

Predmet: Dubinska Analiza Podataka

Student: Masovic Haris

Indeks: 1689/17993

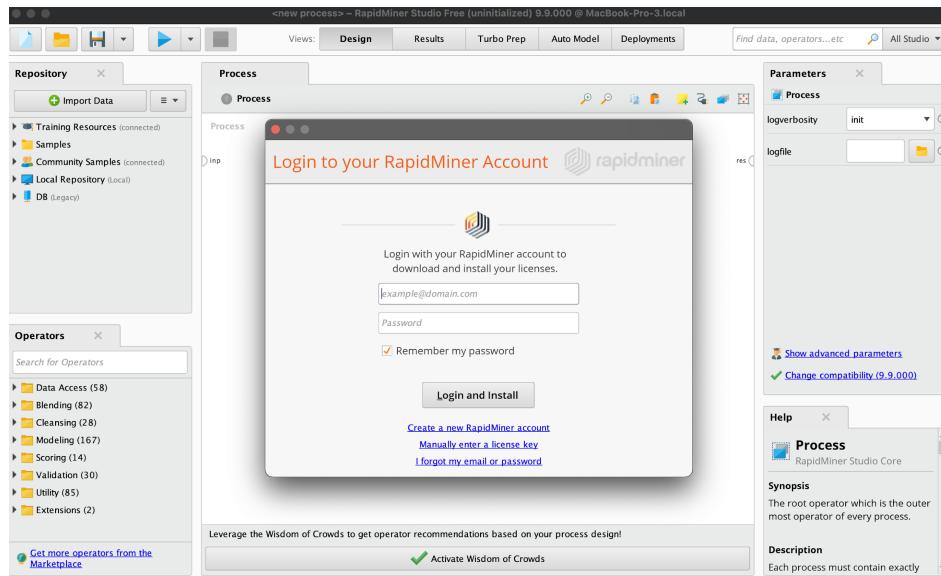
## Laboratorijska vježba 3

# Zadatak 1

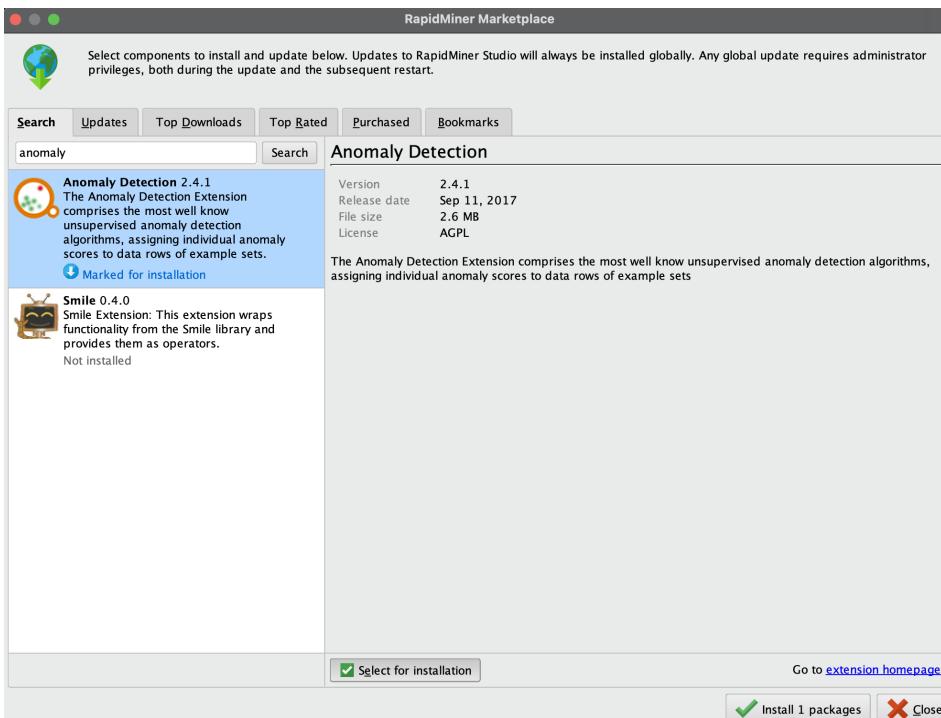
Izvršiti korake koji su prikazani u laboratorijskoj vježbi – instalirati okruženje i ekstenziju za detekciju anomalija, kreirati predefinisani projekat i izvršiti detekciju anomalija nad datasetom.

