

JOBSHEET 2

DATA PREPARATION

1.1. Tujuan Praktikum

Setelah menyelesaikan praktikum ini, mahasiswa mampu:

- Mengetahui tentang Normalisasi Data dan Outlier
- Mengimplementasikan Normalisasi Data dan Outlier pada RapidMiner

1.2. Peralatan yang dibutuhkan

Beberapa peralatan yang dibutuhkan dalam menyelesaikan praktikum ini adalah:

- Aplikasi Notepad atau sejenisnya
- Aplikasi RapidMiner versi 7 ke atas
- Aplikasi Microsoft Excel

1.3. Dasar Teori

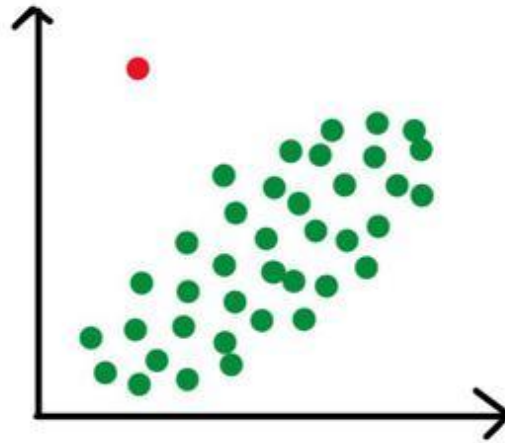
1.3.1. Outlier

Outlier adalah objek data yang menyimpang secara signifikan dari objek data lainnya dan berperilaku berbeda. Outlier adalah suatu objek yang menyimpang secara signifikan dari objek lainnya. Hal ini dapat disebabkan oleh kesalahan pengukuran atau pelaksanaan. Analisis data outlier disebut dengan outlier analysis atau outlier mining. Outlier tidak bisa disebut sebagai noise atau error. Sebaliknya, mereka diduga tidak dihasilkan dengan metode yang sama seperti objek data lainnya. Outlier dapat dibagi menjadi 3 bagian, yaitu:

- Global (or Point) Outliers
- Collective Outliers
- Contextual (or Conditional) Outliers

a) Global Outlier

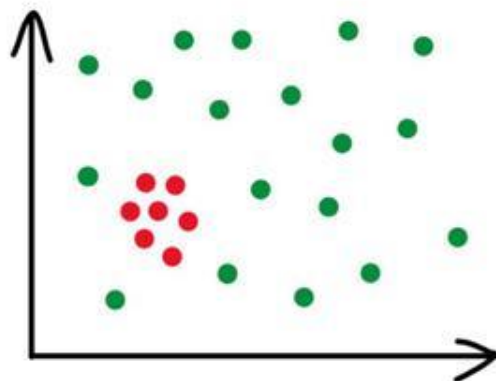
Outlier global adalah titik data yang menyimpang secara signifikan dari keseluruhan distribusi kumpulan data. Kesalahan dalam pengumpulan data, kesalahan pengukuran, atau kejadian yang benar-benar tidak biasa dapat mengakibatkan outlier global. Outlier global dapat mendistorsi hasil analisis data dan memengaruhi performa model pembelajaran mesin.



Gambar 1. 1 Visualisasi Global Outlier

b) Collective Outlier

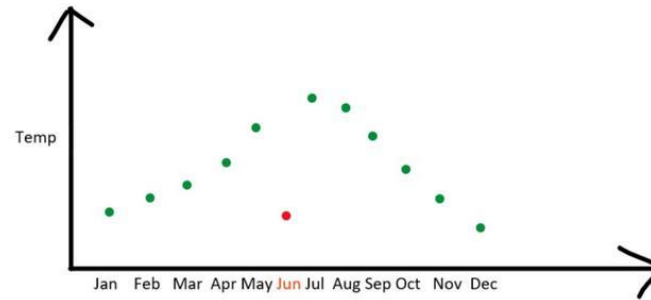
Collective Outliers adalah sekelompok titik data yang secara kolektif menyimpang secara signifikan dari keseluruhan distribusi kumpulan data. Collective Outliers mungkin bukan outlier jika dipertimbangkan secara individu, namun sebagai sebuah kelompok, mereka menunjukkan perilaku yang tidak biasa. Teknik untuk mendeteksi pencilan kolektif mencakup clustering algorithms, density-based methods, dan subspace-based approaches.



