DIGITALTALENTSCHOLARSHIP–KOMINFO(BPPTIKCIKARANG)

VOCATIONAL SCHOOL GRADUATE ACADEMY

ASSOCIATEDATASCIENCE

Gelombang- 12

Pengajar:Mona Arif Muda Batubara

Tugas Akhir/ Project ADS

Nama Peserta : Hendry Imam Sanjaya

Soal:

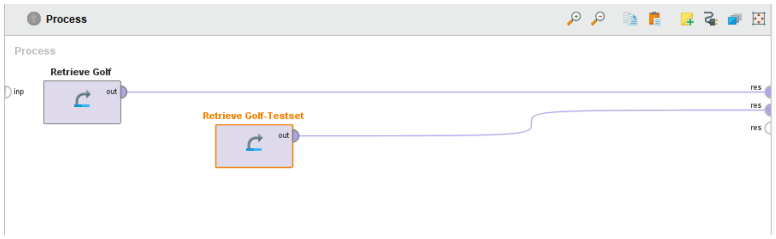
1. Dengan menggunakan dataset latih Golf dan dataset tes Golf-Testset yang ada di data samples di RapidMiner, lakukan:
    a. telaah data tersebut dengan menggunakan RapidMiner
    b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner
    c. bangun Model Classification dengan algoritma K-NN dan evaluasi hasil pemodelannya
    d. buat laporan tertulisnya
2. Dengan menggunakan dataset Facebook-Live-Sellers-in-Thailand_20210128.csv terlampir, lakukan:
    a. telaah data tersebut dengan menggunakan RapidMiner
    b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner
    c. bangun Model Clustering dengan algoritma K-Means dan evaluasi hasil pemodelannya
    d. buat laporan tertulisnya.
3. Dengan menggunakan dataset ToyotaCorolla.csv terlampir, lakukan:
    a. telaah data tersebut dengan menggunakan RapidMiner
    b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner
    c. Bangun Model Regresi Linier dengan 3 variable bebas dan evaluasi hasil pemodelannya
    d. buat laporan tertulisnya.

1. LAPORAN ANALISIS GOLF dan GOLF –TEST (CLASSIFICATION)

Dataset ini berisi data mengenai kondisi cuaca dan keputusan untuk bermain golf. Tujuan dari analisis ini adalah untuk membangun model klasifikasi menggunakan algoritma K-NN dan mengevaluasi performanya.

a. Ini merupaka telaahan data menggunakan RapidMiner yang mana kita

- Mengimpor dataset Golf dan dataset tes Golf-Testset dari data samples di RapidMiner.
- Pilih “Samples” pada repository, dan cari dataset yang dimaksud.
- Drag dan drop dataset ke dalam proses baru lalu run
- Gambar di bawah ini merupakan hasil data yang di dapat



Golf

ExampleSet (Retrieve Golf-Testset) | ExampleSet (Retrieve Golf)

Open in Turbo Prep | Auto Model

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

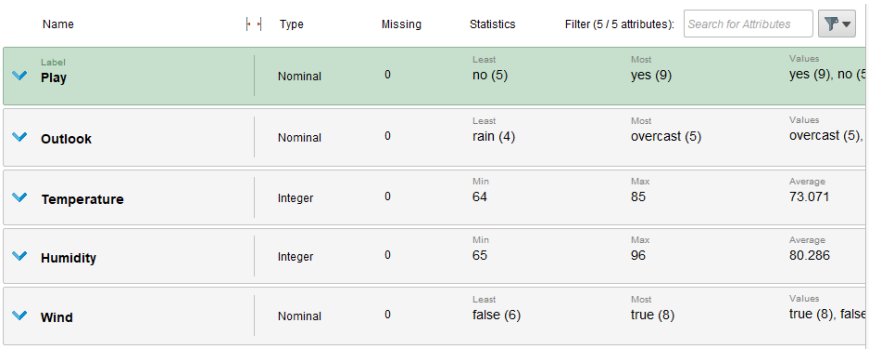
Golf-Testset

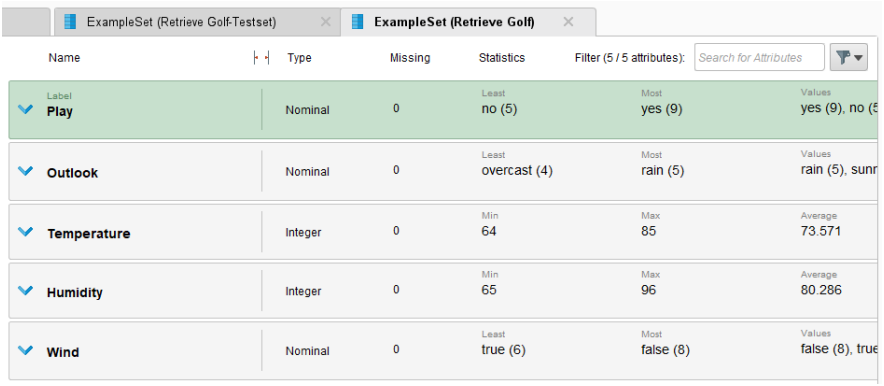
ExampleSet (Retrieve Golf-Testset) | ExampleSet (Retrieve Golf)

Open in Turbo Prep | Auto Model

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	yes	sunny	85	85	false
2	no	overcast	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	true
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	no	sunny	75	80	false
11	yes	sunny	68	70	true
12	yes	overcast	72	90	true
13	no	overcast	81	75	true
14	yes	rain	71	80	true

b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner. Setelah kita lihat datanya di missing values dan outliers tidak ada yang harus di bersihkan.