## 1. Konfiguracija Rapid Miner okruženja

Nakon instalacije i pracenja koraka dobijamo:



I instaliramo anomaly detection:



Restartovanjem studia smo zavrsili instalaciju.

## 2. Odabir dataseta za detekciju anomalija

Nakon skidanja dataset-a i cutanja na 1k elemenata (1k cisto radi specifikacije laptopa), imamo sljedeće stanje:

The screenshot shows the RapidMiner Studio interface. The main window displays a table titled 'ExampleSet // /Local Repository/data/1k-pendigits'. The table has 20 rows and 10 columns, labeled Row No., att1 through att9. The data consists of various numerical values. On the left, there's a sidebar with tabs for Data, Statistics, Visualizations, and Annotations. The 'Data' tab is selected. On the right, there's a 'Repository' panel showing a tree structure with 'Training Resources' (connected), 'Community Samples' (connected), 'Samples', and a 'Local Repository (local)' section containing 'Connections', 'data' (with '1k-pendigits' selected), 'processes', and 'DB (Legacy)'. The status bar at the bottom indicates '[1] Process 13:25 ↗ [1] Local Correlation Integeral (LOCI) 13:21'.

## 3. Kreiranje projekta za detekciju anomalija

Kreiranjem i pokretanjem projekta nad nasim dataset-om imamo rezultate x-means:

# Cluster Model

```
Cluster 0: 211 items
Cluster 1: 200 items
Cluster 2: 256 items
Cluster 3: 333 items
Total number of items: 1000
```

Nakon toga rezultat filtriranja dataseta od anomalija:

Row No.	id	cluster	outlier	att1	att2	att3	att4	att5	att6
1	1	cluster_1	1.035	47.000	100.000	27.000	81.000	57.000	37.000
2	2	cluster_2	1.114	0	89.000	27.000	100.000	42.000	75.000
3	3	cluster_3	1.031	0	57.000	31.000	68.000	72.000	90.000
4	4	cluster_0	1.156	0	100.000	7.000	92.000	5.000	68.000
5	5	cluster_3	1.118	0	67.000	49.000	83.000	100.000	100.000
6	7	cluster_1	1.096	0	100.000	3.000	72.000	26.000	35.000
7	9	cluster_1	0.973	13.000	89.000	12.000	50.000	72.000	38.000
8	10	cluster_0	1.017	57.000	100.000	22.000	72.000	0	31.000
9	11	cluster_3	1.065	74.000	87.000	31.000	100.000	0	69.000
10	12	cluster_1	0.996	48.000	96.000	62.000	65.000	88.000	27.000
11	13	cluster_3	1.110	100.000	100.000	72.000	99.000	36.000	78.000
12	14	cluster_3	1.014	91.000	74.000	54.000	100.000	0	87.000
13	15	cluster_2	1.076	0	85.000	38.000	100.000	81.000	88.000
14	16	cluster_3	1.127	35.000	76.000	57.000	100.000	100.000	92.000
15	17	cluster_3	1.002	50.000	84.000	66.000	100.000	75.000	75.000
16	18	cluster_3	0.998	99.000	80.000	63.000	100.000	25.000	76.000
17	19	cluster_2	1.063	24.000	66.000	43.000	100.000	59.000	65.000
18	20	cluster_2	1.096	0	73.000	19.000	99.000	72.000	100.000
19	21	cluster_1	1.118	12.000	77.000	20.000	62.000	78.000	40.000
20	22	cluster_3	1.117	0	46.000	49.000	64.000	78.000	87.000

ExampleSet (920 examples, 3 special attributes, 17 regular attributes)

I anomalije kao rezultat filtriranja podataka:

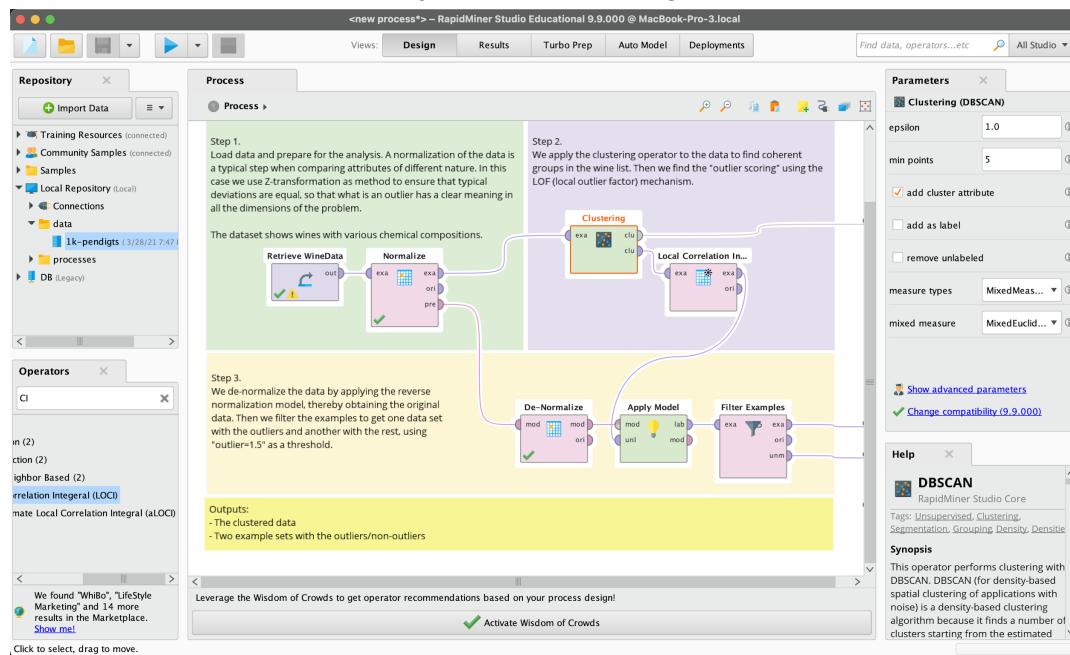
Row No.	id	cluster	outlier	att1	att2	att3	att4	att5	att6
1	6	cluster_2	1.523	100.000	100.000	88.000	99.000	49.000	74.000
2	8	cluster_1	1.716	0	39.000	2.000	62.000	11.000	5.000
3	30	cluster_2	2.256	100.000	84.000	31.000	100.000	0	88.000
4	31	cluster_3	1.669	32.000	59.000	53.000	100.000	100.000	95.000
5	35	cluster_1	2.265	0	0	31.000	15.000	63.000	30.000
6	61	cluster_2	1.544	57.000	63.000	100.000	100.000	0	91.000
7	94	cluster_3	1.784	3.000	96.000	53.000	100.000	81.000	68.000
8	118	cluster_2	1.918	100.000	77.000	92.000	100.000	49.000	76.000
9	135	cluster_0	1.698	100.000	70.000	58.000	100.000	10.000	75.000
10	140	cluster_1	1.985	73.000	85.000	31.000	65.000	0	0
11	151	cluster_3	1.565	29.000	38.000	61.000	60.000	92.000	83.000
12	155	cluster_3	1.896	88.000	80.000	12.000	80.000	100.000	100.000
13	161	cluster_3	1.750	0	100.000	73.000	97.000	95.000	80.000
14	176	cluster_3	1.629	0	70.000	53.000	88.000	64.000	100.000
15	182	cluster_3	1.568	0	91.000	46.000	100.000	85.000	80.000
16	214	cluster_3	1.760	100.000	100.000	92.000	85.000	75.000	71.000
17	236	cluster_3	2.467	14.000	55.000	40.000	61.000	59.000	100.000
18	247	cluster_0	1.837	77.000	100.000	53.000	90.000	0	43.000
19	248	cluster_1	2.201	0	0	38.000	13.000	71.000	34.000
20	257	cluster_3	1.668	100.000	100.000	14.000	90.000	74.000	55.000

ExampleSet (80 examples, 3 special attributes, 17 regular attributes)

## Zadatak 2

Promijeniti metodu detekcije na LOCI, a za model clusteringa odabratи DBSCAN. Usporediti razliku u broju i strukturi detektovanih anomalija – da li je broj podataka pribliжno isti i da li je većina istih instanci oznaчena anomalijama?

