://sites.google.com/site/dataminingforthemasses/>)



3. Data Preparation

1.19

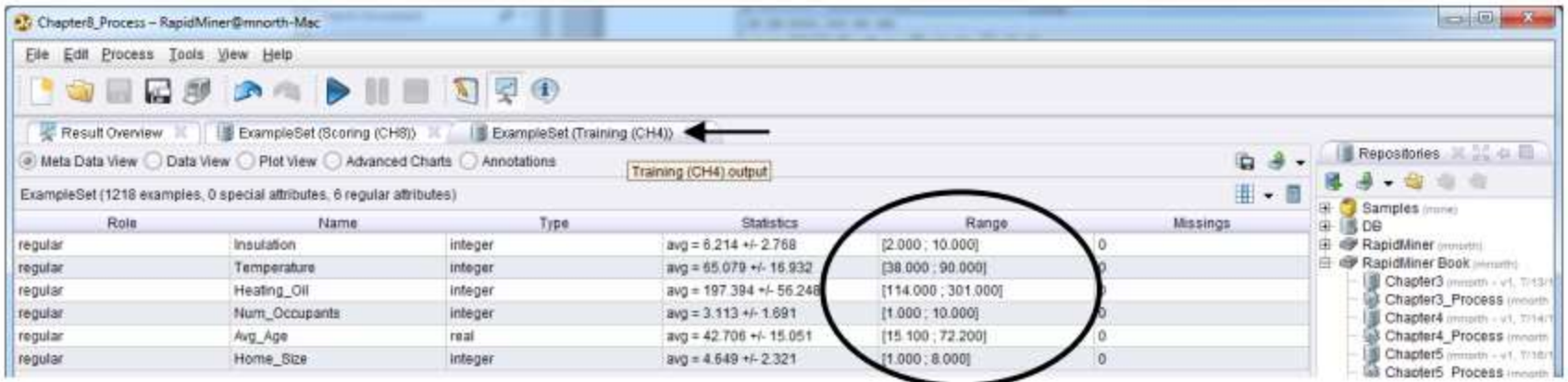


Figure 8-2. Value ranges for the training data set's attributes.

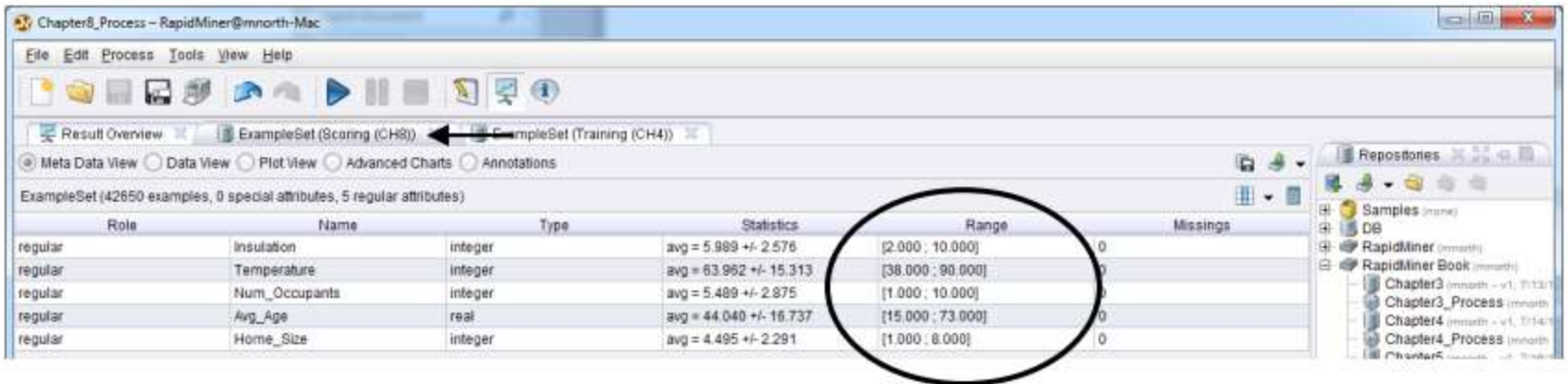


Figure 8-3. Value ranges for the scoring data set's attributes.

