Klasterizacija

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

URL: http://nikola.milikic.info

Klasterizacija

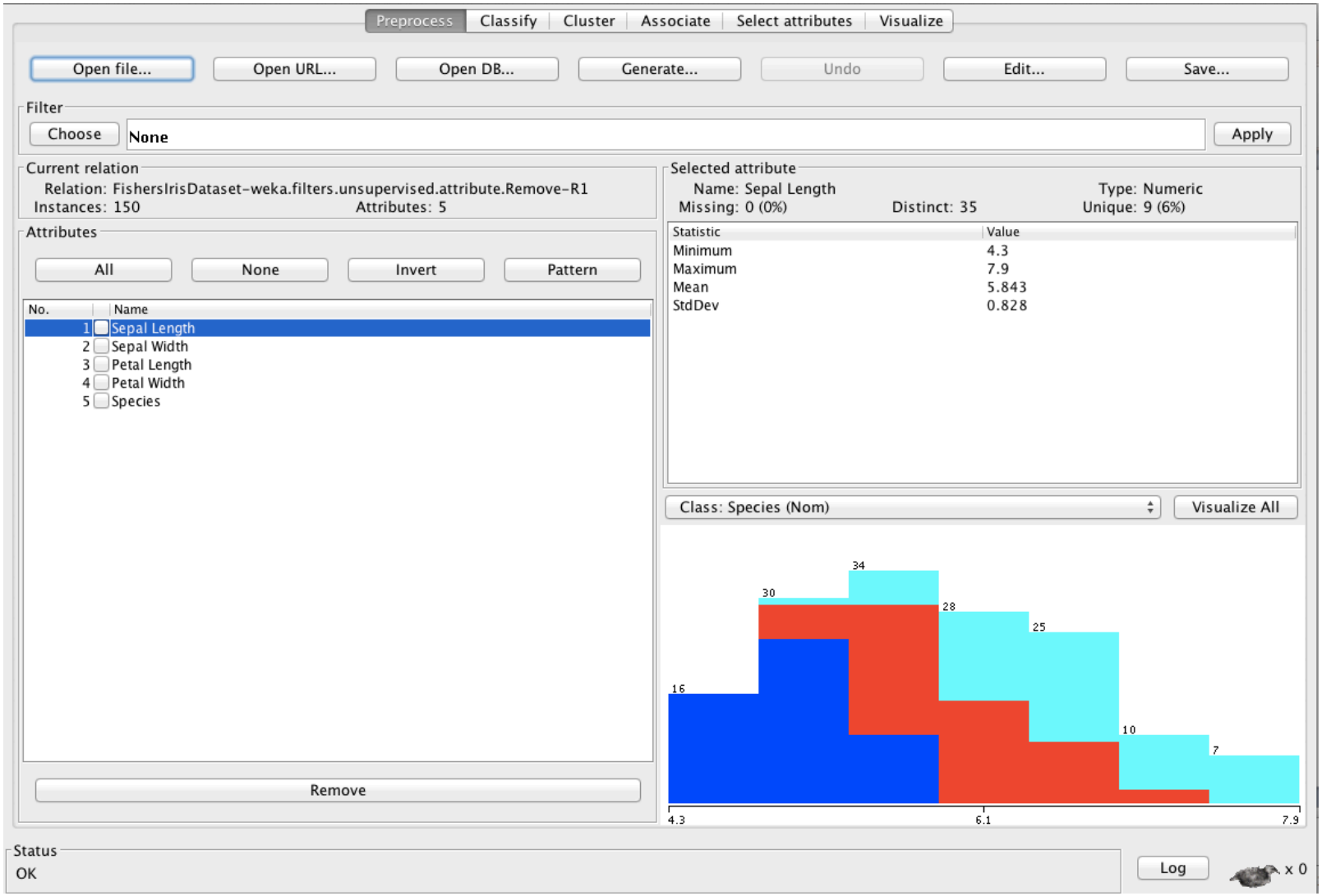
Klasterizacija (eng. Clustering) spada u grupu tehnika nenadgledanog učenja i omogućava grupisanje instanci u grupe, gde unapred ne znamo koje su sve grupe moguće.

Grupe u koje se instance dele se nazivaju **klasteri**.

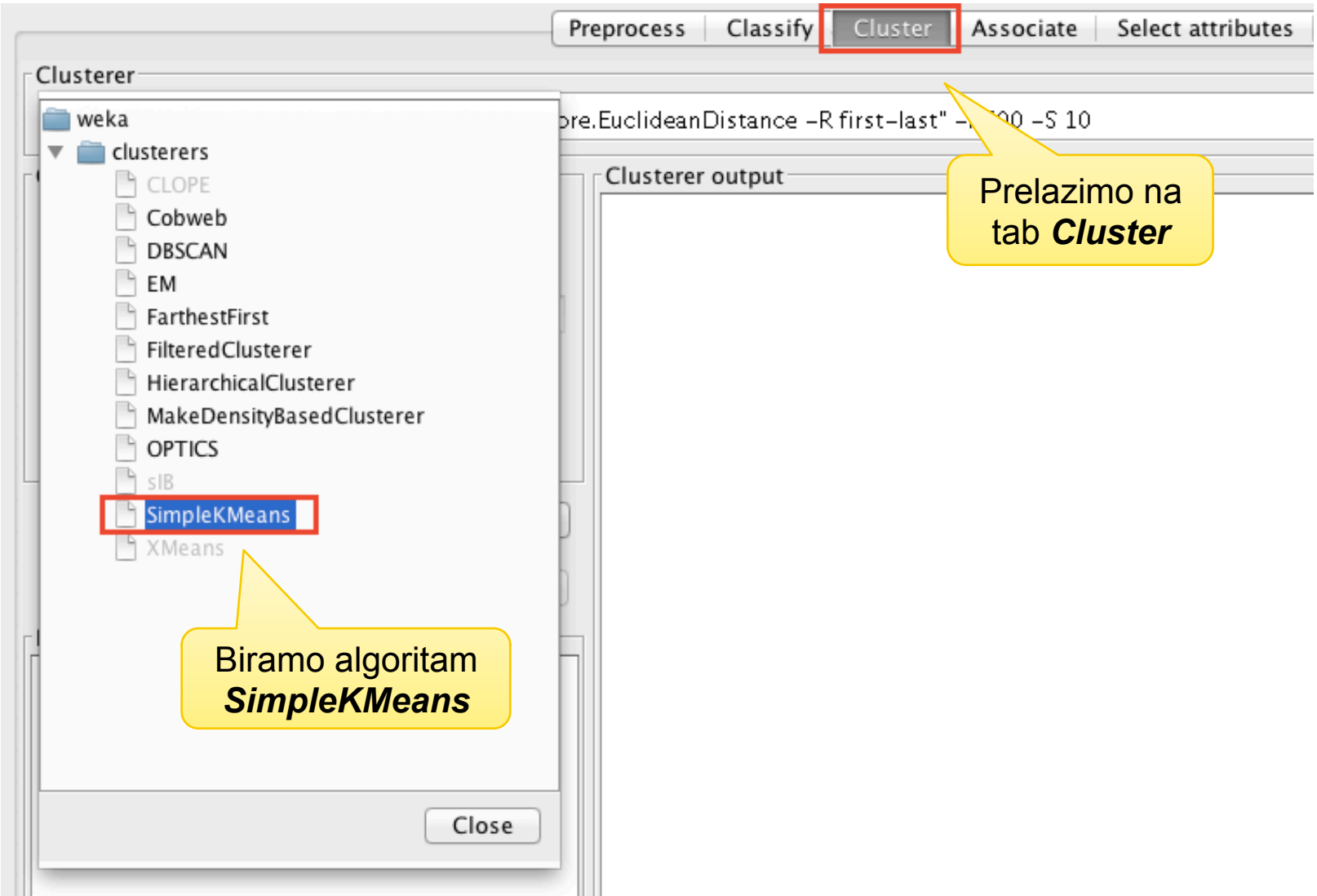
Kao rezultat klasterizacije svakoj instanci je dodeljen novi atribut koji predstavlja klaster kojoj pripada. Može se reći da je klasterovanje uspešno ukoliko su dobijeni klasteri smisleni i ukoliko se mogu imenovati.