Gambar 1. 2 Visualisasi Collective Outlier

c) Contextual Outlier

Contextual Outliers adalah titik data yang menyimpang secara signifikan dari perilaku yang diharapkan dalam konteks atau subkelompok tertentu. Contextual Outliers mungkin bukan outlier jika dipertimbangkan dalam keseluruhan kumpulan data, namun menunjukkan perilaku yang tidak biasa dalam konteks atau subgrup tertentu. Teknik untuk mendeteksi outlier kontekstual mencakup contextual clustering, contextual anomaly detection, context-aware machine learning.



Gambar 1. 3 Visualisasi Contextual Outlier

1.3.2. Normalisasi Data

Teknik Normalisasi dalam Data Mining digunakan untuk mengurangi rentang nilai suatu atribut, seperti -1.0 hingga 1.0. Normalisasi data terutama digunakan untuk mengurangi data yang berlebihan, sehingga membantu mengurangi ukuran data untuk mempercepat pemrosesan informasi. Dalam kebanyakan kasus, Teknik Normalisasi Data dalam Data Mining diimplementasikan dalam model klasifikasi. Beberapa manfaat yang didapatkan dengan melakukan Normalisasi Data, antara lain:

- Menerapkan Data Mining dalam Kumpulan data yang telah ternormalisasi akan lebih mudah
- Teknik Normalisasi dalam Data Mining yang diterapkan pada kumpulan data yang dinormalisasi memberikan hasil yang lebih akurat dan efektif.
- Ekstraksi data dari database menjadi lebih cepat setelah data distandarisasi.
- Pada data yang dinormalisasi, metode analisis data yang lebih khusus dapat digunakan.

Beberapa Teknik normalisasi data yang sering digunakan dalam data mining, antara lain:

- Min-Max Normalization (Linear Scaling)
- Z-Score Normalization (Standardization)
- Decimal Scaling

a) Min-Max Normalization

Min-Max Normalization atau yang bisa disebut dengan linear Scaling adalah teknik yang digunakan dalam prapemrosesan data untuk mengubah data numerik menjadi skala umum, biasanya antara 0 dan 1. Teknik ini sangat berguna ketika menangani fitur yang memiliki rentang nilai berbeda.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Dimana,

- x' adalah nilai ternormalisasi
- x adalah nilai asli
- $\min(X)$ dan $\max(X)$ adalah nilai minimum dan maksimum pada dataset

b) Z-Score Normalization

Z-Score Normalization, disebut juga Standard Score Normalization atau Standardization teknik statistik yang digunakan untuk mengubah data menjadi distribusi normal standar dengan mean 0 dan standar deviasi 1.

$$z = \frac{x - \mu}{\sigma}$$

Dimana,

- z adalah nilai ternormalisasi
- x adalah nilai data asli
- μ adalah nilai Means (rata-rata)
- σ adalah standar deviasi, didapatkan dari $\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

c) Decimal Scaling

Cara kerja Decimal Scaling:

- Determine the Scaling Factor**, Pilih faktor skala, yang biasanya pangkat 10. Faktor skala menentukan seberapa banyak Anda menggeser koma desimal
- Normalize the Data**, untuk setiap titik data, bagilah dengan faktor skala yang dipilih. Tindakan ini menggeser koma desimal ke kiri jika faktor skala lebih besar dari 1 atau ke kanan jika faktor skala kurang dari 1.

