

Modul 3

Data Preparation

Mata Kuliah Data Mining



Disusun Oleh

Rizky Pratama Yudha

2241760020

Kelas SIB 2A

Jurusan Teknologi Informasi

Program Studi D4 Sistem Informasi Bisnis

POLITEKNIK NEGERI MALANG

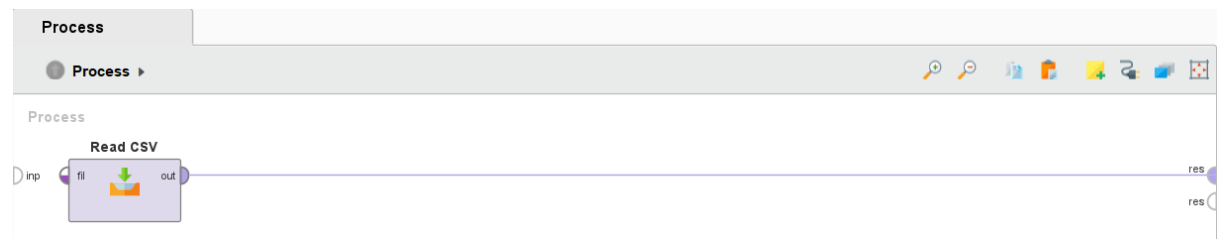
MALANG

2024

Praktikum

- Handling missing value

Menambahkan operator read CSV untuk load data



Hasil run

Result History

ExampleSet (Read CSV)

Open in: Turbo Prep, Auto Model, Interactive Analysis

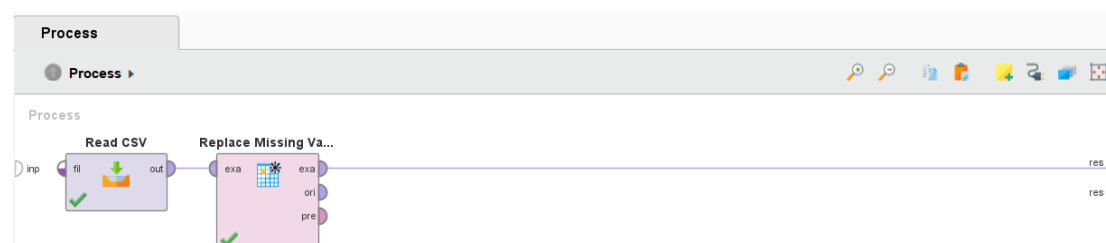
Row No.	ID	Nama	Usia	Jenis Kelamin	Pendapatan
1	1	John	25	Laki-laki	40
2	2	Mary	30	Perempuan	55
3	3	Michael	?	Laki-laki	62
4	4	Sarah	28	Perempuan	?
5	5	David	35	?	70
6	6	Lisa	?	Perempuan	45

Adanya missing pada field usia, jenis kelamin, pendapatan

ExampleSet (Read CSV)

Name	Type	Missing	Statistics	Filter (5 / 5 attributes):
ID	Integer	0	Min: 1, Max: 6, Average: 3.500	
Nama	Nominal	0	Least: Sarah (1), Most: David (1), Values: David (1), John (1), ...[4 more]	
Usia	Integer	2	Min: 25, Max: 35, Average: 29.500	
Jenis Kelamin	Nominal	1	Least: Laki-laki (2), Most: Perempuan (3), Values: Perempuan (3), Laki-laki (2)	
Pendapatan	Real	1	Min: 40, Max: 70, Average: 54.400	

- Menambahkan operator replace missing value



Hasil

ExampleSet (Replace Missing Values)					
Open in Turbo Prep Auto Model Interactive Analysis					
Row No.	ID	Nama	Usia	Jenis Kelamin	Pendapatan
5	5	David	35	Perempuan	70
1	1	John	25	Laki-laki	40
6	6	Lisa	30	Perempuan	45
2	2	Mary	30	Perempuan	55
3	3	Michael	30	Laki-laki	62
4	4	Sarah	28	Perempuan	54.400

ExampleSet (Replace Missing Values)						
Name	Type	Missing	Statistics		Filter (5 / 5 attributes)	Search for Attributes
✓ ID	Integer	0	Min 1	Max 6	Average 3.500	
✓ Nama	Polynomial	0	Least Sarah (1)	Most David (1)	Values David (1), John (1), ...[4 more]	
✓ Usia	Integer	0	Min 25	Max 35	Average 29.667	
✓ Jenis Kelamin	Polynomial	0	Least Laki-laki (2)	Most Perempuan (4)	Values Perempuan (4), Laki-laki (2)	
✓ Pendapatan	Real	0	Min 40	Max 70	Average 54.400	

