Modul 6

Detect Outlier

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

Latihan 1 :

1. Perhatikan dataset Titanic, menurut anda kira-kira atribut mana yang memiliki oulier selain atribut age?

- Selain atribut age adalah no of sibling or spouses on board, no of parents or children on board, passanger fare

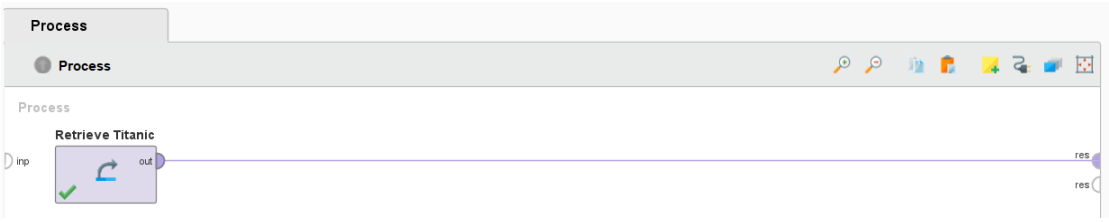
Latihan 2 :

1. Kira-kira atribut mana yang tidak diperlukan dalam menentukan model prediktif penumpang yang selamat?

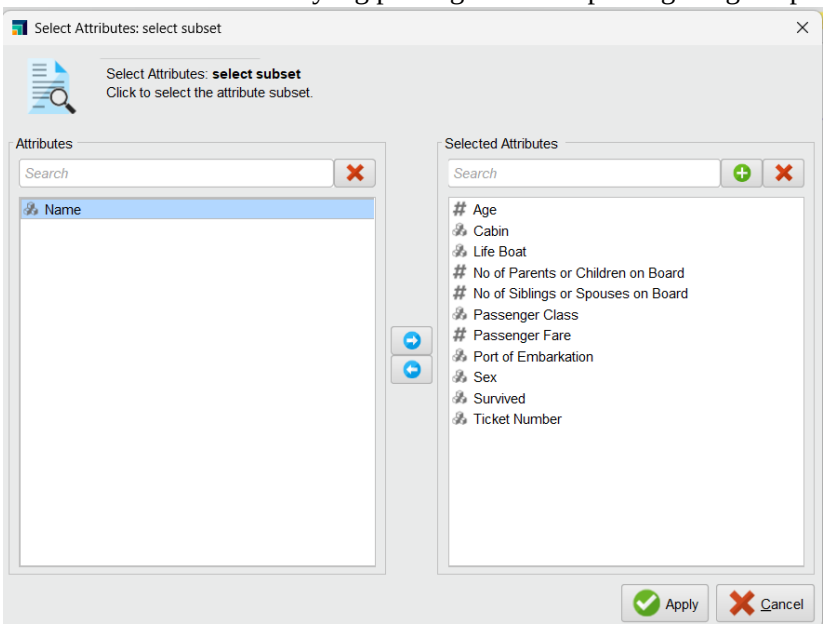
- Atribut nama

Langkah-langkah praktikum

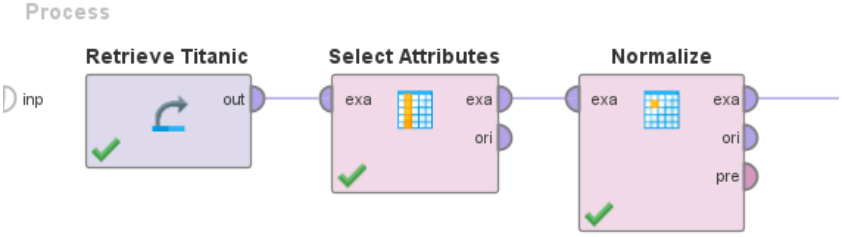
- Menambahkan data titanic



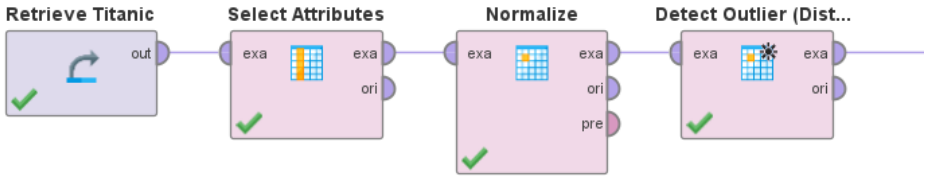
- Menentukan atribut mana yang penting dan tidak penting dengan operator select attributes



