

Mata Kuliah Coding & Machine Learning

# **Laporan Tugas Studi Kasus Modul Hal. 69-70**

**Dosen Pengampu: Sri Wulandari, S.Kom., M.Cs.**



**Disusun oleh:**

**Lathif Ramadhan (5231811022)**

**PROGRAM STUDI SAINS DATA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS TEKNOLOGI YOGYAKARTA  
YOGYAKARTA**

**2025**

# Daftar Isi

<b>Daftar Isi .....</b>	<b>II</b>
<b>1. Modul Hal 69 Studi Kasus 1a: Supervised (Estimasi Harga Rumah).....</b>	<b>1</b>
a. Berikan penjelasan data yang saudara dapatkan. ....	1
Kolom-Kolom Dataset .....	1
Ringkasan Karakteristik Data .....	2
Insight Awal .....	2
b. Jelaskan langkah-langkah saudara dalam mengestimasi harga rumah tersebut mulai dari pengambilan data sampai terbangunnya model. ....	3
Langkah 1: Persiapan Data .....	3
Langkah 2: Import Data ke RapidMiner .....	4
Langkah 3: Preprocessing Data .....	8
Langkah 4: Bagi Data menjadi Training dan Testing .....	12
Langkah 5: Bangun Model Regresi .....	13
Langkah 6: Training Model .....	15
Langkah 7: Evaluasi Model .....	15
Langkah 8: Jalankan Semua Process .....	16
c. Berikan penjelasan model yang saudara dapatkan. ....	27
d. Cobalah model tersebut untuk mengestimasi harga rumah, silahkan saudara gunakan data yang saudara kembangkan sendiri .....	30
<b>2. Modul Hal 70 Studi Kasus 1b: Estimasi (sesuai program studi) .....</b>	<b>32</b>
a. Deskripsikan kasus yang saudara kembangkan, tuliskan sumber datanya, berikan penjelasan data yang saudara peroleh tersebut. ....	32
1. Kasus yang Dikembangkan.....	32
2. Sumber Data.....	32
3. Penjelasan Data .....	32
5. Tujuan Analisis.....	34
b. Jelaskan langkah-langkah saudara dalam mengembangkan model untuk menyelesaikan kasus tersebut. ....	35
1. Mengambil Data (Retrieve Data).....	35
2. Mengubah Data Kategorikal Menjadi Numerik (One-Hot Encoding) .....	36

3. Menetapkan Peran Kolom (Set Role) .....	37
4. Membagi Data (Split Data) .....	38
5. Membangun Model Neural Network (Neural Net).....	39
6. Menerapkan Model pada Data Testing (Apply Model) .....	41
7. Mengevaluasi Performa Model (Performance).....	43
9. Ringkasan Workflow .....	47
c. Tuliskan rekomendasi saudara setelah model saudara dapatkan. ....	48

# 1. Modul Hal 69 Studi Kasus 1a: Supervised (Estimasi Harga Rumah)

Saudara adalah seorang data scientist di perusahaan "Mydata". Saat ini "Mydata" sedang mendapatkan proyek pembuatan model untuk mengestimasi harga rumah. Model yang harus saudara kembangkan tersebut didasarkan pada data yang terdapat pada tautan <https://www.kaggle.com/quantbruce/real-estate-price-prediction> (jika saudara tidak berhasil mengunduh, silahkan hubungi dosen pengampu), dengan menggunakan metode analisis regresi.

- a. Berikan penjelasan data yang saudara dapatkan.
- b. Jelaskan langkah-langkah saudara dalam mengestimasi harga rumah tersebut mulai dari pengambilan data sampai terbangunnya model.
- c. Berikan penjelasan model yang saudara dapatkan.
- d. Cobalah model tersebut untuk mengestimasi harga rumah, silahkan saudara gunakan data yang saudara kembangkan sendiri

## a. Berikan penjelasan data yang saudara dapatkan.

Dataset ini memiliki adalah data tentang properti rumah (real estate) yang berisi informasi terkait berbagai atribut rumah dan harganya. Dataset ini terdiri dari **414 baris** (data rumah) dan **8 kolom** (atribut). Berikut penjelasan detailnya:

### Kolom-Kolom Dataset

#### 1. No

- Kolom ini adalah nomor urut atau ID dari setiap data rumah. Ini hanya sebagai pensaya dan tidak memiliki pengaruh dalam analisis.

#### 2. X1 transaction date

- Kolom ini menunjukkan tanggal transaksi rumah dalam format tahun dan bulan (dalam bentuk desimal). Misalnya, 2012.917 berarti transaksi terjadi pada bulan November 2012 (karena  $0.917 \times 12 \approx 11$ ).

#### 3. X2 house age

- Kolom ini menunjukkan usia rumah dalam tahun. Misalnya, nilai 32.0 berarti rumah tersebut berusia 32 tahun.

#### 4. X3 distance to the nearest MRT station

- Kolom ini menunjukkan jarak rumah ke stasiun MRT (Mass Rapid Transit) terdekat dalam meter. Misalnya, nilai 84.87882 berarti rumah tersebut berjarak sekitar 84,88 meter dari stasiun MRT terdekat.

#### 5. X4 number of convenience stores

- Kolom ini menunjukkan jumlah toko atau minimarket yang ada di sekitar rumah. Misalnya, nilai 10 berarti ada 10 toko di sekitar rumah tersebut.

#### 6. X5 latitude

- Kolom ini menunjukkan koordinat lintang (latitude) dari lokasi rumah. Ini adalah informasi geografis yang menunjukkan posisi rumah di peta.

#### 7. X6 longitude

- Kolom ini menunjukkan koordinat bujur (longitude) dari lokasi rumah. Sama seperti latitude, ini adalah informasi geografis yang menunjukkan posisi rumah di peta.

#### 8. Y house price of unit area

- Kolom ini adalah target atau variabel yang ingin diprediksi. Ini menunjukkan harga rumah per unit area (misalnya, per meter persegi). Nilainya dalam satuan yang belum spesifik (mungkin dalam jutaan atau puluhan juta, tergantung konteks dataset).

### Ringkasan Karakteristik Data:

- **Jumlah Data:** 414 rumah.
- **Tipe Data:**
  - Numerik (float dan int).
  - Tidak ada data kategorikal dalam dataset ini.
- **Tidak Ada Data Hilang:** Semua kolom memiliki 414 nilai (tidak ada missing values).
- **Range Nilai:**
  - Usia rumah (X2 house age) berkisar dari 0 tahun (rumah baru) hingga 43,8 tahun.
  - Jarak ke stasiun MRT (X3 distance to the nearest MRT station) berkisar dari 23,38 meter hingga 6.488,02 meter.
  - Harga rumah (Y house price of unit area) berkisar dari 7,6 hingga 117,5 (satuan belum spesifik).

### Insight Awal:

- Dataset ini cocok untuk analisis **regresi** karena tujuannya adalah memprediksi harga rumah (Y house price of unit area) berdasarkan atribut-atribut lainnya.
- Beberapa fitur yang mungkin berpengaruh besar terhadap harga rumah adalah:
  - Jarak ke stasiun MRT (X3 distance to the nearest MRT station): Semakin dekat ke stasiun MRT, mungkin harga rumah semakin tinggi.
  - Jumlah toko di sekitar (X4 number of convenience stores): Semakin banyak toko, mungkin harga rumah semakin tinggi.
  - Usia rumah (X2 house age): Rumah yang lebih baru mungkin memiliki harga yang lebih tinggi.

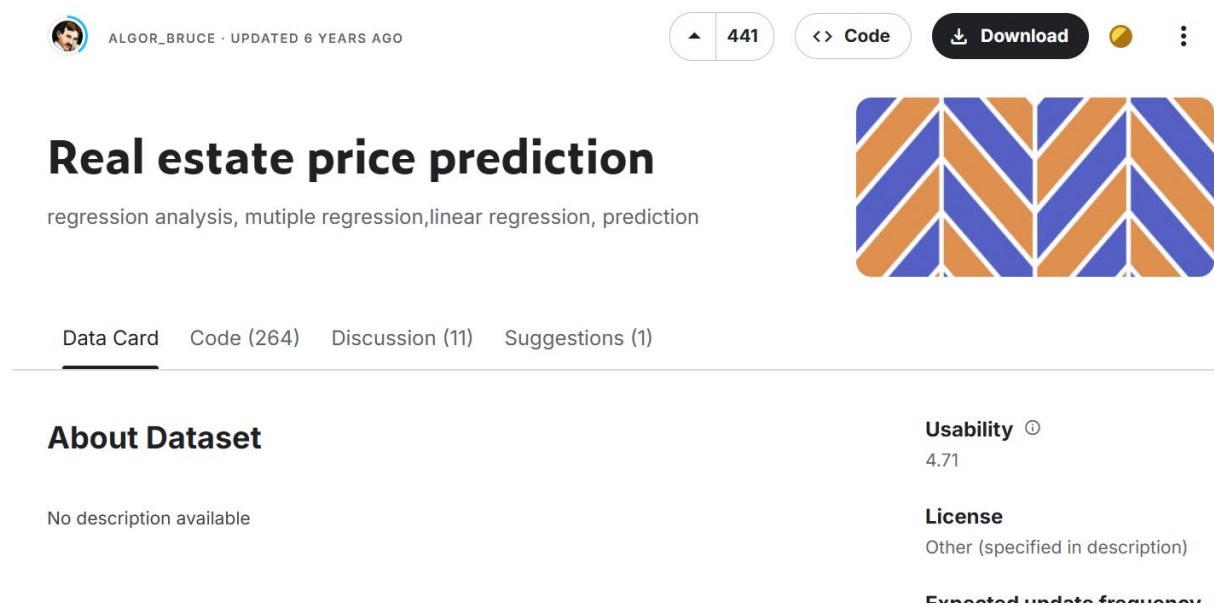
**b. Jelaskan langkah-langkah saudara dalam mengestimasi harga rumah tersebut mulai dari pengambilan data sampai terbangunnya model.**

## **Langkah 1: Persiapan Data**

### **1. Download Dataset:**

Kita download udlu datasetnya dari Kaggle dengan link:

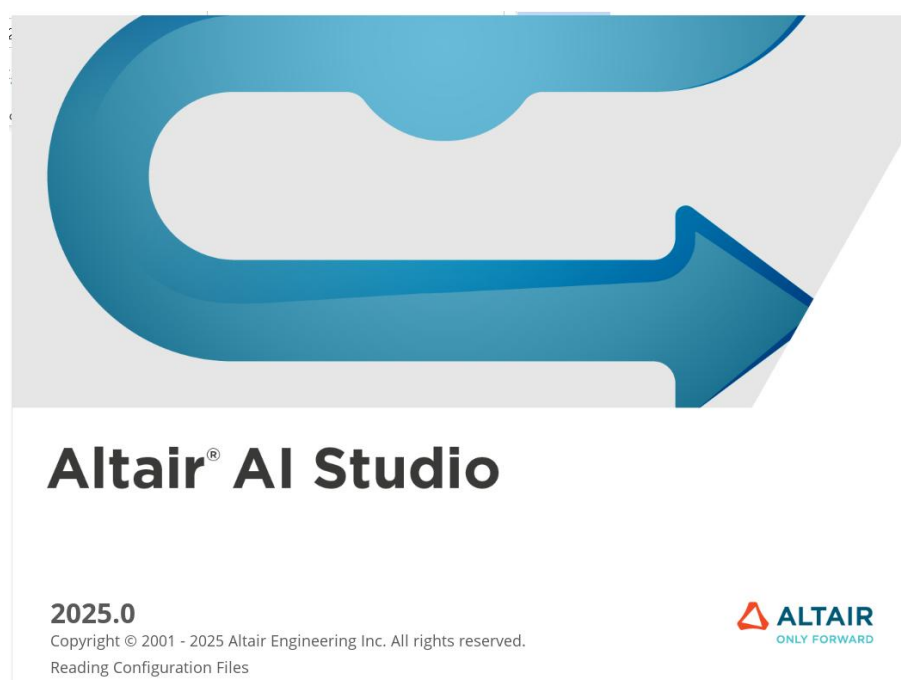
<https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction>



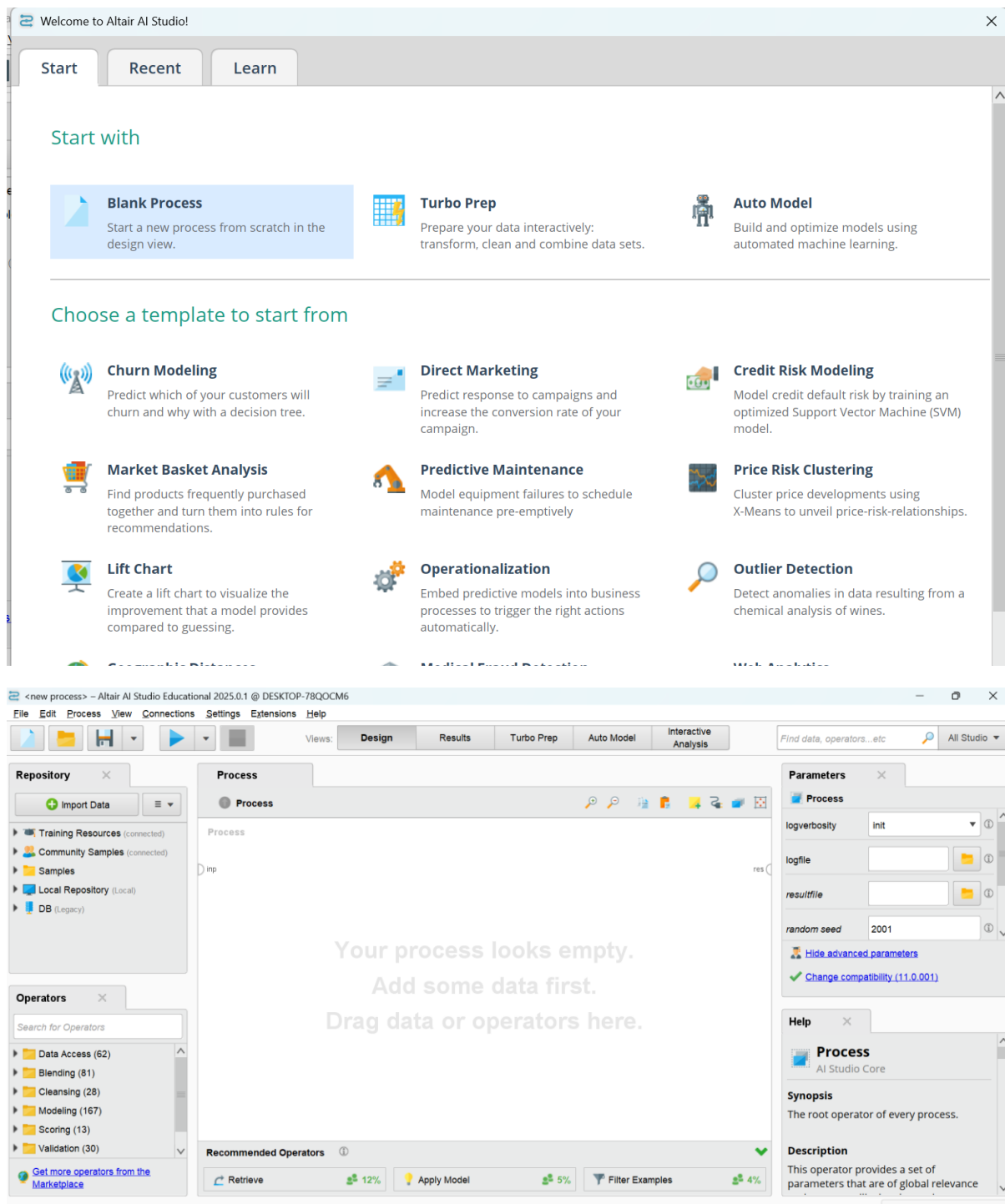
The screenshot shows the Kaggle dataset page for 'Real estate price prediction' by user ALGOR\_BRUCE, updated 6 years ago. It has 441 votes and a 'Download' button. The dataset is described as 'regression analysis, mutiple regression, linear regression, prediction'. Below the title, there are tabs for 'Data Card', 'Code (264)', 'Discussion (11)', and 'Suggestions (1)'. The 'About Dataset' section states 'No description available'. The 'Usability' is 4.71, and the 'License' is 'Other (specified in description)'. The 'Expected update frequency' is also visible.

### **2. Buka RapidMiner:**

Jalankan aplikasi RapidMiner di laptop/komputer.



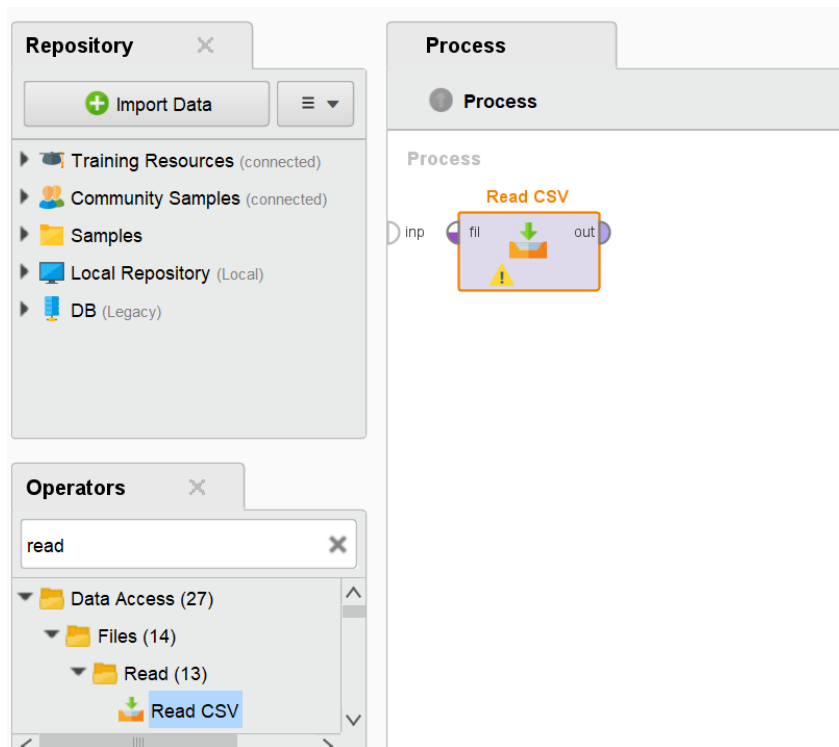
Buat proses baru dengan mengklik **File > New Process**.



