PRAKTIKUM

PENERAPAN METODOLOGI PENGEMBANGAN DATA MINING CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

Studi Kasus: *Heating Oil Consumption*

Dosen Pengampu : **Fitri Ayuning Tyas, M.Kom**.

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS SAINS, TEKNOLOGI, DAN KESEHATAN

UNIVERSITAS MUHAMMADIYAH BREBES

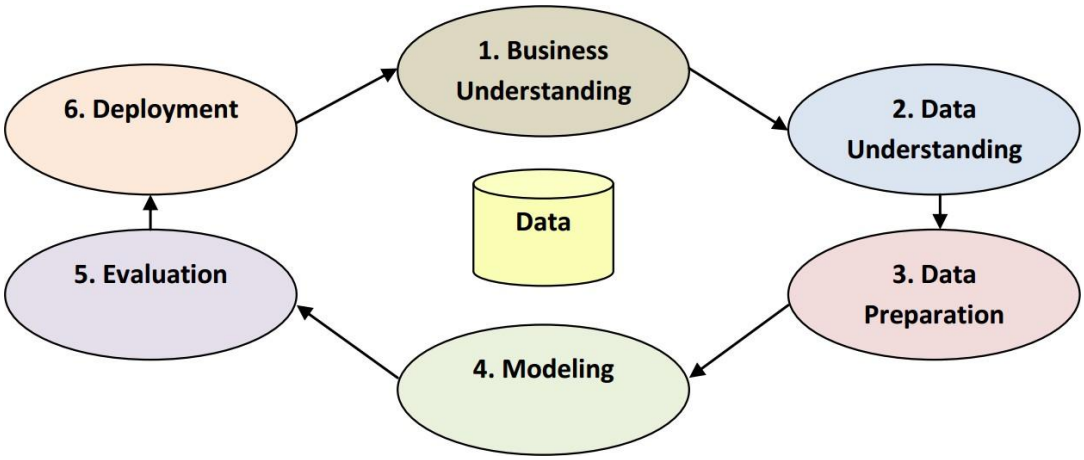
2024

A. Tujuan Praktikum

Mahasiswa mampu melakukan analisis permasalahan *data science* menggunakan tahapan CRISP-DM.

B. Pendahuluan

CRISP-DM adalah salah satu standard proses yang mampu mendukung penggunaan *data mining* untuk menyelesaikan masalah bisnis. Metode ini dirancang untuk membimbing organisasi melalui tahap-tahap yang terlibat dalam proyek *data mining*, mulai dari *business understanding* hingga implementasi solusi. CRISP-DM terdiri dari 6 tahapan, yaitu:



1. ***Business Understanding****:* Memahami pengetahuan bisnis, objek bisnis, dan tujuan pemodelan untuk tujuan bisnis.
2. ***Data Understanding****:* Menentukan jenis data yang diperlukan, sumber data, dan data yang harus diabaikan. Selama tahap ini, tim proyek mengumpulkan data yang diperlukan, mengeksplorasi karakteristiknya, dan memahami kualitasnya.
3. ***Data Preparation****:* Data yang telah dikumpulkan diolah dan disiapkan untuk analisis. Hal ini mencakup pembersihan data, penggabungan data dari berbagai sumber, dan transformasi data.
4. ***Modeling****:* Di tahap ini, model statistik atau model machine learning dikembangkan dan diuji. Berbagai teknik model dapat digunakan tergantung pada tujuan proyek.
5. ***Evaluation****:* Mengevaluasi model yang dihasilkan pada tahap pemodelan untuk mengetahui kualitas dan efektivitasnya sebelum menerapkannya di lapangan serta menentukan apakah model tersebut benar-benar mencapai tujuan bisnis.
6. ***Deployment****:* Menerapkan model yang dihasilkan untuk membantu dalam pengambilan keputusan.

C. Studi Kasus I

Lakukan studi kasus *Heating Oil Consumption - Correlational Methods*

(Matthew North, Data Mining for the Masses 2nd Edition , 2016 Chapter 4 Correlation al Methods , pp. 69 76) dengan *dataset*: *HeatingOil.csv*

1. ***Business Understanding***

a. *Problem*

Sarah adalah seorang manajer penjualan regional untuk pemasok bahan bakar fosil nasional, menghadapi penurunan kinerja pemasarannya sambil mengalami peningkatan pengeluaran pemasaran. Dia merasa perlu memahami faktor-faktor dan perilaku yang dapat memengaruhi permintaan minyak pemanas di pasar domestik. Sarah menyadari kompleksitas faktor yang mempengaruhi konsumsi minyak pemanas dan ingin menyelidiki hubungan di antara mereka. Tujuannya adalah memantau dan merespons permintaan minyak pemanas dengan lebih efektif serta merancang strategi pemasaran yang lebih baik di masa depan.

b. *Objective*

Menyelidiki hubungan antara sejumlah faktor yang mempengaruhi konsumsi minyak pemanas.

