

# Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών



# Παράδειγμα

## Student Questionnaire

# Student Questionnaire

Στόχος της άσκησης είναι η εφαρμογή της μεθόδου της ομαδοποίησης.

Πιο συγκεκριμένα, θα πρέπει να εφαρμοστεί η ομαδοποίηση στο dataset με όνομα Elearning\_Reaction.xlsx που θα βρείτε στα έγγραφα του μαθήματος.

Το dataset περιέχει στοιχεία σχετικά με την συμπεριφορά αλλά και την αλληλεπίδραση των φοιτητών κατά την διάρκεια ενός online μαθήματος.

# Student Questionnaire

Το βασικό ζητούμενο από την δημιουργία των ομάδων είναι να δούμε πως συσχετίζεται η συμπεριφορά των φοιτητών με τις δεξιότητες που κατέχουν εντός των ομάδων που προκύπτουν.

# Student Questionnaire

## Φάση 1

Στην πρώτη φάση της εργασίας θα θέλαμε να μας παρουσιάσετε τα βήματα που πραγματοποιήσατε στο στάδιο της προετοιμασίας των δεδομένων καθώς επίσης και την αρχική (πριν την προετοιμασία) και την τελική μορφή των δεδομένων (μετά το στάδιο της προετοιμασίας).

# Student Questionnaire

## Φάση 2.

Στην δεύτερη φάση θα θέλαμε να εφαρμόσετε τον αλγόριθμο K-means (για  $K=5$ ) και να παράγεται τα αντίστοιχα αποτελέσματα. Σε αυτό το στάδιο θα θέλαμε το “Centroid Table” μαζί και με την γραφική αναπαράσταση των ομάδων (Plot)

# Student Questionnaire

## Φάση 3.

Στην τρίτη φάση θα θέλαμε να αξιολογήσετε τα αποτελέσματα από την εφαρμογή του αλγορίθμου αλλάζοντας κάθε φορά την τιμή του αριθμού των cluster ( $k$ ).

Δηλαδή αν έχετε επιλέξει σαν αρχικό αριθμό  $K=5$  τότε θα θέλαμε να ξανά τρέξετε την ανάλυση για  $K=4$  και  $K=6$ .

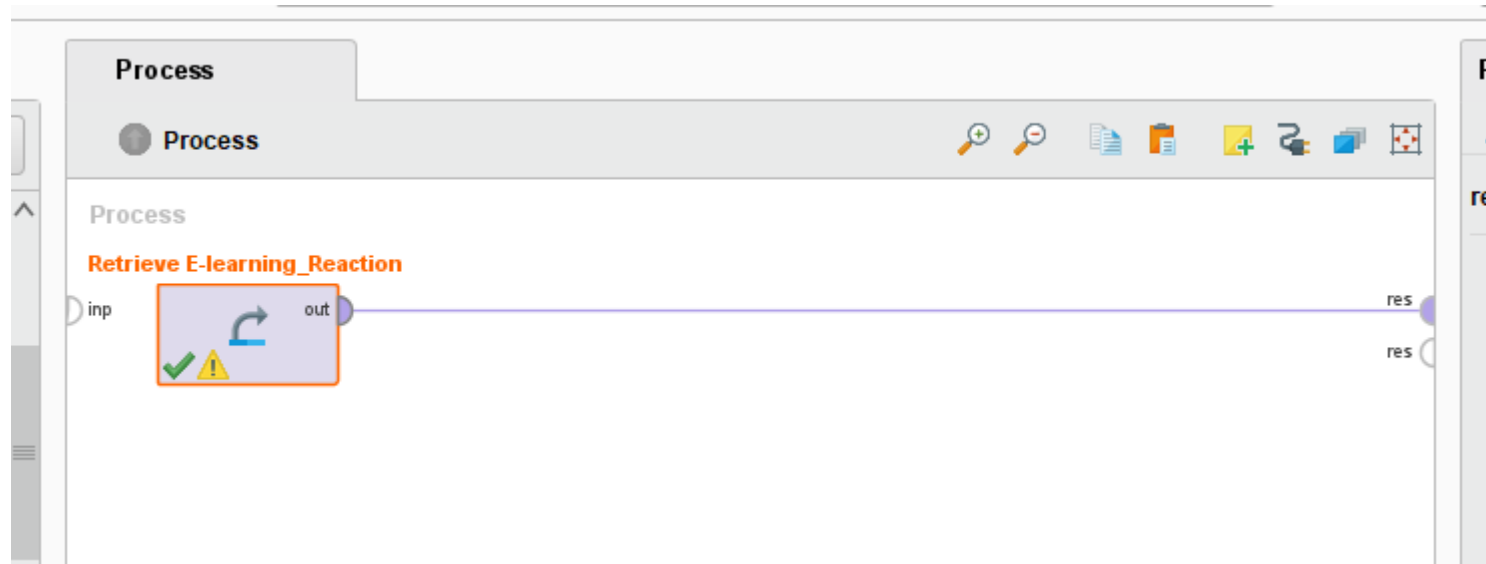
Για κάθε αλλαγή στον αριθμό των cluster θα θέλαμε να μας παραδώσετε τα αποτελέσματα του M.O. των centroid τιμών