## Langkah 2: Import Data ke RapidMiner

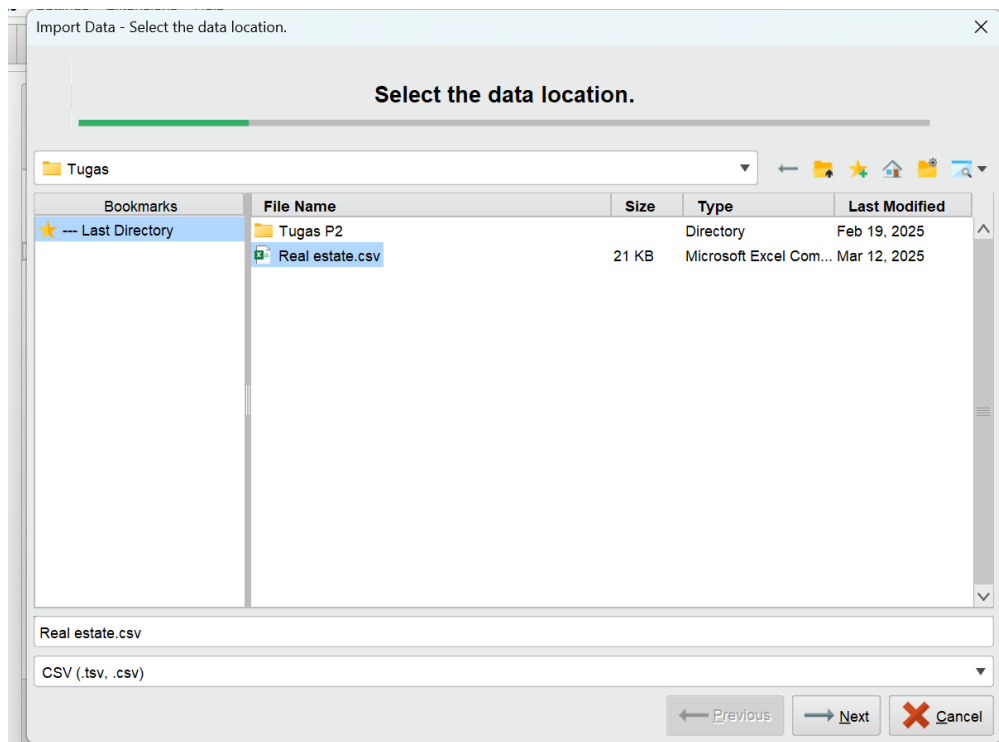
### 1. Tambahkan Operator Read CSV

- Cari operator **Read CSV** di panel operator (bisa ketik di search bar).
- Tarik operator tersebut ke area proses.



## 2. Konfigurasi Operator Read CSV

- Klik operator **Read CSV**.
- Di panel kanan, klik **Import Configuration Wizard**.
- Pilih file CSV yang sudah diunduh.



- Pastikan format file sudah benar (misalnya, delimiter koma, encoding UTF-8).



Import Data - Specify your data format

### Specify your data format

☒ Header Row    1  
 Start Row    1  
 Column Separator    Comma ","

File Encoding    windows-1252  
 Escape Character    \  
 Decimal Character    .

☒ Use Quotes    "  
☒ Skip Comments    #  
☐ Trim Lines    ☐ Multiline Text

1	No	X1 transac...	X2 house a...	X3 distanc...	X4 number ...	X5 latitude	X6 longitude	Y house pr...
2	1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
3	2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
4	3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
5	4	2013.500	13.3	561.9845	5	24.98746	121.54391	54.8
6	5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
7	6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
8	7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
9	8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
10	9	2013.500	31.7	5512.038	1	24.95095	121.48458	18.8
11	10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1

no problems.

Previous Next Cancel

- Klik **Finish** untuk mengimpor data.

Import Data - Format your columns.

### Format your columns.

Date format    Enter value...  
☐ Replace errors with missing values

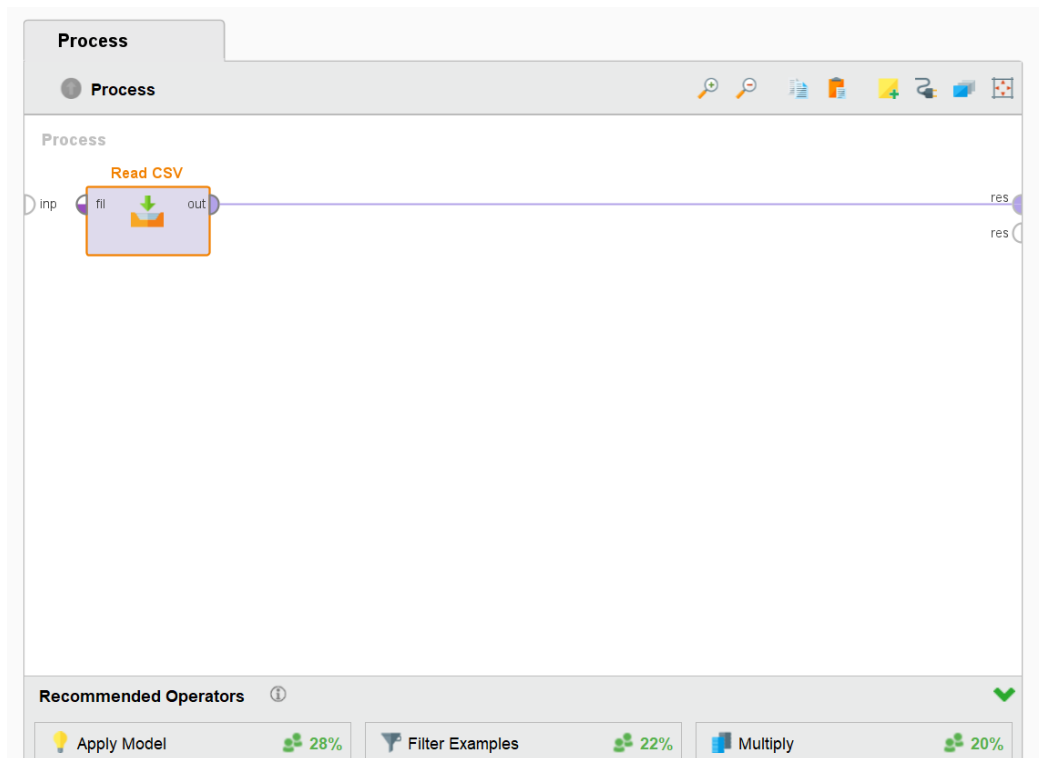
No	X1 transact...	X2 house age	X3 distance...	X4 number ...	X5 latitude	
integer	real	real	real	integer	real	
1	1	2012.917	32.000	84.879	10	24.983
2	2	2012.917	19.500	306.595	9	24.980
3	3	2013.583	13.300	561.985	5	24.987
4	4	2013.500	13.300	561.985	5	24.987
5	5	2012.833	5.000	390.568	5	24.979
6	6	2012.667	7.100	2175.030	3	24.963
7	7	2012.667	34.500	623.473	7	24.979
8	8	2013.417	20.300	287.603	6	24.980
9	9	2013.500	31.700	5512.038	1	24.951
10	10	2013.417	17.900	1783.180	3	24.967
11	11	2013.083	34.800	405.213	1	24.973

no problems.

Previous Finish Cancel

### 3. Cek Data

- Jalankan proses dengan mengklik tombol **Run**.



- Pastikan data sudah terimpor dengan benar. kita bisa melihat pratinjau data di panel **Results**.

The screenshot shows the Orange3 Results panel. The panel title is 'ExampleSet (Read CSV)'. Below the title, there are tabs for 'Open in', 'Turbo Prep', 'Auto Model', and 'Interactive Analysis'. The 'Turbo Prep' tab is selected. The panel shows a table of data with 13 rows and 9 columns. The columns are: Row No., No, X1 transacti..., X2 house age, X3 distance to the nearest MRT station, X4 number of convenience stores, X5 latitude, X6 longitude, and Y house price of unit area. The data is filtered to show 414 examples.

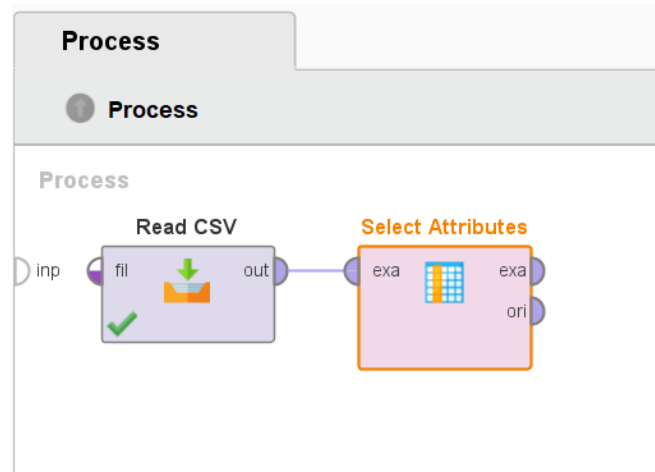
Row No.	No	X1 transacti...	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	1	2012.917	32	84.879	10	24.983	121.540	37.900
2	2	2012.917	19.500	306.595	9	24.980	121.540	42.200
3	3	2013.583	13.300	561.985	5	24.987	121.544	47.300
4	4	2013.500	13.300	561.985	5	24.987	121.544	54.800
5	5	2012.833	5	390.568	5	24.979	121.542	43.100
6	6	2012.667	7.100	2175.030	3	24.963	121.513	32.100
7	7	2012.667	34.500	623.473	7	24.979	121.536	40.300
8	8	2013.417	20.300	287.603	6	24.980	121.542	46.700
9	9	2013.500	31.700	5512.038	1	24.951	121.485	18.800
10	10	2013.417	17.900	1783.180	3	24.967	121.515	22.100
11	11	2013.083	34.800	405.213	1	24.973	121.534	41.400
12	12	2013.333	6.300	90.456	9	24.974	121.543	58.100
13	13	2012.917	13	492.231	5	24.965	121.537	38.300

ExampleSet (414 examples, 0 special attributes, 8 regular attributes)

## Langkah 3: Preprocessing Data

### 1. Hapus Kolom yang Tidak Dibutuhkan

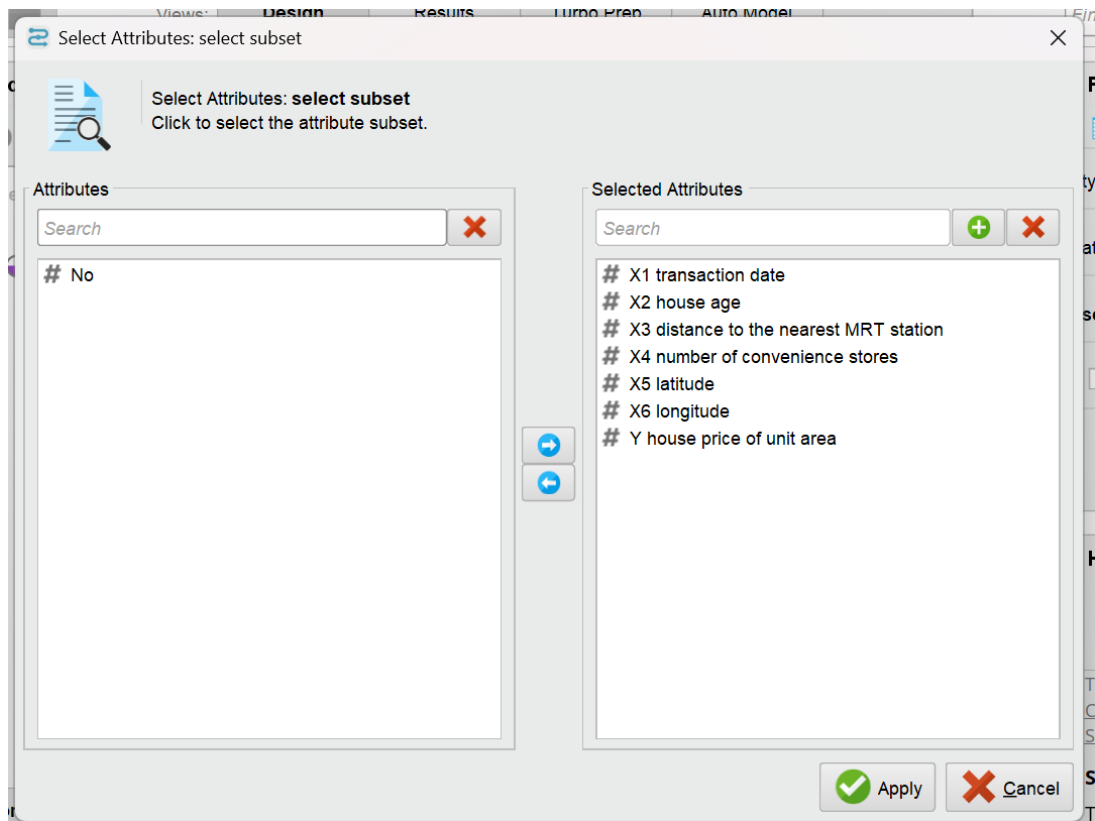
- Kolom **No** tidak diperlukan karena hanya nomor urut. Maka dari itu kita tambahkan operator **Select Attributes**.



- Konfigurasi operator untuk memilih kolom yang relevan (semua kolom kecuali **No**).

The 'Parameters' window for the 'Select Attributes' operator is shown. It contains the following settings:

- type**: include attributes
- attribute filter type**: a subset
- select subset**: Select Attributes...
- ☐ **also apply to special attributes (id, label..)**



ExampleSet (Select Attributes)

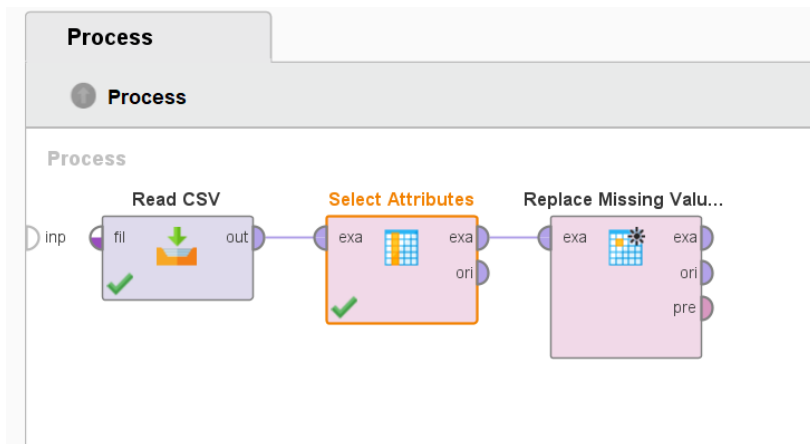
Open in: Turbo Prep Auto Model Interactive Analysis Filter (414 / 414 examples): all

Row No.	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	2012.917	32	84.879	10	24.983	121.540	37.900
2	2012.917	19.500	306.595	9	24.980	121.540	42.200
3	2013.583	13.300	561.985	5	24.987	121.544	47.300
4	2013.500	13.300	561.985	5	24.987	121.544	54.800
5	2012.833	5	390.568	5	24.979	121.542	43.100
6	2012.667	7.100	2175.030	3	24.963	121.513	32.100
7	2012.667	34.500	623.473	7	24.979	121.536	40.300
8	2013.417	20.300	287.603	6	24.980	121.542	46.700
9	2013.500	31.700	5512.038	1	24.951	121.485	18.800
10	2013.417	17.900	1783.180	3	24.967	121.515	22.100
11	2013.083	34.800	405.213	1	24.973	121.534	41.400
12	2013.333	6.300	90.456	9	24.974	121.543	58.100
13	2012.917	13	192.231	5	24.965	121.537	39.300

ExampleSet (414 examples, 0 special attributes, 7 regular attributes)

## 2. Cek Missing Values

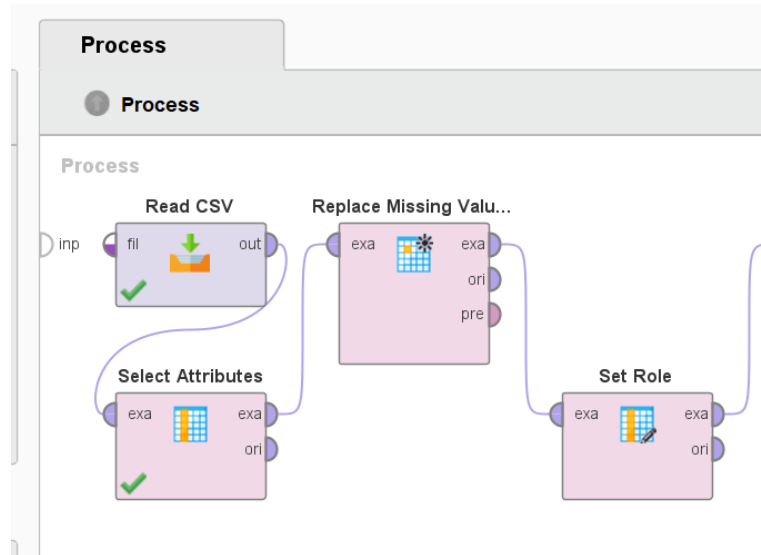
- Tambahkan operator **Replace Missing Values** jika ada data yang hilang.