3. Data Preparation

The screenshot displays the RapidMiner interface with a process flow and the configuration for a 'Set Role' operator.

Process Flow:

- Training (CH4):** The starting point of the process.
- Set Role:** Receives input from 'Training (CH4)'. It is highlighted with a black arrow pointing from the 'Repositories' panel.
- Scoring (CH4):** Receives input from 'Set Role'.
- Filter Examples:** Receives input from 'Scoring (CH4)'. It is highlighted with a black arrow pointing from the 'Repositories' panel.
- Filter Example...:** Receives input from 'Filter Examples'.

Repositories Panel:

- Shows a tree structure with 'Set Role' selected under 'Name and Role Modification'.
- A black arrow points from the 'Set Role' operator in the process flow to this entry in the panel.

Parameters Panel:

- name:** Heating_Oil
- target role:** label
- set additional roles:** Edit List (0)...

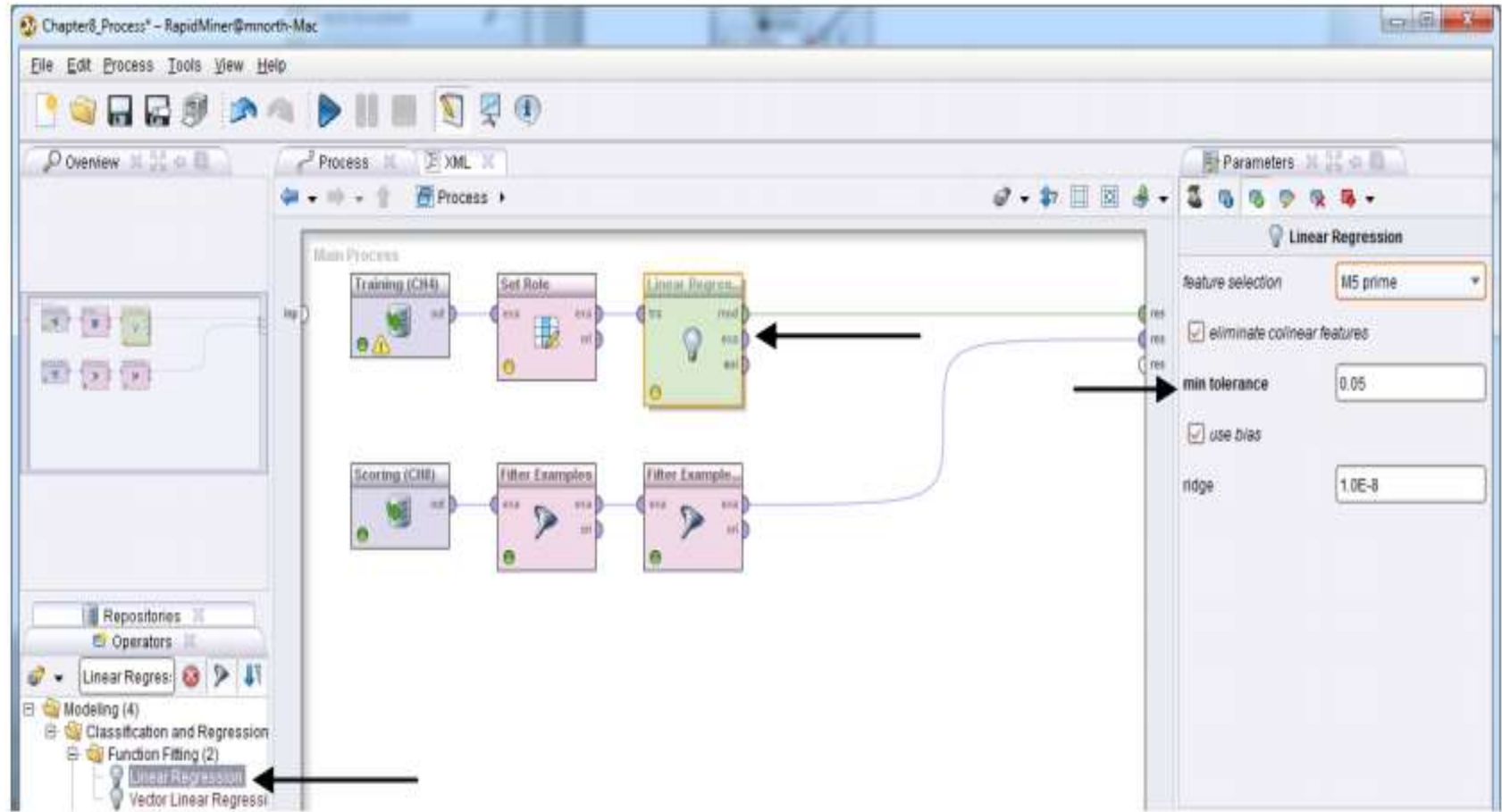
Problems Panel:

| Message | Fixes | Location |
|---|------------------------|----------------|
| Parameter 'repository entry' accesses a repository by name (/RapidMin...) | No quick fix available | Training (CH4) |

Set Role Operator Details:

- Synopsis:** This operator can be used to change the attribute role (regular, special, label, id...).
- Description:** This operator can be used to change the role of an attribute of the input ExampleSet. If you want to change the attribute name you should...

4. Modeling



4. Modeling

The screenshot displays the RapidMiner software interface. The main window shows a process flow diagram titled "Main Process". The workflow consists of the following steps:

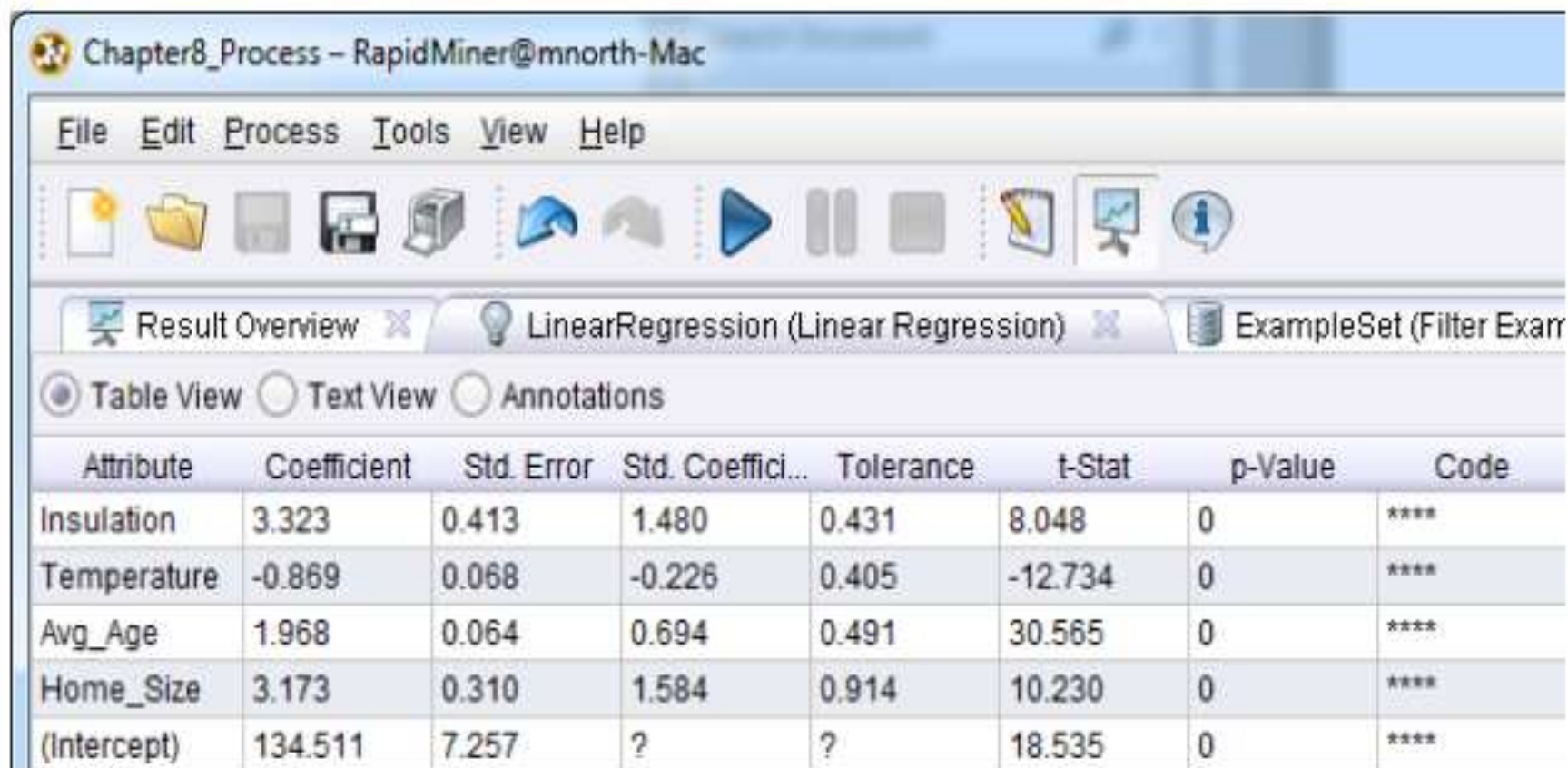
- Training (CH)**: A process node that outputs a model.
- Set Role**: A process node that prepares data for training.
- Linear Regress**: A process node that trains a linear regression model.
- Apply Model**: A process node that applies the trained model to new data. This node is highlighted with a black arrow.
- Scoring (CH)**: A process node that scores data using the model.
- Filter Examples**: A process node that filters the data based on the model's output.
- Filter Example...**: A second filter node.

The "Apply Model" node is selected, and its parameters are visible in the right-hand pane:

- Linear Regression** parameters:
 - feature selection: M5 prime
 - eliminate colinear features
 - min tolerance: 0.05
 - use bias
 - ridge: 1.0E-8

The bottom status bar indicates "One potential problem". The bottom of the window shows the names of the underlying files: "Maccana", "Five", and "Location".