# Student Questionnaire

## Λύση Άσκησης



# Student Questionnaire - Data Preperation



# Student Questionnaire - Data Preperation

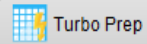
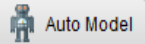
- Replace Missing Values

The screenshot displays the Orange3 data mining software interface. At the top, there are tabs for 'Design', 'Results', 'Turbo Prep', 'Auto Model', and 'Deployments'. A search bar contains the text 'Find data, operators...etc' and a dropdown menu is set to 'All Studio'. The main workspace shows a process flow starting with 'Retrieve E-learning\_...' followed by 'Replace Missing Values'. The 'Replace Missing Values' process is highlighted with an orange border and has a green checkmark. Its parameters are shown in a panel on the right:

- Parameters**
- Replace Missing Values**
- create view
- attribute filter type: all
- invert selection
- include special attributes
- default: average
- columns: Edit List (0)...

# Student Questionnaire - Data Preperation

- Replace Missing Values

Open in  Turbo Prep  Auto Model Filter (71 / 71 examples): all

Row No.	id	total_posts	helpful_post	nice_code_...	collaborativ...	confused_p...	creative_post	bad_post	ama
1	0	1	0	0	0	0	6	0	1
2	1	1	0	0	1	0	2	0	3
3	2	2	4	3	9	0	16	1	8
4	3	5	1	3	9	2	11	0	8
5	4	14	6	15	28	0	50	0	45
6	5	9	3	9	16	7	21	0	17
7	6	15	10	21	21	1	34	0	37
8	7	8	9	21	20	0	31	0	28
9	8	6	3	12	13	0	24	0	19
10	9	4	4	1	11.304	0	15	0	14
11	10	5	1	3	6	0	16	0	20
12	11	4	1	0	1	0	13	0	10
13	12	4	3	1	4	0	21	0	19
14	13	7	2	8	20	0	16	0	29



# Student Questionnaire - Data Preperation

- Detect Outliers

The screenshot displays the Orange Data Mining interface. The main workspace shows a workflow with four processes: 'Retrieve E-learning\_...', 'Replace Missing Val...', 'Remove Duplicates', and 'Detect Outlier (Distances)'. The 'Detect Outlier (Distances)' process is highlighted with an orange border. To the right, the 'Parameters' window for this process is open, showing the following settings:

Parameter	Value
number of neighbors	10
number of outliers	10
distance function	euclidian distan...