2. ***Data Understanding***

Sarah meminta bantuan kami dalam membuat matriks korelasi untuk enam atribut yang relevan. Sumber data yang digunakan sebagian besar berasal dari basis data penagihan perusahaan. Kami telah menyusun kumpulan data yang mencakup atribut-atribut berikut untuk memenuhi permintaannya.

Insulation	Peringkat kepadatan yang menunjukkan ketebalan insulasi di setiap rumah. Range dimulai dari 1-10. Rumah dengan insulasi yang buruk memiliki peringkat kepadatan satu, sedangkan rumah dengan insulasi yang sangat baik memiliki peringkat kepadatan sepuluh.

Temperature	Suhu lingkungan luar ruangan rata-rata di setiap rumah untuk tahun terakhir, diukur dalam derajat Fahrenheit.
Heating_oil	Jumlah total unit minyak pemanas yang dibeli oleh pemilik setiap rumah pada tahun terakhir.

Num_occupants	Jumlah total penghuni yang tinggal di setiap rumah.
Avg_age	Usia rata-rata penghuni rumah tersebut.
Home_size	Peringkat dalam skala 1-8, dari ukuran keseluruhan rumah. Semakin tinggi angkanya, semakin besar rumah tersebut.

3. ***Data Preparation***

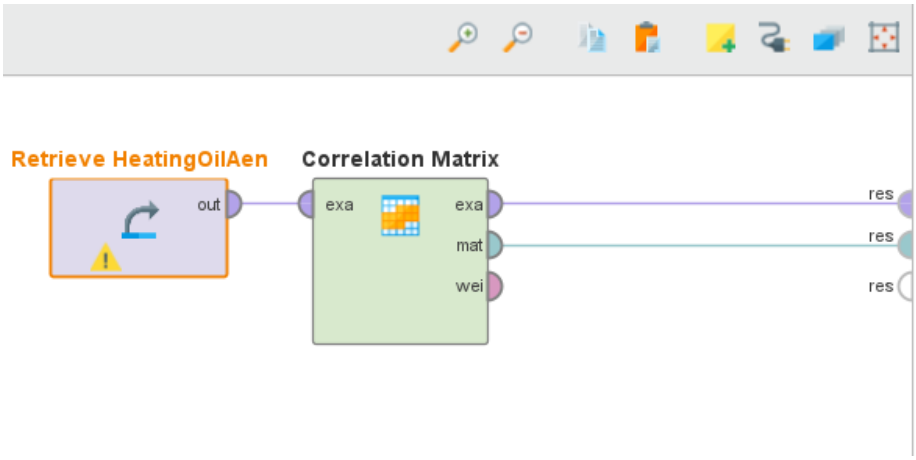
Dataset yang digunakan adalah *HeatingOil*.csv. Pastikan bahwa data ini telah bersih dari *noise* dan dapat digunakan untuk analisis data.

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

ExampleSet (1,218 examples,0 special attributes,6 regular attributes)

Name	Type	Missing	Statistics	Filter (6 / 6 attributes):	Search for Attributes
Insulation	Integer	0	Min 2	Max 10	Average 6.214
Temperature	Integer	0	Min 38	Max 90	Average 65.079
Heating_Oil	Integer	0	Min 114	Max 301	Average 197.394
Num_Occupants	Integer	0	Min 1	Max 10	Average 3.113
Avg_Age	Real	0	Min 15.100	Max 72.200	Average 42.706
Home_Size	Integer	0	Min 1	Max 8	Average 4.649

4. ***Modeling***



Operator *Correlation Matrix* digunakan untuk mengetahui hubungan antaratribut dalam dataset. Hasil dari proses di atas adalah tabel korelasi sebagai berikut:

Correlation Matrix (Correlation Matrix) | ExampleSet (Retrieve

Attributes	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperat...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Si...	0.201	-0.214	0.381	-0.023	0.307	1

5. ***Evaluation***

Tabel korelasi adalah tabel yang menunjukkan kekuatan hubungan antara dua atau lebih variabel. Nilai korelasi berkisar antara -1 hingga 1. Nilai 0 menunjukkan tidak ada hubungan antara dua variabel. Nilai positif menunjukkan hubungan positif antara dua variabel, sedangkan nilai negatif menunjukkan hubungan negatif antara dua variabel. Semakin tinggi nilainya (semakin tebal warna ungu), semakin tinggi tingkat korelasinya.

Berdasarkan tabel korelasi tersebut, dapat disimpulkan bahwa:

a. Atribut (faktor) yang paling signifikan berpengaruh (hubungan positif) pada konsumsi minyak pemanas (*Heating_Oil*) adalah *Avg_Age* (rata-rata umur) penghuni rumah.
b. Atribut (faktor) kedua yang paling berpengaruh adalah *Temperature* (hubungan negatif).
c. Atribut (faktor) ketiga yang paling berpengaruh adalah *Insulation* (hubungan positif).
d. Atribut *Home_Size*, pengaruhnya sangat kecil, sedangkan *Num_Occupant* boleh dikatakan tidak ada pengaruh ke konsumsi minyak pemanas.