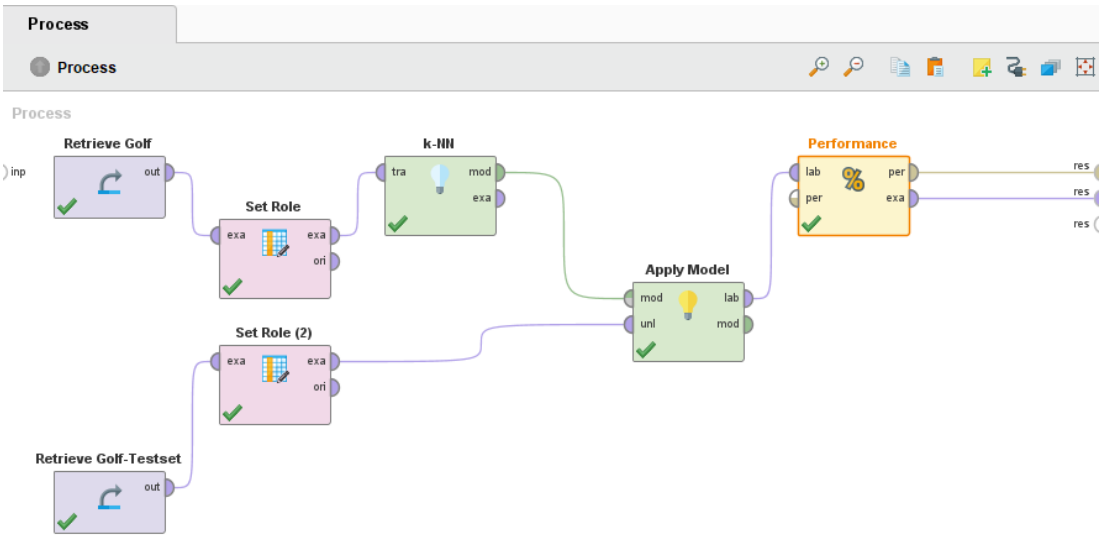


c. bangun Model Classification dengan algoritma K-NN dan evaluasi hasil pemodelannya

- Setup Proses dan Evaluasi Hasil:
  - Hubungkan Data set ke Set Role untuk menetapkan atribut target (disini saya memberikan lable pada artibut Play).
  - Tambahkan "K-NN" untuk membangun model klasifikasi dengan parameter K yang sesuai (misalnya, K=3).
  - Sambungkan "Apply Model" untuk menerapkan model pada dataset tes.
  - Tambahkan "Performance (Classification)" untuk mengevaluasi model.
  - Lalu Run.

Algoritma K-NN dengan parameter K=3 digunakan untuk membangun model. Berikut adalah struktur proses yang digunakan:

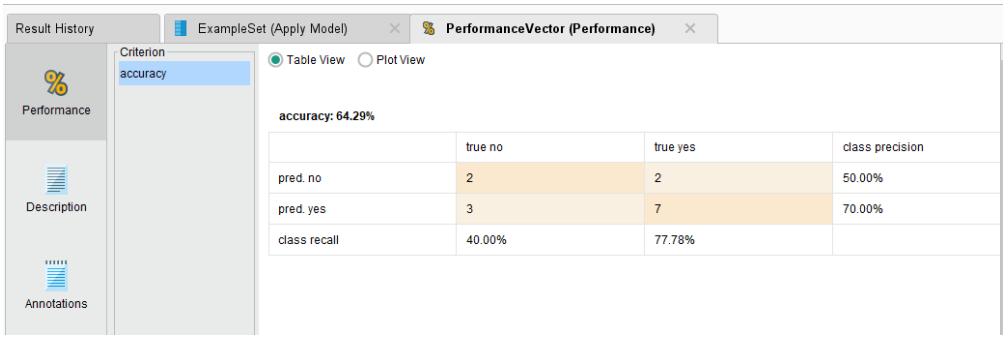
- Data Set-> Set Role -> K-NN -> Apply Model -> Performance (Classification)



ExampleSet (Apply Model) | PerformanceVector (Performance)

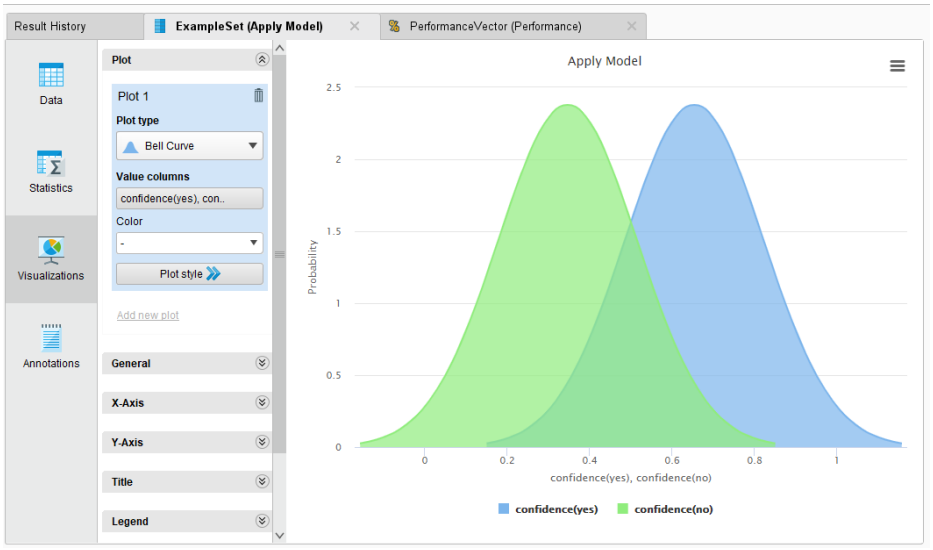
Open in Turbo Prep | Auto Model | Filter (14 / 14 examples): all