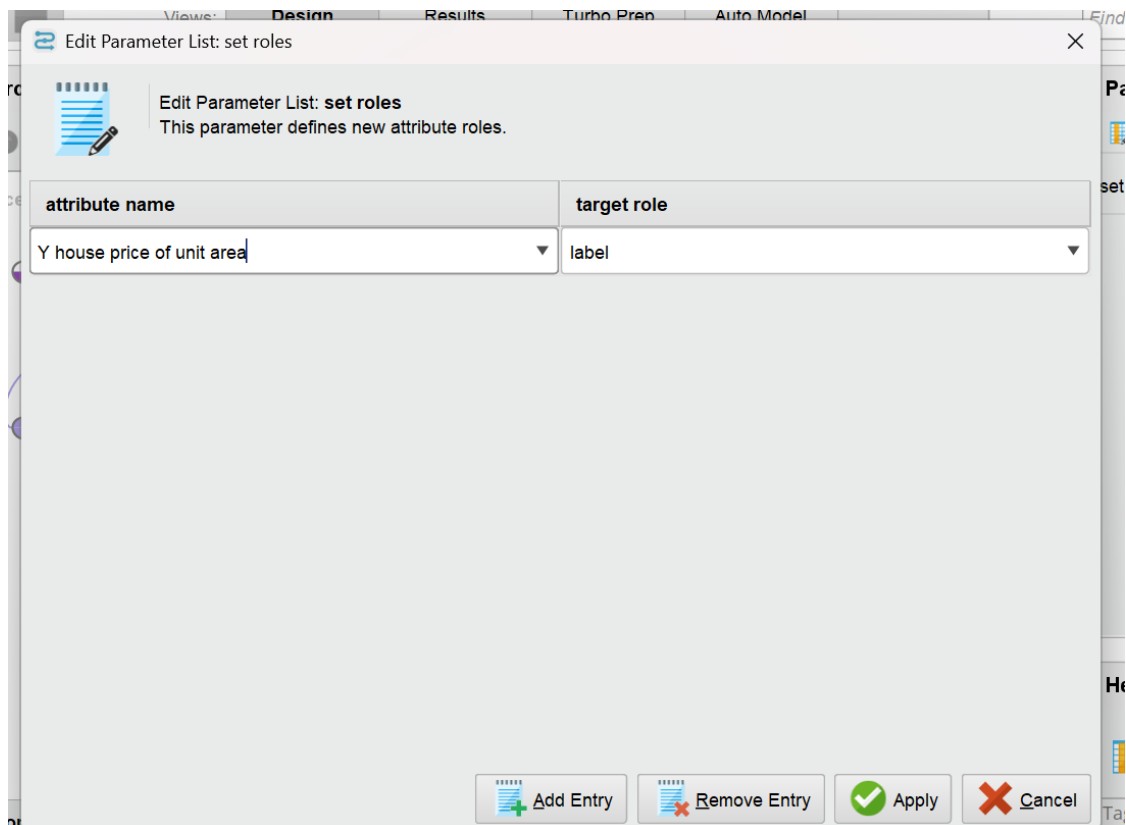
- Konfigurasi operator untuk mengisi missing values dengan nilai rata-rata (mean) atau median.

### 3. Set Role untuk Kolom Label

- Tambahkan operator **Set Role**.



- Klik operator **Set Role**.
- Di panel kanan, pilih kolom **Y house price of unit area** sebagai **label** (target variabel).

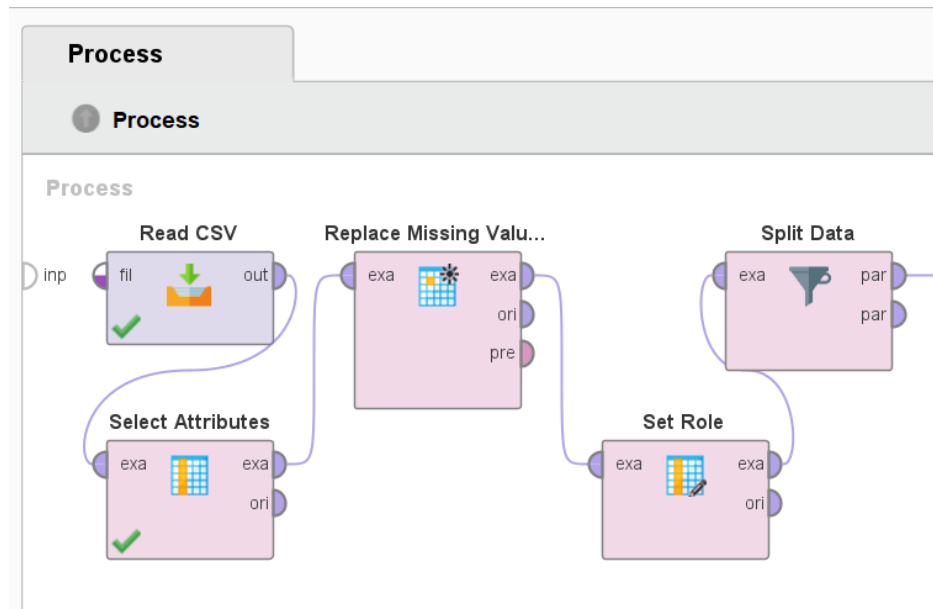


- Caranya:
  - Klik **Edit List** di bagian **attribute filter**.
  - Pilih kolom **Y house price of unit area**.
  - Set **target role** menjadi **label**.
  - Klik **OK**.

## Langkah 4: Bagi Data menjadi Training dan Testing

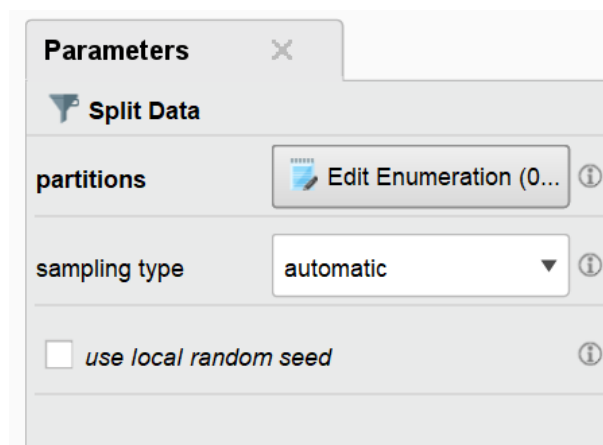
### 1. Tambahkan Operator Split Data

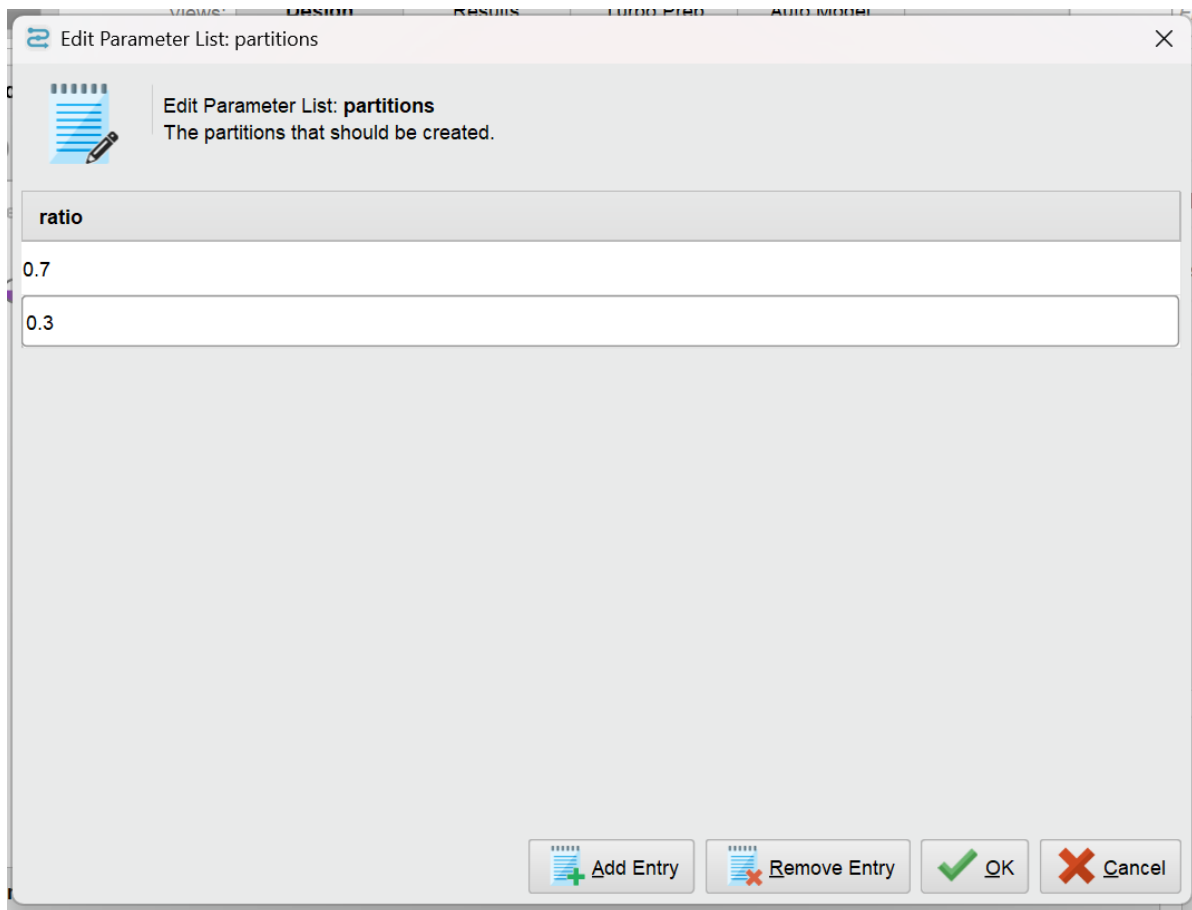
- Operator ini digunakan untuk membagi dataset menjadi data training dan data testing.



### 2. Konfigurasi Operator Split Data

- Set rasio pembagian (misalnya, 0.7 untuk training dan 0.3 untuk testing).

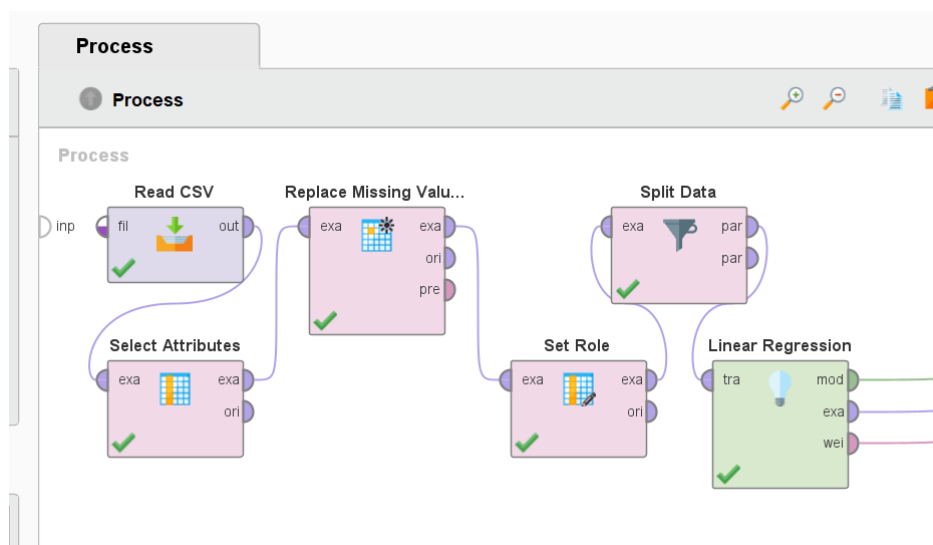




## Langkah 5: Bangun Model Regresi

### 1. Pilih Algoritma Regresi

- Untuk kasus ini, kita bisa menggunakan **Linear Regression**.
- Cari operator **Linear Regression** di panel operator.
- Tarik operator tersebut ke area proses.



### 2. Hubungkan Operator

- Hubungkan output **Split Data (training)** ke input **Linear Regression**.
- Output **Linear Regression** akan menghasilkan model.



AttributeWeights (Linear Regression) <span>×</span>	
attribute	weight
X1 transaction date	4.982
X2 house age	-0.299
X3 distance to the nearest MRT station	-0.004
X4 number of convenience stores	1.245
X5 latitude	208.932
X6 longitude	0

AttributeWeights (Linear Regression)

ExampleSet (Split Data)

LinearRegression (Linear Regression)

Open in

Turbo Prep

Auto Model

Interactive Analysis

Filter (290 / 290 examples): all

Row No.	Y house price of unit area	X1 transaction date	X2 house age	X3 distance to the nearest MRT stati...	X4 number of convenien...	X5 latitude	X6 longitude
1	37.900	2012.917	32	84.879	10	24.983	121.540
2	42.200	2012.917	19.500	306.595	9	24.980	121.540
3	47.300	2013.583	13.300	561.985	5	24.987	121.544
4	32.100	2012.667	7.100	2175.030	3	24.963	121.513
5	40.300	2012.667	34.500	623.473	7	24.979	121.536
6	46.700	2013.417	20.300	287.603	6	24.980	121.542
7	18.800	2013.500	31.700	5512.038	1	24.951	121.485
8	22.100	2013.417	17.900	1783.180	3	24.967	121.515
9	41.400	2013.083	34.800	405.213	1	24.973	121.534
10	58.100	2013.333	6.300	90.456	9	24.974	121.543
11	39.300	2012.917	13	492.231	5	24.965	121.537
12	23.800	2012.667	20.400	2469.645	4	24.961	121.510
13	50.500	2013.583	35.700	579.208	2	24.982	121.546

ExampleSet (290 examples, 1 special attribute, 6 regular attributes)

AttributeWeights (Linear Regression)

ExampleSet (Split Data)

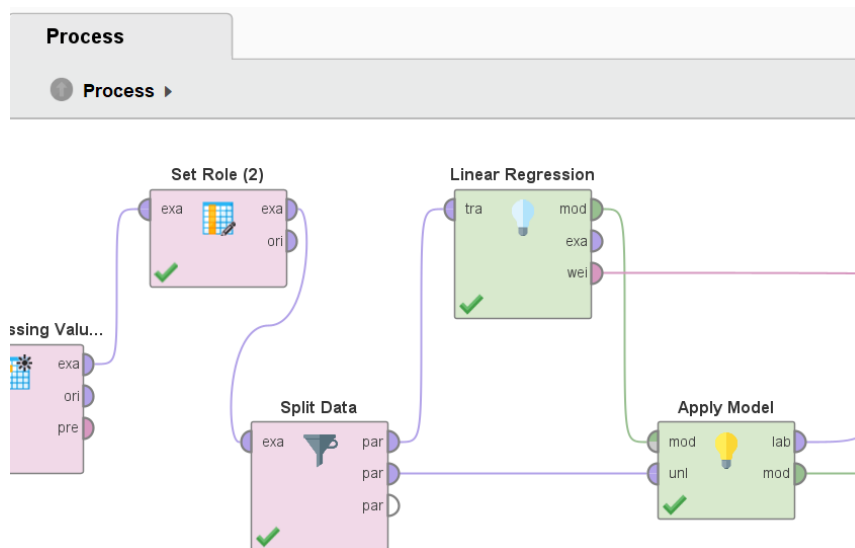
LinearRegression (Linear Regression)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
X1 transaction date	4.982	1.902	0.102	1.000	2.619	0.009	***
X2 house age	-0.299	0.049	-0.243	1.000	-6.127	0.000	****
X3 distance to the nearest MRT station	-0.004	0.001	-0.382	0.500	-6.577	0.000	****
X4 number of convenience stores	1.245	0.240	0.263	0.664	5.180	0.000	****
X5 latitude	208.932	58.633	0.184	0.658	3.563	0.000	****
(Intercept)	-15204.872	4008.404	?	?	-3.793	0.000	****

## Langkah 6: Training Model

### 1. Tambahkan Operator Apply Model:

- Operator ini digunakan untuk menerapkan model ke data training.
- Hubungkan output **Linear Regression (model)** ke input **Apply Model**.
- Hubungkan output **Split Data (training)** ke input **Apply Model**.



### 2. Jalankan Proses:

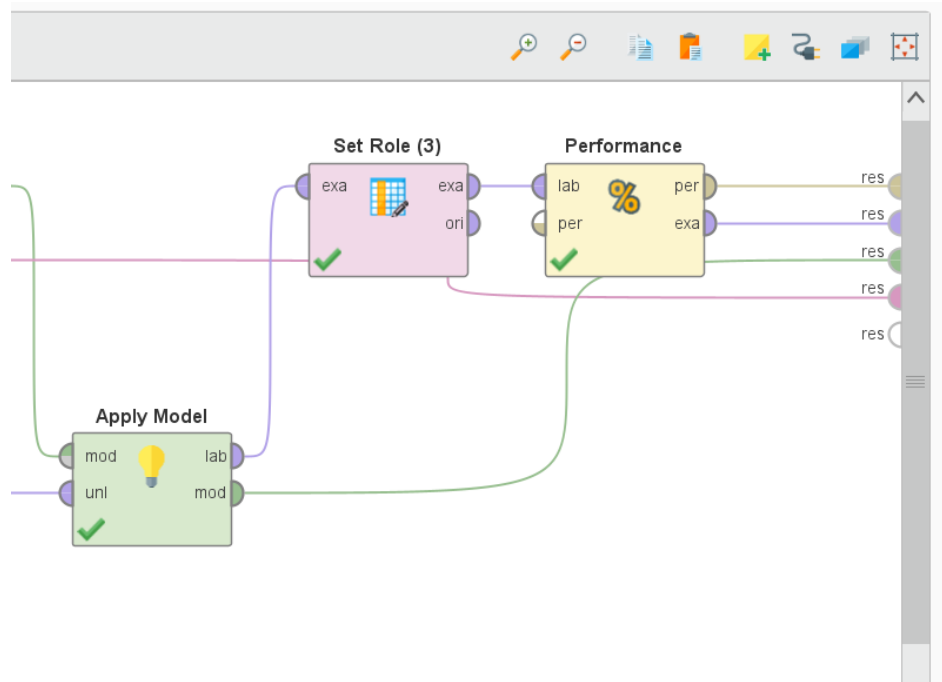
- Klik tombol **Run** untuk melatih model.
- Model akan dipelajari dari data training.