6. ***Deployment***

Atribut *Num_Occupant* tidak memberikan pengaruh yang signifikan terhadap konsumsi *heating oil,* sehingga atribut ini bisa dihapus. Atribut *Insulation* memiliki hubungan positif yang cukup kuat, maka dari itu Sarah bisa berfokus untuk mencari beberapa peluang untuk bermitra dengan perusahaan yang memiliki spesialisasi dalam menambahkan insulasi pada rumah yang sudah ada. Melihat dari atribut *Avg_Age* dan *Temperature*, Sarah dapat memfokuskan upaya pemasaran pada kota dengan suhu rendah dan usia rata-rata penduduk yang tinggi. Selain itu, Sarah juga bisa menambahkan perincian yang lebih besar dalam kumpulan data dengan tujuan mengetahui tentang penggunaan minyak pemanas selama periode waktu yang lebih pendek dari satu tahun (per bulan).

D. Studi Kasus II

Lakukan studi kasus lanjutan dari Studi Kasus I

STUDI KASUS
SARAH: HEATING_OIL AFTER BOOMING

1. ***Business Understanding***

   a. *Problem*

   Bisnis berkembang pesat, tim penjualan sedang merekrut ribuan klien baru, dan Sarah ingin memastikan bahwa perusahaan akan mampu memenuhi tingkat permintaan yang baru ini. Tujuan penggalian data baru Sarah cukup jelas: dia ingin mengantisipasi permintaan untuk produk yang dapat dikonsumsi.

   b. *Objective*

   Memprediksi penggunaan yang akan dibawa oleh 42.650 klien baru ke perusahaannya.

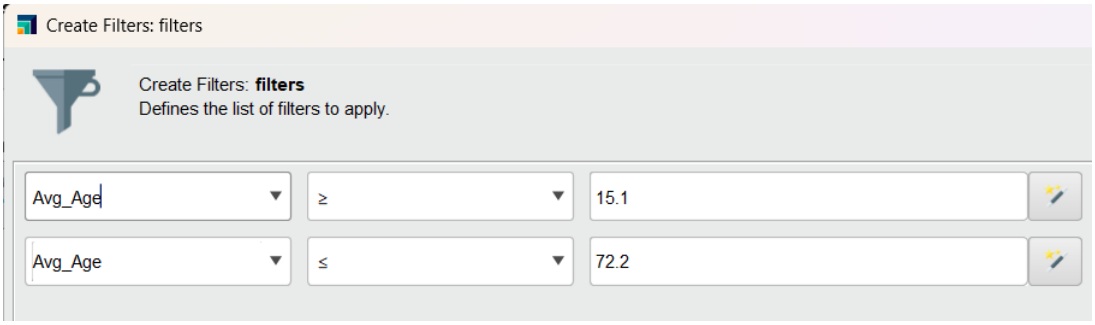
2. ***Data Understanding***

Sarah telah mengumpulkan file *Comma Separated Values* terpisah yang berisi semua atribut yang sama, untuk 42.650 klien barunya. Dataset tersebut terdiri dari beberapa atribut berikut.

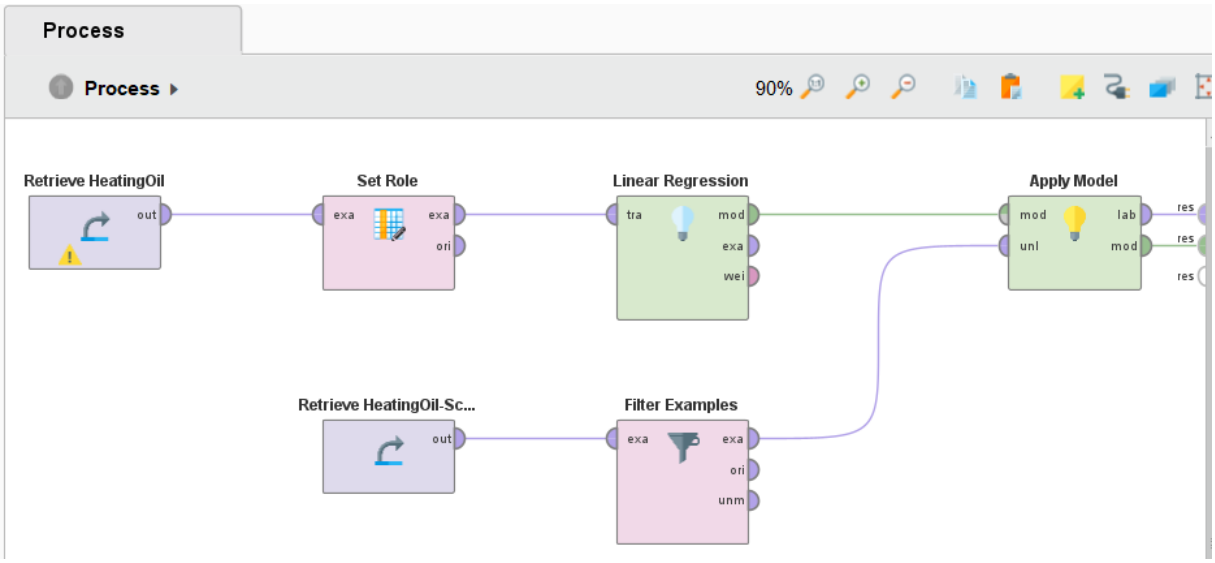
Insulation	Peringkat kepadatan yang menunjukkan ketebalan insulasi di setiap rumah. Range dimulai dari 1-10. Rumah dengan insulasi yang buruk memiliki peringkat kepadatan satu, sedangkan rumah dengan insulasi yang sangat baik memiliki peringkat kepadatan sepuluh.
Temperature	Suhu lingkungan luar ruangan rata-rata di setiap rumah untuk tahun terakhir, diukur dalam derajat Fahrenheit.
Heating_oil	Jumlah total unit minyak pemanas yang dibeli oleh pemilik setiap rumah pada tahun terakhir.
Num_occupants	Jumlah total penghuni yang tinggal di setiap rumah.
Avg_age	Usia rata-rata penghuni rumah tersebut.
Home_size	Peringkat dalam skala 1-8, dari ukuran keseluruhan rumah. Semakin tinggi angkanya, semakin besar rumah tersebut.