Pertanyaan :

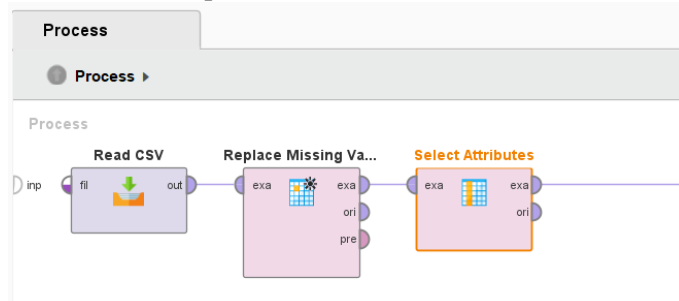
1. Analisis hasil replace Missing value
2. Jelaskan pada hasil replace missing value, data yang kosong digantikan dengan nilai apa!
3. Sebutkan untuk masing-masing kolom yang memiliki missing value (Kolom usia, jenis kelamin, dan pendapatan)
4. Temukan jawabannya berdasarkan informasi pada slide Teori Pertemuan 3!

Jawaban

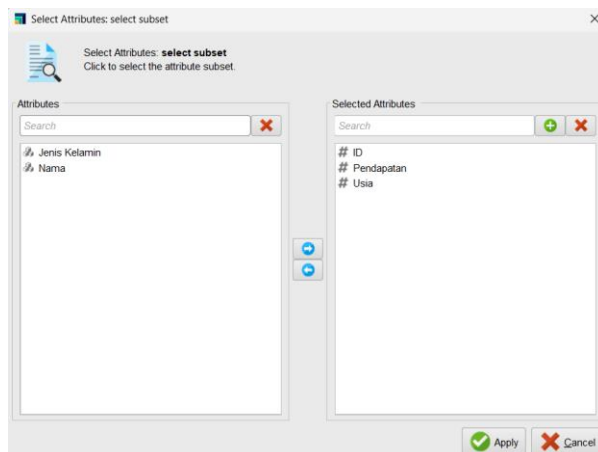
1. Hasil Replace Missing Value mengganti nilai kosong
2. Diangantikan dengan attribute default average yaitu nilai rata rata nilai pada atribut
3. Usia pada row no3 diisi dengan value 30, pendapatan pada row no 4 diisi dengan 54.500
Jenis kelamin pada row no 5 diisi dengan perempuan, usia pada row no 6 diisi dengan 30
Nilai yang dimasukkan berasal dari nilai rata-rata value masing-masing atribut

- Data reduction

Menambahkan operator select attributes untuk melakukan feature selection



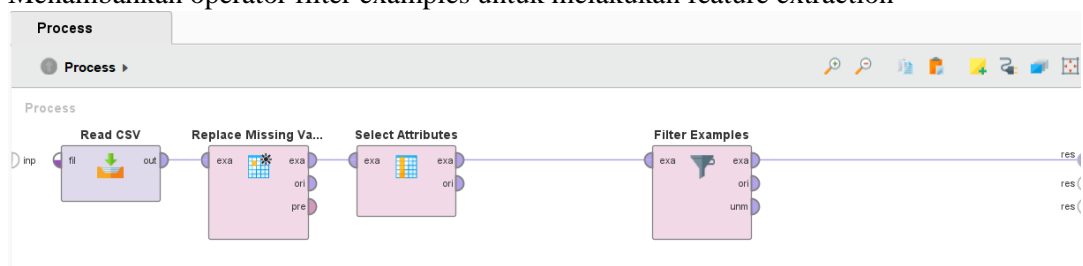
Melakukan pemilihan atribut yang memiliki data berupa angka yakni ID, Pendapatan, Usia