Row No.	Play	prediction(P...	confidence(...	confidence(...	Outlook	Temperature	Humidity	Wind
1	yes	yes	0.451	0.549	sunny	85	85	false
2	no	no	0.630	0.370	overcast	80	90	true
3	yes	yes	0.190	0.810	overcast	83	78	false
4	yes	no	0.544	0.456	rain	70	96	false
5	yes	yes	0.394	0.606	rain	68	80	true
6	no	yes	0.250	0.750	rain	65	70	true
7	yes	yes	0.218	0.782	overcast	64	65	true
8	no	no	0.559	0.441	sunny	72	95	false
9	yes	yes	0.212	0.788	sunny	69	70	false
10	no	yes	0.213	0.787	sunny	75	80	false
11	yes	yes	0.222	0.778	sunny	68	70	true
12	yes	no	0.554	0.446	overcast	72	90	true
13	no	yes	0.164	0.836	overcast	81	75	true
14	yes	yes	0.250	0.750	rain	71	80	true



PerformanceVector

```
PerformanceVector:
accuracy: 64.29%
ConfusionMatrix:
True:    no      yes
no:      2       2
yes:     3       7
```



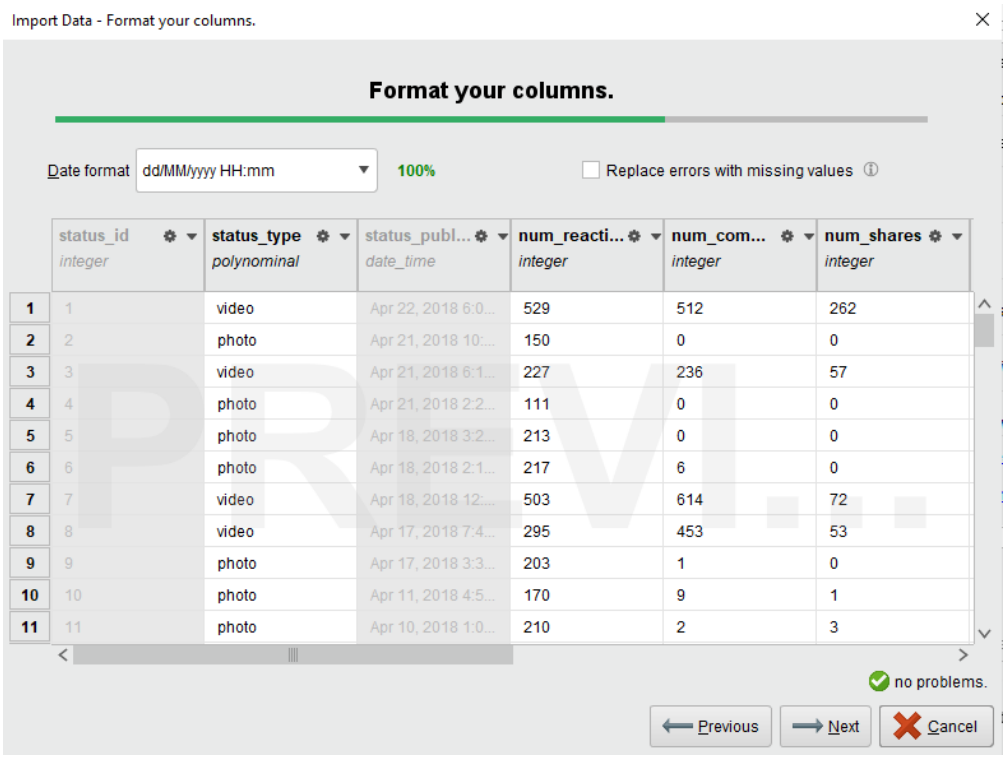
Model K-NN menunjukkan performa yang cukup baik dengan akurasi sebesar 64.29%.

2. LAPORAN ANALISIS FACEBOOK LIVE SELLERS IN THAILAND (CULUSTERING)

Dataset ini berisi data mengenai penjual live di Facebook di Thailand. Tujuan dari analisis ini adalah untuk membangun model clustering menggunakan algoritma K-Means dan mengevaluasi performanya.

a. telaah data tersebut dengan menggunakan RapidMiner

- Buka RapidMiner Studio dan buat proyek baru. Impor dataset Facebook-Live-Sellers-in-Thailand_20210128.csv ke dalam RapidMiner.
- Pilih "Import Data" dan arahkan ke file Facebook-Live-Sellers-in-Thailand_20210128.csv.
- Drag dan drop dataset ke dalam proses baru.
- Lihat distribusi data untuk memahami jenis atribut (categorical atau numerical). Disini saya mengexclud data yag menurut saya tidak di gunakan
- Gunakan operator "Statistics" untuk melihat statistik dasar dari setiap atribut.
- Visualisasikan data dengan "Charts" untuk memahami hubungan antara variabel.