## Langkah 7: Evaluasi Model

**Tujuan:** Mengevaluasi performa model menggunakan data testing.

### 1. Tambahkan Operator Performance (Regression):

- Cari operator Performance (Regression) di panel operator.
- Tarik operator tersebut ke area proses.



## 2. Hubungkan Operator:

- Hubungkan output labelled dari Apply Model ke input Performance.
- Hubungkan output testing dari Split Data ke input Performance.

## 3. Jalankan Proses:

- Klik tombol Run untuk mengevaluasi model.
- Hasil evaluasi akan muncul di panel Results. Perhatikan metrik seperti R-squared, Mean Absolute Error (MAE), dan Root Mean Squared Error (RMSE).

# Langkah 8: Jalankan Semua Process

## 1. Linear Regression

Result History							
AttributeWeights (Linear Regression)							
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
X1 transaction date	4.982	1.902	0.102	1.000	2.619	0.009	***
X2 house age	-0.299	0.049	-0.243	1.000	-6.127	0.000	****
X3 distance to the ne...	-0.004	0.001	-0.382	0.500	-6.577	0.000	****
X4 number of conveni...	1.245	0.240	0.263	0.664	5.180	0.000	****
X5 latitude	208.932	58.633	0.184	0.658	3.563	0.000	****
(Intercept)	-15204.872	4008.404	?	?	-3.793	0.000	****

## Signifikansi Variabel (p-Value)

Hasil output juga menunjukkan **p-value** untuk setiap variabel. **p-value** digunakan untuk menentukan apakah variabel tersebut signifikan secara statistik dalam memprediksi harga rumah. Biasanya:

- Jika **p-value** < **0.05**, variabel dianggap **signifikan**.
- Jika **p-value** ≥ **0.05**, variabel dianggap **tidak signifikan**.

**Penjelasan p-Value:**

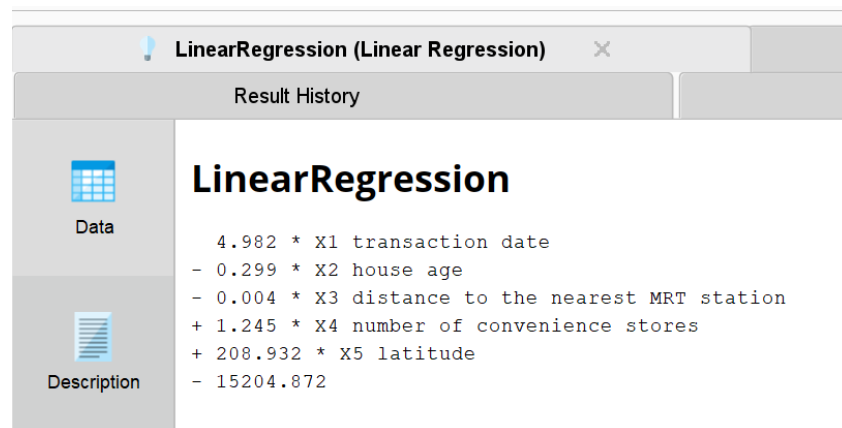
- **X1 (transaction date):** p-value = **0.009** (signifikan).
- **X2 (house age):** p-value = **0.000** (signifikan).
- **X3 (distance to MRT station):** p-value = **0.000** (signifikan).
- **X4 (number of convenience stores):** p-value = **0.000** (signifikan).
- **X5 (latitude):** p-value = **0.000** (signifikan).
- **Intercept:** p-value = **0.000** (signifikan).

**Kesimpulan:** Semua variabel dalam model ini signifikan secara statistik.

### Performa Model

Hasil output tidak menunjukkan metrik evaluasi seperti **RMSE**, **MAE**, atau **R-squared**. Namun, jika Saya memiliki metrik tersebut, berikut cara menilainya:

- **RMSE (Root Mean Squared Error):** Mengukur rata-rata kesalahan prediksi. Semakin kecil, semakin baik.
- **MAE (Mean Absolute Error):** Mengukur rata-rata kesalahan absolut. Semakin kecil, semakin baik.
- **R-squared (R<sup>2</sup>):** Mengukur seberapa baik model menjelaskan variasi data. Nilainya antara 0 dan 1, semakin mendekati 1, semakin baik.



### Persamaan Linear Regression

Hasil output menunjukkan persamaan regresi linear yang dihasilkan oleh model. Persamaan ini menggambarkan hubungan antara variabel independen (fitur) dan variabel dependen (target). Berikut persamaannya:

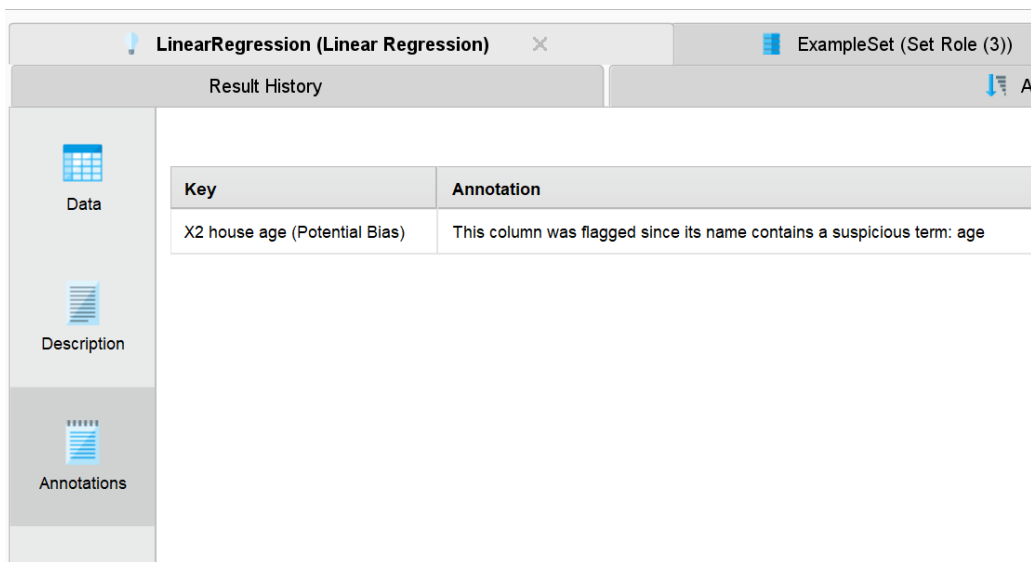
*Harga Rumah*

$$= 4.982 \times X1 - 0.299 \times X2 - 0.004 \times X3 + 1.245 \times X4 + 208.932 \times X5 - 15204.872$$

### Penjelasan Koefisien:

- **X1 (transaction date):** Koefisien **4.982** berarti setiap peningkatan 1 unit pada tanggal transaksi, harga rumah meningkat sebesar **4.982 unit** (asumsi: tanggal transaksi dalam format tahun.bulan).
- **X2 (house age):** Koefisien **-0.299** berarti setiap peningkatan 1 tahun usia rumah, harga rumah **menurun** sebesar **0.299 unit**.

- **X3 (distance to MRT station):** Koefisien **-0.004** berarti setiap peningkatan 1 meter jarak ke stasiun MRT, harga rumah **menurun** sebesar **0.004 unit**.
- **X4 (number of convenience stores):** Koefisien **1.245** berarti setiap penambahan 1 toko di sekitar rumah, harga rumah meningkat sebesar **1.245 unit**.
- **X5 (latitude):** Koefisien **208.932** berarti setiap peningkatan 1 unit pada latitude, harga rumah meningkat sebesar **208.932 unit**.
- **Intercept (-15204.872):** Ini adalah nilai dasar harga rumah ketika semua variabel independen bernilai 0.



The screenshot shows the 'LinearRegression (Linear Regression)' window in RapidMiner. On the left is a sidebar with 'Data', 'Description', and 'Annotations' tabs. The main area displays a 'Result History' table with the following content:

Key	Annotation
X2 house age (Potential Bias)	This column was flagged since its name contains a suspicious term: age

### Potensi Bias dalam Variabel

**Arti:** RapidMiner mensayai variabel ini karena mengandung kata "age" yang mungkin menimbulkan bias. Namun, dalam konteks ini, variabel usia rumah memang relevan untuk memprediksi harga rumah, jadi Saya bisa mengabaikan peringatan ini.

### Kesimpulan

- Model Linear Regression ini telah berhasil dibangun dengan persamaan yang jelas.
- Semua variabel signifikan secara statistik.
- Variabel **X2 (house age)** ditsayai karena potensi bias, tetapi dalam konteks ini bisa diabaikan.
- Untuk menilai performa model secara lengkap, Saya perlu melihat metrik evaluasi seperti **RMSE**, **MAE**, dan **R-squared**.

## 2. Example Set

LinearRegression (Linear Regression)								
ExampleSet (Set Role (3))								
PerformanceVector (Performance)								
Result History								
AttributeWeights (Linear Regression)								
<div> <div>Open in</div> <div>Turbo Prep</div> <div>Auto Model</div> <div>Interactive Analysis</div> </div> <div>Filter (124 / 124 examples): all</div>								
...	Y house price of unit area	prediction(Y house price of unit area)	X1 transaction date	X2 house age	X3 distance to ...	X4 number of conveni...	X5 latitude	X6 longitude
6	22.100	12.786	2013.500	25.900	4519.690	0	24.948	121.496
7	25	39.927	2012.750	29.600	769.403	7	24.983	121.534
8	27.300	32.118	2013.500	13.900	4079.418	0	25.015	121.518
9	47.700	44.756	2012.667	3.100	577.962	6	24.972	121.547
10	15.900	14.445	2013	13.600	4082.015	0	24.942	121.504
11	34.700	35.758	2013.417	36.100	519.462	5	24.963	121.538
12	34.100	39.320	2012.750	34.400	512.787	6	24.987	121.543
13	38.900	41.467	2013.083	13.300	492.231	5	24.965	121.537
14	51.700	45.025	2013.083	16.100	289.325	5	24.982	121.543
15	13.700	22.402	2012.833	31.700	1160.632	0	24.950	121.530
16	27.700	29.335	2012.917	17.200	2175.877	3	24.963	121.513
17	50	54.570	2013.583	5.500	60.156	0	24.974	121.543

ExampleSet (124 examples, 2 special attributes, 6 regular attributes)

LinearRegression (Linear Regression)								
ExampleSet (Set Role (3))								
PerformanceVector (Performance)								
Result History								
AttributeWeights (Linear Regression)								
<div> <div>Name</div> <div>Type</div> <div>Missing</div> <div>Statistics</div> </div> <div>Filter (8 / 8 attributes): Search for Attributes</div>								
Label	Y house price of unit area	Real	0	Min 12.800	Max 78	Average 38.561		
Prediction	prediction(Y house price of unit ...)	Real	0	Min 7.341	Max 55.093	Average 37.382		
	X1 transaction date	Real	0	Min 2012.667	Max 2013.583	Average 2013.130		
	X2 house age	Real	0	Min 0	Max 43.800	Average 18.387		
	X3 distance to the nearest MRT ...	Real	0	Min 49.661	Max 6306.153	Average 1062.462		
	X4 number of convenience stores	Integer	0	Min 0	Max 10	Average 3.911		
	X5 latitude	Real	0	Min 24.942	Max 25.015	Average 24.970		

Showing attributes 1 - 8

Examples: 124 Special Attributes: 2 Regular Attributes: 6

X3 distance to the nearest MRT ...	Real	0	Min 49.661	Max 6306.153	Average 1062.462
X4 number of convenience stores	Integer	0	Min 0	Max 10	Average 3.911
X5 latitude	Real	0	Min 24.942	Max 25.015	Average 24.970
X6 longitude	Real	0	Min 121.475	Max 121.560	Average 121.533

### 1. Struktur Example Set

Example Set terdiri dari:

- **124 contoh (rows):** Ini adalah jumlah data yang digunakan dalam proses.
- **2 special attributes:** Biasanya ini adalah **label** (target variabel) dan **prediction** (hasil prediksi model).
- **6 regular attributes:** Ini adalah fitur-fitur yang digunakan untuk memprediksi harga rumah.

### 2. Kolom-Kolom dalam Example Set

Berikut adalah kolom-kolom yang muncul di Example Set:

**a. Y house price of unit area (Label)**

- Ini adalah **target variabel** yang ingin diprediksi (harga rumah per unit area).
- **Statistik:**
  - **Min:** 7.341
  - **Max:** 78.000
  - **Average:** 38.661

**b. prediction(Y house price of unit area) (Prediction)**

- Ini adalah **hasil prediksi** dari model Linear Regression.
- **Statistik:**
  - **Min:** 7.341
  - **Max:** 55.093
  - **Average:** 37.382

**c. X1 transaction date**

- Tanggal transaksi rumah dalam format tahun.bulan.
- **Statistik:**
  - **Min:** 2012.687
  - **Max:** 2013.583
  - **Average:** 2013.130

**d. X2 house age**

- Usia rumah dalam tahun.
- **Statistik:**
  - **Min:** 0
  - **Max:** 43.800
  - **Average:** 18.387

**e. X3 distance to the nearest MRT station**

- Jarak ke stasiun MRT terdekat dalam meter.
- **Statistik:**
  - **Min:** 49.661
  - **Max:** 6306.153
  - **Average:** 1062.462

**f. X4 number of convenience stores**

- Jumlah toko di sekitar rumah.
- **Statistik:**
  - **Min:** 0
  - **Max:** 10
  - **Average:** 3.911

**g. X5 latitude**

- Koordinat lintang (latitude) lokasi rumah.
- **Statistik:**
  - **Min:** 24.942
  - **Max:** 25.015
  - **Average:** 24.970

#### **h. X6 longitude**

- Koordinat bujur (longitude) lokasi rumah.
- **Statistik:**
  - **Min:** 121.475
  - **Max:** 121.560
  - **Average:** 121.533

### **3. Analisis Example Set**

#### **a. Perbandingan Label dan Prediksi**

- **Label (Y house price of unit area):** Nilai aktual harga rumah.
- **Prediction:** Nilai prediksi harga rumah dari model.
- **Contoh:**
  - Pada baris pertama, nilai aktual = **22.100**, prediksi = **12.786**.
  - Pada baris kedua, nilai aktual = **25.000**, prediksi = **39.927**.
- **Kesimpulan:** Terlihat ada perbedaan antara nilai aktual dan prediksi. Ini menunjukkan bahwa model tidak selalu akurat, tetapi secara rata-rata, prediksi mendekati nilai aktual (average label = 38.661, average prediction = 37.382).

#### **b. Variasi Data**

- **X2 house age:** Usia rumah berkisar dari 0 tahun (rumah baru) hingga 43.8 tahun.
- **X3 distance to MRT station:** Jarak ke stasiun MRT berkisar dari 49.661 meter hingga 6306.153 meter.
- **X4 number of convenience stores:** Jumlah toko di sekitar rumah berkisar dari 0 hingga 10.

#### **c. Koordinat Lokasi**

- **Latitude:** Lokasi rumah berada di antara 24.942 hingga 25.015.
- **Longitude:** Lokasi rumah berada di antara 121.475 hingga 121.560.

### **4. Kesimpulan**

- Example Set menunjukkan data yang digunakan untuk melatih dan menguji model, serta hasil prediksi dari model Linear Regression.
- Terlihat bahwa model mampu memprediksi harga rumah dengan cukup baik, meskipun ada beberapa perbedaan antara nilai aktual dan prediksi.
- Untuk menilai performa model secara lengkap, Saya perlu melihat metrik evaluasi seperti **RMSE**, **MAE**, dan **R-squared**.