3. ***Data Preparation***

Operator *Filter Examples* akan digunakan untuk melakukan *filtering* pada atribut *Avg_Age*. Filter nilai atribut digunakan untuk menghapus data dengan usia rata-rata di bawah 15,1 tahun dan di atas 72,2 tahun. Data-data tersebut dianggap sebagai *noise* karena kemungkinan besar merupakan data yang salah atau tidak lengkap.



4. ***Modeling***



Model regresi linier akan membantu Sarah membuat prediksi yang diinginkan. Sarah memiliki 1.218 data hasil observasi yang memberikan profil atribut untuk setiap rumah, bersama dengan konsumsi minyak pemanas tahunan rumah-rumah tersebut. Ia ingin menggunakan kumpulan data ini sebagai data pelatihan untuk memprediksi penggunaan yang akan dilakukan oleh 42.042 klien baru di perusahaannya.

5. ***Evaluation***

Berikut adalah model regresi yang dihasilkan:

LinearRegression (Linear Regression) | ExampleSet (Apply Model)

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

LinearRegression (Linear Regression) | ExampleSet (Apply Model)

LinearRegression

```
  3.323 * Insulation
- 0.869 * Temperature
+ 1.968 * Avg_Age
+ 3.173 * Home_Size
+ 134.511
```

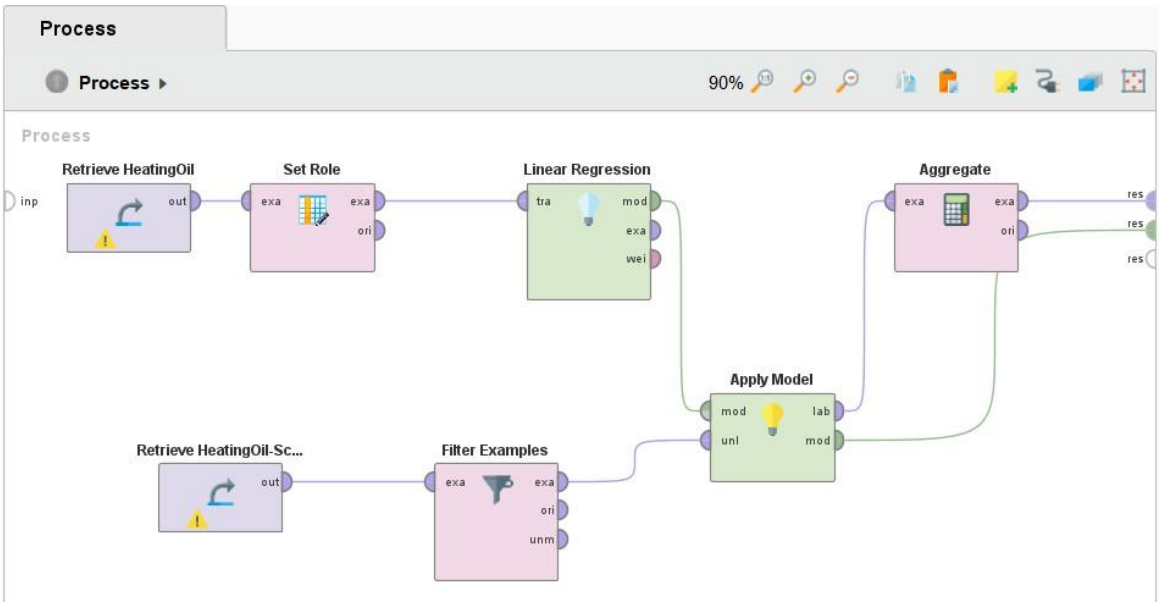
Prediksi yang dihasilkan dari model adalah sebagai berikut:

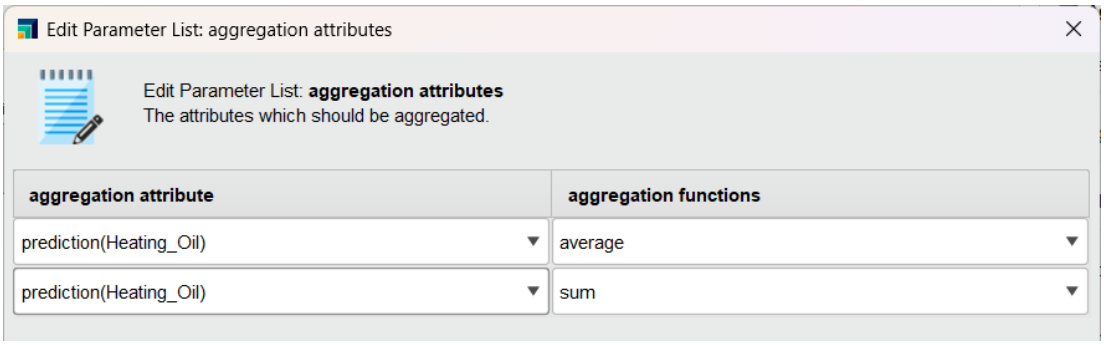
LinearRegression (Linear Regression) | ExampleSet (Apply Model)

Open in Turbo Prep | Auto Model | Filter (42,042 / 42,042 e

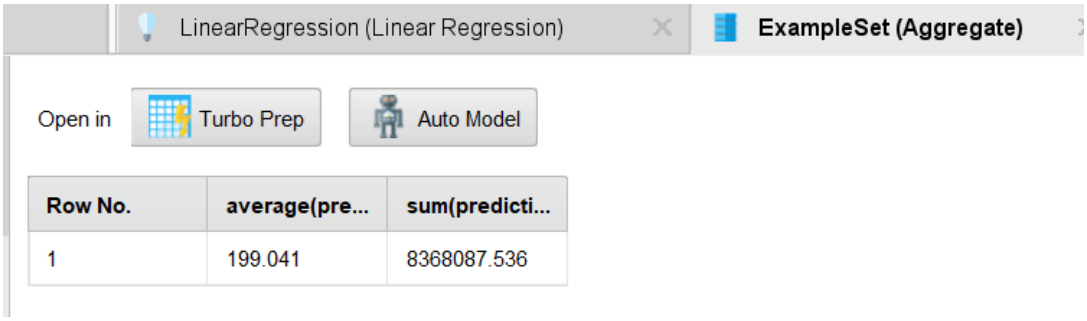
Row No.	prediction(Heating_Oil)	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	89	9	67.800	7
4	209.529	7	81	9	52.400	6
5	164.669	4	58	8	22.900	7
6	180.512	4	58	6	37.400	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4
9	264.712	9	39	1	56.900	7
10	221.364	8	84	5	64.500	2
11	221.328	10	74	6	58.300	1
12	262.580	5	49	6	68.600	6
13	214.082	8	45	2	33.900	8
14	212.392	3	49	4	49.700	4
15	253.199	9	66	6	66.200	5
16	275.043	9	57	10	70.100	7
17	190.837	9	66	10	32.900	6
18	234.624	4	47	3	55.200	6

6. ***Deployment***





Operator *aggregate* di atas digunakan untuk menghitung nilai rata-rata dan jumlah dari variabel *predictionHeating_Oil*. Nilai rata-rata dari atribut *predictionHeating_Oil* akan menunjukkan nilai rata-rata dari prediksi harga minyak pemanas. Nilai ini dapat digunakan untuk menilai kinerja model pembelajaran mesin yang digunakan untuk memprediksi harga minyak pemanas. Jumlah dari atribut *predictionHeating_Oil* akan menunjukkan total dari prediksi harga minyak pemanas. Nilai ini dapat digunakan untuk menganalisis tren harga minyak pemanas.



E. Studi Kasus III

Lakukan studi kasus lanjutan dari Studi Kasus II

STUDI KASUS
SARAH: HEATING_OIL AFTER PROMOTION

Sarah mendapatkan promosi menjadi VP *marketing*, yang mengelola ratusan *marketer*. Sarah ingin para *marketer* dapat memprediksi pelanggan potensial mereka masing-masing secara mandiri. Masalahnya, data *HeatingOil*.csv hanya boleh diakses oleh level VP (Sarah), dan tidak diperbolehkan diakses oleh *marketer* secara langsung.

Sarah ingin masing-masing *marketer* membuat proses yang dapat mengestimasi kebutuhan konsumsi minyak dari *client* yang mereka *approach*, dengan menggunakan model yang sebelumnya dihasilkan oleh Sarah, meskipun tanpa mengakses data *training* (*HeatingOil*.csv).