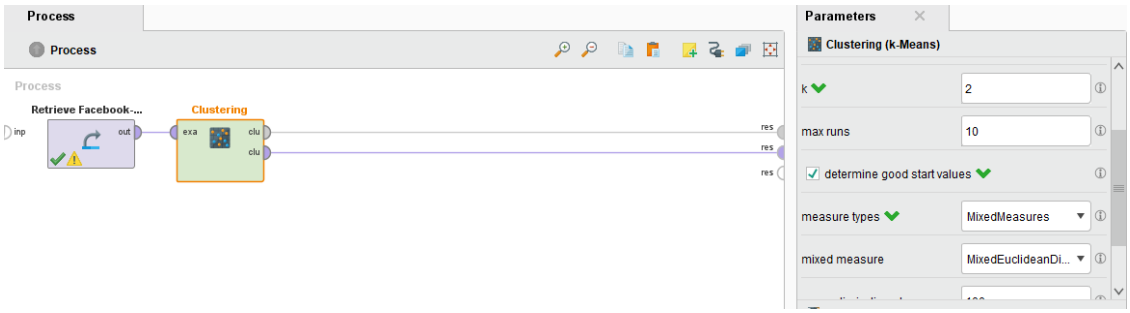
	status_id integer	status_type polynominal	status_publ... date_time	num_reacti... integer	num_com... integer	num_shares integer
1	1	video	Apr 22, 2018 6:0...	529	512	262
2	2	photo	Apr 21, 2018 10:...	150	0	0
3	3	video	Apr 21, 2018 6:1...	227	236	57
4	4	photo	Apr 21, 2018 2:2...	111	0	0
5	5	photo	Apr 18, 2018 3:2...	213	0	0
6	6	photo	Apr 18, 2018 2:1...	217	6	0
7	7	video	Apr 18, 2018 12:...	503	614	72
8	8	video	Apr 17, 2018 7:4...	295	453	53
9	9	photo	Apr 17, 2018 3:3...	203	1	0
10	10	photo	Apr 11, 2018 4:5...	170	9	1
11	11	photo	Apr 10, 2018 1:0...	210	2	3

b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner. Setelah kita lihat datanya di missing values dan outliers tidak ada yang harus di bersihkan.

Name	Type	Missing	Statistics			Filter (12 / 12 attributes):	Search for Attributes
status_id	Integer	0	Min 1	Max 7050	Average 3525.500		
status_type	Nominal	0	Least link (63)	Most photo (4288)	Values photo (428		
status_published	Date-time	0	Earliest date Jul 15, 2012 2:51 AM	Latest date Jun 13, 2018 1:12 AM	Duration 2158d 22h		
num_reactions	Integer	0	Min 0	Max 4710	Average 230.117		
num_comments	Integer	0	Min 0	Max 20990	Average 224.356		
num_shares	Integer	0	Min 0	Max 3424	Average 40.023		
num_likes	Integer	0	Min 0	Max 4710	Average 215.043		

c. Model Clustering dengan algoritma K-Means dan evaluasi hasil pemodelannya.

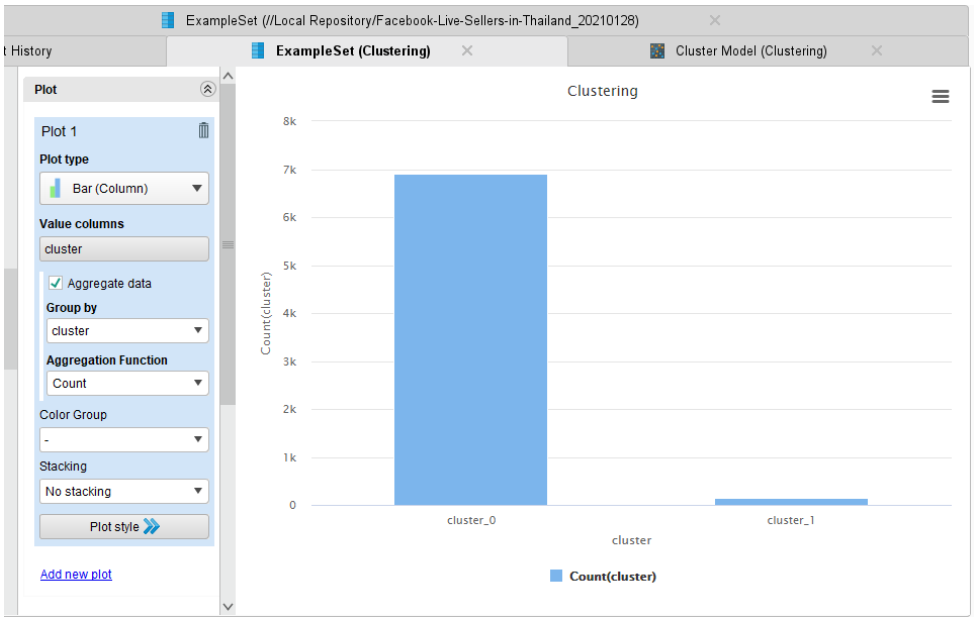
- **Setup Proses**:
  - Drag operator “Read CSV” untuk membaca dataset.
  - Hubungkan “Read CSV” ke “Set Role” untuk menetapkan atribut yang akan digunakan dalam clustering (pastikan atribut non-numerik diabaikan atau diubah ke bentuk numerik jika perlu).
  - Tambahkan “K-Means” untuk membangun model clustering dengan jumlah cluster yang sesuai (misalnya, k=2).
  - Sambungkan “Apply Model” untuk menerapkan model pada dataset.
  - Tambahkan “Performance (Clustering)” untuk mengevaluasi model.
  - Lihat metrik seperti Silhouette Coefficient, Within-Cluster Sum of Squares (WCSS), dan Elbow Method untuk menentukan jumlah cluster yang optimal.