K-Means algoritam u Weka-i

FishersIrisDataset.arff



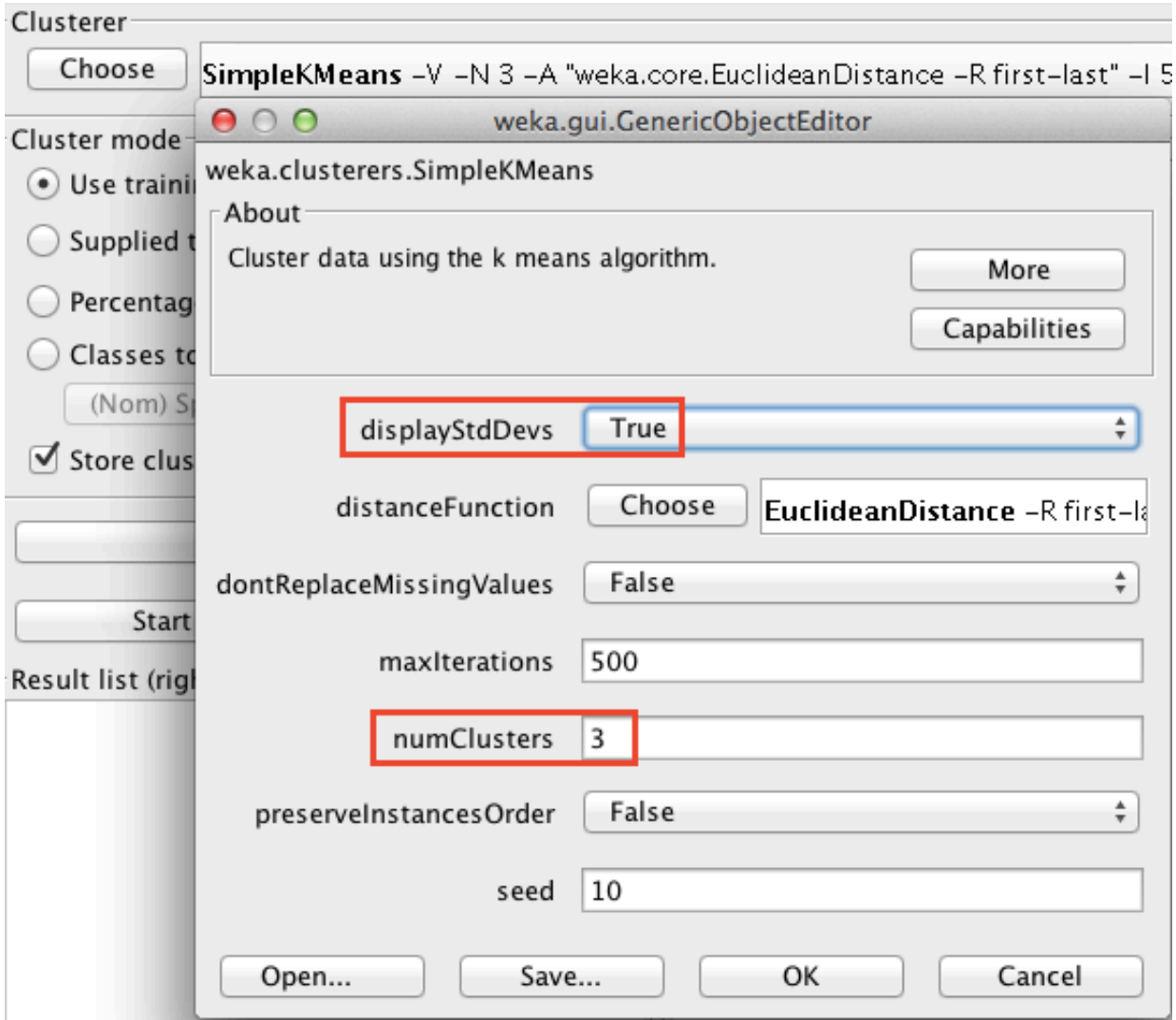
Odabir algoritma za klasterovanje



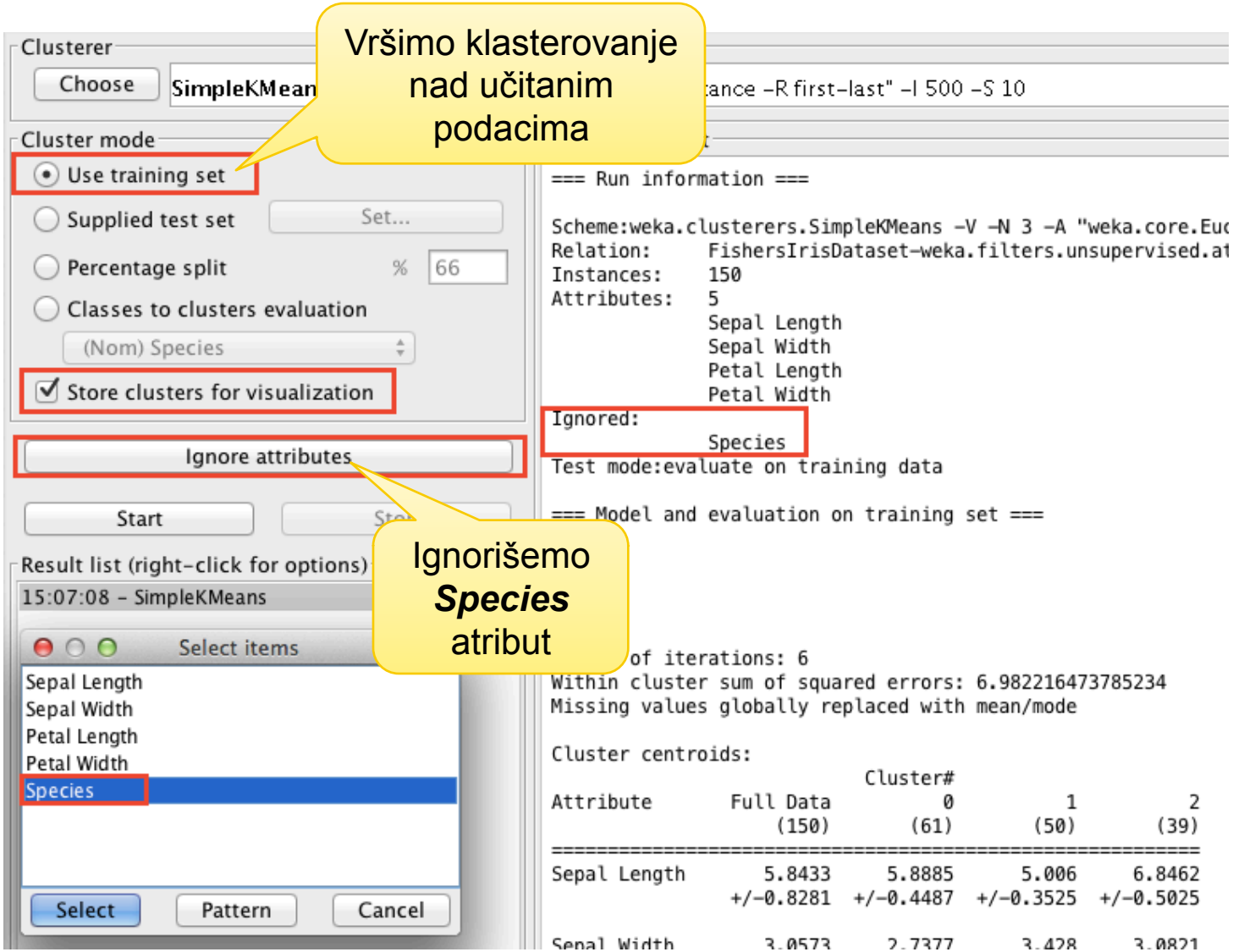
Podešavanje parametara

numClusters – broj željenih klastera; postavljamo na 3 jer imamo tri vrste

displayStdDevs– ako je *true*, onda će se ispisati vrednosti standardne devijacije