# Student Questionnaire - Data Preperation

Open in  Turbo Prep  Auto Model

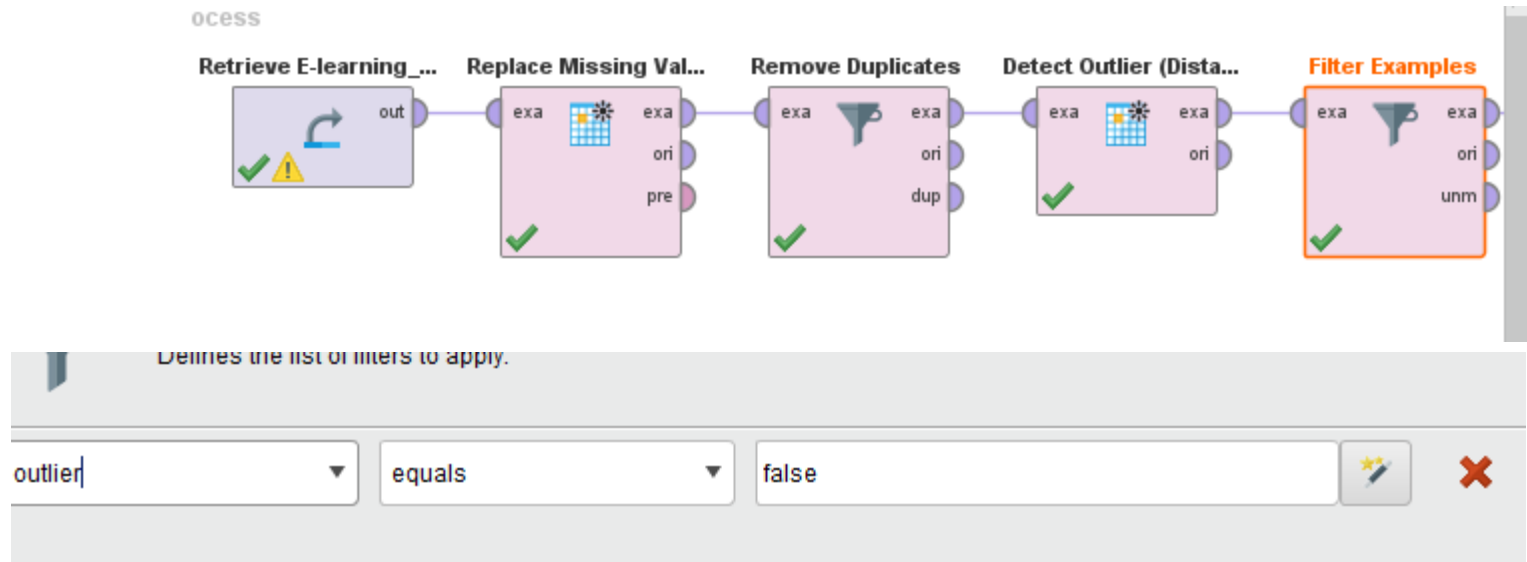
Filter (71 / 71 examples):

Row No.	outlier	id	total_posts	helpful_post	nice_code_...	collaborativ...	confused_p...	creative_post	bad_
1	false	0	1	0	0	0	0	6	0
2	false	1	1	0	0	1	0	2	0
3	false	2	2	4	3	9	0	16	1
4	false	3	5	1	3	9	2	11	0
5	true	4	14	6	15	28	0	50	0
6	false	5	9	3	9	16	7	21	0
7	true	6	15	10	21	21	1	34	0
8	false	7	8	9	21	20	0	31	0
9	true	8	6	3	12	13	0	24	0
10	false	9	4	4	1	11.304	0	15	0
11	false	10	5	1	3	6	0	16	0
12	false	11	4	1	0	1	0	13	0
13	false	12	4	3	1	4	0	21	0
14	false	13	7	2	8	20	0	16	0

ExampleSet (71 examples, 1 special attribute, 16 regular attributes)