5. Evaluation



The screenshot shows the RapidMiner software interface. The title bar reads "Chapter8_Process - RapidMiner@mnorth-Mac". The menu bar includes "File", "Edit", "Process", "Tools", "View", and "Help". The toolbar contains various icons for file operations and process execution. The main window displays three tabs: "Result Overview", "LinearRegression (Linear Regression)", and "ExampleSet (Filter Exarr)". Below the tabs, there are radio buttons for "Table View" (selected), "Text View", and "Annotations". The main content area shows a table with the following data:

| 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

Chapter8_Process - RapidMiner@mnorth-Mac

File Edit Process Tools View Help

Result Overview LinearRegression (Linear Regression) ExampleSet (Filter Examples (2))

Meta Data View Data View Plot View Advanced Charts Annotations

ExampleSet (42042 examples, 1 special attribute, 5 regular attributes)

| 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 | 15 | 2 | 22.000 | 8 |

6. Deployment


The screenshot displays the RapidMiner interface with a workflow titled "Main Process". The workflow consists of the following operators: Training (CH), Set Role, Linear Regres..., Scoring (CH), Filter Examples, Filter Example..., and Apply Model. An Aggregate operator is connected to the output of the Apply Model operator. Two black arrows point to the Aggregate operator: one from the top toolbar and another from the Parameters panel on the right.