Pokretanje procesa klasterovanja



Rezultat klasterovanja

Cluster mode

- Use training set
- Supplied test set — Set...
- Percentage split — % 66

Ignore attributes

Start — Stop

Result list (right-click for options)

15:07:08 - SimpleKMeans

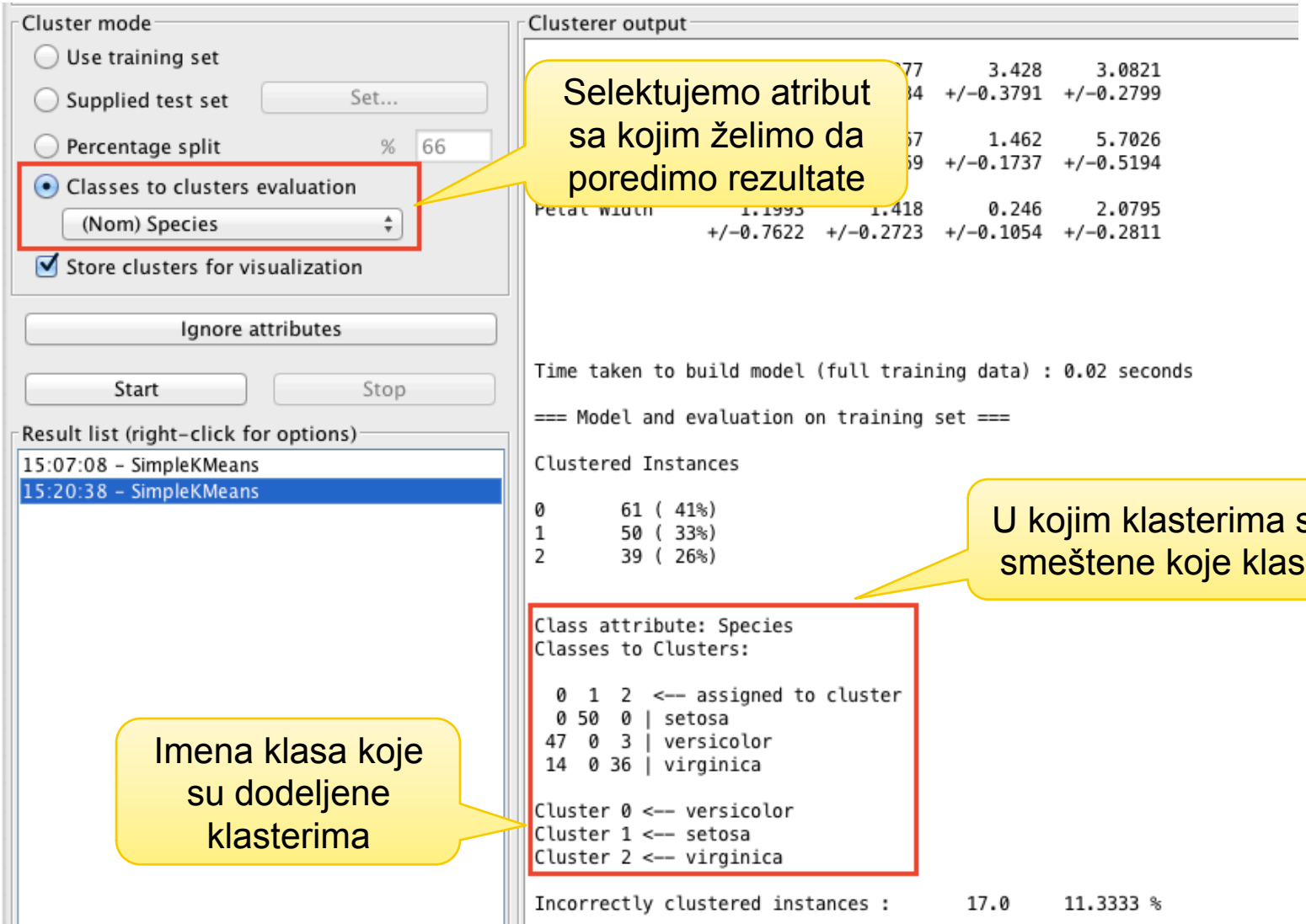
Centroidi svakog klastera i njihove standardne devijacije