Odnos podataka i anomalija je u prethodnom zadatku 920 - 80 (broj podataka - broj anomalija). Postavimo sada metodu detekcije i model clusteringa:



Odnos broja podataka i anomalija nije isti, i sad iznosi 693 - 307. Clustering model je prikazan na sljedecoj slici:

### Cluster Model

```
Cluster 0: 643 items
Cluster 1: 25 items
Cluster 2: 77 items
Cluster 3: 67 items
Cluster 4: 6 items
Cluster 5: 19 items
Cluster 6: 5 items
Cluster 7: 62 items
Cluster 8: 18 items
Cluster 9: 16 items
Cluster 10: 12 items
Cluster 11: 7 items
Cluster 12: 8 items
Cluster 13: 5 items
Cluster 14: 7 items
Cluster 15: 5 items
Cluster 16: 7 items
Cluster 17: 5 items
Cluster 18: 6 items
Total number of items: 1000
```

## Rezultat filtriranja dataset:

**Result History**

ExampleSet (Filter Examples)    ExampleSet (Filter Examples)    Cluster Model (Clustering)

Views: Design    Results    Turbo Prep    Auto Model    Deployments

Find data, operators...etc    All Studio

Data    Statistics    Visualizations    Annotations

Show the data in a table

	cluster	outlier	att1	att2	att3	att4	att5	att6
1	cluster_0	1.369	47.000	100.000	27.000	81.000	57.000	37.000
2	cluster_0	0.013	0	89.000	27.000	100.000	42.000	75.000
3	cluster_15	0.628	0	57.000	31.000	68.000	72.000	90.000
4	cluster_0	1.252	0	100.000	7.000	92.000	5.000	68.000
5	cluster_0	0.920	0	67.000	49.000	83.000	100.000	100.000
6	cluster_0	1.053	0	100.000	3.000	72.000	26.000	35.000
7	cluster_1	1.401	13.000	89.000	12.000	50.000	72.000	38.000
8	cluster_0	0.980	57.000	100.000	22.000	72.000	0	31.000
9	cluster_0	1.426	74.000	87.000	31.000	100.000	0	69.000
10	cluster_0	1.494	100.000	100.000	72.000	99.000	36.000	78.000
11	cluster_0	1.167	91.000	74.000	54.000	100.000	0	87.000
12	cluster_0	0.889	0	85.000	38.000	100.000	81.000	88.000
13	cluster_2	1.364	35.000	76.000	57.000	100.000	100.000	92.000
14	cluster_2	0.318	50.000	84.000	66.000	100.000	75.000	75.000
15	cluster_0	0.922	99.000	80.000	63.000	100.000	25.000	76.000
16	cluster_3	0.930	24.000	66.000	43.000	100.000	59.000	65.000
17	cluster_0	0.962	0	73.000	19.000	99.000	72.000	100.000
18	cluster_0	1.455	12.000	77.000	20.000	62.000	78.000	40.000
19	cluster_0	1.499	0	46.000	49.000	64.000	78.000	87.000
20	cluster_0	1.243	10.000	86.000	34.000	66.000	68.000	34.000

ExampleSet (693 examples, 3 special attributes, 17 regular attributes)

## I anomalije kao rezultat:

**Result History**

ExampleSet (Filter Examples)    ExampleSet (Filter Examples)    Cluster Model (Clustering)

Views: Design    Results    Turbo Prep    Auto Model    Deployments

Find data, operators...etc    All Studio

Data    Statistics    Visualizations    Annotations

Show the data in a table

Row No.	id	cluster	outlier	att1	att2	att3	att4	att5	att6
1	6	cluster_0	1.935	100.000	100.000	88.000	99.000	49.000	74.000
2	8	cluster_0	3.826	0	39.000	2.000	62.000	11.000	5.000
3	12	cluster_0	1.538	48.000	96.000	62.000	65.000	88.000	27.000
4	25	cluster_0	1.726	54.000	100.000	34.000	75.000	6.000	43.000
5	30	cluster_0	2.122	100.000	84.000	31.000	100.000	0	88.000
6	31	cluster_0	1.805	32.000	59.000	53.000	100.000	100.000	95.000
7	34	cluster_6	2.617	27.000	76.000	1	42.000	16.000	0
8	35	cluster_0	14.221	0	0	31.000	15.000	63.000	30.000
9	38	cluster_0	1.528	77.000	97.000	40.000	100.000	0	59.000
10	39	cluster_0	1.885	64.000	93.000	0	67.000	97.000	67.000
11	59	cluster_0	1.609	18.000	73.000	0	73.000	10.000	24.000
12	61	cluster_0	2.014	57.000	63.000	100.000	100.000	0	91.000
13	63	cluster_0	5.418	12.000	61.000	0	20.000	44.000	0
14	70	cluster_0	2.107	0	82.000	75.000	56.000	100.000	12.000
15	72	cluster_6	2.550	35.000	76.000	0	42.000	11.000	0
16	73	cluster_0	1.604	34.000	100.000	21.000	76.000	0	35.000
17	77	cluster_8	1.510	38.000	94.000	4.000	63.000	0	20.000
18	84	cluster_0	1.739	32.000	100.000	12.000	76.000	0	48.000
19	87	cluster_0	1.608	0	100.000	54.000	98.000	91.000	85.000
20	88	cluster_0	1.623	47.000	93.000	1	67.000	0	16.000