### 3. Performance Vector (Performance)

## root\_mean\_squared\_error

root\_mean\_squared\_error: 8.290 +/- 0.000

## PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 8.290 +/- 0.000
absolute_error: 6.101 +/- 5.613
relative_error: 16.98% +/- 14.67%
relative_error_lenient: 14.98% +/- 11.91%
relative_error_strict: 20.52% +/- 21.12%
normalized_absolute_error: 0.599
root_relative_squared_error: 0.635
squared_error: 68.728 +/- 148.191
correlation: 0.778
squared_correlation: 0.606
prediction_average: 38.561 +/- 13.061
spearman_rho: 0.766
kendall_tau: 0.581
```

### 1. Root Mean Squared Error (RMSE)

- **Nilai:** 8.290
- **Arti:** RMSE mengukur rata-rata kesalahan prediksi model. Semakin kecil nilai RMSE, semakin akurat model tersebut.
- **Analisis:**
  - RMSE = 8.290 berarti rata-rata kesalahan prediksi adalah **8.290 unit** dari harga rumah aktual.
  - Jika rata-rata harga rumah adalah **38.661** (seperti yang terlihat di Example Set), maka kesalahan prediksi sekitar **21.4%** dari rata-rata harga:  
$$\left( \frac{8.290}{38.661} \times 100 \right).$$
  - Secara umum, RMSE dianggap **baik** jika nilainya kurang dari 10% dari rata-rata target. Dalam kasus ini, RMSE = 8.290 masih **cukup baik**, tetapi bisa ditingkatkan.

### 2. Absolute Error

- **Nilai:** 6.101 +/- 5.613
- **Arti:** Absolute Error mengukur rata-rata kesalahan absolut prediksi model.
- **Analisis:**
  - Rata-rata kesalahan absolut adalah **6.101 unit**, dengan deviasi stsayar **5.613**.

- Ini berarti sebagian besar prediksi memiliki kesalahan sekitar **6.101 unit**.

### 3. Relative Error

- **Nilai:** 16.98% +/- 14.67%
- **Arti:** Relative Error mengukur kesalahan prediksi relatif terhadap nilai aktual.
- **Analisis:**
  - Rata-rata kesalahan relatif adalah **16.98%**, dengan deviasi standar **14.67%**.
  - Ini berarti prediksi model memiliki kesalahan sekitar **16.98%** dari nilai aktual.

### 4. Root Relative Squared Error (RRSE)

- **Nilai:** 0.635
- **Arti:** RRSE mengukur kesalahan prediksi relatif terhadap variasi data.
- **Analisis:**
  - $RRSE = 0.635$  berarti model memiliki kesalahan sekitar **63.5%** dari variasi data.
  - Semakin kecil RRSE, semakin baik. Nilai ini menunjukkan bahwa model masih memiliki ruang untuk perbaikan.

### 5. Correlation

- **Nilai:** 0.778
- **Arti:** Correlation mengukur seberapa kuat hubungan antara prediksi dan nilai aktual.
- **Analisis:**
  - Nilai korelasi **0.778** menunjukkan hubungan yang **kuat** antara prediksi dan nilai aktual.
  - Semakin mendekati 1, semakin baik.

### 6. Squared Correlation (R-squared)

- **Nilai:** 0.606
- **Arti:** R-squared mengukur seberapa baik model menjelaskan variasi data.
- **Analisis:**
  - $R\text{-squared} = 0.606$  berarti model menjelaskan **60.6%** variasi data.
  - Nilai ini dianggap **cukup baik**, tetapi masih bisa ditingkatkan.

### 7. Spearman's Rho

- **Nilai:** 0.766
- **Arti:** Spearman's Rho mengukur korelasi rank antara prediksi dan nilai aktual.
- **Analisis:**
  - Nilai **0.766** menunjukkan hubungan rank yang **kuat** antara prediksi dan nilai aktual.

### 8. Kendall's Tau

- **Nilai:** 0.581
- **Arti:** Kendall's Tau mengukur kesesuaian antara prediksi dan nilai aktual.
- **Analisis:**
  - Nilai **0.581** menunjukkan kesesuaian yang **cukup baik** antara prediksi dan nilai aktual.

## 9. Prediction Average

- **Nilai:** 38.561 +/- 13.061
- **Arti:** Ini adalah rata-rata nilai prediksi dari model.
- **Analisis:**
  - Rata-rata prediksi adalah **38.561 unit**, dengan deviasi stsayar **13.061**.
  - Nilai ini mendekati rata-rata nilai aktual (**38.661**), yang menunjukkan bahwa model memiliki bias yang rendah.

## 10. Kesimpulan

- **Kelebihan Model:**
  - Model memiliki korelasi yang kuat antara prediksi dan nilai aktual (correlation = 0.778).
  - R-squared = 0.606 menunjukkan bahwa model menjelaskan **60.6%** variasi data.
  - Kesalahan prediksi relatif (relative error) sekitar **16.98%**, yang masih bisa diterima.
- **Kekurangan Model:**
  - RMSE = 8.290 masih relatif besar (sekitar **21.4%** dari rata-rata harga rumah).
  - RRSE = 0.635 menunjukkan bahwa model masih memiliki kesalahan yang signifikan.
- **Rekomendasi:**
  - Coba lakukan **feature engineering** untuk menambahkan fitur yang lebih relevan.
  - Coba algoritma lain seperti **Decision Tree Regression**, **Random Forest Regression**, atau **Gradient Boosting**.
  - Lakukan **hyperparameter tuning** untuk meningkatkan performa model.

## 4. Attribute Weights (Linear Regression)

LinearRegression (Linear Regression)		ExampleSet (Set Role (3))		Performan
Result History		AttributeWeights (Linear Regression)		
attribute	weight			
X1 transaction date	4.982			
X2 house age	-0.299			
X3 distance to the nearest MRT station	-0.004			
X4 number of convenience stores	1.245			
X5 latitude	208.932			
X6 longitude	0			

tribute Weights ini menunjukkan **koefisien regresi** dari setiap fitur dalam model Linear Regression. Koefisien ini menggambarkan seberapa besar pengaruh setiap fitur terhadap prediksi harga rumah. Berikut penjelasan rinci tentang setiap koefisien:

## 1. Arti Koefisien Regresi

- **Koefisien regresi** menunjukkan seberapa besar perubahan pada target variabel (**Y house price of unit area**) ketika fitur tersebut berubah sebesar 1 unit, dengan asumsi fitur lainnya tetap.
- **Koefisien positif**: Peningkatan fitur tersebut akan meningkatkan harga rumah.
- **Koefisien negatif**: Peningkatan fitur tersebut akan menurunkan harga rumah.

## 2. Penjelasan Koefisien untuk Setiap Fitur

Berikut adalah koefisien regresi untuk setiap fitur:

### a. X1 transaction date

- **Koefisien**: 4.982
- **Arti**: Setiap peningkatan 1 unit pada tanggal transaksi, harga rumah meningkat sebesar **4.982 unit**.
- **Interpretasi**: Tanggal transaksi yang lebih baru (misalnya, tahun 2013 vs 2012) cenderung meningkatkan harga rumah.

### b. X2 house age

- **Koefisien**: -0.299
- **Arti**: Setiap peningkatan 1 tahun usia rumah, harga rumah **menurun** sebesar **0.299 unit**.
- **Interpretasi**: Rumah yang lebih tua cenderung memiliki harga yang lebih rendah.

### c. X3 distance to the nearest MRT station

- **Koefisien**: -0.004
- **Arti**: Setiap peningkatan 1 meter jarak ke stasiun MRT, harga rumah **menurun** sebesar **0.004 unit**.
- **Interpretasi**: Rumah yang lebih dekat ke stasiun MRT cenderung memiliki harga yang lebih tinggi.

### d. X4 number of convenience stores

- **Koefisien**: 1.245
- **Arti**: Setiap penambahan 1 toko di sekitar rumah, harga rumah meningkat sebesar **1.245 unit**.
- **Interpretasi**: Semakin banyak toko di sekitar rumah, semakin tinggi harga rumah.

### e. X5 latitude

- **Koefisien**: 208.932
- **Arti**: Setiap peningkatan 1 unit pada latitude, harga rumah meningkat sebesar **208.932 unit**.

- **Interpretasi:** Lokasi rumah dengan latitude yang lebih tinggi (misalnya, lebih utara) cenderung memiliki harga yang lebih tinggi.

#### f. X6 longitude

- **Koefisien:** 0
- **Arti:** Koefisien 0 berarti longitude tidak memiliki pengaruh terhadap harga rumah.
- **Interpretasi:** Lokasi rumah berdasarkan longitude tidak signifikan dalam memprediksi harga rumah.

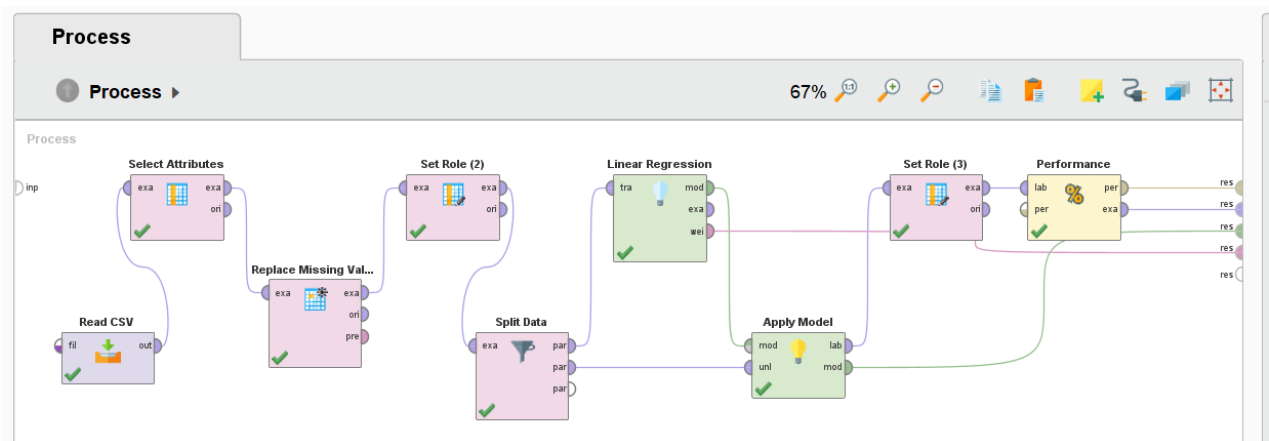
### 3. Analisis Koefisien

- **Fitur Paling Berpengaruh:**
  - **X5 latitude** memiliki koefisien terbesar (**208.932**), yang berarti latitude memiliki pengaruh yang sangat besar terhadap harga rumah.
  - **X4 number of convenience stores** juga memiliki pengaruh yang signifikan (**1.245**).
- **Fitur yang Mengurangi Harga Rumah:**
  - **X2 house age** dan **X3 distance to the nearest MRT station** memiliki koefisien negatif, yang berarti fitur ini cenderung menurunkan harga rumah.
- **Fitur yang Tidak Berpengaruh:**
  - **X6 longitude** memiliki koefisien 0, yang berarti fitur ini tidak berpengaruh terhadap harga rumah.

### 4. Kesimpulan

1. Model Linear Regression Saya telah mengidentifikasi fitur-fitur yang signifikan dalam memprediksi harga rumah.
2. **Latitude (X5)** dan **jumlah toko di sekitar rumah (X4)** adalah fitur yang paling berpengaruh dalam meningkatkan harga rumah.
3. **Usia rumah (X2)** dan **jarak ke stasiun MRT (X3)** cenderung menurunkan harga rumah.
4. **Longitude (X6)** tidak berpengaruh terhadap harga rumah.

### c. Berikan penjelasan model yang saudara dapatkan.



#### 1. Jenis Model

Model yang Saya dapatkan adalah **Linear Regression**. Ini adalah model statistik yang digunakan untuk memprediksi nilai numerik (dalam kasus ini, harga rumah) berdasarkan hubungan linear antara variabel independen (fitur) dan variabel dependen (target).

#### 2. Persamaan Model

Model Linear Regression menghasilkan persamaan berikut:

*Harga Rumah*

$$= 4.982 \times X1 - 0.299 \times X2 - 0.004 \times X3 + 1.245 \times X4 + 208.932 \times X5 - 15204.872$$

- **X1 (transaction date):** Tanggal transaksi rumah.
- **X2 (house age):** Usia rumah.
- **X3 (distance to MRT station):** Jarak ke stasiun MRT terdekat.
- **X4 (number of convenience stores):** Jumlah toko di sekitar rumah.
- **X5 (latitude):** Koordinat lintang (latitude) lokasi rumah.
- **Intercept (-15204.872):** Nilai dasar harga rumah ketika semua fitur bernilai 0.

#### 3. Arti Koefisien

- **Koefisien positif:** Menunjukkan bahwa peningkatan fitur tersebut akan meningkatkan harga rumah.