$$x' = \frac{x}{10^j}$$

Dimana

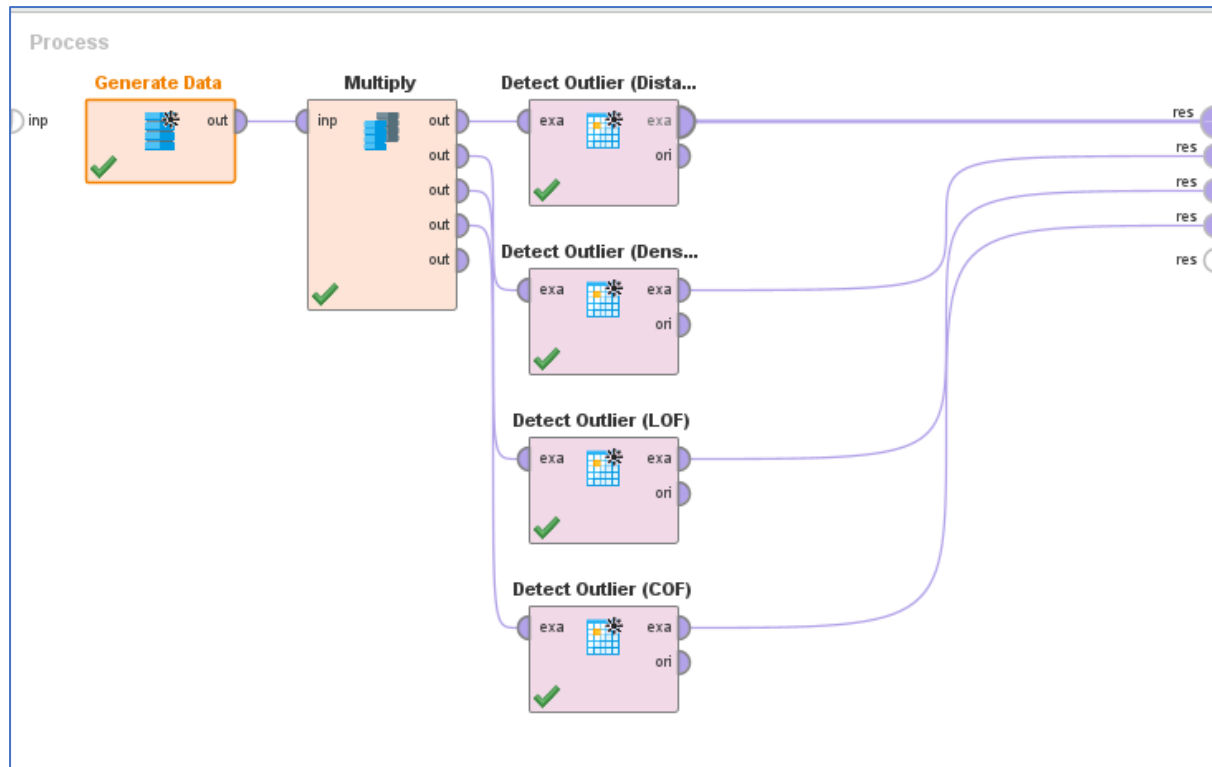
- x' adalah nilai ternormalisasi
- x adalah nilai data asli
- j adalah Jumlah tempat desimal yang diperlukan untuk mewakili nilai terbesar dalam kumpulan data.

1.4. Praktikum

1.4.1. Outlier

Lakukan praktikum sesuai tahapan berikut:

- Buka aplikasi RapidMiner
- Susun Operator sesuai dengan Gambar. Gunakan Operator Detect Outlier (Distances), Detect Outlier (Densities), Detect Outlier (LOF), dan Detect Outlier (COF)