# Student Questionnaire - Data Preperation

## Filter Examples



# Student Questionnaire

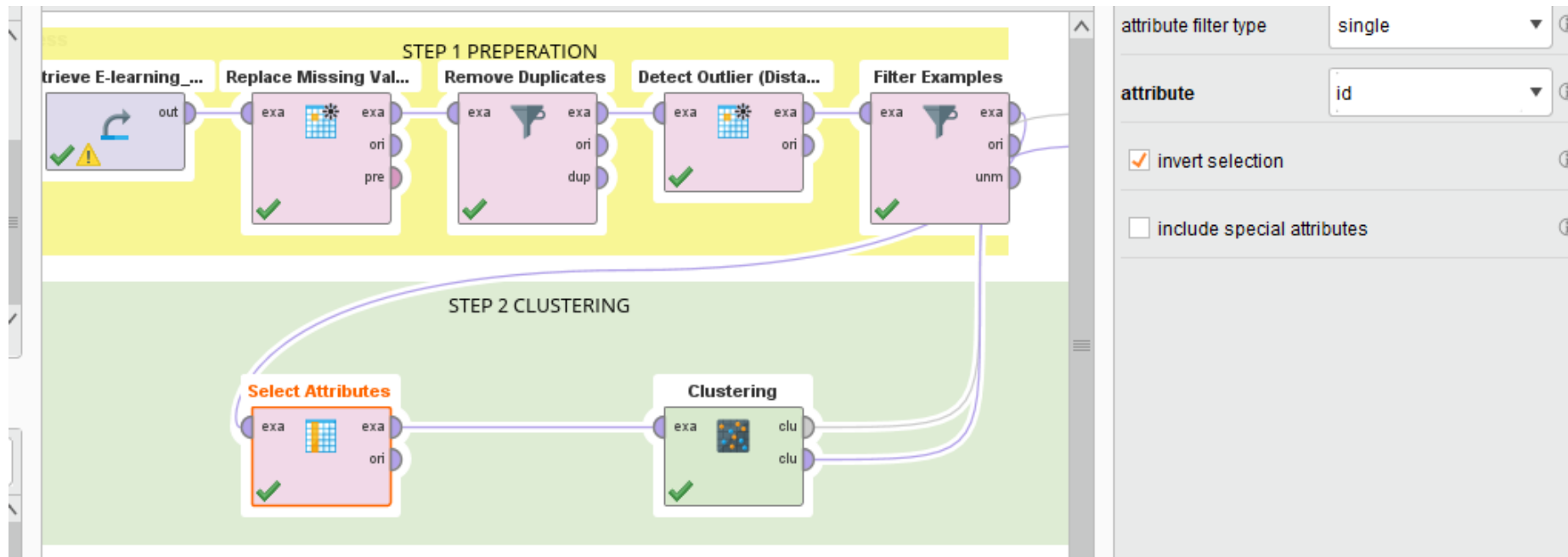
ExampleSet (Filter Examples) ExampleSet (//Local Repository/New-clustering/Data/E-learning\_Reaction)

Open in Turbo Prep Auto Model Filter (61 / 61 examples): all

Row No.	outlier	id	total_posts	helpful_post	nice_code_...	collaborativ...	confused_p...	creative_post	bad_...
1	false	0	1	0	0	0	0	6	0
2	false	1	1	0	0	1	0	2	0
3	false	2	2	4	3	9	0	16	1
4	false	3	5	1	3	9	2	11	0
5	false	5	9	3	9	16	7	21	0
6	false	7	8	9	21	20	0	31	0
7	false	9	4	4	1	11.304	0	15	0
8	false	10	5	1	3	6	0	16	0
9	false	11	4	1	0	1	0	13	0
10	false	12	4	3	1	4	0	21	0
11	false	13	7	2	8	20	0	16	0
12	false	14	5	1	0	4	0	8	0
13	false	15	2	2	1	7	1	9	0

# Student Questionnaire

## Clustering





# Student Questionnaire

ExampleSet (Clustering) Cluster Model (Clustering)