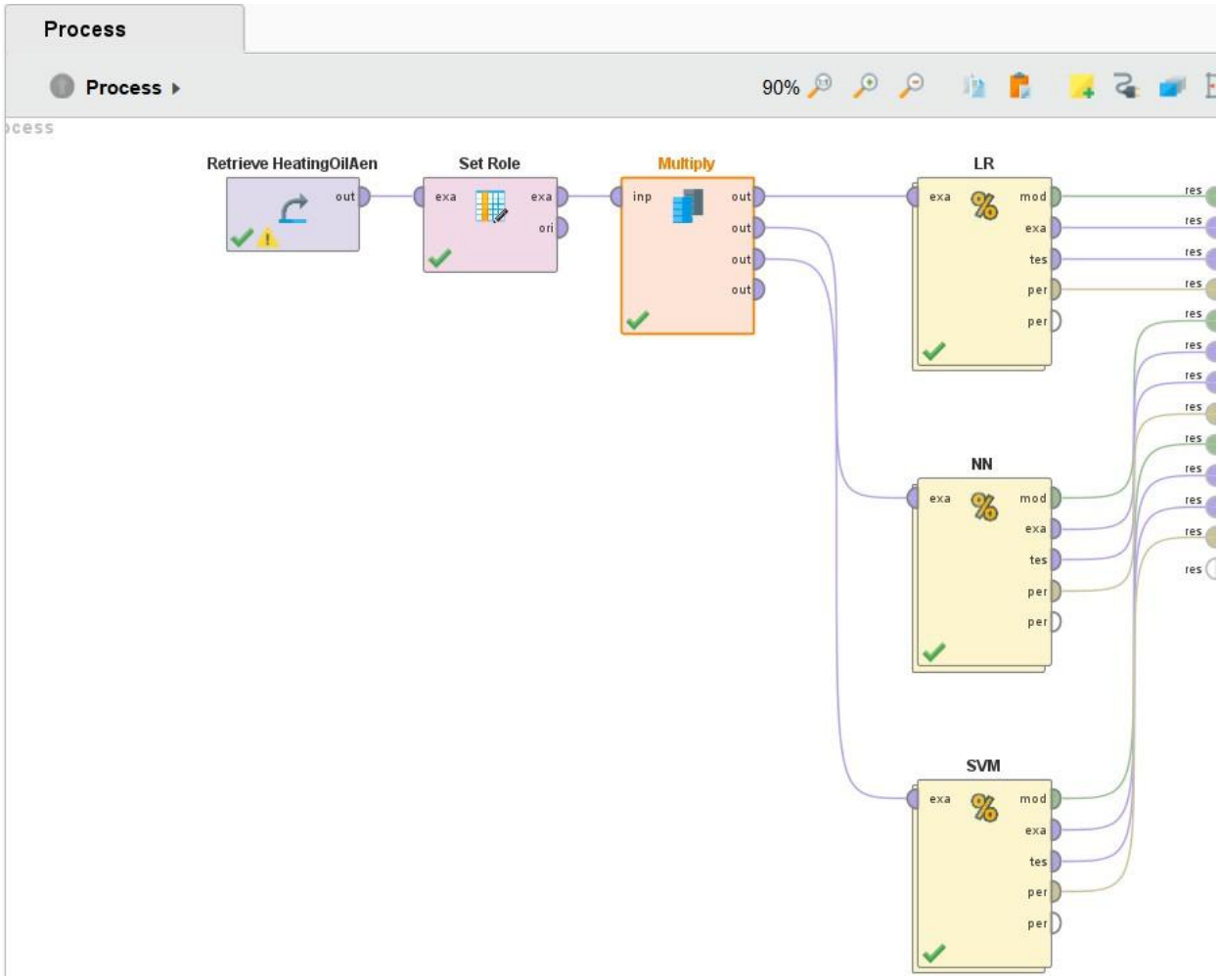
Asumsikan bahwa data *HeatingOil-Marketing*.csv adalah data calon pelanggan yang berhasil di *approach* oleh salah satu marketingnya. Yang harus dilakukan Sarah adalah membuat proses untuk:

a. Mengkomparasi algoritma yang menghasilkan model yang memiliki akurasi tertinggi (LR, NN, SVM), gunakan 10 *Fold X Validation*
b. Menyimpan model terbaik ke dalam suatu file (operator *Store*)

Sementara itu, yang harus dilakukan *marketer* adalah membuat proses untuk:

a. Membaca model yang dihasilkan Sarah (operator *Retrieve*)
b. Menerapkannya di data *HeatingOil-Marketing*.csv yang mereka miliki