Broj instanci u svakom klasteru

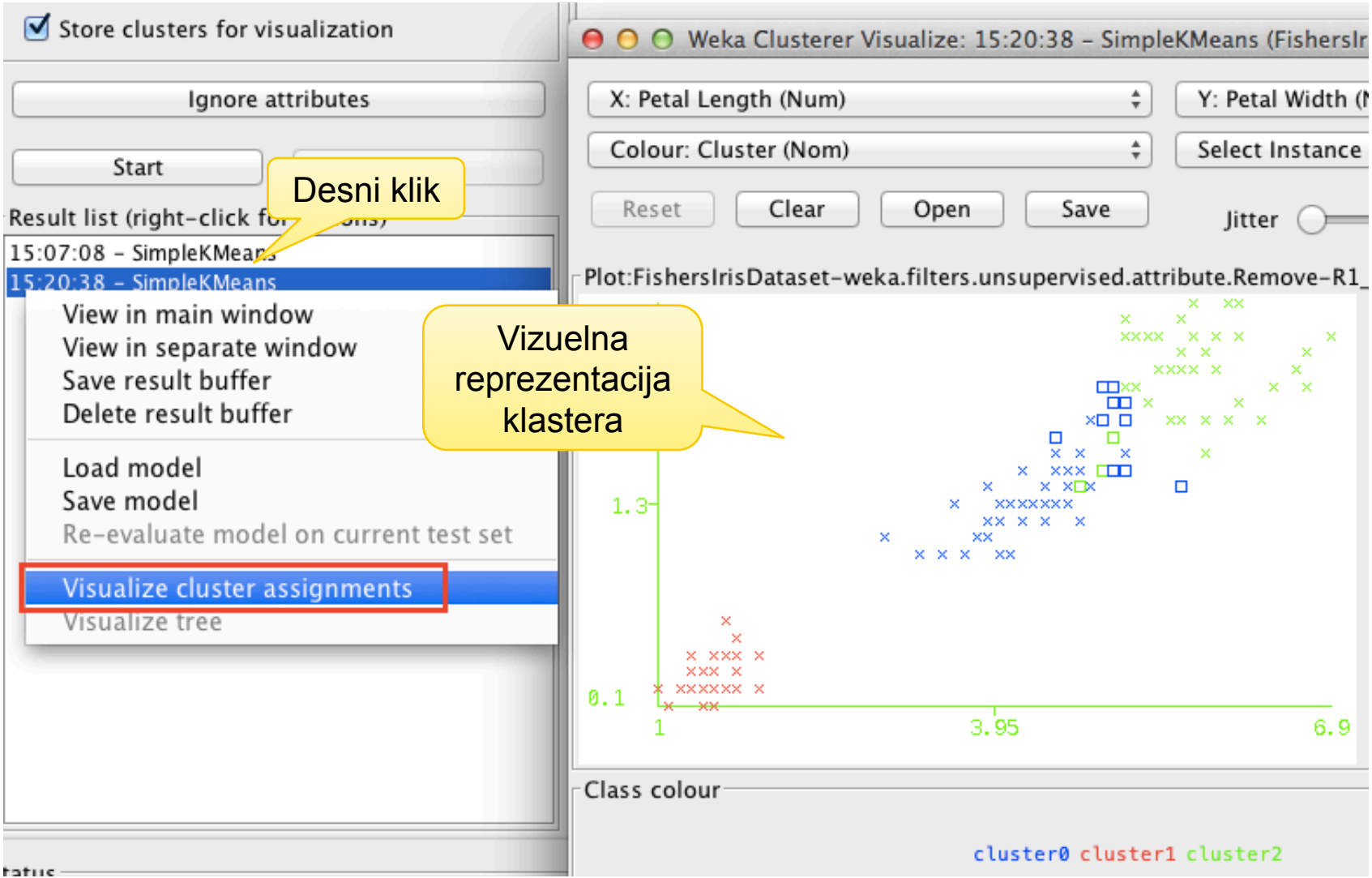
Clusterer output

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.982216473785234
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute       Full Data          0          1          2
                    (150)       (61)       (50)       (39)
==========================================================
Sepal Length       5.8433     5.8885      5.006     6.8462
                 +/-0.8281  +/-0.4487  +/-0.3525  +/-0.5025

Sepal Width        3.0573     2.7377      3.428     3.0821
                 +/-0.4359  +/-0.2934  +/-0.3791  +/-0.2799

Petal Length        3.758     4.3967      1.462     5.7026
                 +/-1.7653  +/-0.5269  +/-0.1737  +/-0.5194

Petal Width        1.1993      1.418      0.246     2.0795
                 +/-0.7622  +/-0.2723  +/-0.1054  +/-0.2811