Open in Turbo Prep Auto Model Filter (61 / 61 examples): all

Row No.	id	label	outlier	total_posts	helpful_post	nice_code_...	collaborativ...	confused_p...	crea
1	1	cluster_0	false	1	0	0	0	0	6
2	2	cluster_0	false	1	0	0	1	0	2
3	3	cluster_0	false	2	4	3	9	0	16
4	4	cluster_2	false	5	1	3	9	2	11
5	5	cluster_1	false	9	3	9	16	7	21
6	6	cluster_2	false	8	9	21	20	0	31
7	7	cluster_3	false	4	4	1	11.304	0	15
8	8	cluster_4	false	5	1	3	6	0	16
9	9	cluster_2	false	4	1	0	1	0	13
10	10	cluster_2	false	4	3	1	4	0	21
11	11	cluster_4	false	7	2	8	20	0	16
12	12	cluster_4	false	5	1	0	4	0	8
13	13	cluster_3	false	2	2	1	7	1	9
14	14	cluster_3	false	2	3	3	10	0	16

ExampleSet (61 examples, 3 special attributes, 15 regular attributes)

# Student Questionnaire




<new process\*> - RapidMiner Studio Educational 9.10.001 @ DESKTOP-8H00FBS

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

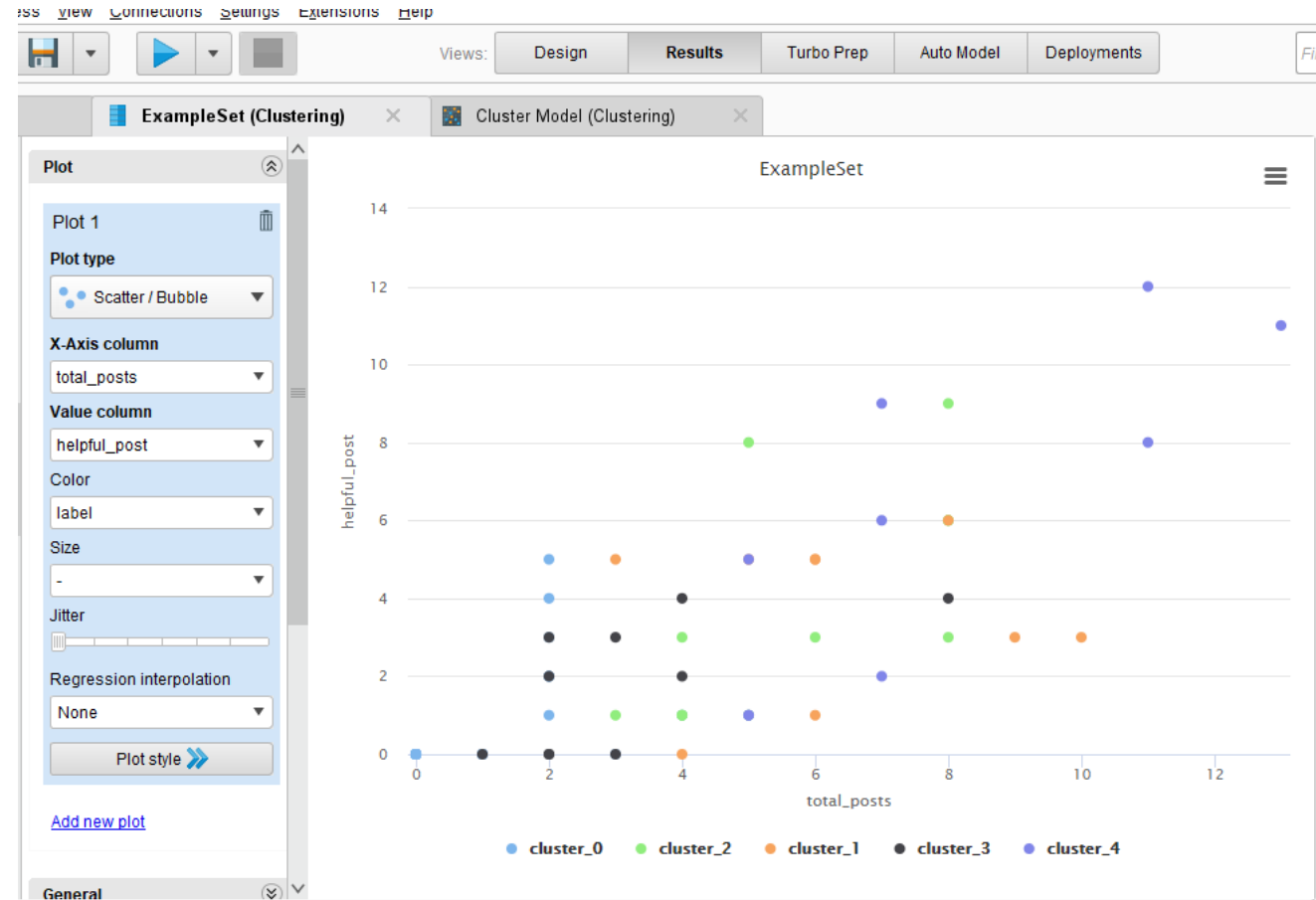
Result History ExampleSet (Clustering) Cluster Model (Clustering)