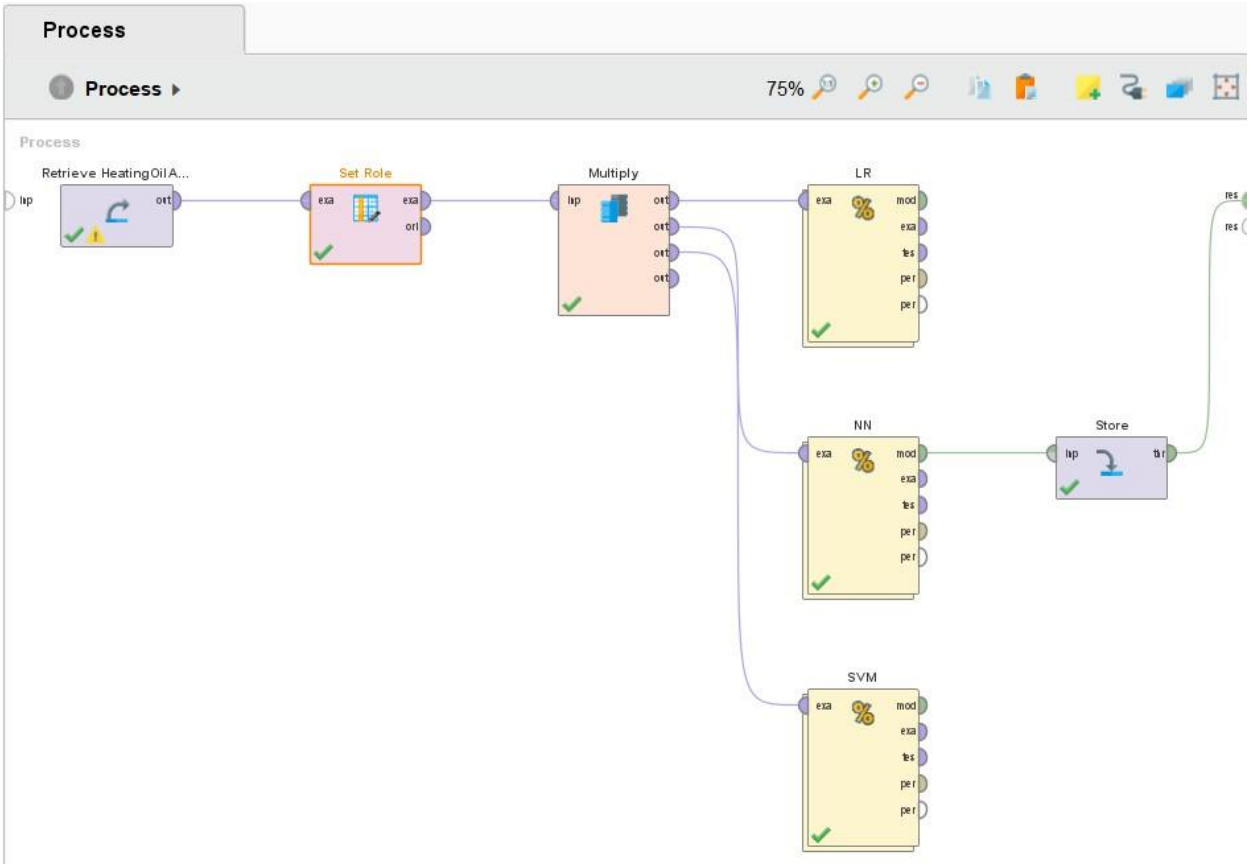
Berikut ini merupakan pemodelan yang dilakukan oleh Sarah dengan cara mengkomparasikan algoritma untuk mendapatkan model yang memiliki akurasi tertinggi. Algoritma yang digunakan antara lain *Linear Regression, Neural Network,* dan *Support Vector Machine* (SVM).