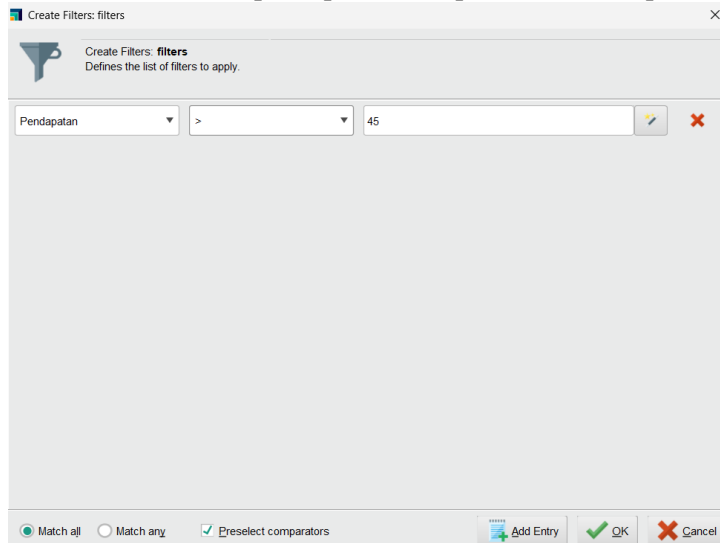
Result berupa atribut yang telah di select berupa angka

ExampleSet (Select Attributes)			
Open in Turbo Prep Auto Model Interactive Analysis			
Row No.	ID	Usia	Pendapatan
1	1	25	40
2	2	30	55
3	3	30	62
4	4	28	54.400
5	5	35	70
6	6	30	45

Menambahkan operator filter examples untuk melakukan feature extraction



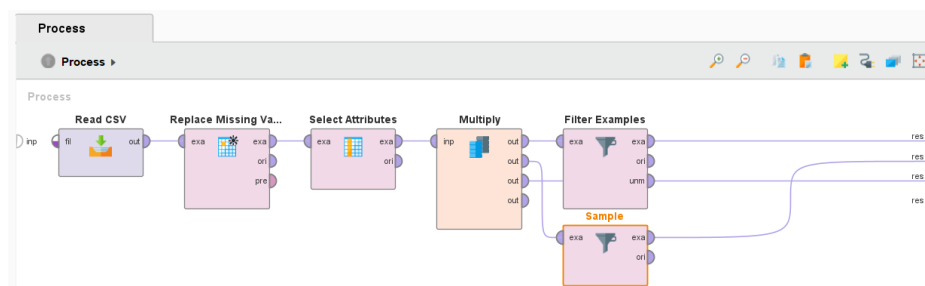
Menambahkan filter pendapatan > 45 pada filter examples



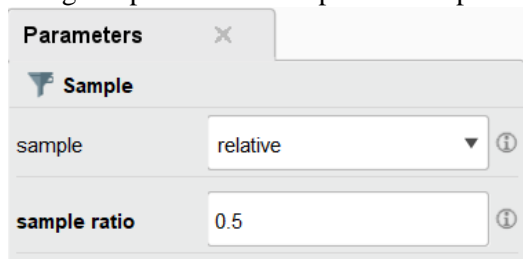
Maka hasil field pendapatan bernilai diatas 45

ExampleSet (Filter Examples)			
Open in Turbo Prep Auto Model Interactive Analysis			
Row No.	ID	Usia	Pendapatan
1	2	30	55
2	3	30	62
3	4	28	54.400
4	5	35	70

Menambahkan operator sample untuk melakukan Sampling



Mengatur parameter dan operator sample



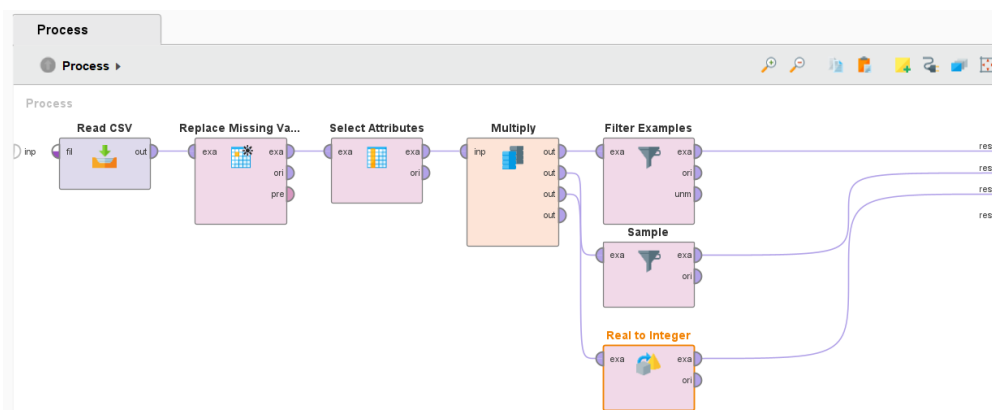
Result

ExampleSet (Select Attributes)ExampleSet (Sample)