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

Evaluacija rezultata

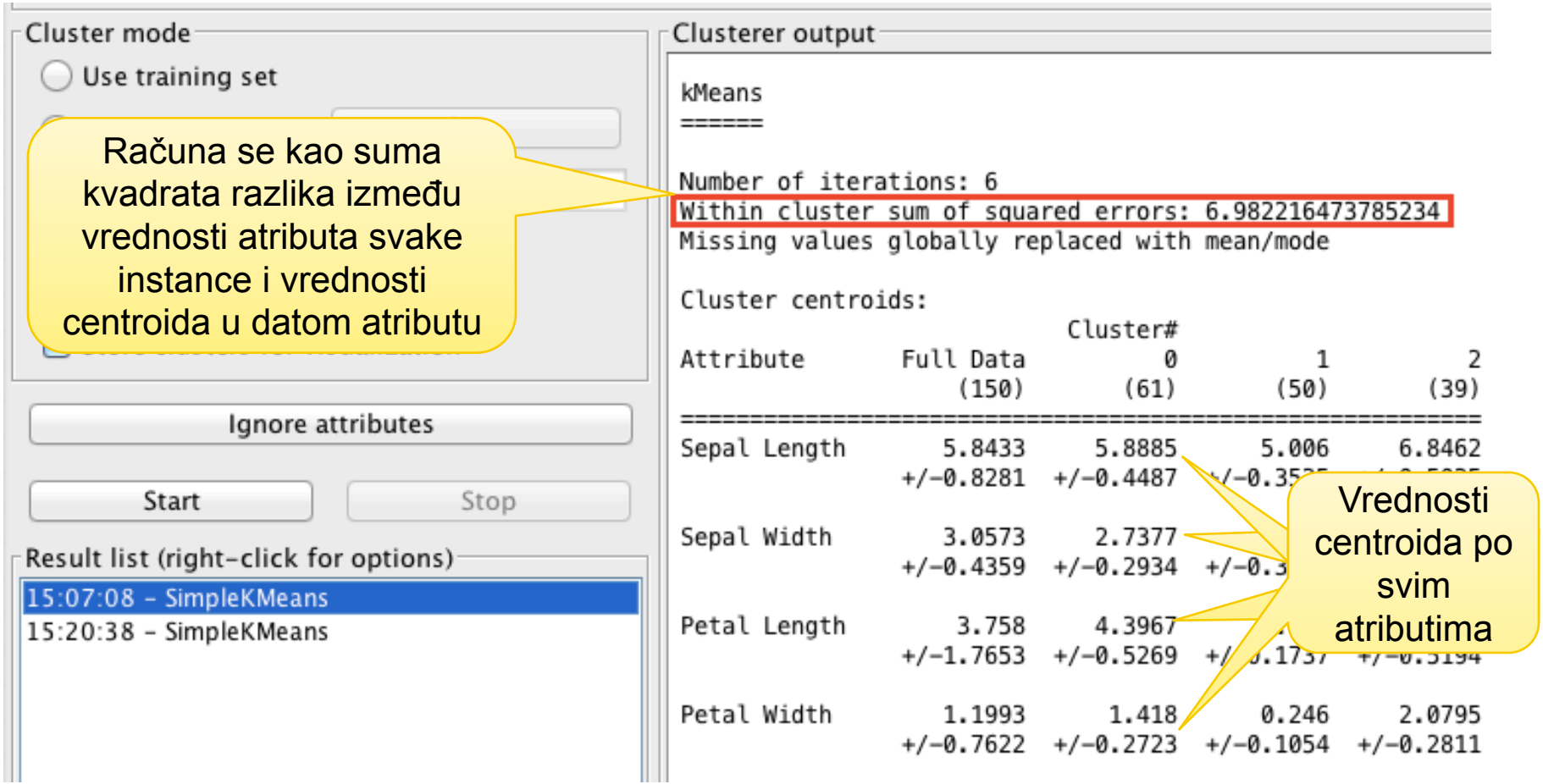


Vizuelizacija klastera



Procena uspešnosti klasterovanja

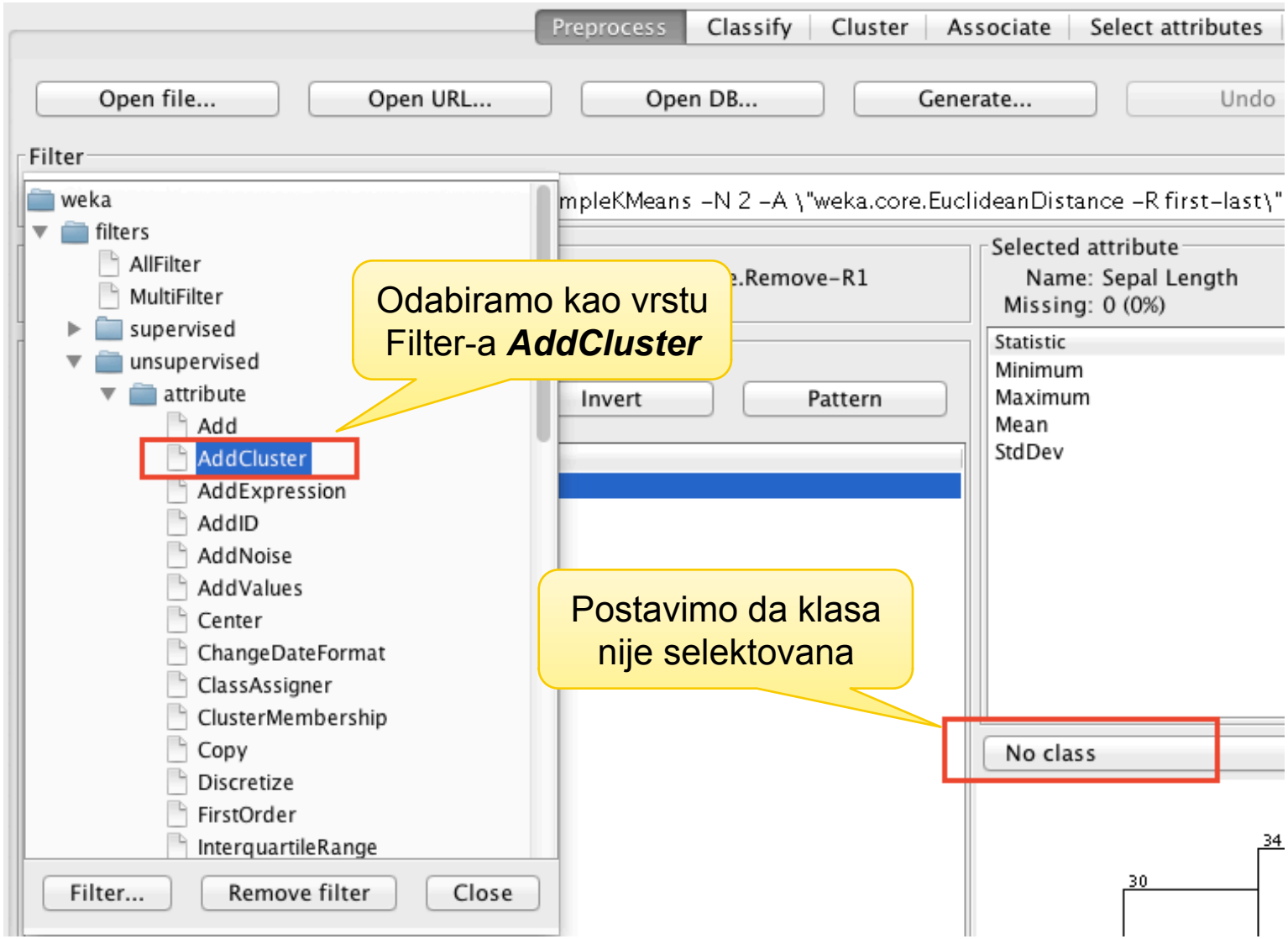
Within cluster sum of squared error (suma kvadrata greške unutar klastera) daje procenu kvaliteta dobijenih klastera