Name Type Missing Statistics Filter (18 / 18 attributes): Search for Attributes

Name	Type	Missing	Statistics	Filter (18 / 18 attributes):
<b>id</b>	Integer	0	 Min 1 Max 61	
<b>label</b>	Nominal	0	 Least cluster_4 (9) Most cluster_...	
<b>outlier</b>	Binominal	0	 Negative false Positive true	
<b>total_posts</b>	Real	0	Min 0 Max 13 Average 4	
<b>helpful post</b>	Real	0	Min 0 Max 12 Average 2.721	

Showing attributes 1 - 18 Examples: 61 Special Attributes: 3 Regular Attributes: 15

# Student Questionnaire



# Student Questionnaire

## Clusters που έχουν φτιαχτεί: k=5

### Cluster Model

```
Cluster 0: 8 items  
Cluster 1: 21 items  
Cluster 2: 25 items  
Cluster 3: 5 items  
Cluster 4: 2 items  
Total number of items: 61
```

## Μέση απόσταση μεταξύ αντικειμένων: k=5

### Avg. within centroid distance

```
Avg. within centroid distance: -3.844
```

# Student Questionnaire

**Για  $k=4$**

**Avg. within centroid distance**

Avg. within centroid distance: -4.464

# Student Questionnaire

Για  $k=4$

**Avg. within centroid distance\_cluster\_0**

Avg. within centroid distance\_cluster\_0: -4.208

**Avg. within centroid distance\_cluster\_1**

Avg. within centroid distance\_cluster\_1: -2.238

**Avg. within centroid distance\_cluster\_2**

Avg. within centroid distance\_cluster\_2: -4.435

**Avg. within centroid distance\_cluster\_3**

Avg. within centroid distance\_cluster\_3: -8.245

# Student Questionnaire

**Για  $k=6$**

# Student Questionnaire

Για  $k=6$

**Avg. within centroid distance**

Avg. within centroid distance: -3.133



# Student Questionnaire

**Για k=6**

**Avg. within centroid distance\_cluster\_0**

Avg. within centroid distance\_cluster\_0: -2.742

**Avg. within centroid distance\_cluster\_1**

Avg. within centroid distance\_cluster\_1: -4.435

**Avg. within centroid distance\_cluster\_2**

Avg. within centroid distance\_cluster\_2: -2.238

# Student Questionnaire

**Για k=6**

**Avg. within centroid distance\_cluster\_3**

Avg. within centroid distance\_cluster\_3: -6.324

**Avg. within centroid distance\_cluster\_4**

Avg. within centroid distance\_cluster\_4: -2.115

**Avg. within centroid distance\_cluster\_5**

Avg. within centroid distance\_cluster\_5: -3.995

ΤΕΛΟΣ