Open in Turbo PrepAuto ModelInteractive Analysis

Row No.	ID	Usia	Pendapatan
1	2	30	55
2	4	28	54.400
3	6	30	45

Menambahkan operator real to integer untuk melakukan handling inconsistent data





Result


ExampleSet (Real to Integer)

ExampleSet (Sa

Open in

 Turbo Prep

 Auto Model

 Interactive Analysis

Row No.	Pendapatan	ID	Usia
1	40	1	25
2	55	2	30
3	62	3	30
4	54	4	28
5	70	5	35
6	45	6	30

Pertanyaan 2

1. Analisis setiap hasil dari praktikum Data Reduction
2. Jelaskan hasil analisis anda terhadap hasil feature selection!
3. Jelaskan hasil analisis anda terhadap hasil Feature Extraction!
4. Jelaskan hasil analisis anda terhadap hasil Sampling!
5. Jelaskan hasil analisis anda terhadap hasil Feature Selection!
6. Jelaskan hasil analisis anda terhadap hasil handling Inconsistent Data!

Jawaban

1. Praktikum Data Reduction melakukan proses mengurangi jumlah dan kompleksitas data untuk memudahkan pemahaman, dan juga pengambilan keputusan seperti selection untuk

memilih atribut mana yang akan digunakan, feature extraction filter samples untuk memfilter suatu fields, sampling, real to integer

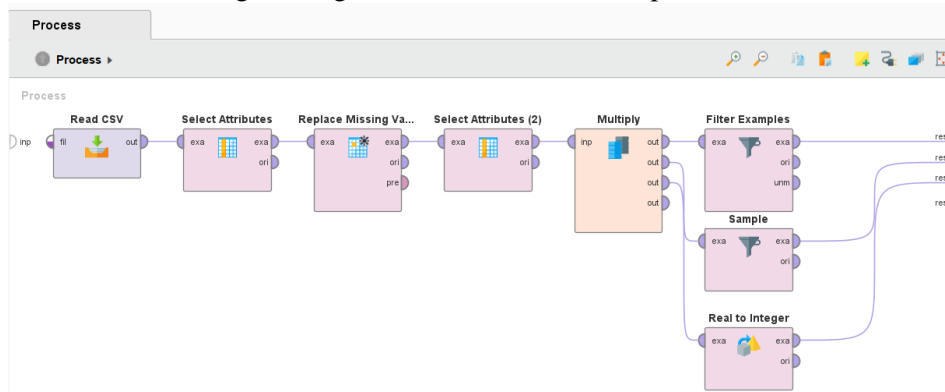
2. Feature selection – selection attributes digunakan untuk seleksi sesuai dengan parameter yang di adjust, dalam praktikum memilih atribut yang memiliki value angka yakni id, pendapatan, usia. Sehingga result yang ditampilkan berupa berupa field yang ditentukan.
3. Feature extraction-filter examples untuk memfilter value dari sesuai fields atau atribut dari pendapatan
4. Sampling untuk memilih Sebagian kecil dari suatu populasi yang mewakili keseluruhan bergantung pada tujuan da jenis sampel yang digunakan.
5. Digunakan untuk seleksi dengan parameter yang di adjust
6. Real to integer untuk melakukan handling consistent data pada valuenya sehingga tidak ada nilai ribuan pada value pendepaan

TUGAS PRAKTIKUM

1. Carilah data yang berhubungan dengan bisnis di internet!
2. Lakukan handling missing value dan/atau data reduction pada data tersebut!
3. Jelaskan dan gambarkan hasil dari masing-masing proses missing value dan/atau data reduction yang anda lakukan
4. Lakukan hal yang sesuai pada materi kali ini pada dataset titanic!

Jawaban

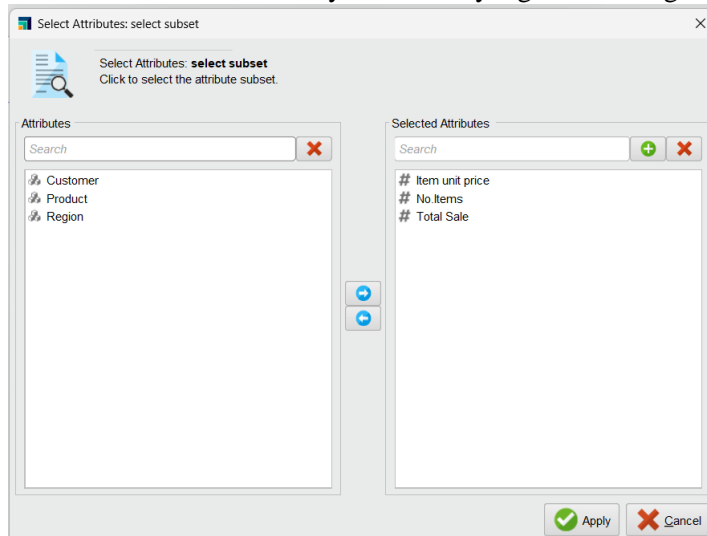
1. Menggunakan data CarDistributionSales
2. Melakukan handling missing value dan data reduction pada CarDistributionSales.csv