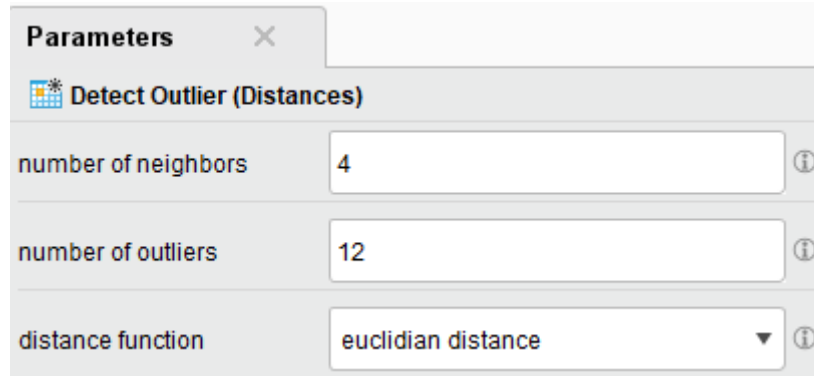
Gambar 1 Penggunaan Operator Detect Outlier

- c. Atur Parameter dari Operator Generate Data sesuai dengan Gambar


Parameters	
Generate Data	
target function	gaussian mixture clusters
number examples	200
number of attributes	2
attributes lower bound	-10.0
attributes upper bound	10.0




Gambar 1. 4 Pengaturan Parameter pada Operator Generate Data

- d. Atur Parameter dari Operator Detect Outlier (Distances) sesuai pada Gambar



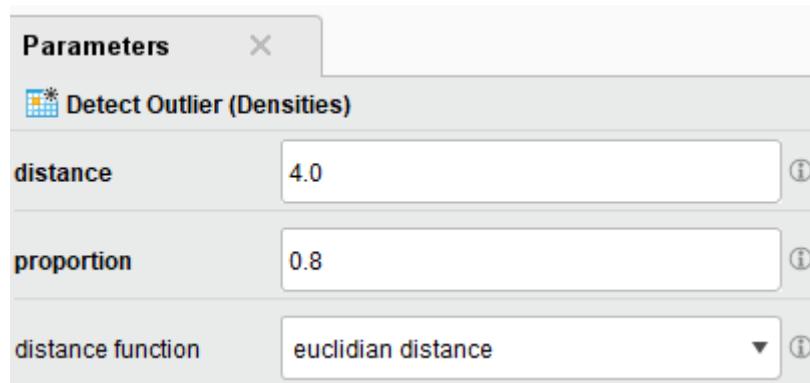
Parameters X

 **Detect Outlier (Distances)**


number of neighbors	4	
number of outliers	12	
distance function	euclidian distance	




Gambar 1. 5 Pengaturan Parameter dari Operator Detect Outlier (Distances)

- e. Atur Parameter dari Operator Detect Outlier (Densities)



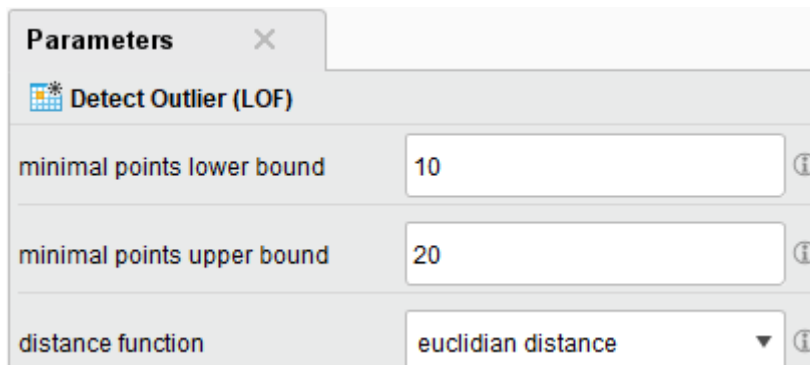
Parameters X

 **Detect Outlier (Densities)**


distance	4.0	
proportion	0.8	
distance function	euclidian distance	




Gambar 1. 6 Pengaturan Parameter Operator Detect Outlier (Densities)

- f. Atur Parameter dari Operator Detect Outlier (LOF)



Parameters X

 **Detect Outlier (LOF)**


minimal points lower bound	10	
minimal points upper bound	20	
distance function	euclidian distance	

Gambar 1. 7 Pengaturan Parameter dari Operator Detect Outlier (LOF)

- g. Atur Parameter dari Operator Detect Outlier (COF)