Open in Turbo Prep Auto Model

Filter (7,050 / 7,050 examples): all

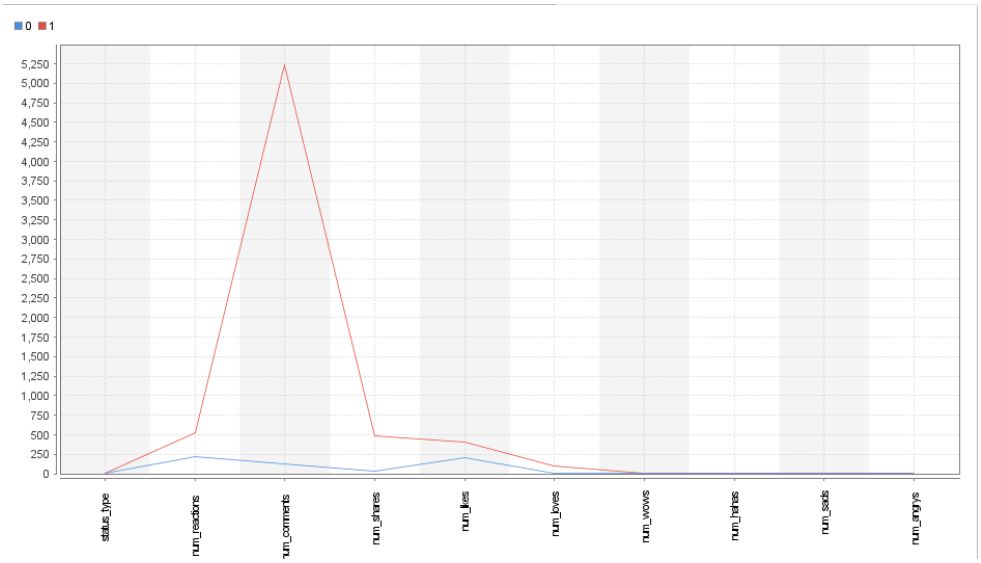
Row No.	id	cluster	status_type	num_reacti...	num_comm...	num_shares	num_likes	num_loves
1	1	cluster_0	video	529	512	262	432	92
2	2	cluster_0	photo	150	0	0	150	0
3	3	cluster_0	video	227	236	57	204	21
4	4	cluster_0	photo	111	0	0	111	0
5	5	cluster_0	photo	213	0	0	204	9
6	6	cluster_0	photo	217	6	0	211	5
7	7	cluster_0	video	503	614	72	418	70
8	8	cluster_0	video	295	453	53	260	32
9	9	cluster_0	photo	203	1	0	198	5
10	10	cluster_0	photo	170	9	1	167	3
11	11	cluster_0	photo	210	2	3	202	7
12	12	cluster_0	photo	222	4	0	213	5
13	13	cluster_0	photo	313	4	2	305	6



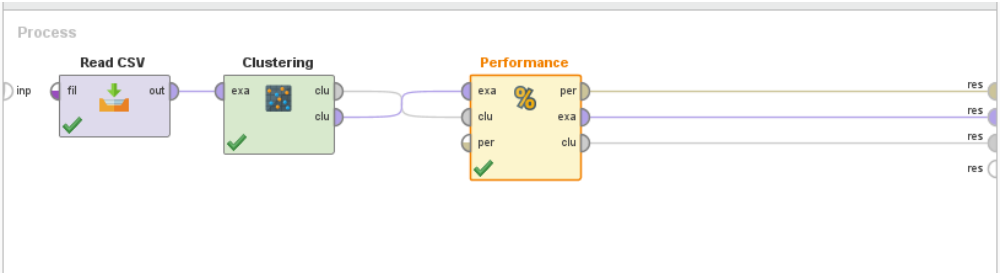
Cluster Model

```
Cluster 0: 6909 items
Cluster 1: 141 items
Total number of items: 7050
```

