Diskretizacija i selekcija atributa. Klasterovanje

NIKOLA MILIKIĆ

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

URL: http://nikola.milikic.info

Osobine Naivnog Bajesa

- Namenjen primarno za rad sa nominalnim atributima
- U slučaju numeričkih atributa:
  - koristiti raspodelu verovatnoća atributa (tipično Normalna raspodela) za procenu verovatnoće svake od vrednosti atributa
  - uraditi diskretizaciju vrednosti atributa

Diskretizacija atributa

Diskretizacija je proces transformacije numeričkih podataka u nominalne tako što se numeričke vrednosti smeštaju u odgovarajuće grupe kojih ima konačan broj.

Najčešći pristupi diskretizacije su:

- Nenadgledani pristupi:
  - Jednake širine opsega (Equal-width binning)
  - Jednaka pojavljivanja u opsezima (Equal-frequency binning)
- Nadgledani pristupi – uzimaju u obzir klase

Jednake širine opsega

Jednake širine opsega (eng. Equal-width binning) deli opseg mogućih vrednosti na **N** podopsega iste širine.

širina = (maks. vrednost – min. vrednost) / N

Primer: Ako je opseg posmatranih vrednosti između 0 – 100, možemo kreirati 5 podopsega na sledeći način:

Širina = (100 – 0) / 5 = 20

Opsezi su: [0-20], (20-40], (40-60], (60-80], (80-100]

Obično se prvi i poslednji opsezi proširuju kako bi uključili vrednosti van opsega.

Jednaka pojavljivanja u opsezima

Jednaka pojavljivanja u opsezima (eng. Equal-frequency ili equal-height binning) deli opseg mogućih vrednosti na **N** podopsega gde svaki podopseg sadrži isti broj instanci.

Primer: Pretpostavimo da želimo da smestimo u 5 podopsega vrednosti:

5, 7, 12, 35, 65, 82, 84, 88, 90, 95

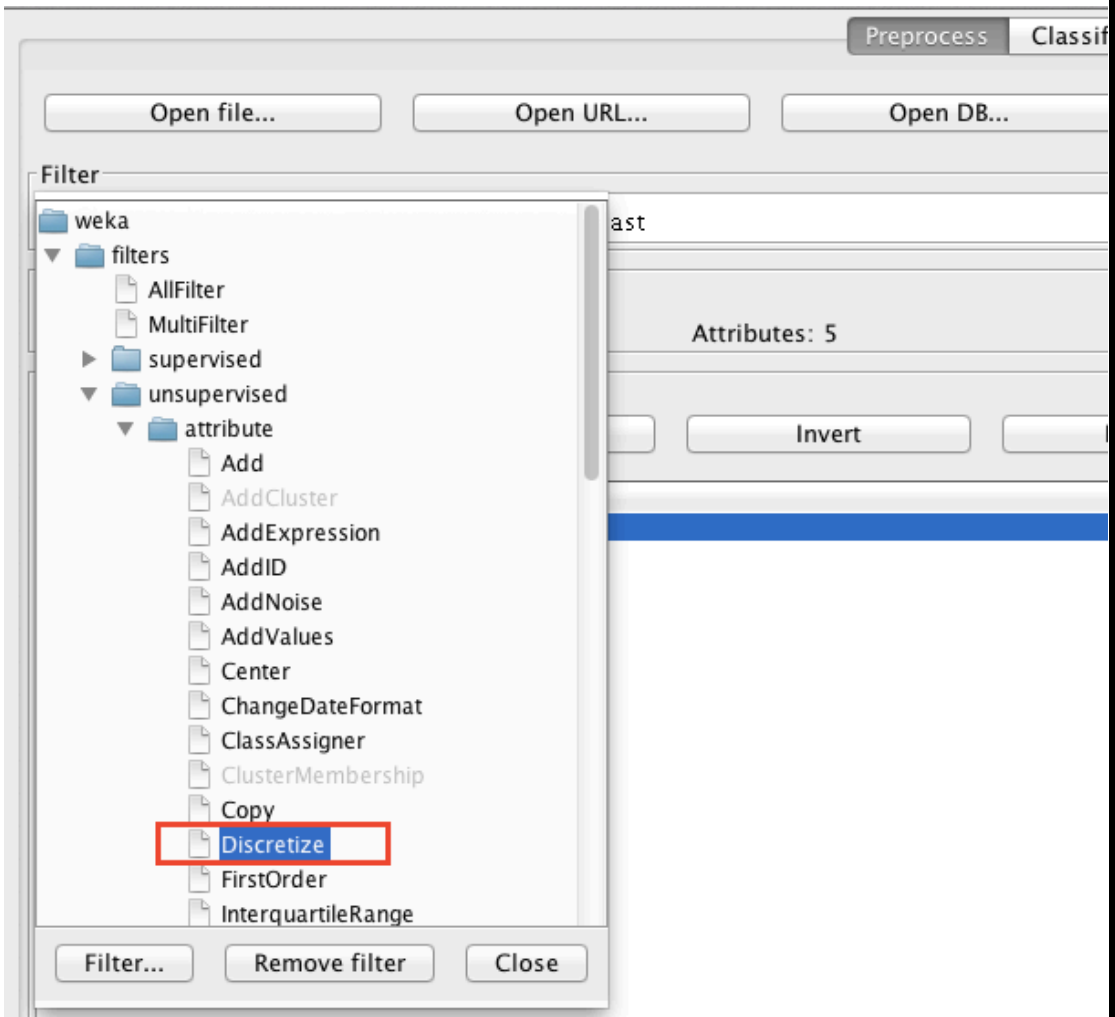
Podopsege ćemo podeliti tako što će svaki sadržati po dve instance:

5, 7, | 12, 35, | 65, 82, | 84, 88, | 90, 95

Diskretizacija u Weka-i

Atributi se diskretizuju tako što se nad njihovim vrednostima primeni odgovarajući *Filter*.

Na *Preprocess* tabu se bira opcija *Choose -> Filter* i u folderu *filters/unsupervised/attribute* se odabira filter ***Discretize***.

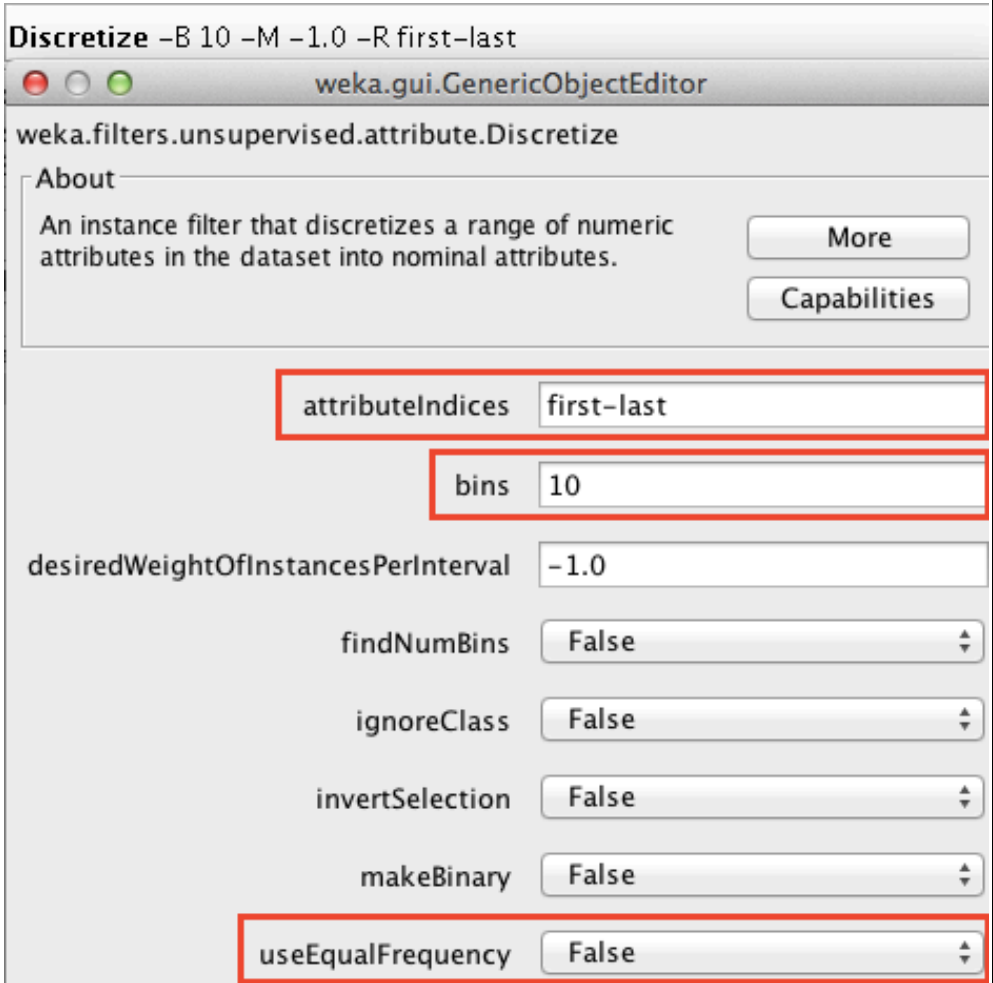


FishersIrisDataset.arff

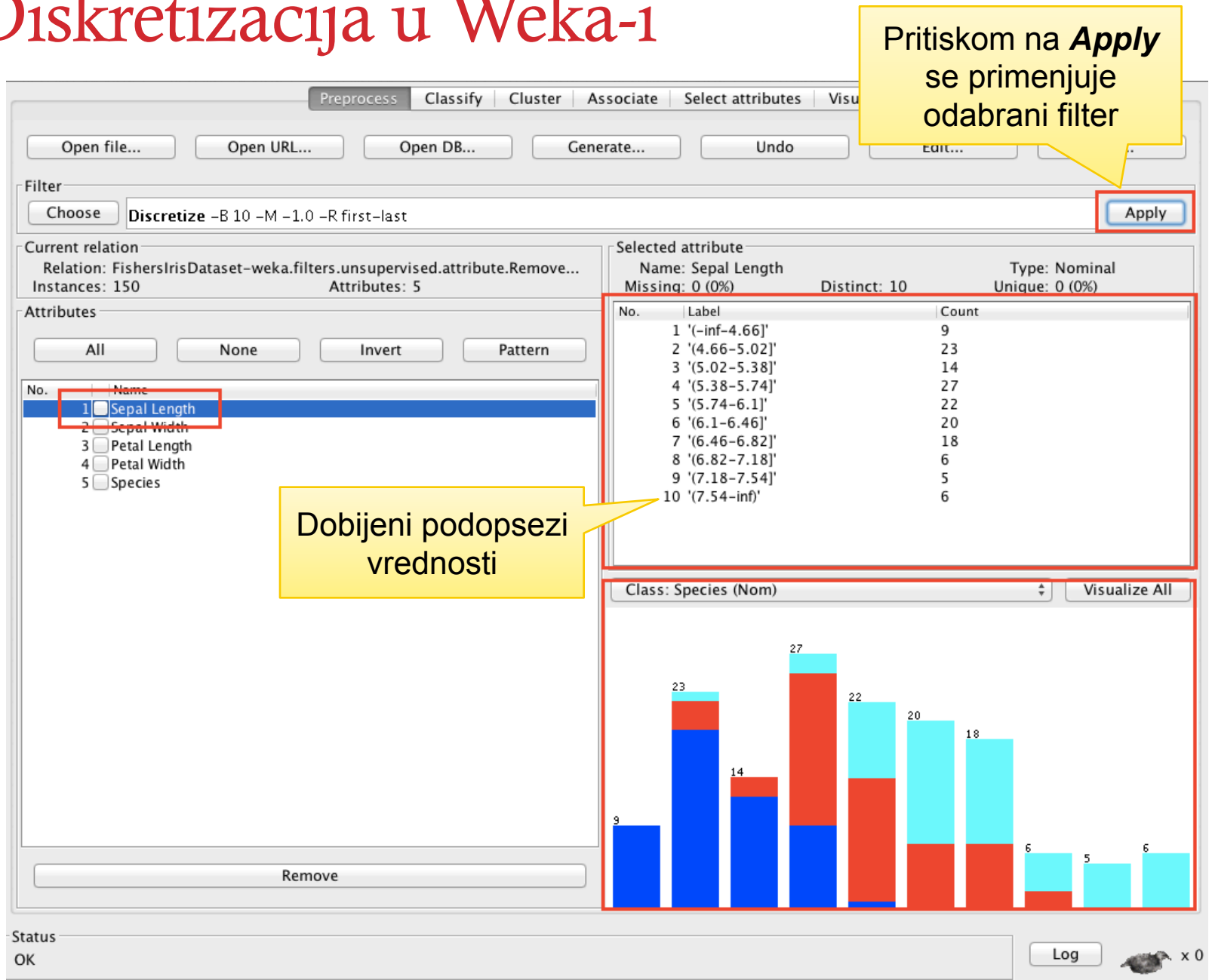
Diskretizacija u Weka-i

Po defaultu se primenjuje Diskretizacija sa jednakim širinama opsega.

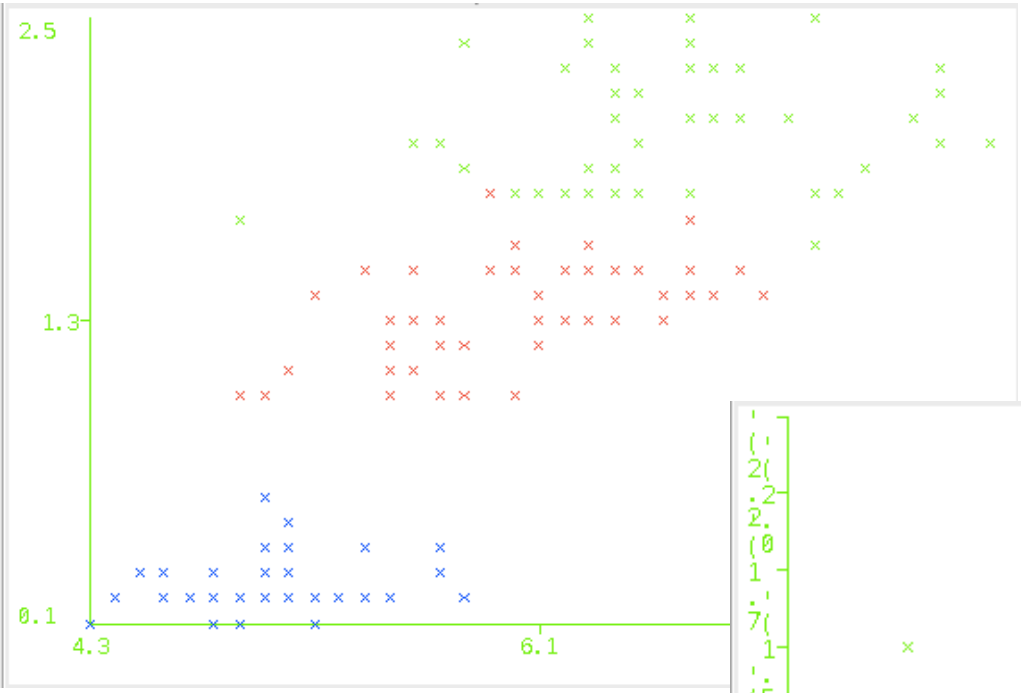
- ***attributeIndices*** - vrednost *first-last* označava da diskretizjemo sve atribute. Mogu se navesti i redni brojevi atributa
- ***bins*** - željeni broj opsega
- ***useEqualFrequency*** – *true* ako se koristi diskretizacija sa jednakim pojavljivanjima u opsezima, *false* ako se koristi Diskretizacija sa jednakim širinama opsega