The Parameters panel for the Aggregate operator is visible on the right side of the interface. It includes the following settings:

- use default aggregation
- aggregation attributes: [Edit List \(0\)...](#)
- group by attributes: [Select Attributes...](#)
- count all combinations
- only distinct
- ignore missings
- Compatibility level: 5.2.008





On the left side, the Repositories pane shows the Operators section expanded to "Aggregate", with a black arrow pointing to it.

6. Deployment

 Edit Parameter List: aggregation attributes X

 Edit Parameter List: **aggregation attributes**
The attributes which should be aggregated.

| aggregation attribute | aggregation functions |
|---------------------------|-----------------------|
| prediction(Heating_Oil) ▼ | sum ▼ |
| prediction(Heating_Oil) ▼ | average ▼ |

 Add Entry  Remove Entry  Ok  Cancel

6. Deployment

The screenshot shows the RapidMiner interface with the following components:

- Window title: Chapter8_Process - RapidMiner@mnorth-Mac
- Menu bar: File, Edit, Process, Tools, View, Help
- Toolbar: Includes icons for file operations, navigation, execution (play, pause, stop), and help.
- Process View: Shows a workflow with three nodes: Result Overview, LinearRegression (Linear Regression), and ExampleSet (Aggregate).
- View Options: Radio buttons for Meta Data View, Data View (selected), Plot View, Advanced Charts, and Annotations.
- ExampleSet Summary: ExampleSet (1 example, 0 special attributes, 2 regular attributes)
- Table of Results:

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

28. TIME SERIES FORECASTING



Time Series Forecasting

- Time series forecasting adalah **salah satu teknik analisis prediktif tertua**.
 - Ini telah ada dan digunakan secara luas bahkan sebelum istilah “analisis prediktif” diciptakan
- Variabel independen atau prediktor tidak sepenuhnya diperlukan untuk peramalan deret waktu **univariat**, namun sangat disarankan untuk deret waktu **multivariat**

- Metode **Time series forecasting** :
 1. **Metode Berdasarkan Data**: Tidak ada perbedaan antara prediktor dan target. Teknik seperti rata-rata atau pemulusan deret waktu dianggap sebagai pendekatan berbasis data dalam peramalan deret waktu
 2. **Metode Berbasis Model**: Mirip dengan model prediktif “**konvensional**”, yang memiliki variabel independen dan dependen, namun dengan perbedaan: variabel independen adalah waktu sekarang

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

- Dalam metode berbasis model, **waktu adalah variabel prediktor** atau independen dan nilai deret waktu adalah variabel dependen
- Metode berbasis model umumnya lebih disukai jika rangkaian waktu tampak memiliki pola “**global**”.
- Idenya adalah bahwa parameter model akan mampu menangkap pola-pola ini
 - Dengan demikian, kita dapat membuat prediksi untuk setiap langkah ke depan di masa depan dengan asumsi bahwa pola ini akan terulang
- Untuk deret waktu dengan pola lokal dan bukan pola global, penggunaan pendekatan berbasis model memerlukan penentuan bagaimana dan kapan pola tersebut berubah, dan hal ini sulit dilakukan.

Model Driven Methods

- **Linear Regression**
- Polynomial Regression
- Linear Regression with Seasonality
- Autoregression Models and **ARIMA**