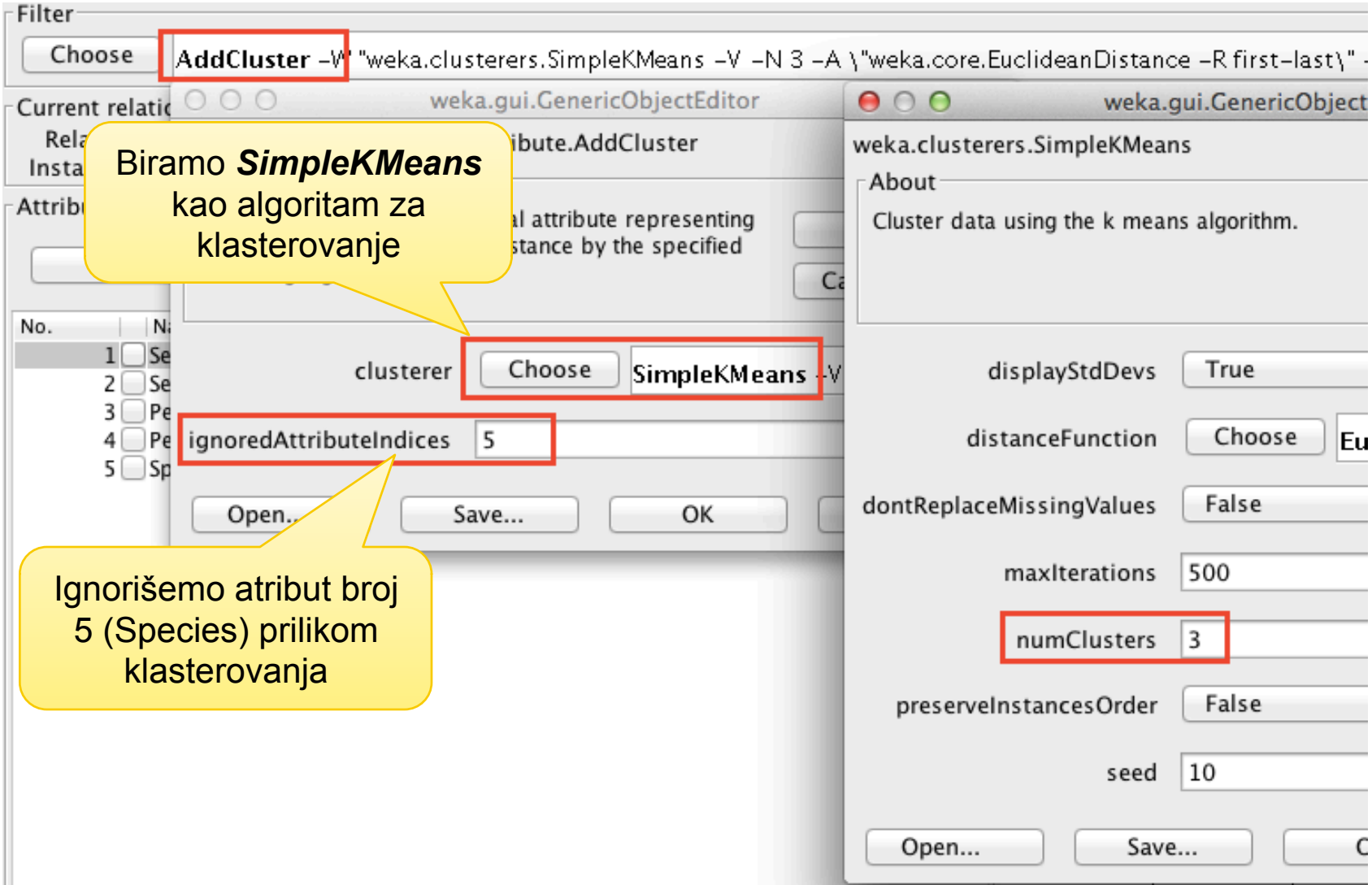
Kako proceniti dobar broj klastera?

Klasteri	Greška
1	55.6
2	12.1
3	7.0
4	5.5
5	5.0
6	4.8
7	4.7
8	4.2
9	4.1
10	3.6
20	1.7
50	0.6

Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose: AddCluster -W "weka.clusterers.SimpleKMeans -V -N 3 -A \"weka.core.EuclideanDistance -R first-last\" -I 500 -S 10" -I 5 — Apply

Current relation

Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1-wek...

Instances: 150 Attributes: 6

Attributes

All | None | Invert | Pattern

No.	Name
1	Sepal Length
2	Sepal Width
3	Petal Length
4	Petal Width
5	Species
6	cluster