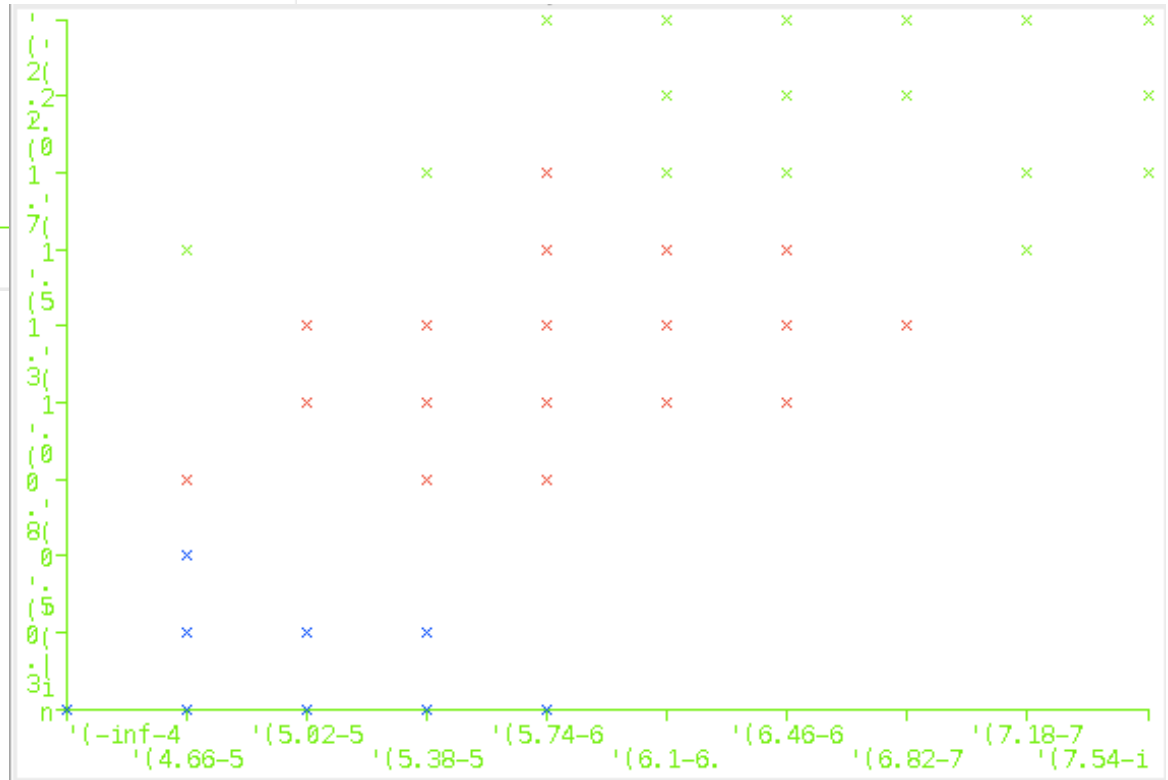
Diskretizacija u Weka-i



Podaci pre i posle diskretizacije



Podaci pre diskretizacije



Podaci nakon diskretizacije

Selekcija atributa

Selekcija atributa (eng. Attribute Selection ili Feature Selection) je proces odabira podskupa relevantnih atributa koji će se koristiti.

Primenjuje se u slučajevima kada se u datasetu nalaze atributi koji su redundantni ili nerelevantni.

- Redundanti atributi su oni koji ne pružaju nikakve dodatne informacije u odnosu na već selektovane atribute.
- Nerelevantni atributi su oni koji ne pružaju nikakve informacije u datom kontekstu.

Prednosti primene selekcije atributa

Suvišni atributi mogu degradirati performanse modela.

Prednosti selekcije atributa:

- Poboljšava čitljivost modela time što se model sastoji samo iz relevantnih atributa
- Kraće vreme treniranja
- Povećana generalizacija time što smanjuje mogućnosti za overfitting

Najbolji način za selekciju atributa je ručno ukoliko se dobro poznaje problem koji se rešava. I automatizovani pristupi selekcije daju dobre rezultate.