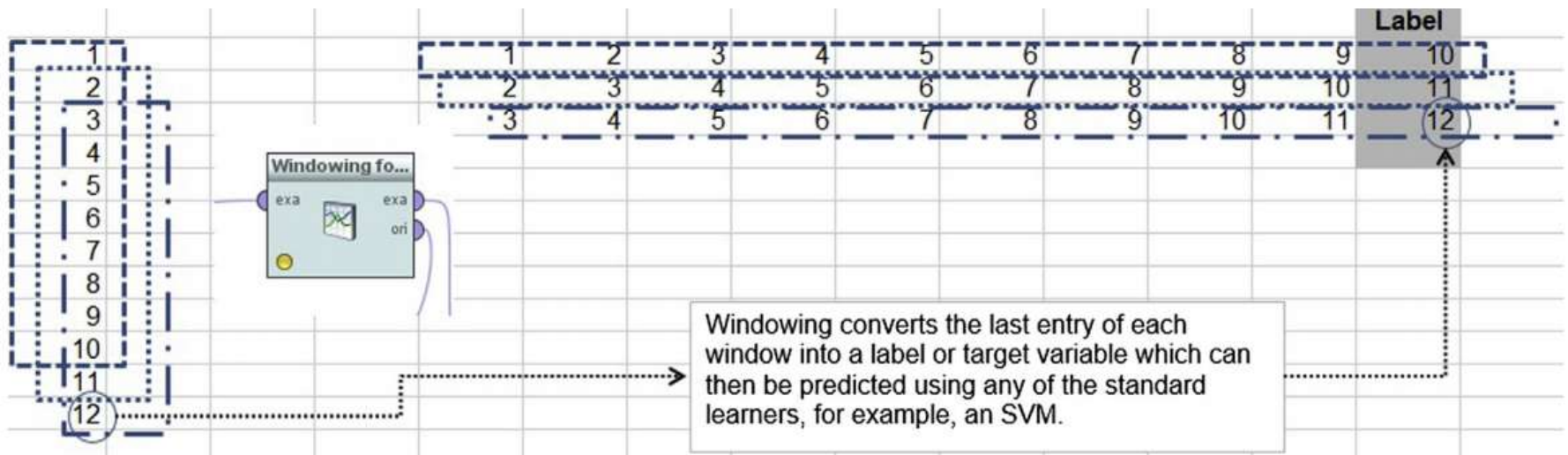


How to Implement

- Pendekatan RapidMiner terhadap deret waktu didasarkan pada dua proses transformasi data utama
- Yang pertama adalah melakukan **windowing** untuk mengubah data deret waktu menjadi kumpulan data umum:
 - Langkah ini akan mengubah baris terakhir jendela dalam deret waktu menjadi label atau variabel target
- Terapkan salah satu " learners " atau algoritme untuk **memprediksi variabel target** dan dengan demikian memprediksi langkah waktu berikutnya dalam rangkaian tersebut

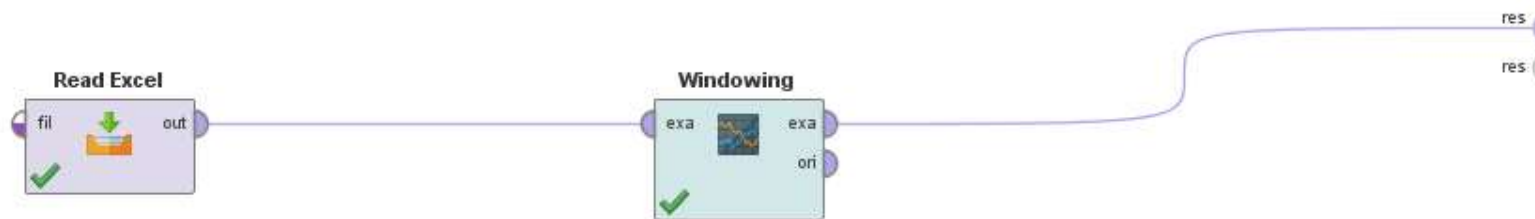
Windowing Concept

- Parameter dari operator **Windowing** memungkinkan perubahan ukuran windows, tumpang tindih antara windows yang berurutan (ukuran langkah), dan prediksi horizontal, yang digunakan untuk forecasting
- Prediksi Horizontal mengontrol baris mana dalam rangkaian data mentah yang berakhir sebagai variabel label dalam rangkaian yang diubah