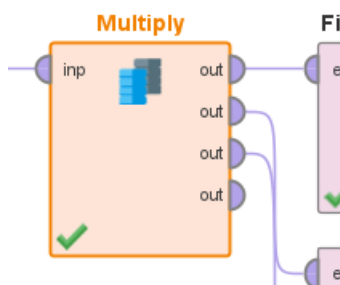
3. Replace missing value dengan parameter default untuk menggantikan data null



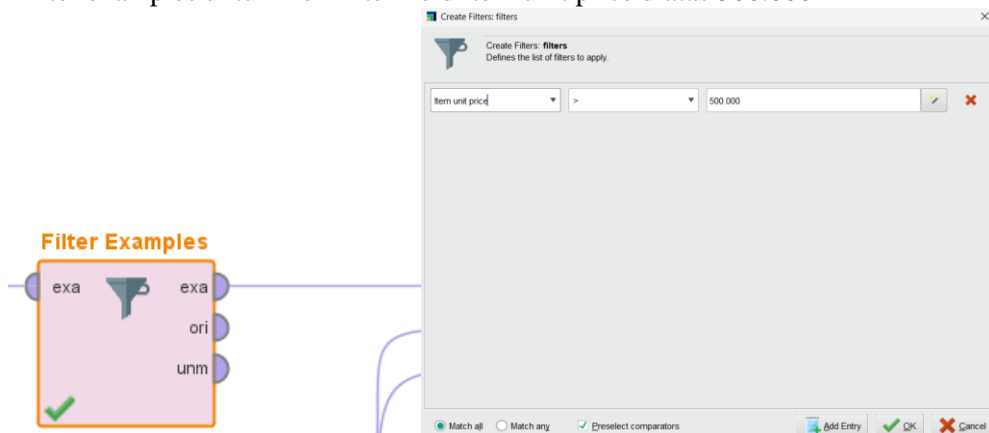
Select attributes untuk menyeleksi data yang bernilai angka



Multiply untuk membuat cabang dari output yang dihasilkan



Filter examples untuk memfilter field item unit price diatas 500.000



ExampleSet (Real to Integer) × ExampleSet (Sample) × ExampleSet (Filter Examples) ×			
Open in Turbo Prep Auto Model Interactive Analysis			
Row No.	Item unit pri...	No.Items	Total Sale
1	799.950	1	799.950
2	799.950	7	5599.650
3	799.950	8	6399.600
4	799.950	2	1599.900
5	799.950	9	7199.550
6	799.950	14	11199.300