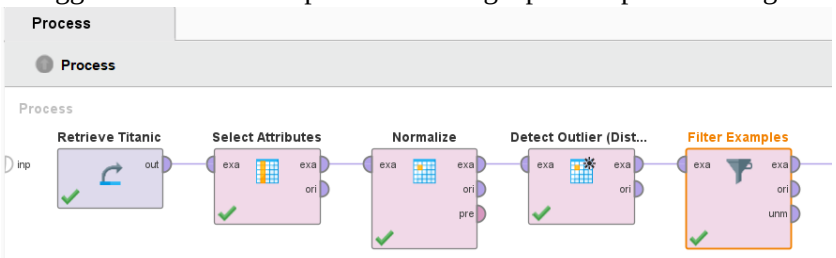
- Menambahkan operator normalize



- Menambahkan operator detect outliers (distance)



- Menggunakan filter examples untuk menghapus sample data dengan outlier

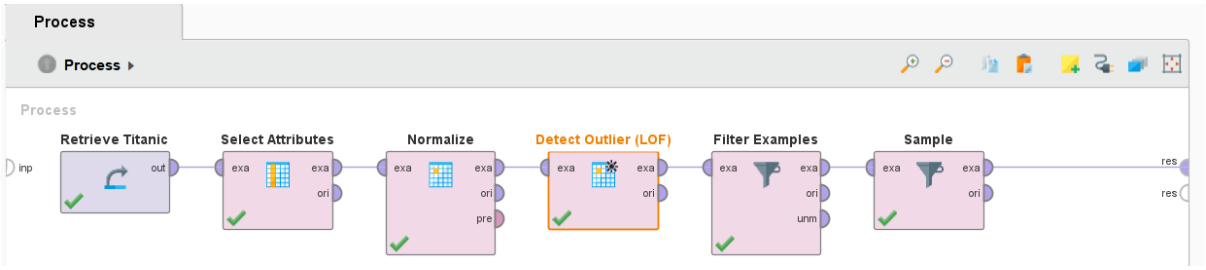


Open in Turbo Prep Auto Model

Row No.	outlier	Age
1	false	1.188
2	false	-0.824
3	false	-0.408
4	false	-0.269
5	false	3.477
6	false	?
7	false	-0.408
8	false	1.396
9	false	0.147
10	false	0.425
11	false	0.494
12	false	1.188
13	false	-0.269
14	false	0.841
15	false	-0.061
16	false	-0.339
17	false	-0.339

Latihan Praktikum 1

1. Pada data titanic



ExampleSet (Sample)

Open in Turbo Prep Auto Model Interactive Analysis Filter (20 / 20 examples): all

Row No.	outlier	Age	No of Siblin...	No of Parent...	Passenger F...	Passenger ...	Sex	Ticket Numb...	Cabin	Port of Em
15	2.535	0.008	-0.479	-0.445	1.163	First	Female	12749	B73	Southampto
20	2.077	-0.408	-0.479	1.866	-0.321	Third	Female	PP 9549	G6	Southampto
19	1.761	-0.339	-0.479	-0.445	-0.504	Third	Male	2654	F E57	Cherbourg
16	1.647	0.910	-0.479	0.710	3.440	First	Female	24160	B3	Southampto
18	1.282	0.286	-0.479	-0.445	-0.440	Second	Female	C.A. 34260	F33	Southampto
4	1.223	0.980	-0.479	-0.445	-0.108	First	Female	PC 17610	B4	Cherbourg
5	1.202	0.425	-0.479	0.710	9.255	First	Male	PC 17755	B51 B53 B55	Cherbourg
14	1.163	0.563	0.481	-0.445	1.096	First	Male	19943	C93	Southampto
11	1.116	0.355	0.481	-0.445	0.383	First	Female	113803	C123	Southampto
3	1.102	1.951	-0.479	-0.445	-0.130	First	Female	113783	C103	Southampto
2	1.071	0.008	-0.479	-0.445	2.542	First	Female	36928	C7	Southampto
10	1.053	2.090	0.481	4.176	4.438	First	Female	19950	C23 C25 C27	Southampto
12	1.042	-0.755	-0.479	-0.445	-0.064	First	Female	112053	B42	Southampto
7	1.042	0.425	0.481	1.866	1.675	First	Female	113760	B96 B98	Southampto
6	1.040	-1.102	0.481	1.866	1.675	First	Female	113760	B96 B98	Southampto
8	1.030	1.049	-0.479	-0.445	-0.069	First	Male	PC 17594	A9	Cherbourg
13	1.029	1.049	-0.479	0.710	0.581	First	Female	PC 17759	D10 D12	Cherbourg