Remove

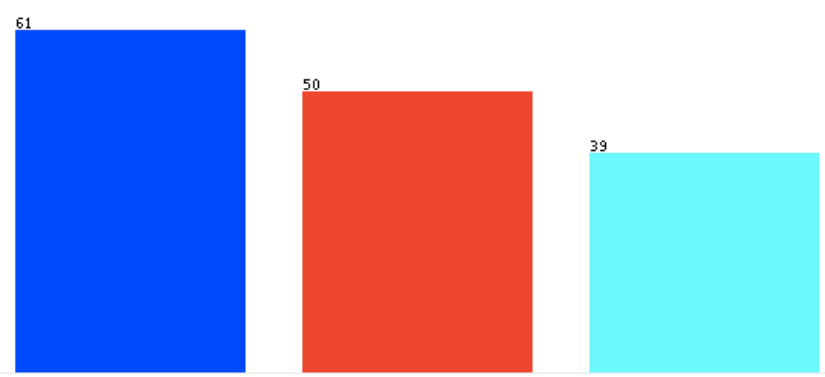
Selected attribute

Name: cluster Type: Nominal

Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

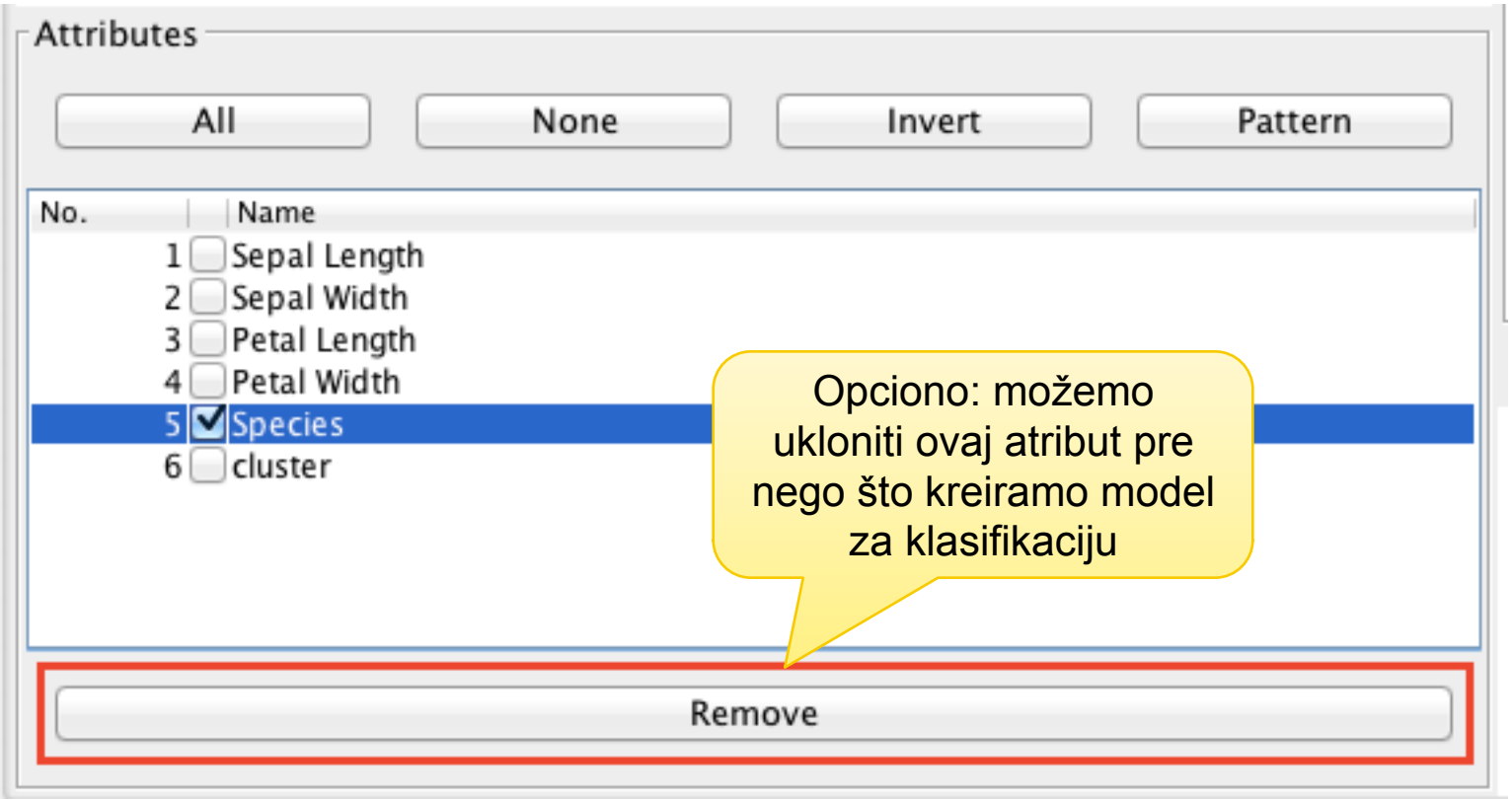
No.	Label	Count
1	cluster1	61
2	cluster2	50
3	cluster3	39

Class: cluster (Nom) | Visualize All



Nakon primene filtera (***Apply***) dodat je novi atribut pod nazivom ***cluster***

Korišćenje klastera za klasifikaciju



Opciono: možemo ukloniti ovaj atribut pre nego što kreiramo model za klasifikaciju

Korišćenje klastera za klasifikaciju

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose NaiveBayes

Koristimo ***NaiveBayes*** klasifikator

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) cluster

Klasifikujemo prema atributu ***cluster***

Start | Stop

Result list (right-click for options)

16:18:49 - bayes.NaiveBayes

```
  mean            1.4108   0.2766   2.0806
  std. dev.       0.2837   0.1074   0.2717
  weight sum          61       50       39
  precision       0.1143   0.1143   0.1143


Time taken

=== Summary

Correctly C                                    98.6667 %
Incorrectly                2                    1.3333 %
Kappa statistic                          0.9796
Mean absolute error                      0.0206
Root mean squared error                  0.0851
Relative absolute error                  4.7192 %
Root relative squared error             18.209  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                  1         0.022      0.968                                   cluster1
                  1         0          1                                       cluster2
                  0.949     0          1                                       cluster3
Weighted Avg.     0.987     0.009      0.987

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  0  0 |  a = cluster1
  0 50  0 |  b = cluster2
  2  0 37 |  c = cluster3
```

Matrica konfuzije za klasifikovane instance

Expectation Maximization (EM)

Sastoji se iz dva koraka:

- E (expectation) korak – u ovom koraku podrazumevamo da znamo vrednosti parametara modela i na osnovu njih, za svaku instancu, računamo verovatnoću pripadanja svakom od klastera
- M (maximization) korak – na osnovu datih instanci, računamo (ponovo) vrednosti parametara modela; maksimizacija se odnosi na usklađivanje (parametara) modela sa datim podacima

Expectation Maximization (EM)

Postupak prilikom klasterovanja:

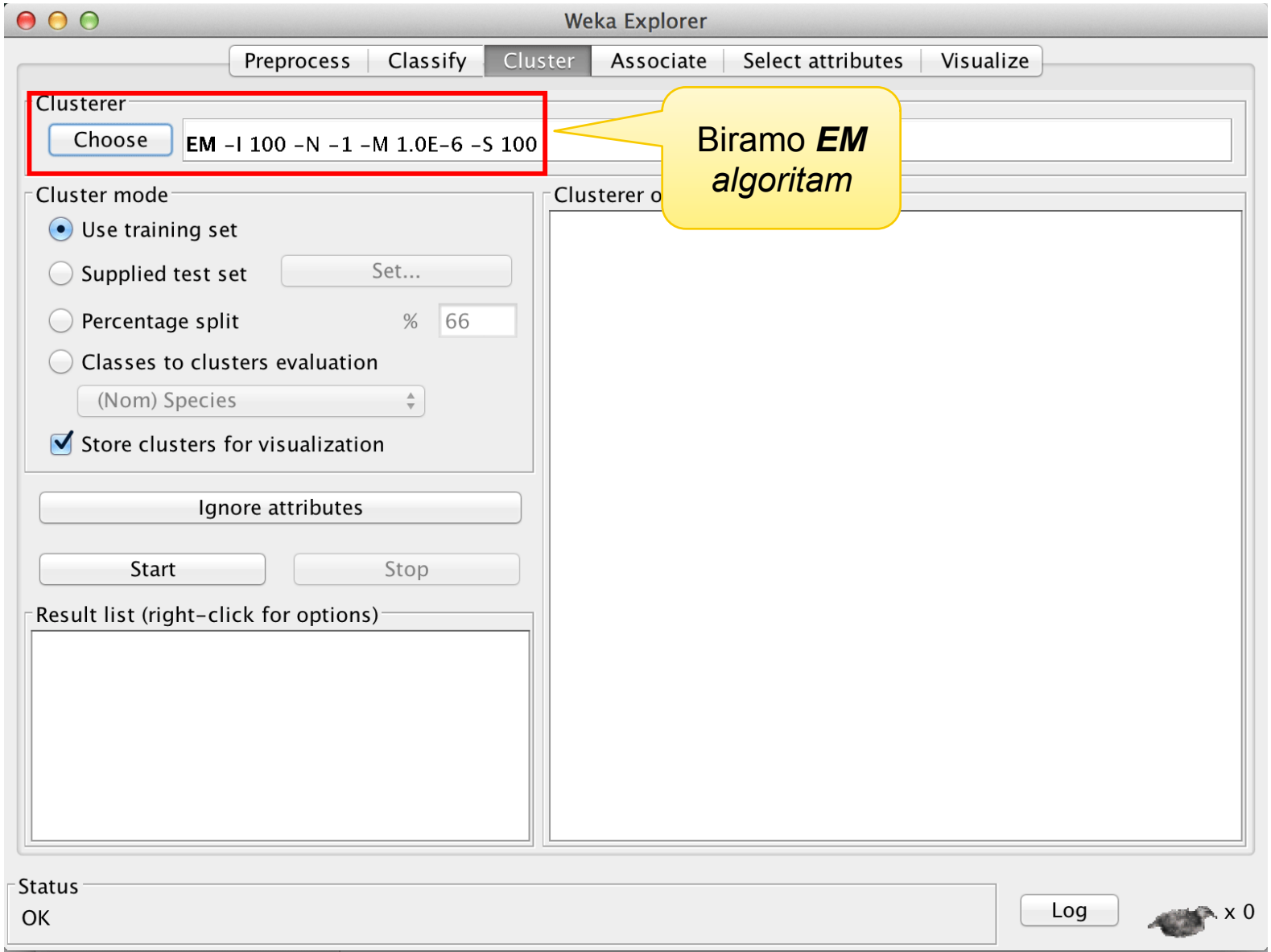
1) Inicijalno, definisati broj klastera (k) I nasumice izabrati vrednosti parametara modela (μ_i, σ_i, p_i, i=1,k)