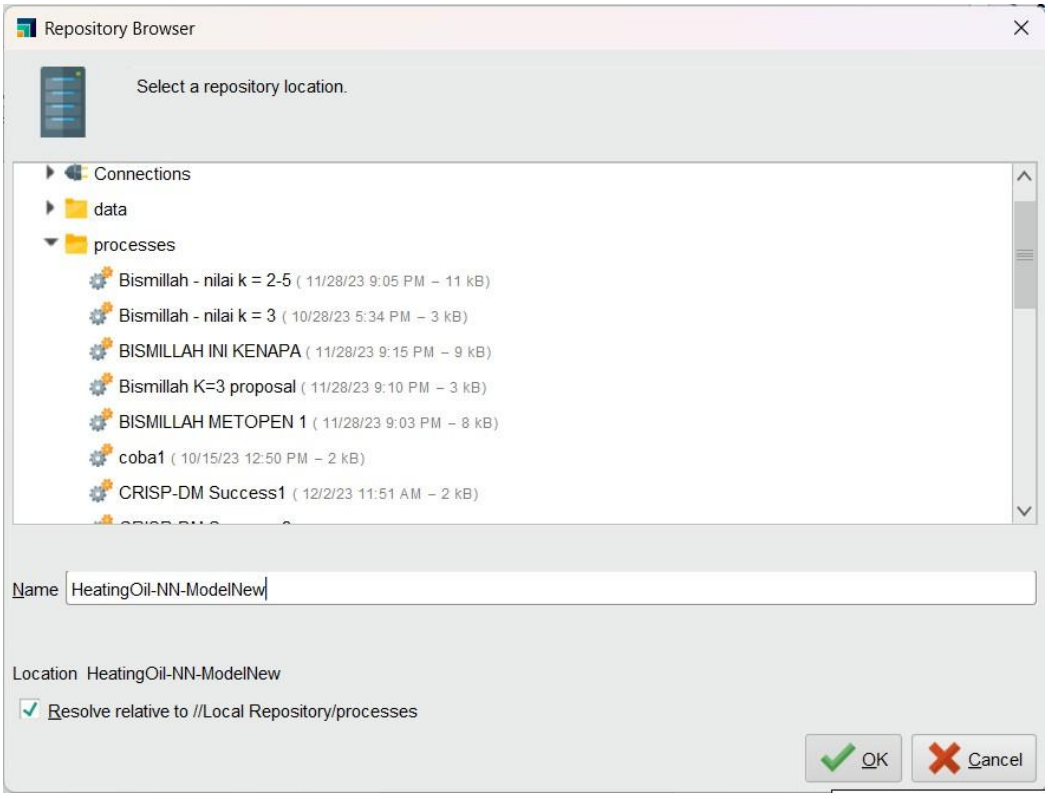


Nilai *Root Mean Squared Error* (RMSE) dari ketiga algoritma di atas dapat dilihat pada tabel di bawah ini.

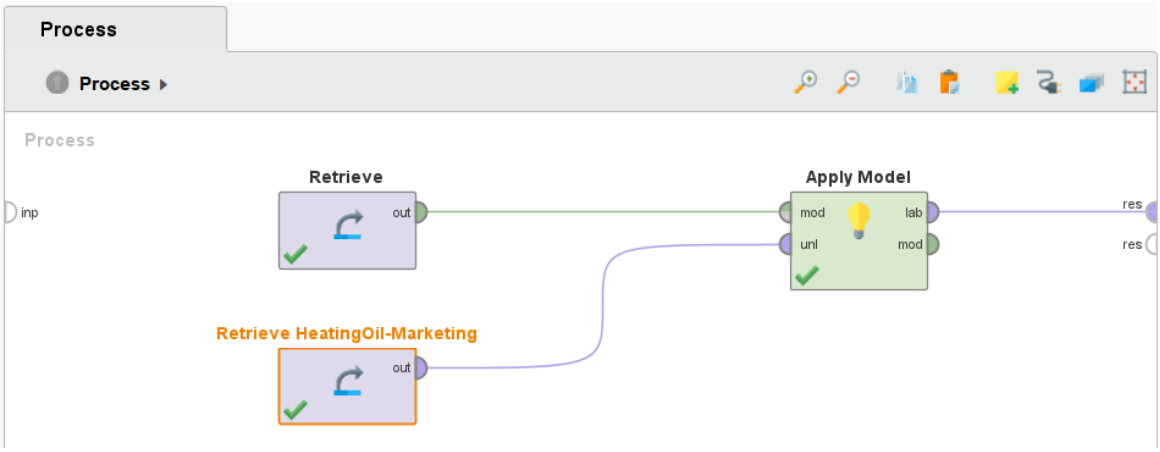
Algoritma	**Nilai *Root Mean Squared Error* (RMSE)**
Linear Regression	23.962 +/- 2.299
Neural Network	14.230 +/- 2.647
Support Vector Machine	24.978 +/- 4.551

Algoritma *Neural Network* (NN) memiliki nilai akurasi (RMSE) yang terkecil sehingga pemodelan dengan algoritma inilah yang akan disimpan dan digunakan. Operator *Store* dapat digunakan untuk menyimpan pemodelan algoritma NN ini.


Process
Process
75%
Process
Retrieve HeatingOil A...
Set Role
Multiply
LR
NN
Store
SVM


Repository Browser
Select a repository location.
Connections
data
processes
Bismillah - nilai k = 2-5 (11/28/23 9:05 PM – 11 kB)
Bismillah - nilai k = 3 (10/28/23 5:34 PM – 3 kB)
BISMILLAH INI KENAPA (11/28/23 9:15 PM – 9 kB)
Bismillah K=3 proposal (11/28/23 9:10 PM – 3 kB)
BISMILLAH METOPEN 1 (11/28/23 9:03 PM – 8 kB)
coba1 (10/15/23 12:50 PM – 2 kB)
CRISP-DM Success1 (12/2/23 11:51 AM – 2 kB)
Name HeatingOil-NN-ModelNew
Location HeatingOil-NN-ModelNew
Resolve relative to //Local Repository/processes
OK
Cancel

Selanjutnya, para marketer dapat memanfaatkan pemodelan algoritma NN yang tadi telah disimpan untuk menganalisis dataset yang lain, contoh dalam kasus ini menggunakan data baru dari dataset *HeatingOil-Marketing*.csv.



Hasil dari pemodelan di atas adalah sebagai berikut.

ExampleSet (Apply Model)

Open in Turbo Prep Auto Model Filter (4 / 4 exa

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