ExampleSet (20 examples,1 special attribute,11 regular attributes)

Pada data mahasiswa :

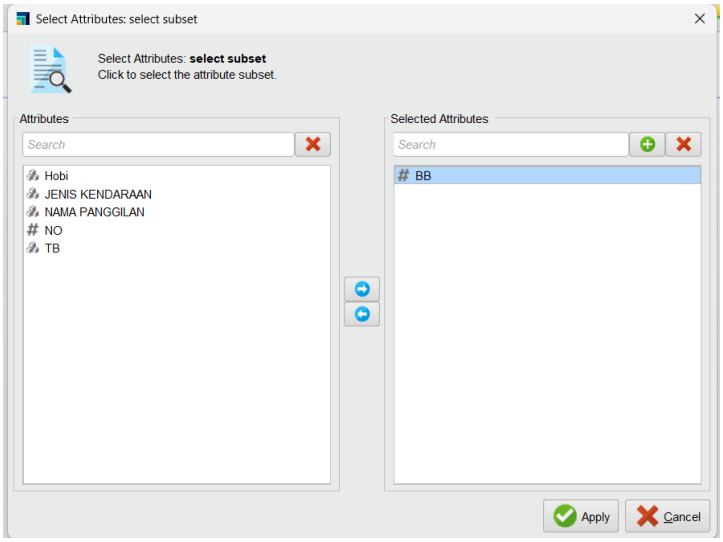
ExampleSet (Read CSV)

Open in Turbo Prep Auto Model Interactive Analysis

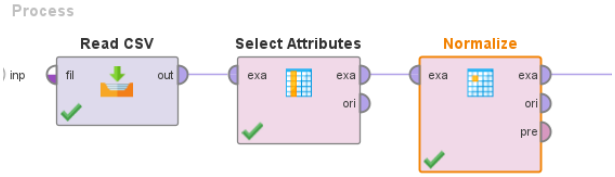
Row No.	NO	NAMA PANG...	Hobi	TB	JENIS KEN...	BB
1	1	Danni	Art, Game Dev	-	Jalan Kaki	700
2	2	Adam	Game	170	Sepeda Motor	50
3	3	Afin	?	?	?	?
4	4	Ifi	?	?	?	?
5	5	Albi	?	?	?	?
6	6	Alo	Menonton Film	170	Sepeda Motor	63
7	7	Angel	Mendengarkan...	154	Sepeda motor	43
8	8	Mita	Memasak	155	Sepeda motor	55
9	9	Aryod	Game	170	Sepeda Motor	48
10	10	Nasya	Bermain puzzle	160	Sepeda Motor	44
11	11	Daffa	Memasak	173	Sepeda Motor	69
12	12	Fia	Memasak	15	Sepeda Motor	533
13	13	Iqbal	Olahraga	170	Sepeda Motor	63
14	14	Fannisa	Mendengarkan...	160	Sepeda Motor	50
15	15	Khoirul	Game	152	Sepeda Motor	42
16	16	Fikri	Lari	170	Sepeda Motor	82
17	17	Dona	?	?	?	?
18	18	bima	?	?	?	?