Parameters ✕

 **Detect Outlier (COF)**

number of neighbors 7 ⓘ

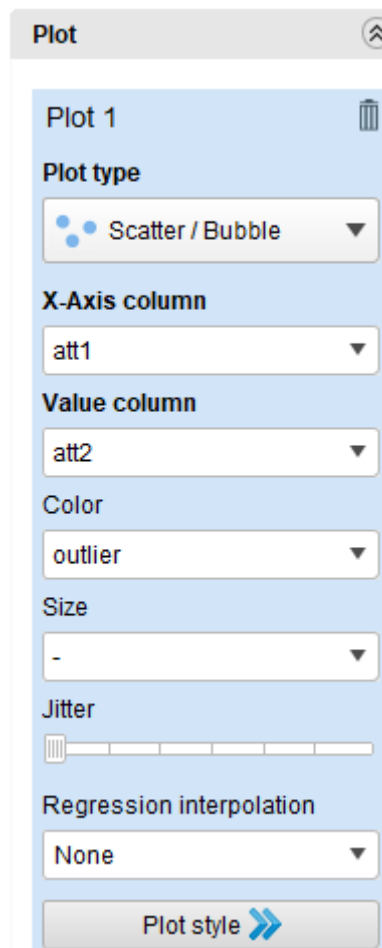
number of class outliers 7 ⓘ

measure types MixedMeasures ⓘ

mixed measure MixedEuclideanDistance ⓘ

Gambar 1. 8 Pengaturan Parameter dari Operator Detect Outlier (COF)


- h. Lakukan Eksekusi dari susunan operator yang telah dilakukan
- i. Untuk setiap hasil (Bagian Result), Pilih Visualizations lalu lakukan pengaturan sesuai pada Gambar untuk **SEMUA HASIL OPERATOR DETECT OUTLIER!**



Plot ⬆

Plot 1 🗑

Plot type

 Scatter / Bubble ▾

X-Axis column

att1 ▾

Value column

att2 ▾


Color

outlier ▾

Size

- ▾

Jitter

 | | | | |

Regression interpolation

None ▾

Plot style ➡➡

Gambar 1. 9 Pengaturan Hasil dari semua operator Detect Outlier

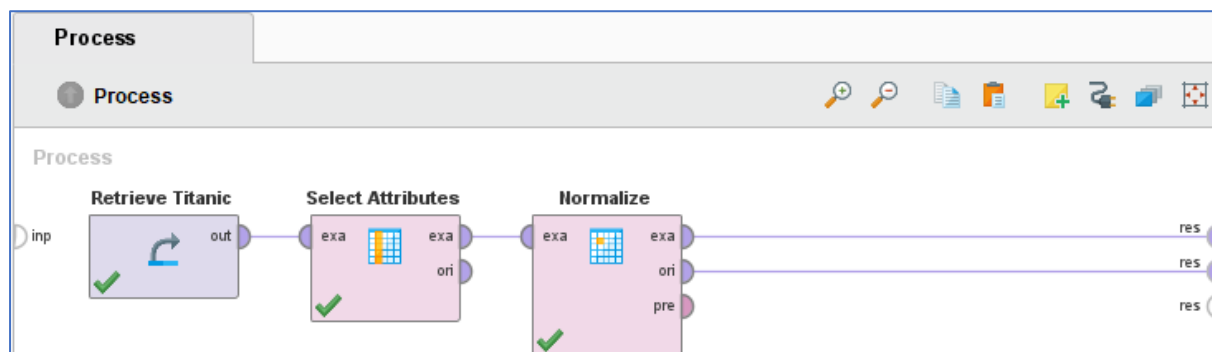
Pertanyaan 1

1. Sertakan gambar grafik Scatter Plot dari masing-masing Detect Oulier!
2. Deskripsikan makna dari masing-masing hasil yang didapatkan berdasarkan dari gambar Scatter Plot!