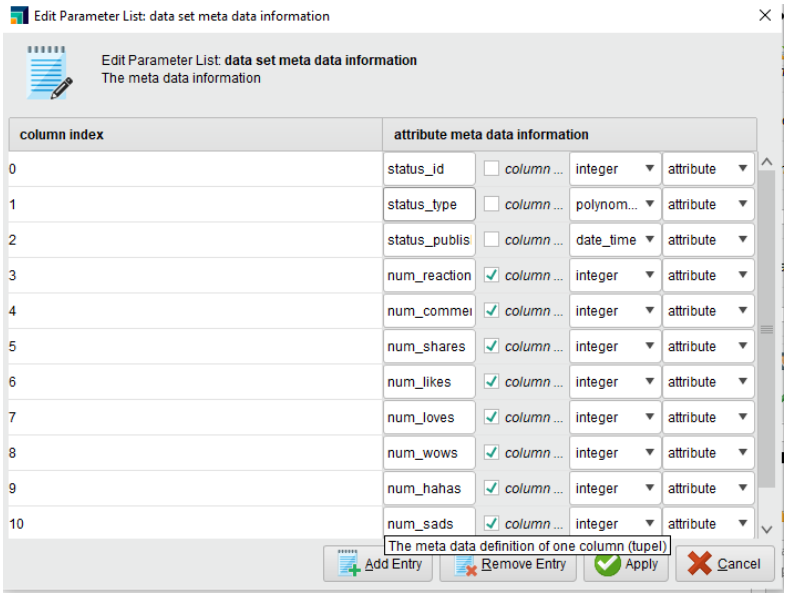
Attribute	cluster_0	cluster_1
status_type	1.797	1.007
num_reactions	224.142	522.887
num_comments	122.110	5234.418
num_shares	30.806	491.624
num_likes	211.228	402.007
num_loves	10.908	101.922
num_wows	1.153	7.972
num_hahas	0.557	7.539
num_sads	0.197	2.511
num_angrys	0.096	0.936



Penambahan Performance



Karena ada data polynominal kita hapus check list status_type



Ini merupakan hasil performace vector yang mana kita dapat mengujinya di excel atau aplikasi lainnya

Avg. within centroid distance_cluster_0

```
Avg. within centroid distance_cluster_0: 549396.694
```

Avg. within centroid distance_cluster_1

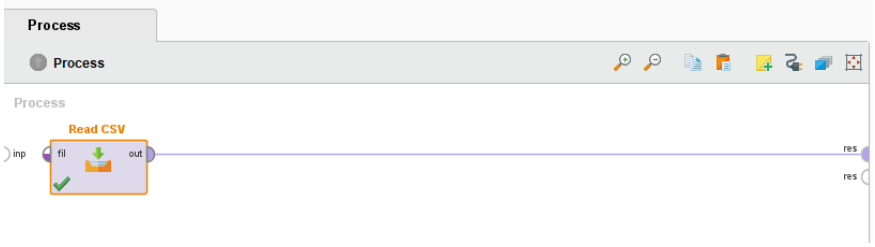
```
Avg. within centroid distance_cluster_1: 8444043.201
```

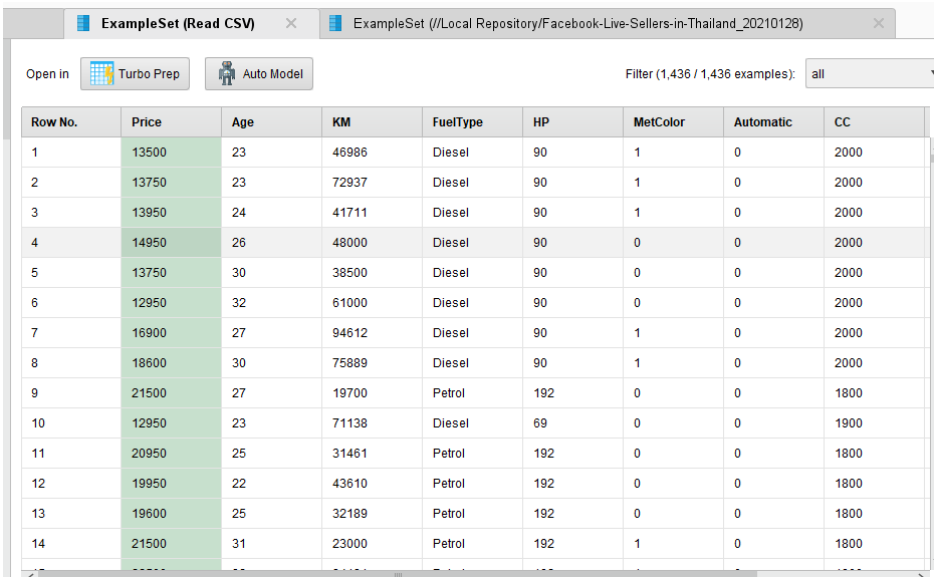
3. LAPORAN ANALISIS TOYOTA COROLLA (REGRESI LINIER)

Dataset ini berisi data mengenai berbagai atribut mobil Toyota Corolla, termasuk usia, jarak tempuh, tenaga kuda, dan harga. Tujuan dari analisis ini adalah untuk membangun model regresi linier untuk memprediksi harga mobil berdasarkan tiga variabel bebas.

a. telaah data tersebut dengan menggunakan RapidMiner

- Buka RapidMiner Studio dan buat proyek baru.
- Impor dataset ToyotaCorolla.csv ke dalam RapidMiner.
- Pilih “Import Data” dan arahkan ke file ToyotaCorolla.csv.
- Drag dan drop dataset ke dalam proses baru.
- Buka dataset ToyotaCorolla di RapidMiner.
- Lihat distribusi data untuk memahami jenis atribut (categorical atau numerical).
- Gunakan operator “Statistics” untuk melihat statistik dasar dari setiap atribut.
- Visualisasikan data dengan “Charts” untuk memahami hubungan antara variabel.