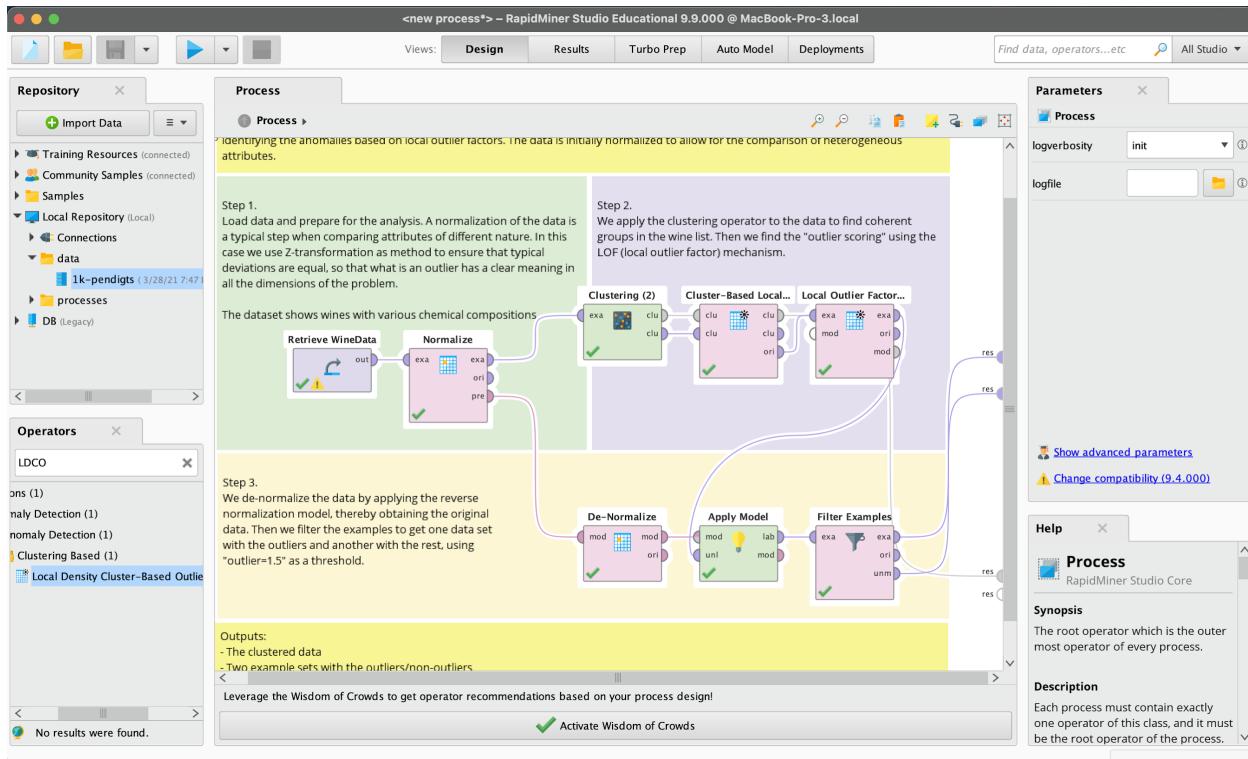
ExampleSet (307 examples, 3 special attributes, 17 regular attributes)

Broj podataka nije približno isti i većina istih instanci je označena anomalijama, ali drugi nacin ima vise anomalija oznacenih.

## Zadatak 3

Promjeniti metodu detekcije na CBLOF, a zatim na LDCOF. Koristiti po dvije metode clusteringa po želji (npr. k-means, k-medoids). Izvršiti istu analizu za broj i strukturu detektovanih anomalija kao u prethodnom zadatku.

CBLOF sa k-means varijanta:



Odnos je 943 - 57.