1.4.2. Normalization Data

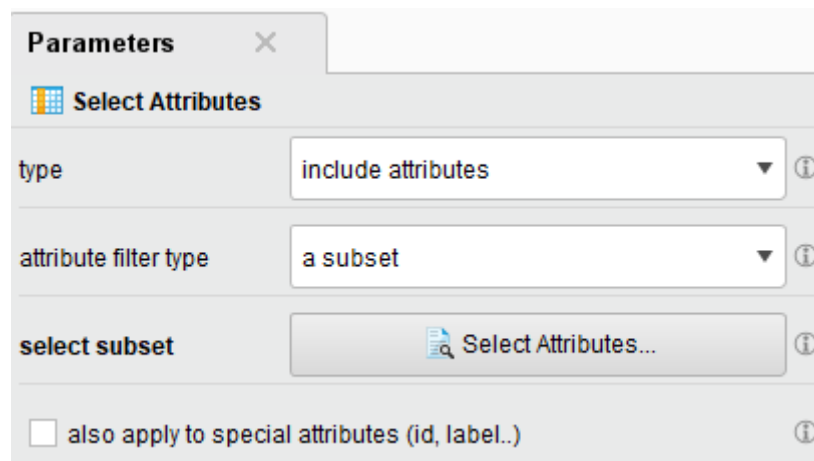
Lakukan Praktikum dengan mengikuti tahapan berikut:

- a. Buka Rapidminer
- b. Susun Operator sesuai pada Gambar.