- Manipulasi data yang dilakukan sebanyak 6 data

- Atribut yang hanya ditambahkan BB



- Menambahkan normalize



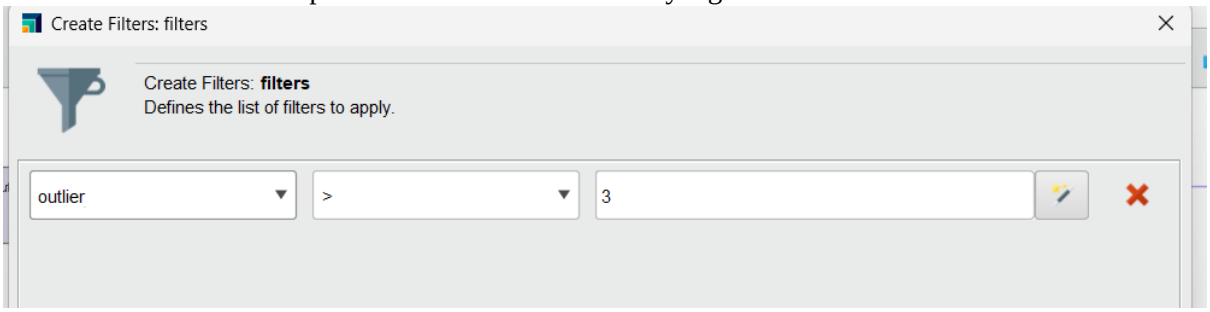
Row No.	BB
1	2.971
2	-0.376
3	?
4	?
5	?
6	-0.309
7	-0.412
8	-0.351
9	-0.387
10	-0.407
11	-0.279
12	2.111
13	-0.309
14	-0.376
15	-0.418
16	-0.212
17	?
18	?

ExampleSet (29 examples,0 special attri

- Menambahkan detect outlier (LOF)

Row No.	outlier	BB
1	50.175	2.971
2	0.946	-0.376
3	0	?
4	0	?
5	0	?
6	1.157	-0.309
7	1.244	-0.412
8	1.103	-0.351
9	0.956	-0.387
10	1.102	-0.407
11	2.079	-0.279
12	36.805	2.111
13	1.157	-0.309
14	0.946	-0.376
15	1.340	-0.418
16	3.766	-0.212
17	0	?
18	0	?

- Terlihat outlier dengan nilai diatas 3 merupakan nilai yang tidak semestinya, maka menambahkan filter examples untuk seleksi data outlier yang lebih dari 3



Hasil :

Row No.	outlier	BB
1	50.175	2.971
2	36.805	2.111
3	3.766	-0.212
4	3.665	-0.593
5	3.988	-0.613
6	34.342	1.941
7	3.988	-0.613

2. Deteksi outlier mengunakan pyhton

Mendefinisikan data yang akan digunakan

+ Code + Text

```
[ ]  data = [5,7,9,3,4,-20,3,8,8,6,90,7,56]
```

+ Code + Text

yang termasuk outlier adalah -20, 90, 56

menampilkan data yang termasuk outlier

```
low_out = []
high_out = []

for i in data:
    if (i < min_IQR):
        low_out.append(i)
    if (i > max_IQR):
        high_out.append(i)

print('Low outlier : ', low_out)
print('High outlier : ', high_out)
```

```
Low outlier :  [-20]
High outlier :  [90, 56]
```

maka ditemukan nilai outlier

Link collab :
https://github.com/rizkypratamayudha/dataMining/blob/main/detectOutliers.ipynb

Link metode yang dipakai untuk menemukan oulier:
https://ilmudatapy.com/menemukan-outlier-dengan-python/#google_vignette

kesimpulan metode atau cara yang digunakan :

- Cara yang digunakan dalam menentukan outlier dalam teks tersebut adalah dengan menggunakan metode interquartile range (IQR) metode ini melibatkan beberapa langkah, antara lain :
  1. Menghitung nilai Q1 (kuartil pertama) dan Q3 (kuartil ketiga) dari data menggunakan fungsi quantile() dari Numpy.
  2. Menghitung selisih antara Q3 dan Q1 untuk mendapatkan nilai IQR.
  3. Menghitung nilai IQR minimum dan maksimum dengan mengalikan IQR dengan 1.5 (konstanta untuk menemukan outliers).
  4. Mencari nilai minimum dan maksimum dari data.
  5. Membuat kondisi untuk mendefinisikan outlier berdasarkan perbandingan antara nilai minimum dan maksimum data dengan nilai IQR minimum dan maksimum.
  6. Mengidentifikasi dan menampilkan data yang termasuk dalam kategori outlier berdasarkan kondisi yang telah ditentukan.