Rapidminer Windowing Operator

inp



| 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 $h = 1$

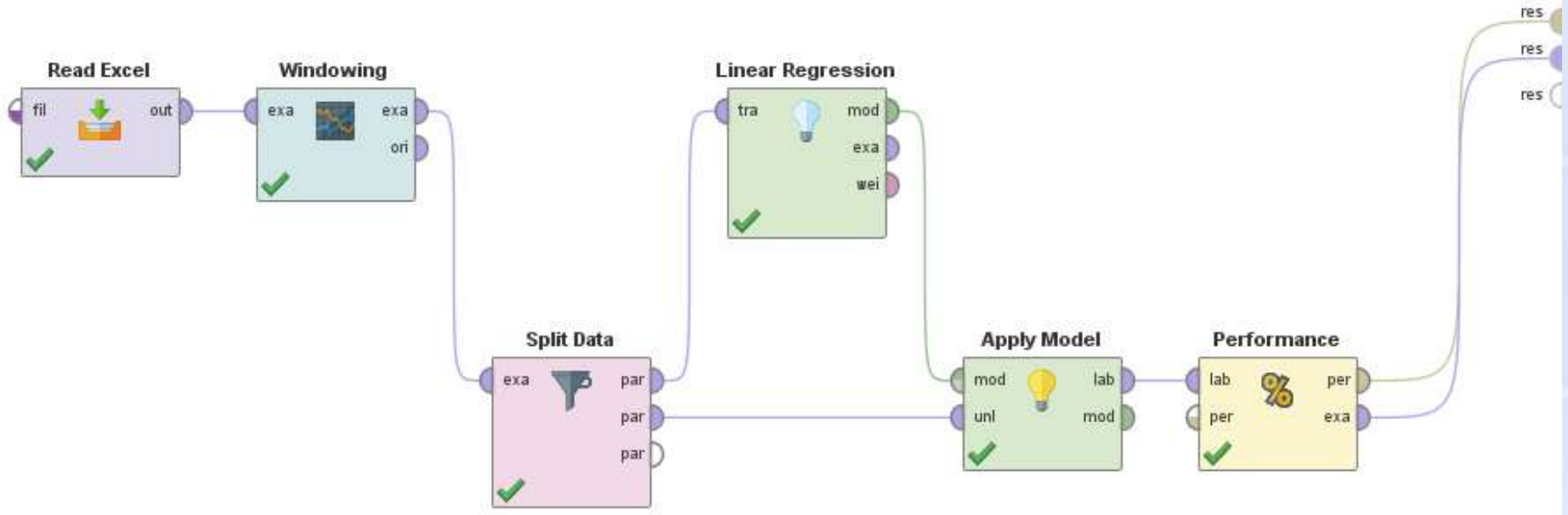
Windowing Operator Parameters

- **Ukuran Windows:** Menentukan berapa banyak “atribut” yang dibuat untuk data cross-sectional
 - Setiap baris deret waktu asli dalam lebar windows akan menjadi atribut baru
 - Misal $w = 6$
- **Step size :** Menentukan cara memajukan **window**
 - Gunakan $s = 1$
- **Horizon:** Menentukan seberapa jauh perkiraan dibuat
 - Jika ukuran **window** adalah 6 dan **horizon** adalah 1, maka baris ketujuh dari deret waktu asli menjadi sampel pertama untuk variabel “label”
 - Gunakan $h = 1$