- **Koefisien negatif:** Menunjukkan bahwa peningkatan fitur tersebut akan menurunkan harga rumah.

Berikut penjelasan koefisien untuk setiap fitur:

##### 1. X1 (transaction date):

- Koefisien = **4.982**

- Arti: Setiap peningkatan 1 unit pada tanggal transaksi, harga rumah meningkat sebesar **4.982 unit**.
  - Contoh: Rumah yang dijual pada tahun 2013 cenderung lebih mahal daripada rumah yang dijual pada tahun 2012.
2. **X2 (house age):**
- Koefisien = **-0.299**
  - Arti: Setiap peningkatan 1 tahun usia rumah, harga rumah **menurun** sebesar **0.299 unit**.
  - Contoh: Rumah yang berusia 10 tahun cenderung lebih murah daripada rumah yang berusia 5 tahun.
3. **X3 (distance to MRT station):**
- Koefisien = **-0.004**
  - Arti: Setiap peningkatan 1 meter jarak ke stasiun MRT, harga rumah **menurun** sebesar **0.004 unit**.
  - Contoh: Rumah yang berjarak 500 meter dari stasiun MRT cenderung lebih mahal daripada rumah yang berjarak 1000 meter.
4. **X4 (number of convenience stores):**
- Koefisien = **1.245**
  - Arti: Setiap penambahan 1 toko di sekitar rumah, harga rumah meningkat sebesar **1.245 unit**.
  - Contoh: Rumah dengan 5 toko di sekitarnya cenderung lebih mahal daripada rumah dengan 2 toko.
5. **X5 (latitude):**
- Koefisien = **208.932**
  - Arti: Setiap peningkatan 1 unit pada latitude, harga rumah meningkat sebesar **208.932 unit**.
  - Contoh: Rumah yang terletak di latitude yang lebih tinggi (misalnya, lebih utara) cenderung lebih mahal.
6. **Intercept (-15204.872):**
- Ini adalah nilai dasar harga rumah ketika semua fitur bernilai 0.
  - Arti: Jika semua fitur bernilai 0, harga rumah diperkirakan **-15204.872 unit**. Nilai negatif ini tidak memiliki arti praktis, tetapi diperlukan untuk melengkapi persamaan regresi.

#### 4. Performa Model

Model ini memiliki beberapa metrik evaluasi yang menunjukkan seberapa baik model tersebut bekerja:

- **RMSE (Root Mean Squared Error): 8.290**
  - Arti: Rata-rata kesalahan prediksi adalah **8.290 unit** dari harga rumah aktual.
  - Jika rata-rata harga rumah adalah **38.661**, maka kesalahan prediksi sekitar **21.4%** dari rata-rata harga.
  - Ini dianggap **cukup baik**, tetapi masih bisa ditingkatkan.
- **R-squared ( $R^2$ ): 0.606**
  - Arti: Model menjelaskan **60.6%** variasi data.
  - Ini menunjukkan bahwa model cukup baik dalam memprediksi harga rumah, tetapi masih ada **39.4%** variasi data yang tidak dijelaskan oleh model.
- **Korelasi: 0.778**
  - Arti: Ada hubungan yang **kuat** antara prediksi dan nilai aktual.
  - Semakin mendekati 1, semakin baik.

## 5. Kelebihan Model

- Model ini **sederhana** dan mudah diinterpretasikan.
- Model mampu menjelaskan **60.6%** variasi data ( $R\text{-squared} = 0.606$ ).
- Ada hubungan yang kuat antara prediksi dan nilai aktual ( $korelasi = 0.778$ ).

## 6. Kekurangan Model

- **RMSE = 8.290** masih relatif besar (sekitar **21.4%** dari rata-rata harga rumah).
- Model tidak menjelaskan **39.4%** variasi data ( $R\text{-squared} = 0.606$ ).
- Beberapa fitur (seperti longitude) tidak berpengaruh terhadap harga rumah.

## 7. Rekomendasi untuk Meningkatkan Model

- **Feature Engineering:** Tambahkan fitur baru yang mungkin berpengaruh terhadap harga rumah (misalnya, luas tanah, jumlah kamar, dll.).
- **Coba Algoritma Lain:** Jika Linear Regression tidak memberikan hasil yang memuaskan, coba algoritma lain seperti **Decision Tree Regression**, **Random Forest Regression**, atau **Gradient Boosting**.
- **Hyperparameter Tuning:** Jika menggunakan algoritma lain, lakukan tuning hyperparameter untuk meningkatkan performa model.

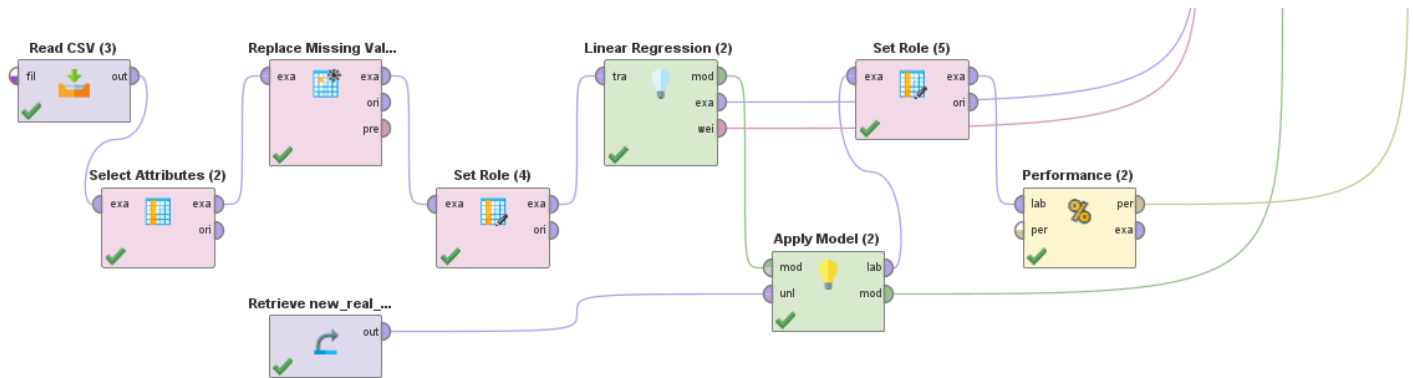
## Kesimpulan

Model Linear Regression ini dapatkan sudah cukup baik dalam memprediksi harga rumah, tetapi masih ada ruang untuk perbaikan. Mungkin dengan beberapa langkah peningkatan, seperti menambahkan fitur baru atau mencoba algoritma lain, kita bisa mendapatkan model yang lebih akurat.



**d. Cobalah model tersebut untuk mengestimasi harga rumah, silahkan saudara gunakan data yang saudara kembangkan sendiri**

Dan berikut modelnya dengan data yang saya kembangkan sendiri:



Berikut hasilnya ketika saya run:

AttributeWeights (Linear Regression (2))		
Result History		
<div> <div>Data</div> <div>Weight Visualizations</div> <div>Plot view</div> </div>	attribute ↑	weight
	X1 transaction date	0.011
	X2 house age	-0.249
	X3 distance to the nearest MRT station	0
	X4 number of convenience stores	2.135
	X5 latitude	-0.000
	X6 longitude	0

AttributeWeights (Linear Regression (2))			ExampleSet (Set Role (5))			PerformanceVector (Performance (2))		
Result History			ExampleSet (Apply Model (2))					
Open in			<a href="#">Turbo Prep</a> <a href="#">Auto Model</a> <a href="#">Interactive Analysis</a>			Filter (30 / 30 examples): <a href="#">all</a>		
Row No.	Y house price of unit area	prediction(Y house price of unit area)	X1 transacti...	X2 house age	X3 distance ...	X4 number ...	X5 latitude	X6 longitude
1	25.500	31.431	2013417	17.400	9957554	0	2496305	12154915
2	45.900	43.153	2013417	13.100	5619845	5	2498746	12154391
3	46.100	34.555	2012667	15.600	2893248	5	2498203	12154348
4	21.400	35.419	2013083	12.800	1449722	3	2497289	12151728
5	44	49.787	2013250	22.200	3795575	10	2498343	12153762
6	43.500	58.518	2013083	0	2740144	1	249748	12153059
7	31.100	34.732	2013333	17.600	1805665	2	2498672	12152091
8	20.900	39.423	2013583	18.100	1783.180	3	2496731	12151486
9	39.700	26.693	2013250	37.800	5909292	1	2497153	12153559
10	38.500	32.607	2013333	9	1402016	0	2498569	1215276
11	12.800	27.241	2013000	16.500	4082015	0	2494155	12150381
12	40.200	39.433	2012917	32.400	2650609	8	2498059	12153986

ExampleSet (30 examples,2 special attributes,6 regular attributes)

# root\_mean\_squared\_error

root\_mean\_squared\_error: 14.030 +/- 0.000

## 2. Modul Hal 70 Studi Kasus 1b: Estimasi (sesuai program studi)

Kembangkan sebuah kasus model estimasi, dengan menggunakan data publik sesuai dengan program studi saudara.

- a. Deskripsikan kasus yang saudara kembangkan, tuliskan sumber datanya, berikan penjelasan data yang saudara peroleh tersebut.
- b. Jelaskan langkah-langkah saudara dalam mengembangkan model untuk menyelesaikan kasus tersebut.
- c. Tuliskan rekomendasi saudara setelah model saudara dapatkan.

### a. Deskripsikan kasus yang saudara kembangkan, tuliskan sumber datanya, berikan penjelasan data yang saudara peroleh tersebut.

#### 1. Kasus yang Dikembangkan

Saya ingin mengembangkan sebuah model **estimasi harga berlian** berdasarkan berbagai karakteristik fisik dan kualitas berlian tersebut. Model ini akan membantu memprediksi harga berlian (dalam satuan dolar) berdasarkan fitur-fitur seperti berat (carat), potongan (cut), warna (color), kejernihan (clarity), dan dimensi fisik (panjang, lebar, kedalaman). Prediksi ini dapat digunakan oleh pedagang berlian, pembeli, atau ahli gemologi untuk memperkirakan harga berlian berdasarkan karakteristiknya.

#### 2. Sumber Data

Dataset ini adalah dataset publik yang tersedia di **Kaggle** dengan judul "**Diamonds**". Dataset ini berisi informasi tentang berbagai berlian, termasuk karakteristik fisik dan harganya. Kita dapat mengunduhnya langsung dari link berikut:

- [Diamonds Dataset](https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction?select=diamonds.csv) (<https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction?select=diamonds.csv>)

#### 3. Penjelasan Data

Dataset ini berisi informasi tentang 53.940 berlian dengan 10 kolom yang mendeskripsikan berbagai karakteristik berlian. Berikut adalah kolom-kolom yang ada dalam dataset:

##### 1. carat:

- Berat berlian dalam satuan carat.
- Nilainya berkisar antara **0,2** hingga **5,01**.
- Semakin besar carat, semakin berat dan biasanya semakin mahal berlian tersebut.

##### 2. cut:

- Kualitas potongan berlian.
- Nilainya adalah kategorikal dengan urutan kualitas: **Fair, Good, Very Good, Premium, Ideal**.
- Potongan yang baik memengaruhi kilau dan harga berlian.

### 3. color:

- Warna berlian, dari yang paling tidak berwarna (terbaik) hingga yang paling berwarna.
- Nilainya adalah kategorikal dengan urutan: **D** (terbaik), **E, F, G, H, I, J** (paling berwarna).
- Warna yang lebih jernih (D) biasanya lebih mahal.

### 4. clarity:

- Kejernihan berlian, yang mengukur seberapa bersih berlian dari inklusi (cacat internal) atau blemishes (cacat permukaan).
- Nilainya adalah kategorikal dengan urutan: **I1** (paling tidak jernih), **SI2, SI1, VS2, VS1, VVS2, VVS1, IF** (paling jernih).
- Kejernihan yang lebih tinggi meningkatkan harga berlian.

### 5. depth:

- Persentase kedalaman berlian, yang dihitung dari rumus:

$$\text{depth} = \frac{2 \times z}{x + y} \times 100$$

di mana  $x$ ,  $y$ , dan  $z$  adalah dimensi fisik berlian.

- Nilainya berkisar antara **43** hingga **79**.
- Kedalaman yang optimal memengaruhi kilau berlian.

### 6. table:

- Lebar permukaan atas berlian relatif terhadap titik terlebarnya.
- Nilainya berkisar antara **43** hingga **95**.
- Table yang terlalu lebar atau terlalu sempit dapat memengaruhi harga.

### 7. price:

- Harga berlian dalam satuan dolar.
- Ini adalah **target variabel** yang ingin Saya prediksi.
- Nilainya berkisar antara **326** hingga **18.823** dolar.