2) Za date vrednosti parametara, za svaku instancu iz dataset-a, izračunati verovatnoću pripadanja svakom od klastera

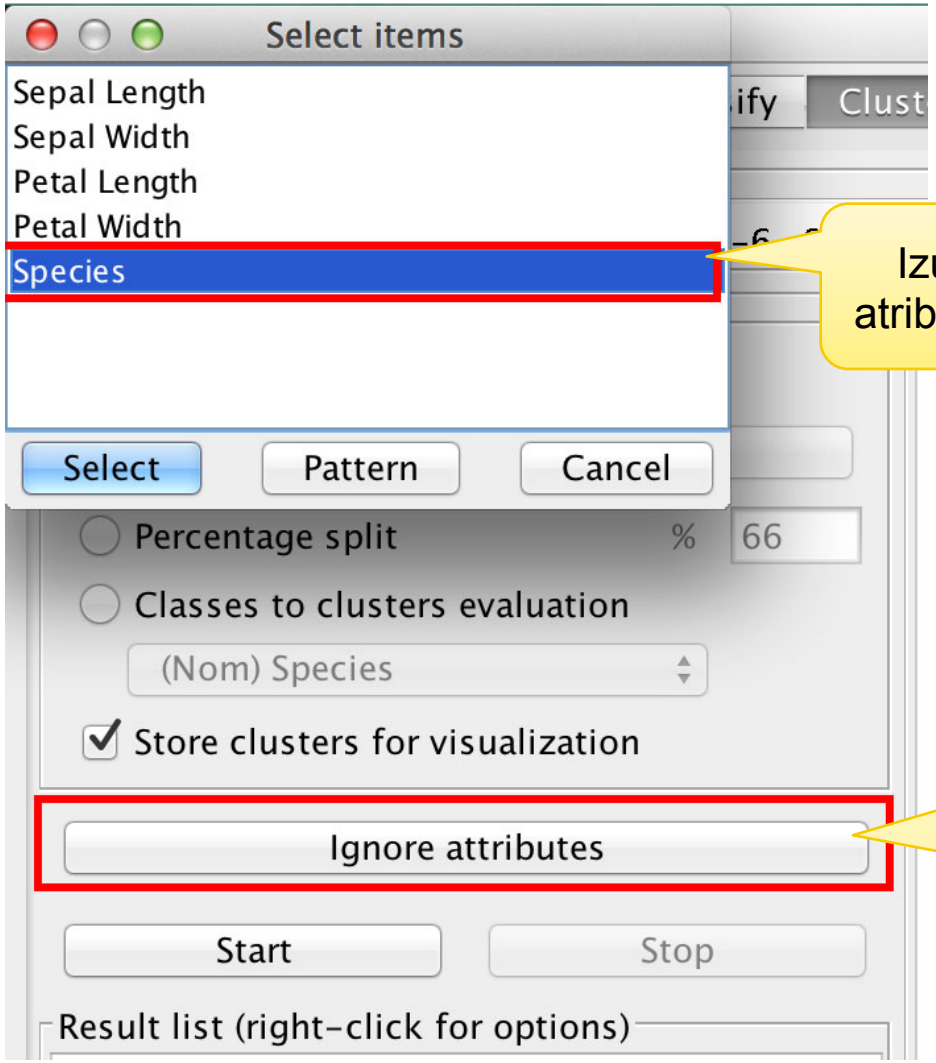
3) Na osnovu verovatnoća pripadnosti klasterima (instanci iz dataset-a), odrediti nove vrednosti parametara modela

Iterativno ponavljati korake 2) i 3) dok vrednosti parametara ne počnu da konvergiraju.

Korišćenje EM algoritma



Ne uzimanje u obzir klase



Izuzimamo atribut **Species**

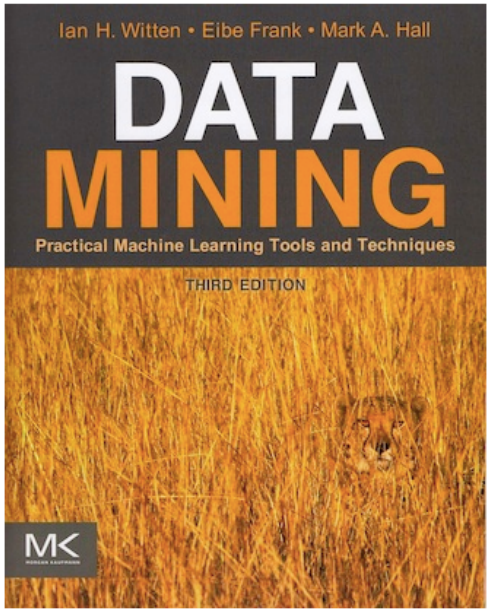
Selektovanje atributa koji neće biti korišćeni prilikom klasterovanja

Preporuke i zahvalnice

Weka Tutorials and Assignments @ The Technology Forge

- Link: http://www.technologyforge.net/WekaTutorials/

Witten, Ian H., Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.



(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I

Pitanja?

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

URL: http://nikola.milikic.info