Latihan

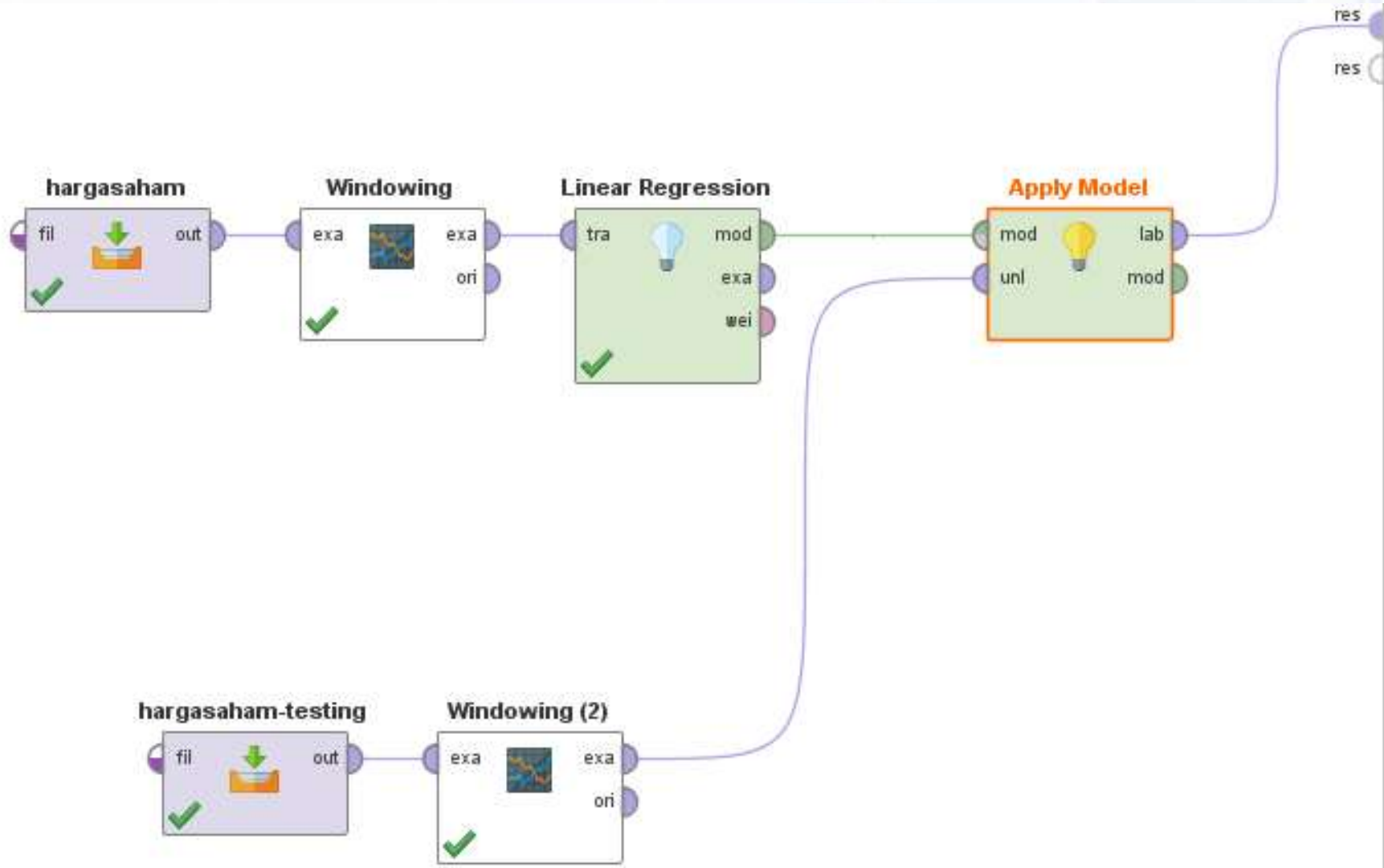
- Lakukan training dengan menggunakan **linear regression** pada dataset **hargasaham-training-uni.xls**
- Gunakan Split Data untuk memisahkan dataset di atas, 90% training dan 10% untuk testing
- Harus dilakukan proses **Windowing** pada dataset
- **Plot grafik** antara label dan hasil prediksi dengan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

I.40



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



Review dan Latihan

☺ **END** ☺