Clustering je prikazan sljedećem slike:

## Cluster Model

```

Cluster 0: 233 items
Cluster 1: 133 items
Cluster 2: 116 items
Cluster 3: 201 items
Cluster 4: 317 items
Total number of items: 1000

```

## Rezultat filtriranja dataset:

Result History    Cluster Model (Clustering (2))    ExampleSet (Filter Examples)    ExampleSet (Filter Examples)

Views: Design    Results    Turbo Prep    Auto Model    Deployments

Find data, operators...etc    All Studio

Data    Statistics    Visualizations    Annotations

Open in: Turbo Prep    Auto Model

Filter (943 / 943 examples): all

Row No.	id	cluster	outlier	att1	att2	att3	att4	att5	att6
1	1	cluster_1	0.997	47.000	100.000	27.000	81.000	57.000	37.000
2	2	cluster_0	1.090	0	89.000	27.000	100.000	42.000	75.000
3	3	cluster_4	1.036	0	57.000	31.000	68.000	72.000	90.000
4	4	cluster_3	1.072	0	100.000	7.000	92.000	5.000	68.000
5	5	cluster_4	1.076	0	67.000	49.000	83.000	100.000	100.000
6	6	cluster_3	1.281	100.000	100.000	88.000	99.000	49.000	74.000
7	7	cluster_2	1.093	0	100.000	3.000	72.000	26.000	35.000
8	8	cluster_2	1.491	0	39.000	2.000	62.000	11.000	5.000
9	9	cluster_1	1.002	13.000	89.000	12.000	50.000	72.000	38.000
10	10	cluster_2	1.003	57.000	100.000	22.000	72.000	0	31.000
11	11	cluster_4	1.011	74.000	87.000	31.000	100.000	0	69.000
12	12	cluster_1	0.988	48.000	96.000	62.000	65.000	88.000	27.000
13	13	cluster_3	1.106	100.000	100.000	72.000	99.000	36.000	78.000
14	14	cluster_3	1.071	91.000	74.000	54.000	100.000	0	87.000
15	15	cluster_0	1.068	0	85.000	38.000	100.000	81.000	88.000
16	16	cluster_4	1.070	35.000	76.000	57.000	100.000	100.000	92.000
17	17	cluster_4	0.994	50.000	84.000	66.000	100.000	75.000	75.000
18	18	cluster_4	1.015	99.000	80.000	63.000	100.000	25.000	76.000
19	19	cluster_0	1.088	24.000	66.000	43.000	100.000	59.000	65.000
20	20	cluster_0	1.095	0	73.000	19.000	99.000	72.000	100.000

ExampleSet (943 examples, 3 special attributes, 17 regular attributes)

## I anomalije kao rezultat:

Result History    Cluster Model (Clustering (2))    ExampleSet (Filter Examples)    ExampleSet (Filter Examples)