Sample

Parameters

Sample

sample: relative

sample ratio: 0.5

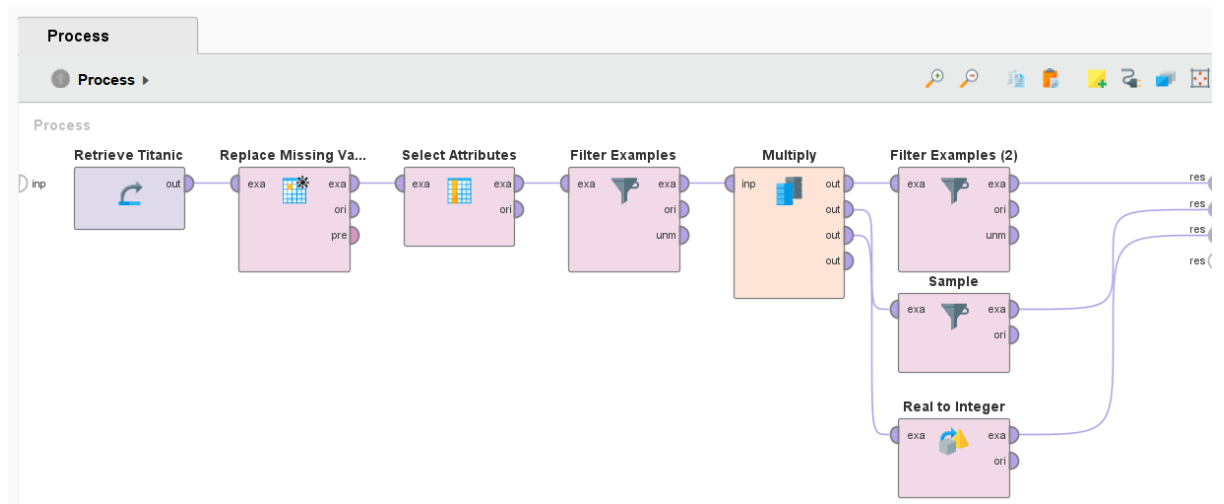
ExampleSet (Real to Integer) × ExampleSet (Sample) ×			
Open in Turbo Prep Auto Model Interactive Analysis			
Row No.	Item unit pri...	No.Items	Total Sale
1	340.950	6	2045.700
2	168.950	9	1520.550
3	799.950	7	5599.650
4	168.950	15	2534.250
5	168.950	15	2534.250
6	168.950	14	2385.300
7	340.950	2	681.900
8	340.950	7	2386.650
9	340.950	4	1363.800

Real to integer untuk mengubah nilai menjadi integer

Real to Integer

ExampleSet (Real to Integer)			
ExampleSet (\$			
Open in Turbo Prep Auto Model Interactive Analysis			
Row No.	Item unit pri...	Total Sale	No.Items
1	340	2045	6
2	799	799	1
3	168	1182	7
4	168	337	2
5	168	1520	9
6	799	5599	7
7	799	6399	8
8	168	2534	15
9	168	2534	15
10	799	1599	2
11	168	2365	14
12	340	681	2
13	340	2386	7
14	799	7199	9
15	168	1689	10
16	340	1363	4
17	799	11199	14
18	168	2027	12

4. Data Titanic



ExampleSet (Real to Integer)

ExampleSet (Sample)

Open in

Turbo Prep

Auto Model

Interactive Analysis

Row No.	Age	No of Sibilin...	No of Parent...	Passenger F...
1	29	0	0	211.338
2	2	1	2	151.550
3	30	1	2	151.550
4	25	1	2	151.550
5	48	0	0	26.550
6	63	1	0	77.958
7	53	2	0	51.479
8	71	0	0	49.504
9	47	1	0	227.525
10	18	1	0	227.525
11	24	0	0	69.300
12	26	0	0	78.850
13	80	0	0	30
14	29	0	0	25.925
15	24	0	1	247.521
16	50	0	1	247.521
17	32	0	0	76.292
18	36	0	0	75.242

ExampleSet (1,270 examples, 0 special attributes, 4 regular attributes)

ExampleSet (1,270 examples,0 special attributes,4 regular attributes)

Open in		Turbo Prep	Auto Model	Interactive Analysis
Row No.	Age	No of Sibilin...	No of Parent...	Passenger F...
1	26	0	0	78.850
2	29.881	0	0	25.925
3	25	1	0	91.079
4	22	0	1	55
5	29.881	0	0	26.550
6	36	1	2	120
7	29.881	0	1	55
8	27	0	0	30.500
9	54	1	1	81.858
10	53	0	0	28.500
11	19	0	0	30
12	38	0	0	80
13	42	0	0	26.550
14	30	0	0	93.500
15	52	0	0	30.500
16	33	0	0	26.550
17	23	1	0	82.267
18	39	1	1	110.883

ExampleSet (Real to Integer)		ExampleSet (Sample)		ExampleSet (Filter Examples (2))	
Open in		Turbo Prep	Auto Model	Interactive Analysis	Filter (1,270 /
Row No.	Age	No of Sibilin...	No of Parent...	Passenger F...	
1	29	0	0	211.338	
2	2	1	2	151.550	
3	30	1	2	151.550	
4	25	1	2	151.550	
5	48	0	0	26.550	
6	63	1	0	77.958	
7	53	2	0	51.479	
8	71	0	0	49.504	
9	47	1	0	227.525	
10	18	1	0	227.525	
11	24	0	0	69.300	
12	26	0	0	78.850	
13	80	0	0	30	
14	29.881	0	0	25.925	
15	24	0	1	247.521	
16	50	0	1	247.521	
17	32	0	0	76.292	
18	36	0	0	75.242	

ExampleSet (1,270 examples,0 special attributes,4 regular attributes)