### 8. x:

- Panjang berlian dalam milimeter.
- Nilainya berkisar antara **0** hingga **10,74**.

### 9. y:

- Lebar berlian dalam milimeter.
- Nilainya berkisar antara **0** hingga **58,9**.

10. **z:**

- Kedalaman berlian dalam milimeter.
- Nilainya berkisar antara **0** hingga **31,8**.

#### 4. Ringkasan Dataset

- **Jumlah Baris:** 53.940 (setiap baris mewakili satu berlian).
- **Jumlah Kolom:** 10 kolom (9 fitur + 1 target).
- **Target Variabel:** price (harga berlian).
- **Fitur:** carat, cut, color, clarity, depth, table, x, y, z.

#### 5. Tujuan Analisis

Saya akan menggunakan model **Neural Network** untuk memprediksi price (harga berlian) berdasarkan fitur-fitur seperti carat, cut, color, clarity, depth, table, dan dimensi fisik (x, y, z). Model ini akan membantu memahami hubungan antara karakteristik berlian dengan harganya.

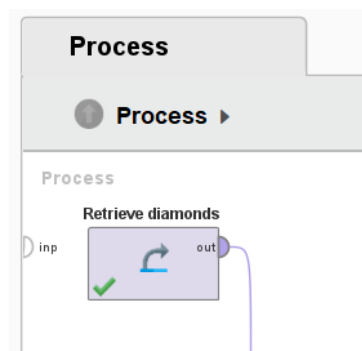
#### 6. Catatan untuk Laporan

- **Judul:** "Estimasi Harga Berlian Menggunakan Model Neural Network".
- **Tujuan:** Membangun model prediktif untuk memperkirakan harga berlian berdasarkan karakteristik fisik dan kualitasnya.
- **Dataset:** Diamonds Dataset dari Kaggle.
- **Metode:** Neural Network.
- **Target Variabel:** price (harga berlian).
- **Fitur:** carat, cut, color, clarity, depth, table, x, y, z.

## b. Jelaskan langkah-langkah saudara dalam mengembangkan model untuk menyelesaikan kasus tersebut.

### 1. Mengambil Data (Retrieve Data)

- **Operator:** Retrieve
- **Tujuan:** Mengimpor dataset **Diamonds** dari Kaggle ke dalam RapidMiner.
- **Penjelasan:** Dataset ini berisi informasi tentang 53.940 berlian dengan 10 kolom, termasuk fitur-fitur seperti carat, cut, color, clarity, depth, table, x, y, z, dan target variabel price.
- **Output:** Dataset siap diproses.



Result History

ExampleSet (//Local Repository/data/diamonds)

Open in: Turbo Prep, Auto Model, Interactive Analysis

	att1	carat	cut	color	clarity	depth	table	price	x	y	z
...	1	0.230	Ideal	E	SI2	61.500	55	326	3.950	3.980	2.430
...	2	0.210	Premium	E	SI1	59.800	61	326	3.890	3.840	2.310
...	3	0.230	Good	E	VS1	56.900	65	327	4.050	4.070	2.310
...	4	0.290	Premium	I	VS2	62.400	58	334	4.200	4.230	2.630
...	5	0.310	Good	J	SI2	63.300	58	335	4.340	4.350	2.750
...	6	0.240	Very Good	J	VVS2	62.800	57	336	3.940	3.960	2.480
...	7	0.240	Very Good	I	VVS1	62.300	57	336	3.950	3.980	2.470
...	8	0.260	Very Good	H	SI1	61.900	55	337	4.070	4.110	2.530
...	9	0.220	Fair	E	VS2	65.100	61	337	3.870	3.780	2.490
...	10	0.230	Very Good	H	VS1	59.400	61	338	4	4.050	2.390
...	11	0.300	Good	J	SI1	64	55	339	4.250	4.280	2.730
...	12	0.230	Ideal	J	VS1	62.800	56	340	3.930	3.900	2.460
...	13	0.220	Premium	F	SI1	60.400	61	342	3.880	3.840	2.330

ExampleSet (53,940 examples,0 special attributes,11 regular attributes)

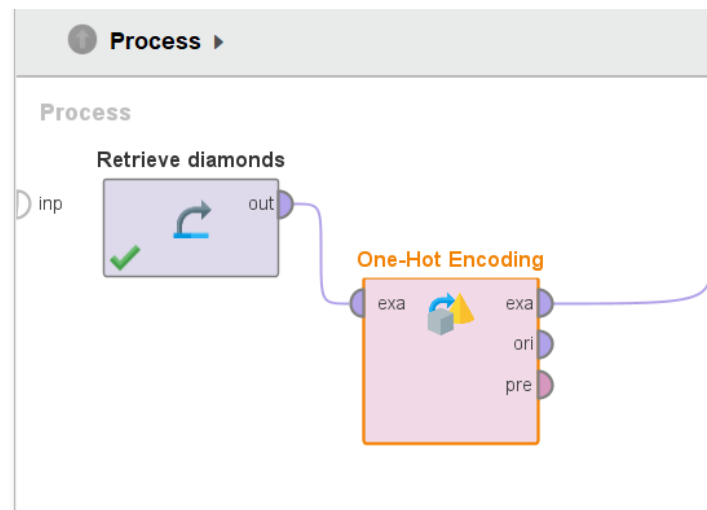
Result History						
ExampleSet (//Local Repository/data/diamonds)						
	Name	Type	Missing	Statistics		Filter (11 / 11 attributes): <input type="text" value="Search for Attributes"/>
Data	att1	Integer	0	Min	Max	Average
				1	53940	26970.500
Statistics	carat	Real	0	Min	Max	Average
				0.200	5.010	0.798
Visualizations	cut	Nominal	0	Least	Most	Values
				Fair (1610)	Ideal (21551)	Ideal (21551), Premium (13791), ...[3 more]
Annotations	color	Nominal	0	Least	Most	Values
				J (2808)	G (11292)	G (11292), E (9797), ...[5 more]
	clarity	Nominal	0	Least	Most	Values
				I1 (741)	SI1 (13065)	SI1 (13065), VS2 (12258), ...[6 more]
	depth	Real	0	Min	Max	Average
				43	79	61.749
	table	Real	0	Min	Max	Average
				43	95	57.457
Showing attributes 1 - 11						
Examples: 53.940 Special Attributes: 0 Regular Attributes: 11						

Statistics	table	Real	0	Min	Max	Average
				43	95	57.457
Visualizations	price	Integer	0	Min	Max	Average
				326	18823	3932.800
Annotations	x	Real	0	Min	Max	Average
				0	10.740	5.731
	y	Real	0	Min	Max	Average
				0	58.900	5.735
	z	Real	0	Min	Max	Average
				0	31.800	3.539
Showing attributes 1 - 11						
Examples: 53.940 Special Attributes: 0 Regular Attributes: 11						

## 2. Mengubah Data Kategorikal Menjadi Numerik (One-Hot Encoding)

- **Operator:** One-Hot Encoding
- **Tujuan:** Mengubah kolom kategorikal (cut, color, clarity) menjadi numerik agar bisa diproses oleh model Neural Network.
- **Penjelasan:** Kolom kategorikal seperti cut (Fair, Good, Very Good, Premium, Ideal) diubah menjadi kolom numerik biner (0 atau 1). Misalnya, kolom cut akan dipecah menjadi beberapa kolom baru seperti cut\_Fair, cut\_Good, cut\_Very Good, dst.
- **Output:** Dataset dengan semua fitur dalam bentuk numerik.



Parameters

One-Hot Encoding

attribute filter type
all

☐ invert selection

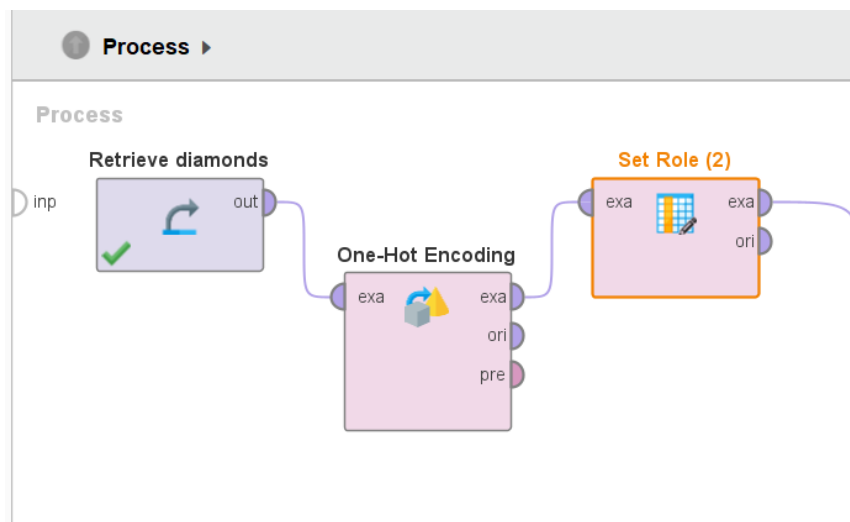
☐ include special attributes

☐ remove with too many values

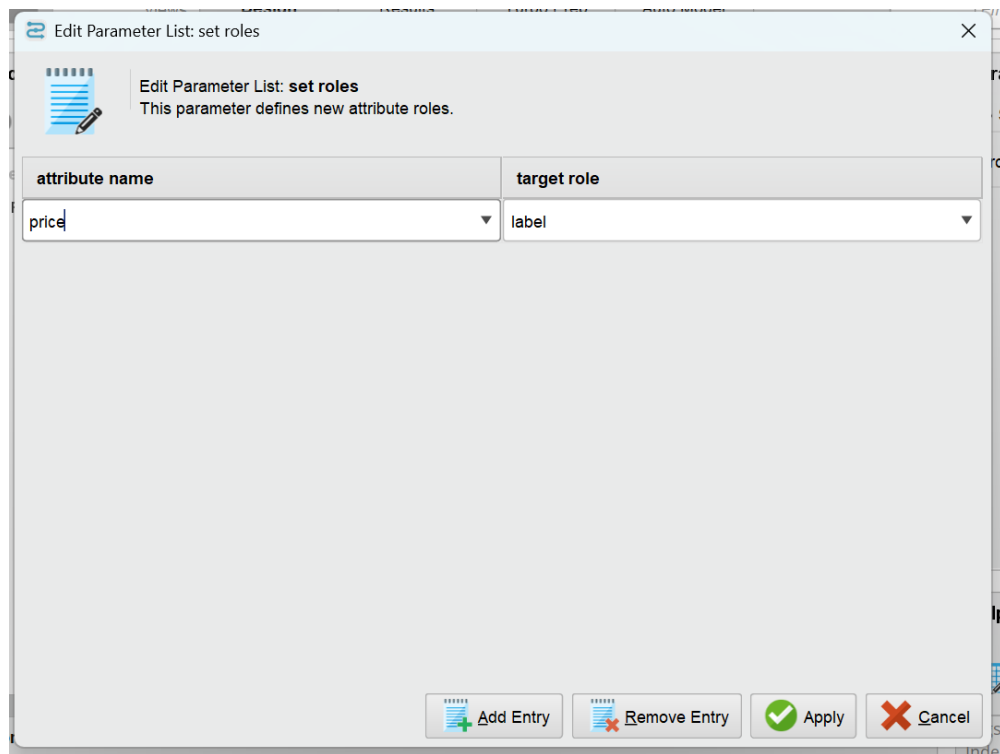
☒ perform encoding

### 3. Menetapkan Peran Kolom (Set Role)

- **Operator:** Set Role
- **Tujuan:** Menetapkan kolom price sebagai **label/target** yang akan diprediksi oleh model.
- **Penjelasan:** Kolom price diatur sebagai target, sedangkan kolom lainnya (carat, cut, color, clarity, depth, table, x, y, z) diatur sebagai fitur.
- **Output:** Dataset dengan kolom target dan fitur yang sudah ditetapkan.

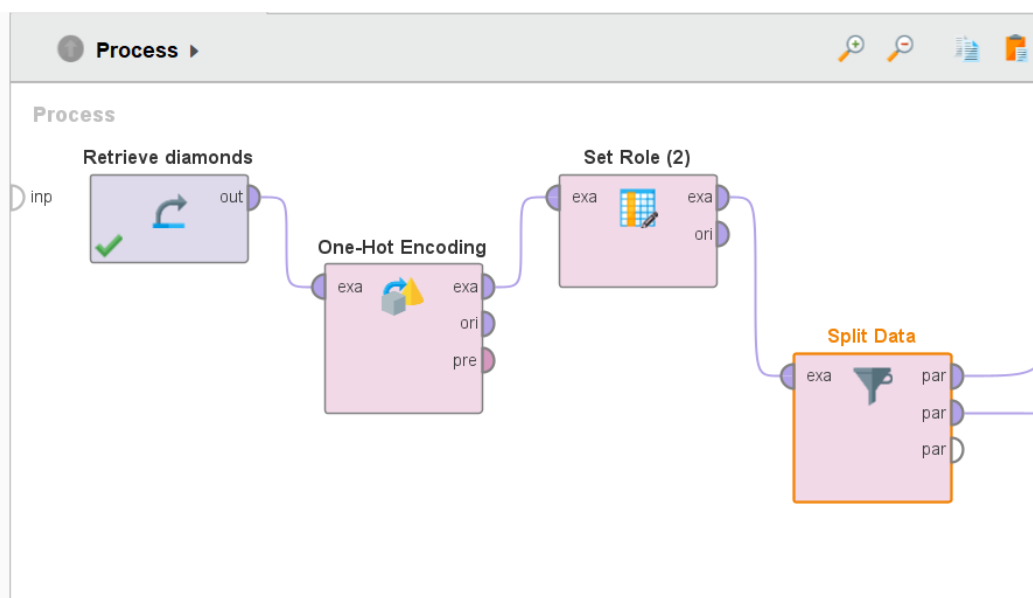


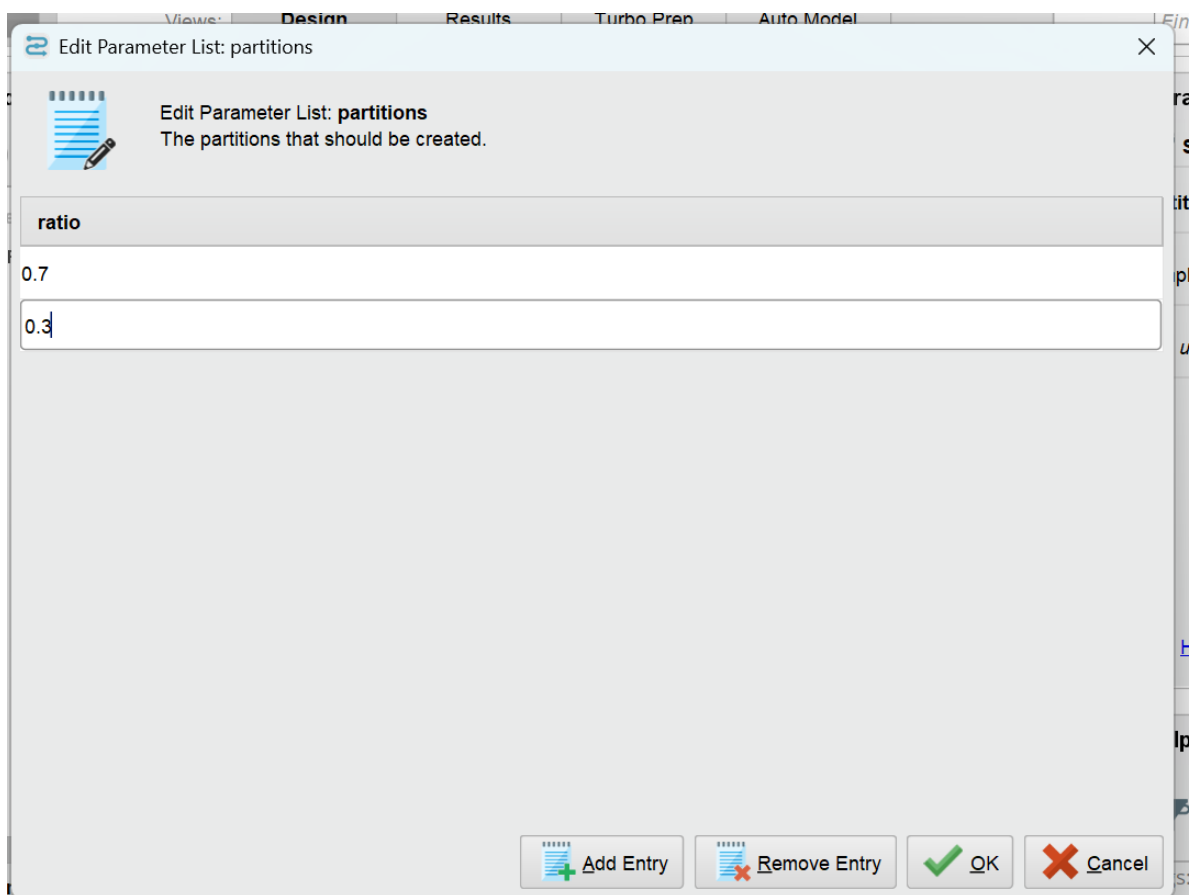
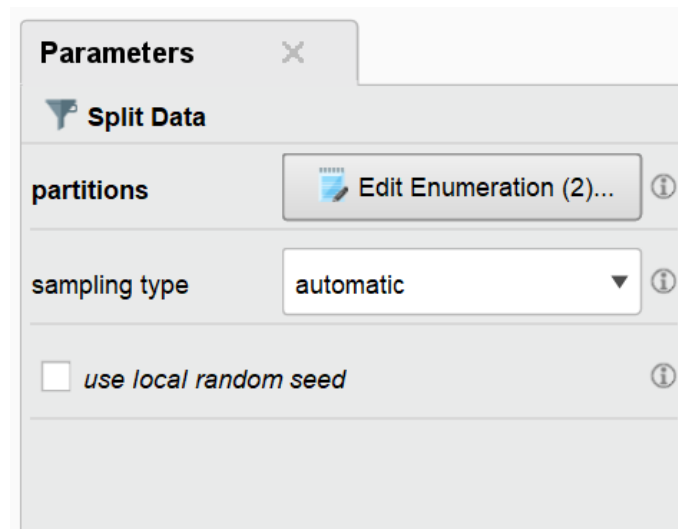




#### 4. Membagi Data (Split Data)

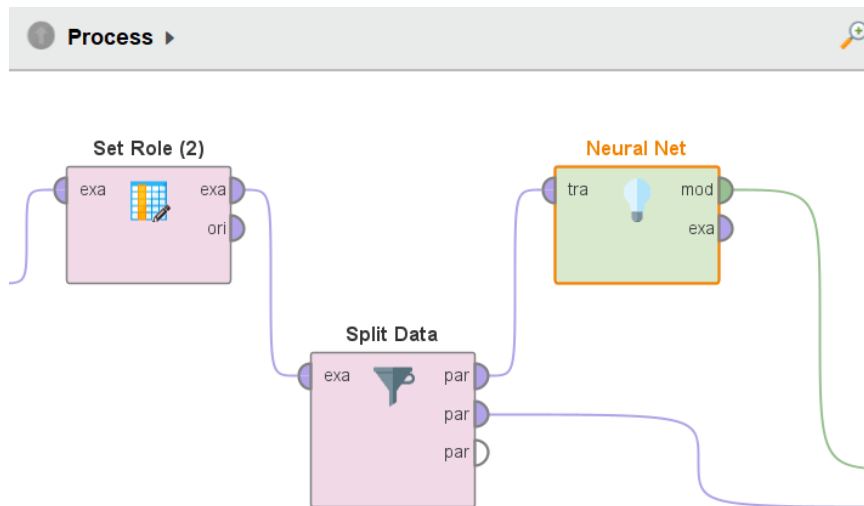
- **Operator:** Split Data
- **Tujuan:** Membagi dataset menjadi dua bagian: **data training** dan **data testing**.
- **Penjelasan:** Misalnya, 70% data digunakan untuk training model, dan 30% data digunakan untuk testing model. Pembagian ini penting untuk mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya.
- **Output:** Dua dataset terpisah, yaitu data training dan data testing.





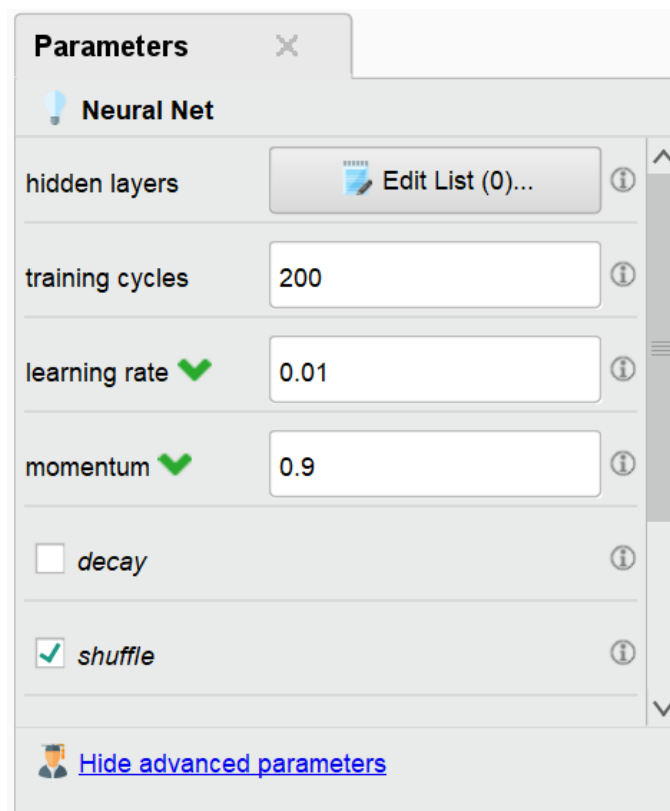
## 5. Membangun Model Neural Network (Neural Net)

- **Operator:** Neural Net
- **Tujuan:** Membangun model Neural Network untuk memprediksi harga berlian.
- **Penjelasan:** Neural Network adalah model machine learning yang meniru cara kerja otak manusia dengan menggunakan lapisan-lapisan neuron. Pada tahap ini, model akan belajar dari data training untuk menemukan pola hubungan antara fitur-fitur (carat, cut, color, dll.) dan target (price).

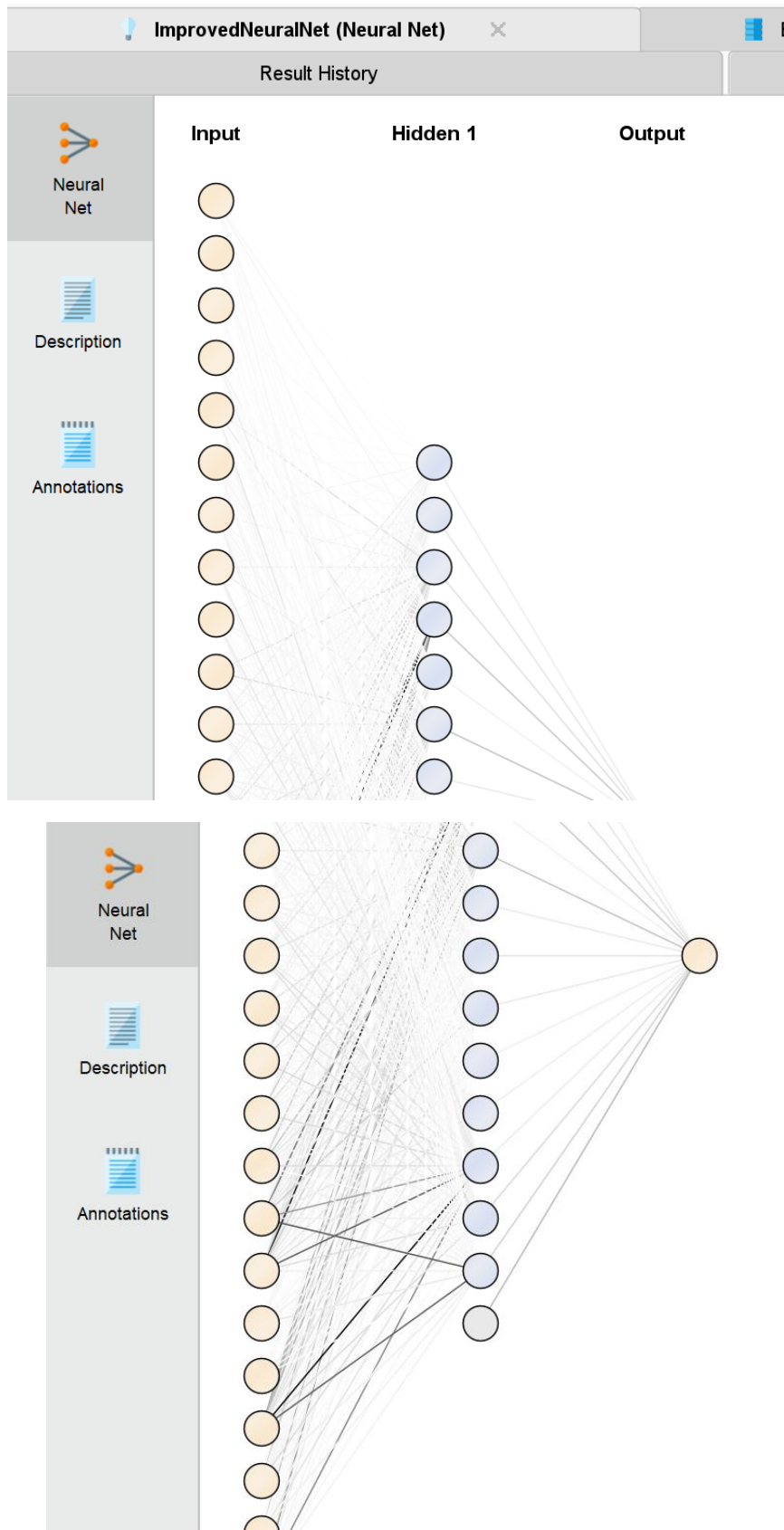


- **Parameter yang Diatur:**

- Jumlah lapisan tersembunyi (hidden layers).
- Jumlah neuron di setiap lapisan.
- Fungsi aktivasi (misalnya, ReLU, sigmoid).
- Learning rate (laju pembelajaran).



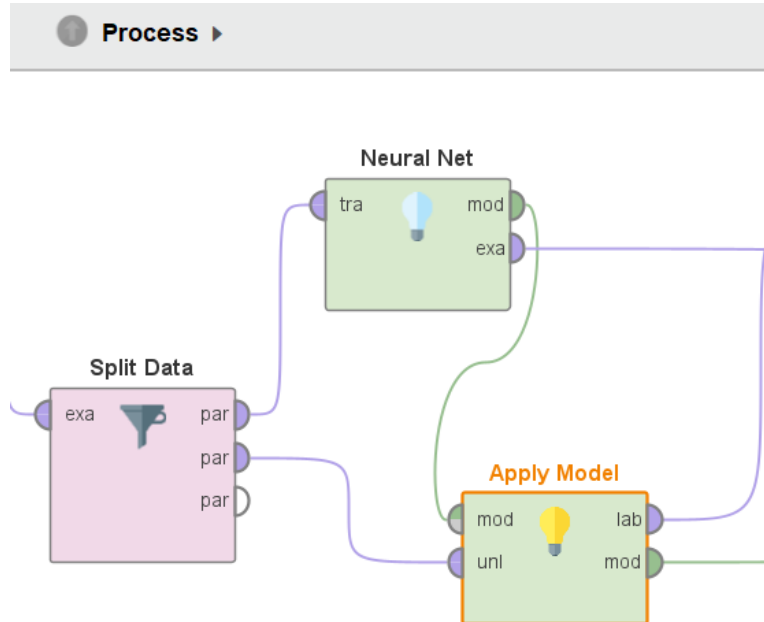
- **Output:** Model Neural Network yang sudah dilatih.



## 6. Menerapkan Model pada Data Testing (Apply Model)

- **Operator:** Apply Model

- **Tujuan:** Menerapkan model yang sudah dilatih pada data testing untuk membuat prediksi.
- **Penjelasan:** Model akan menggunakan fitur-fitur dari data testing untuk memprediksi harga berlian (price). Hasil prediksi ini akan dibandingkan dengan nilai aktual untuk mengevaluasi performa model.



- **Output:** Prediksi harga berlian untuk data testing.

Result History

ExampleSet (Apply Model)

Open in: Turbo Prep, Auto Model, Interactive Analysis

Filter (16,182 / 16,182 examples): all

Row No.	price	prediction(p...	cut = Ideal	cut = Premi...	cut = Good	cut = Very ...	color = E	color = I	color = H	color = F	colc
1	326	-9	0	1	0	0	1	0	0	0	0
2	334	831	0	1	0	0	0	1	0	0	0
3	335	40	0	0	1	0	0	0	0	0	0
4	336	754	0	0	0	1	0	1	0	0	0
5	340	-2	1	0	0	0	0	0	0	0	0
6	345	-323	0	1	0	0	1	0	0	0	0
7	348	410	1	0	0	0	0	1	0	0	0
8	353	502	0	0	0	1	0	0	1	0	0
9	353	348	0	0	0	1	0	0	0	0	0
10	354	793	0	0	0	1	0	0	0	0	1
11	357	673	0	0	0	1	0	0	0	0	0
12	357	806	0	0	0	1	0	0	0	1	0

ExampleSet (16,182 examples,2 special attributes,24 regular attributes)

ImprovedNeuralNet (Neural Net)

ExampleSet (Set Role (3))

PerformanceVector (Performance)

Result History

ExampleSet (Apply Model)

Data

Statistics

Visualizations

Annotations

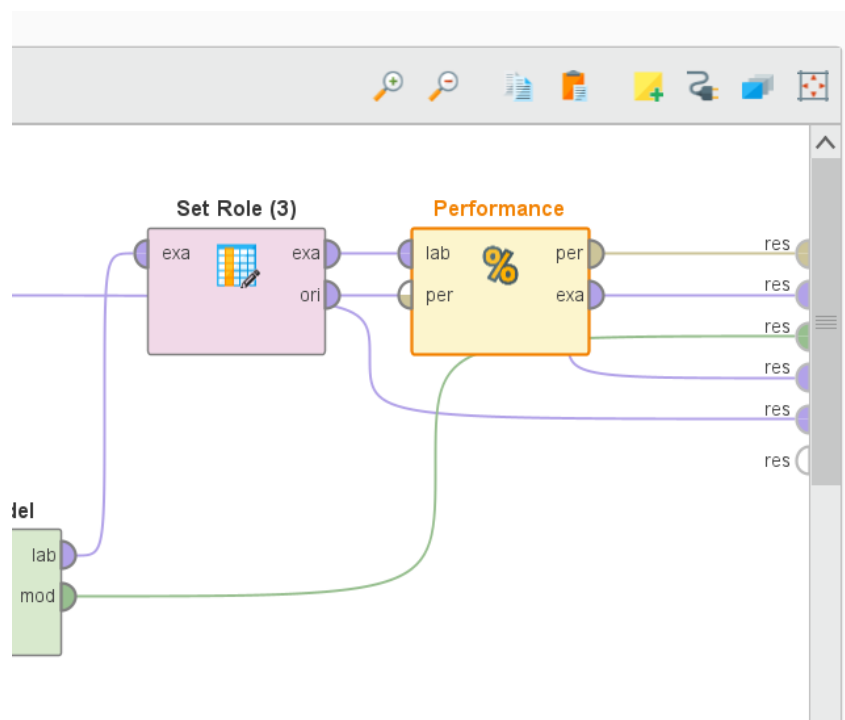
Name	Type	Missing	Statistics		Filter (26 / 26 attributes):	Search for Attributes
Label price	Integer	0	Min 326	Max 18823	Average 3950.598	
Prediction prediction(price)	Integer	0	Min -740	Max 19436	Average 4049.623	
cut = Ideal	Real	0	Min 0	Max 1	Average 0.407	
cut = Premium	Real	0	Min 0	Max 1	Average 0.258	
cut = Good	Real	0	Min 0	Max 1	Average 0.089	
cut = Very Good	Real	0	Min 0	Max 1	Average 0.218	
color = F	Real	0	Min 0	Max 1	Average 0.180	

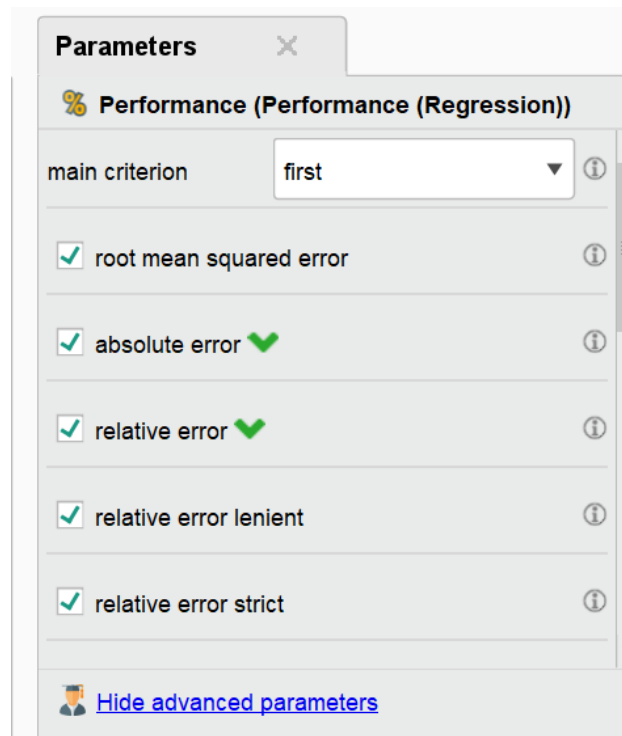
Showing attributes 1 - 26

Examples: 16,182    Special Attributes: 2    Regular Attributes: 24

## 7. Mengevaluasi Performa Model (Performance)

- **Operator:** Performance
- **Tujuan:** Mengevaluasi seberapa baik model Neural Network bekerja dengan menghitung metrik performa seperti **RMSE (Root Mean Squared Error)**.
- **Penjelasan:** RMSE mengukur rata-rata kesalahan prediksi model. Semakin kecil nilai RMSE, semakin baik performa model.





- **Output:** Nilai RMSE dan metrik evaluasi lainnya.

## root\_mean\_squared\_error

root\_mean\_squared\_error: 308.618 +/- 0.000

## PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 308.618 +/- 0.000
absolute_error: 219.369 +/- 217.076
relative_error: 15.40% +/- 22.11%
relative_error_lenient: 11.85% +/- 15.39%
relative_error_strict: 25.65% +/- 368.90%
```