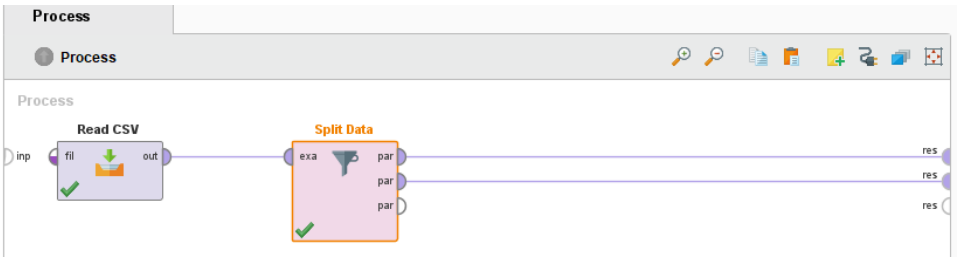
Row No.	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC
1	13500	23	46986	Diesel	90	1	0	2000
2	13750	23	72937	Diesel	90	1	0	2000
3	13950	24	41711	Diesel	90	1	0	2000
4	14950	26	48000	Diesel	90	0	0	2000
5	13750	30	38500	Diesel	90	0	0	2000
6	12950	32	61000	Diesel	90	0	0	2000
7	16900	27	94612	Diesel	90	1	0	2000
8	18600	30	75889	Diesel	90	1	0	2000
9	21500	27	19700	Petrol	192	0	0	1800
10	12950	23	71138	Diesel	69	0	0	1900
11	20950	25	31461	Petrol	192	0	0	1800
12	19950	22	43610	Petrol	192	0	0	1800
13	19600	25	32189	Petrol	192	0	0	1800
14	21500	31	23000	Petrol	192	1	0	1800

b. validasi, pembersihan, dan/atau konstruksi data bila diperlukan, dengan RapidMiner. Setelah kita lihat datanya di missing values dan outliers tidak ada yang harus di bersihkan.

Name	Type	Missing	Statistics	Filter (10 / 10 attributes):	Search for Attributes
FuelType	Nominal	0	Least: CNG (17)	Most: Petrol (1264)	Values: Petrol (12
HP	Integer	0	Min: 69	Max: 192	Average: 101.502
MetColor	Integer	0	Min: 0	Max: 1	Average: 0.675
Automatic	Integer	0	Min: 0	Max: 1	Average: 0.056
CC	Integer	0	Min: 1300	Max: 2000	Average: 1566.828
Doors	Integer	0	Min: 2	Max: 5	Average: 4.033
Weight	Integer	0	Min: 1000	Max: 1615	Average: 1072.460

c. telaah data tersebut dengan menggunakan RapidMiner dan Model Regresi Linier

- Setup Proses:
  - Drag operator “Toyota Corolla” untuk membaca dataset.
  - Hubungkan “Toyota Corolla” ke “Set Role” untuk menetapkan atribut target (misalnya, Price).
  - Pilih tiga variabel bebas yang akan digunakan dalam model (misalnya, Age, KM, HP).
  - Tambahkan “Linear Regression” untuk membangun model regresi linier.
  - Tambahkan “Apply Model” untuk menerapkan model pada dataset.
  - Tambahkan “Performance (Regression)” untuk mengevaluasi model.



Data Traning

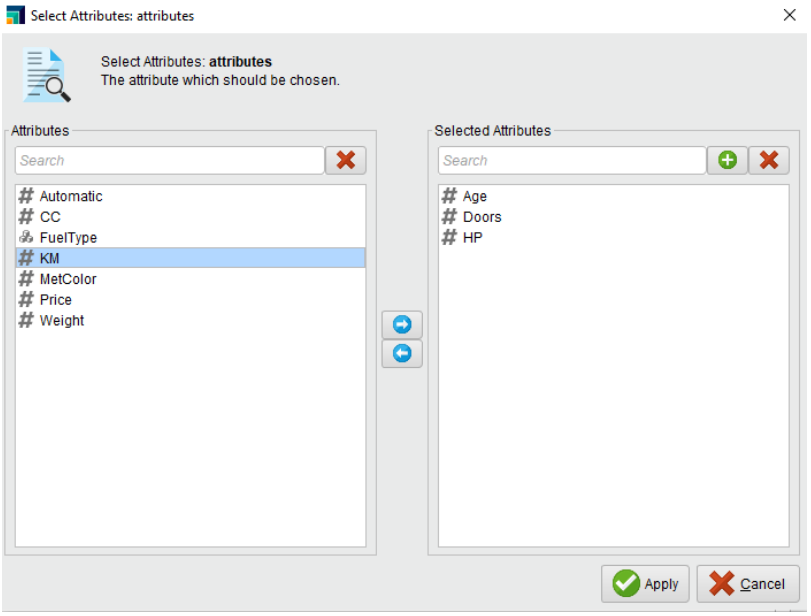
Open in Turbo Prep Auto Model Filter (1,005 / 1,005 examples): all