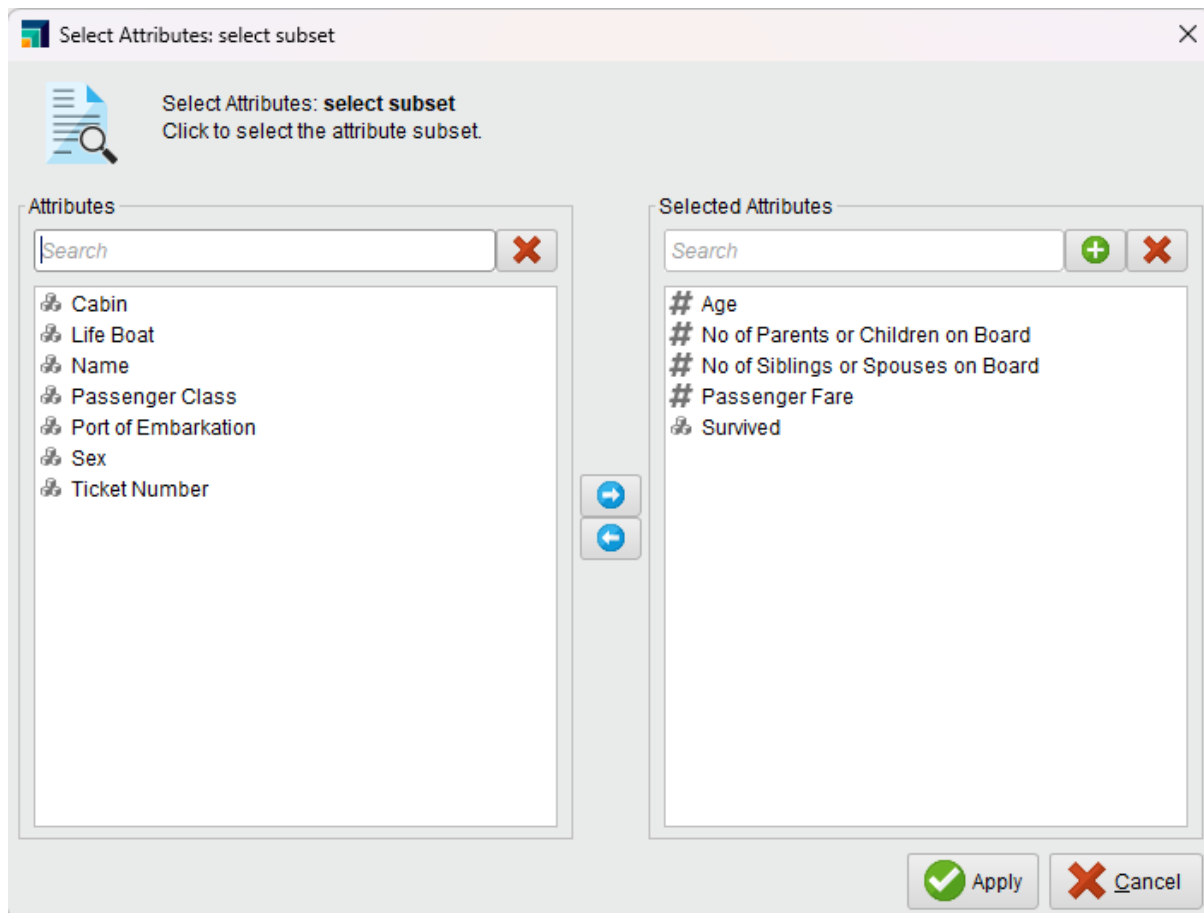
Gambar 1. 10 Operator yang digunakan

- c. Operator Retrieve Titanic didapatkan pada bagian **Repositories** → **Samples** → **Data** → **Titanic**
- d. Operator Select Attributes dan Normalize dapat dicari pada bagian **Operators**
- e. Lakukan pengaturan Parameter pada Operator Select Attribute sesuai pada Gambar



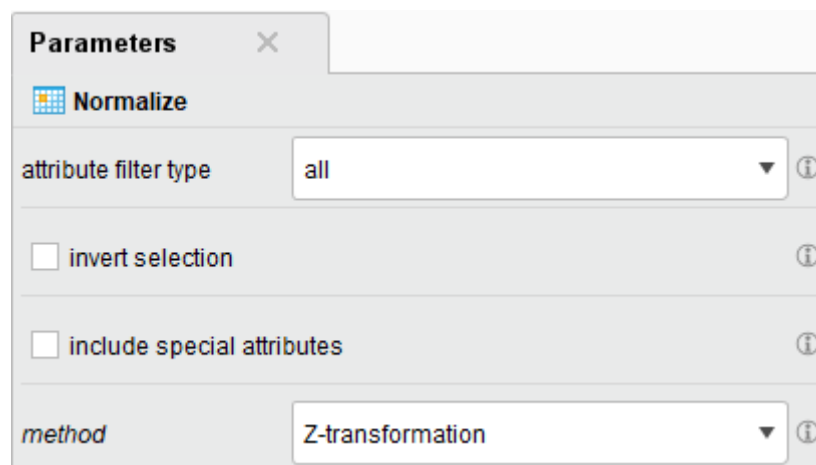
Gambar 1. 11 Pengaturan Parameter pada Operator Select Attribute

f. Tekan pada Select Attributes, lakukan pengaturan sesuai pada Gambar



Gambar 1. 12 Pemilihan Atribut pada Bagian Select Attributes

g. Lakukan pengaturan parameter pada operator Normalize sesuai pada Gambar



Gambar 1. 13 Pengaturan Parameter pada Operator Normalize

h. Eksekusi hasil penyusunan keseluruhan operator.

i. Pada bagian Result, pilih tampilan pada bagian Statistics

Pertanyaan 2

1. Analisis dari luaran dari penyusunan operator yang dilakukan, jelaskan perbedaan antara sebelum dan sesudah dilakukan Normalize berdasarkan hasil pada bagian Statistics! Sertakan gambar dari hasil yang didapatkan!
2. Lakukan percobaan dengan mengubah Parameter dari Operator Normalize dengan method **range tranformation**, jelaskan hasil perbedaan yang didapatkan yang disertai dengan gambar!

1.4.3. Tugas Praktikum

1. Carilah data secara daring yang berhubungan dengan Bisnis!
2. Lakukan Outlier Detection dan Normalize pada Data tersebut!
3. Jelaskan dan Gambarkan hasil dari masing-masing proses Outlier Detection dan Normalize yang anda lakukan!