### 1. Nilai RMSE (Root Mean Squared Error):

- **RMSE: 308,618**
  - **Apa Artinya?**  
RMSE mengukur rata-rata kesalahan prediksi model saya. Nilai RMSE **308,618** berarti rata-rata kesalahan prediksi harga berlian adalah sekitar **\$308,62**.
  - **Bagaimana Menilai RMSE Ini?**

- **Rata-rata Harga Berlian (price): \$3.950,60.**
- **RMSE%:**

$$\text{RMSE\%} = \left( \frac{308,618}{3.950,60} \right) \times 100 \approx 7,81\%$$

Kesalahan prediksi sebesar **7,81%** dari rata-rata harga berlian tergolong **sangat baik**. Biasanya, RMSE di bawah **10%** dari rata-rata target dianggap baik.

- **Rentang Harga Berlian: 326\*\*hingga\*\*326\*\*hingga\*\*18.823.**
- **RMSE% Rentang:**

$$\text{RMSE\% Rentang} = \left( \frac{308,618}{18.823 - 326} \right) \times 100 \approx 1,67\%$$

Kesalahan prediksi sebesar **1,67%** dari rentang harga berlian juga tergolong **sangat baik**.

## 2. Nilai Absolute Error:

- **Absolute Error: 219,369 +/- 217,076**
  - **Apa Artinya?**  
Absolute Error adalah selisih absolut antara harga prediksi dan harga aktual. Nilai rata-rata absolute error adalah **219,37\*\*,dengan deviasistayar \*\*217,08**.
  - **Bagaimana Menilai Ini?**
    - Nilai ini menunjukkan bahwa sebagian besar prediksi memiliki kesalahan sekitar **219,37\*\*,dengan variasikesalahansekitar \*\*217,08**.
    - Ini konsisten dengan RMSE yang rendah, menunjukkan bahwa model saya cukup akurat.

## 3. Nilai Relative Error:

- **Relative Error: 15,40% +/- 22,11%**
  - **Apa Artinya?**  
Relative Error mengukur kesalahan prediksi relatif terhadap harga aktual. Rata-rata relative error adalah **15,40%**, dengan deviasi stsayar **22,11%**.
  - **Bagaimana Menilai Ini?**
    - Ini berarti, rata-rata, prediksi saya meleset sekitar **15,40%** dari harga aktual.
    - Deviasi stsayar yang tinggi (**22,11%**) menunjukkan bahwa ada beberapa prediksi yang memiliki kesalahan lebih besar, tetapi secara umum, nilai ini masih tergolong baik.

## 4. Nilai Relative Error Lenient dan Strict:

- **Relative Error Lenient: 11,85% +/- 15,39%**
  - **Apa Artinya?**  
Ini adalah versi lebih "ringan" dari relative error, yang mungkin mengabaikan



beberapa outlier. Rata-rata kesalahan adalah **11,85%**, dengan deviasi stsayar **15,39%**.

- **Bagaimana Menilai Ini?**
  - Nilai ini menunjukkan bahwa sebagian besar prediksi memiliki kesalahan sekitar **11,85%**, yang tergolong sangat baik.
- **Relative Error Strict: 25,65% +/- 368,90%**
  - **Apa Artinya?**

Ini adalah versi lebih "ketat" dari relative error, yang mungkin lebih sensitif terhadap outlier. Rata-rata kesalahan adalah **25,65%**, dengan deviasi stsayar yang sangat tinggi (**368,90%**).
  - **Bagaimana Menilai Ini?**
    - Deviasi stsayar yang sangat tinggi menunjukkan adanya beberapa prediksi yang meleset jauh dari harga aktual. Namun, karena rata-rata relative error lenient dan RMSE rendah, ini mungkin hanya terjadi pada beberapa kasus tertentu.

## 5. Perbandingan Harga Aktual dan Prediksi:

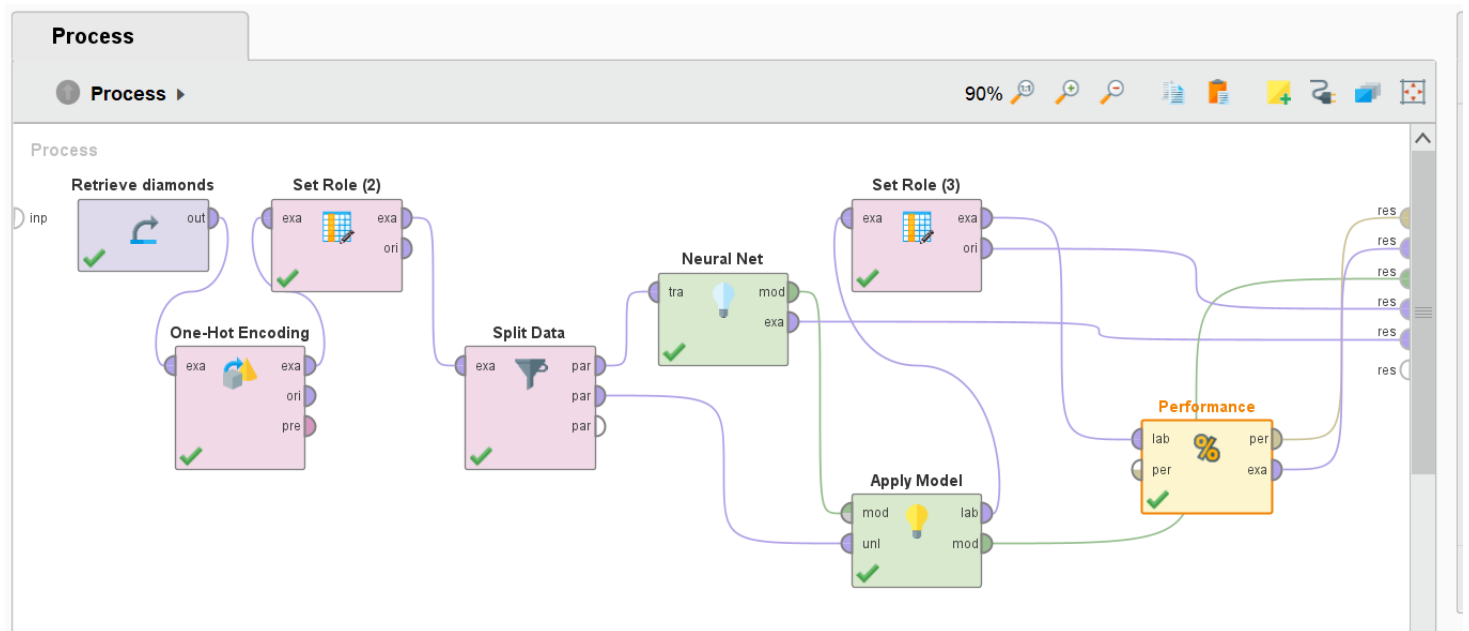
- **Harga Aktual (price):**
  - **Rata-rata: \$3.950,60**
  - **Rentang: 326\*\*hingga\*\*18.823**
- **Harga Prediksi (prediction(price)):**
  - **Rata-rata: \$4.049,62**
  - **Rentang: -740\*\*hingga\*\*19.436**
  - **Apa Artinya?**
    - Rata-rata harga prediksi (**4.049,62\*\***) *sangat dekat dengan rata-rata harga aktual (\*\*3.950,60)*, menunjukkan bahwa model saya secara umum cukup akurat.
    - Namun, ada beberapa prediksi yang meleset jauh, seperti prediksi negatif (**-740\*\***) *atau prediksi yang sangat tinggi (\*\*19.436)*. Ini mungkin disebabkan oleh outlier dalam data atau ketidakmampuan model untuk menangani kasus-kasus ekstrem.

## 6. Kesimpulan Performa Model:

- **Keunggulan Model:**
  - RMSE **308,618** dan relative error **15,40%** menunjukkan bahwa model saya secara umum cukup akurat dalam memprediksi harga berlian.
  - Rata-rata harga prediksi (**4.049,62\*\***) *sangat dekat dengan rata-rata harga aktual (\*\*3.950,60)*, menunjukkan bahwa model saya mampu menangkap pola umum dalam data.
- **Kekurangan Model:**
  - Ada beberapa prediksi yang meleset jauh, seperti prediksi negatif atau prediksi yang sangat tinggi. Ini mungkin disebabkan oleh outlier dalam data atau ketidakmampuan model untuk menangani kasus-kasus ekstrem.

- Deviasi stsayar yang tinggi pada relative error strict (**368,90%**) menunjukkan bahwa model saya mungkin masih perlu ditingkatkan untuk menangani kasus-kasus yang lebih kompleks.

## 9. Ringkasan Workflow



1. **Ambil Data:** Mengimpor dataset Diamonds.
2. **Handle Missing Values:** Menangani data yang hilang.
3. **One-Hot Encoding:** Mengubah data kategorikal menjadi numerik.
4. **Set Role:** Menetapkan kolom target dan fitur.
5. **Split Data:** Membagi data menjadi training dan testing.
6. **Neural Net:** Membangun dan melatih model Neural Network.
7. **Apply Model:** Menerapkan model pada data testing.
8. **Performance:** Mengevaluasi performa model dengan RMSE.

### **c. Tuliskan rekomendasi saudara setelah model saudara dapatkan.**

Setelah melakukan analisis dan evaluasi terhadap model Neural Network yang telah saya buat, berikut adalah beberapa rekomendasi yang dapat saya pertimbangkan untuk meningkatkan performa model dan memastikan bahwa model tersebut dapat digunakan secara efektif dalam memprediksi harga berlian:

#### **1. Handle Outlier:**

- **Apa Itu Outlier?**  
Outlier adalah data yang sangat berbeda dari data lainnya, seperti harga berlian yang sangat tinggi atau sangat rendah.
- **Mengapa Perlu Dihandle?**  
Outlier dapat memengaruhi performa model, terutama pada prediksi harga yang ekstrem (misalnya, prediksi negatif atau prediksi yang terlalu tinggi).
- **Cara Handle Outlier:**
  - Identifikasi outlier menggunakan visualisasi (seperti boxplot) atau metode statistik (seperti Z-score atau IQR).
  - Pertimbangkan untuk menghapus outlier atau mengubah nilainya dengan metode tertentu (misalnya, mengganti dengan nilai median).

#### **2. Feature Engineering:**

- **Apa Itu Feature Engineering?**  
Feature engineering adalah proses menciptakan fitur baru atau memodifikasi fitur yang sudah ada untuk meningkatkan performa model.
- **Mengapa Perlu Dilakukan?**  
Fitur yang lebih informatif dapat membantu model memahami pola dalam data dengan lebih baik.
- **Contoh Feature Engineering:**
  - Tambahkan fitur baru seperti rasio dimensi (x/y, z/depth) atau interaksi antar fitur (misalnya, carat \* cut).
  - Lakukan transformasi pada fitur numerik, seperti logaritmik atau normalisasi, untuk membuat distribusi data lebih baik.

#### **3. Tuning Hyperparameter:**

- **Apa Itu Hyperparameter?**  
Hyperparameter adalah parameter yang diatur sebelum melatih model, seperti jumlah lapisan tersembunyi, jumlah neuron, atau learning rate pada Neural Network.
- **Mengapa Perlu Dilakukan?**  
Hyperparameter yang optimal dapat meningkatkan akurasi model dan mengurangi overfitting.
- **Cara Tuning Hyperparameter:**

- Gunakan metode seperti **Grid Search** atau **Random Search** untuk mencari kombinasi hyperparameter terbaik.
- Contoh hyperparameter yang bisa di-tuning:
  - Jumlah lapisan tersembunyi (hidden layers).
  - Jumlah neuron di setiap lapisan.
  - Learning rate.
  - Fungsi aktivasi (misalnya, ReLU, sigmoid).

#### 4. Coba Model Lain:

- **Mengapa Perlu Mencoba Model Lain?**

Neural Network adalah model yang kompleks dan mungkin tidak selalu menjadi pilihan terbaik untuk dataset tertentu. Mencoba model lain dapat membantu saya menemukan model yang lebih cocok.

- **Model Alternatif yang Bisa Dicoba:**

- **Random Forest Regression:** Model ini kuat terhadap outlier dan mudah diinterpretasi.
- **Gradient Boosting Regression (XGBoost, LightGBM):** Model ini sering memberikan performa yang sangat baik pada dataset tabular.
- **Support Vector Regression (SVR):** Cocok untuk dataset dengan jumlah fitur yang tidak terlalu besar.

- **Cara Mencoba Model Lain:**

- Bandingkan performa model-model tersebut dengan Neural Network menggunakan metrik yang sama (misalnya, RMSE).

#### 5. Validasi Silang (Cross-Validation):

- **Apa Itu Cross-Validation?**

Cross-validation adalah teknik untuk mengevaluasi performa model dengan membagi data menjadi beberapa subset dan melatih model pada subset yang berbeda.

- **Mengapa Perlu Dilakukan?**

Cross-validation membantu memastikan bahwa performa model konsisten dan tidak overfitting pada data training.

- **Cara Melakukan Cross-Validation:**

- Gunakan operator **Cross-Validation** di RapidMiner.
- Bagi data menjadi 5 atau 10 subset (fold) dan evaluasi performa model pada setiap fold.

#### 6. Analisis Residual:

- **Apa Itu Residual?**

Residual adalah selisih antara harga prediksi dan harga aktual.

- **Mengapa Perlu Dilakukan?**

Analisis residual dapat membantu saya memahami di mana model melakukan kesalahan dan mengidentifikasi pola kesalahan yang sistematis.

- **Cara Melakukan Analisis Residual:**

- Plot residual terhadap harga aktual atau fitur-fitur tertentu.

- Identifikasi pola tertentu (misalnya, model cenderung salah prediksi pada harga yang sangat tinggi atau sangat rendah).

## **7. Deployment dan Monitoring:**

- **Apa Itu Deployment?**

Deployment adalah proses mengimplementasikan model ke dalam sistem atau aplikasi yang dapat digunakan oleh pengguna.

- **Mengapa Perlu Dilakukan?**

Setelah model dianggap cukup baik, model dapat digunakan untuk memprediksi harga berlian secara real-time.

- **Cara Deployment:**

- Ekspor model yang sudah dilatih dari RapidMiner.
- Integrasikan model dengan aplikasi atau sistem yang ada (misalnya, menggunakan API).

- **Monitoring:**

- Pantau performa model secara berkala setelah deployment.
- Jika performa menurun (misalnya, karena perubahan pola data), lakukan retraining model.