Views: Design    Results    Turbo Prep    Auto Model    Deployments

Find data, operators...etc    All Studio

Data    Statistics    Visualizations    Annotations

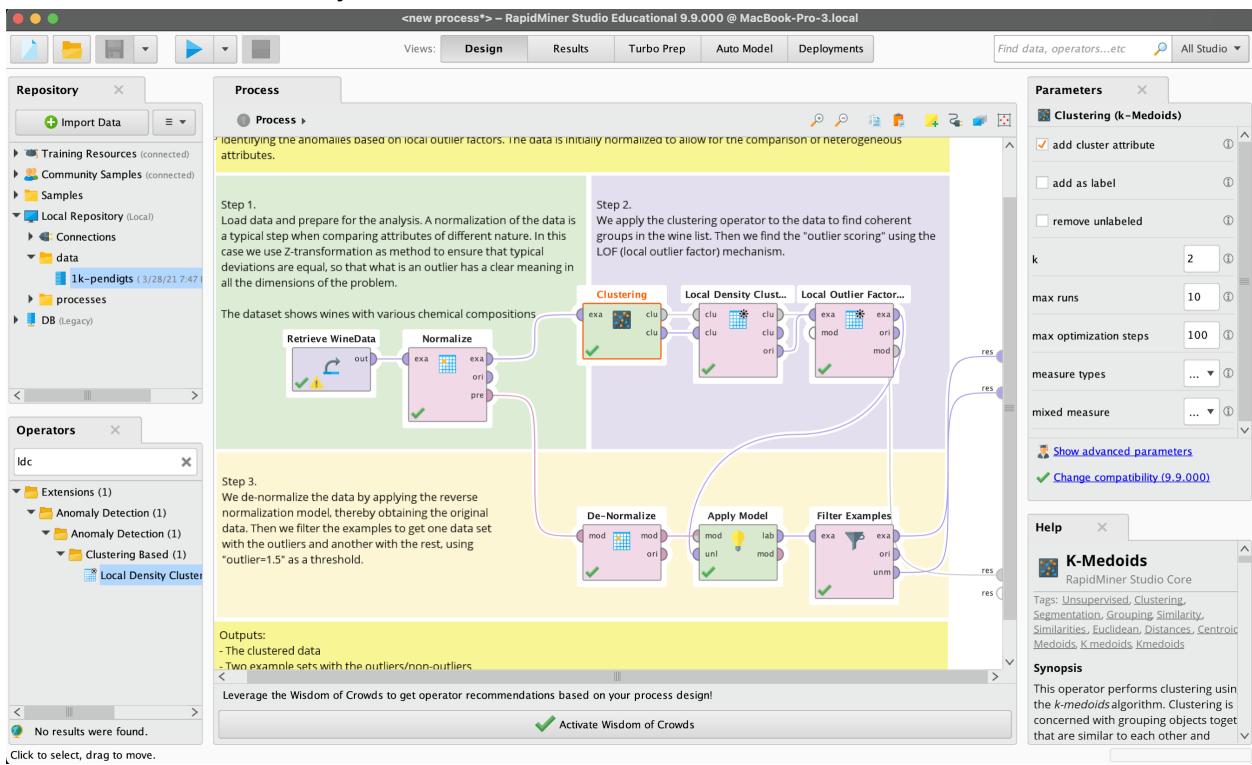
Open in: Turbo Prep    Auto Model

Filter Examples.unmatched example set  
Filter (57 / 57 examples): all

Row No.	id	cluster	outlier	att1	att2	att3	att4	att5	att6
1	30	cluster_3	1.953	100.000	84.000	31.000	100.000	0	88.000
2	31	cluster_4	1.693	32.000	59.000	53.000	100.000	100.000	95.000
3	35	cluster_2	2.129	0	0	31.000	15.000	63.000	30.000
4	94	cluster_4	1.673	3.000	96.000	53.000	100.000	81.000	68.000
5	118	cluster_3	1.615	100.000	77.000	92.000	100.000	49.000	76.000
6	135	cluster_3	1.643	100.000	70.000	58.000	100.000	10.000	75.000
7	140	cluster_2	1.794	73.000	85.000	31.000	65.000	0	0
8	155	cluster_4	1.635	88.000	80.000	12.000	80.000	100.000	100.000
9	161	cluster_4	1.725	0	100.000	73.000	97.000	95.000	80.000
10	163	cluster_4	1.520	2.000	85.000	31.000	100.000	76.000	74.000
11	182	cluster_4	1.503	0	91.000	46.000	100.000	85.000	80.000
12	214	cluster_4	1.704	100.000	100.000	92.000	85.000	75.000	71.000
13	236	cluster_4	2.242	14.000	55.000	40.000	61.000	59.000	100.000
14	247	cluster_3	1.651	77.000	100.000	53.000	90.000	0	43.000
15	248	cluster_2	2.062	0	0	38.000	13.000	71.000	34.000
16	257	cluster_4	1.522	100.000	100.000	14.000	90.000	74.000	55.000
17	264	cluster_3	1.643	85.000	100.000	56.000	72.000	42.000	39.000
18	297	cluster_0	1.767	41.000	53.000	60.000	100.000	100.000	89.000
19	319	cluster_0	1.699	21.000	82.000	13.000	55.000	54.000	100.000
20	330	cluster_3	1.633	38.000	76.000	58.000	100.000	21.000	65.000

ExampleSet (57 examples, 3 special attributes, 17 regular attributes)

## LDCOF sa k-medoids varijanta:



Odnos je 943 - 57.

Clustering je prikazan sljedećem slikom:

## Cluster Model

**Cluster 0: 565 items**  
**Cluster 1: 435 items**  
**Total number of items: 1000**

## Rezultat filtriranja dataset:

## I anomalije kao rezultat:

Broj podataka je (pribлизно) isti i većina (haman sve) instanci su označene anomalijama, samo broj clustera je razlicit izmedju dvije metode.