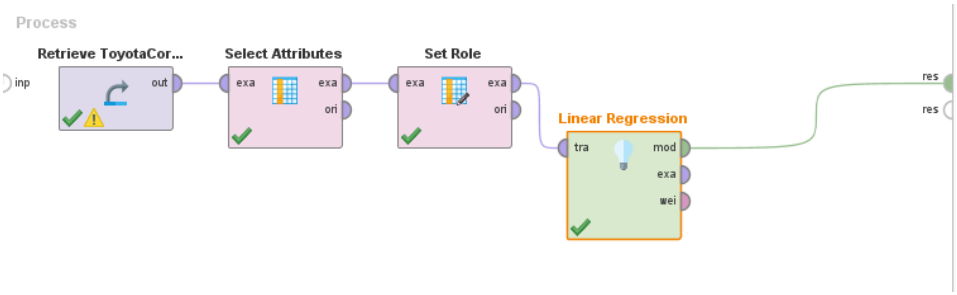
Row No.	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC
1	13500	23	46986	Diesel	90	1	0	2000
2	14950	26	48000	Diesel	90	0	0	2000
3	13750	30	38500	Diesel	90	0	0	2000
4	12950	32	61000	Diesel	90	0	0	2000
5	16900	27	94612	Diesel	90	1	0	2000
6	19950	22	43610	Petrol	192	0	0	1800
7	21500	31	23000	Petrol	192	1	0	1800
8	22500	32	34131	Petrol	192	1	0	1800
9	22000	28	18739	Petrol	192	0	0	1800
10	15950	30	67660	Petrol	110	1	0	1600
11	15950	28	56349	Petrol	110	1	0	1600
12	16950	28	32220	Petrol	110	1	0	1600
13	15950	25	28450	Petrol	110	1	0	1600

Data Testing

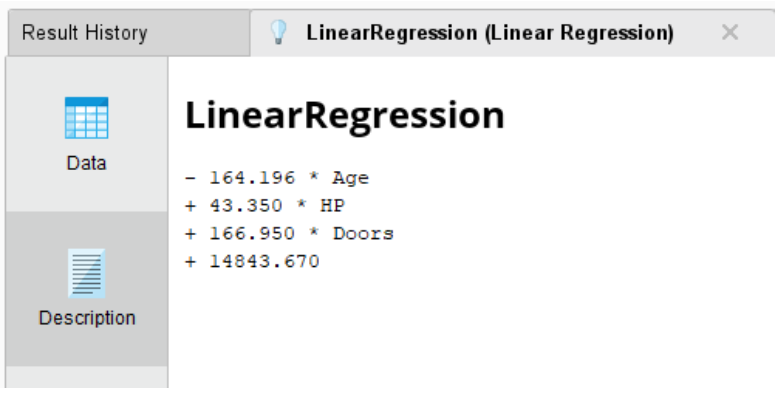
Open in Turbo Prep Auto Model Filter (431 / 431 examples): all

Row No.	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC
1	13750	23	72937	Diesel	90	1	0	2000
2	13950	24	41711	Diesel	90	1	0	2000
3	18600	30	75889	Diesel	90	1	0	2000
4	21500	27	19700	Petrol	192	0	0	1800
5	12950	23	71138	Diesel	69	0	0	1900
6	20950	25	31461	Petrol	192	0	0	1800
7	19600	25	32189	Petrol	192	0	0	1800
8	22750	30	34000	Petrol	192	1	0	1800
9	17950	24	21716	Petrol	110	1	0	1600
10	16750	24	25563	Petrol	110	0	0	1600
11	16950	30	64359	Petrol	110	1	0	1600
12	16950	29	43905	Petrol	110	0	1	1600
13	16250	29	25813	Petrol	110	1	0	1600

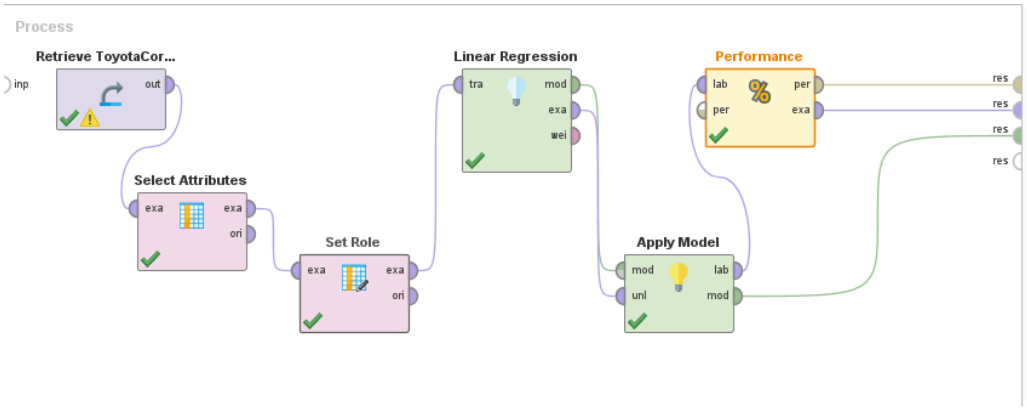




Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Age	-164.196	2.340	-0.842	0.960	-70.183	0	****
HP	43.350	2.885	0.179	0.974	15.026	0	****
Doors	166.950	45.307	0.044	0.975	3.685	0.000	****
(Intercept)	14843.670	384.891	?	?	38.566	0	****



Hasil setelah di performance



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Age	-164.196	2.340	-0.842	0.960	-70.183	0	****
HP	43.350	2.885	0.179	0.974	15.026	0	****
Doors	166.950	45.307	0.044	0.975	3.685	0.000	****
(Intercept)	14843.670	384.891	?	?	38.566	0	****