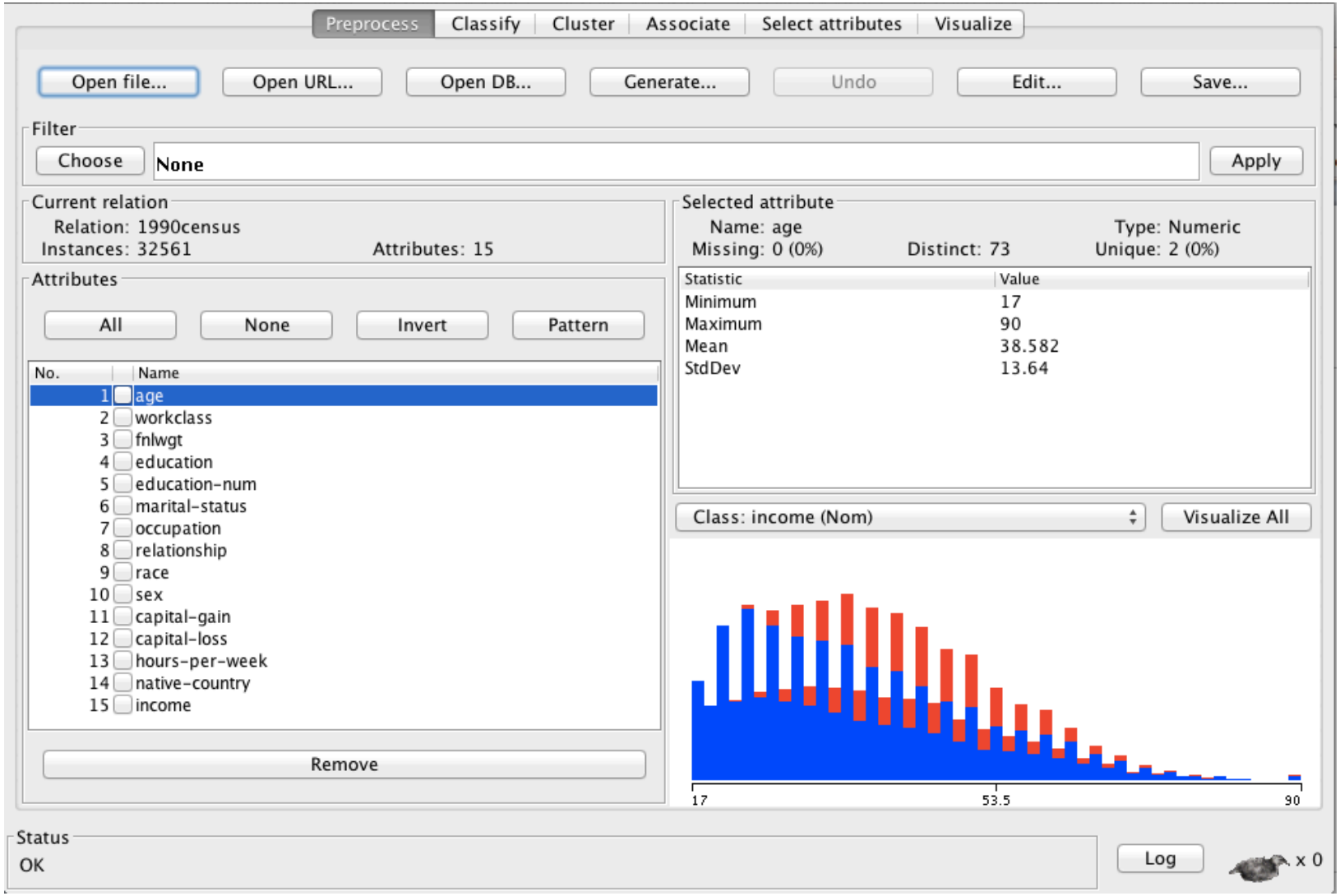
Pristupi selekcije atributa

Postoje dva pristupa:

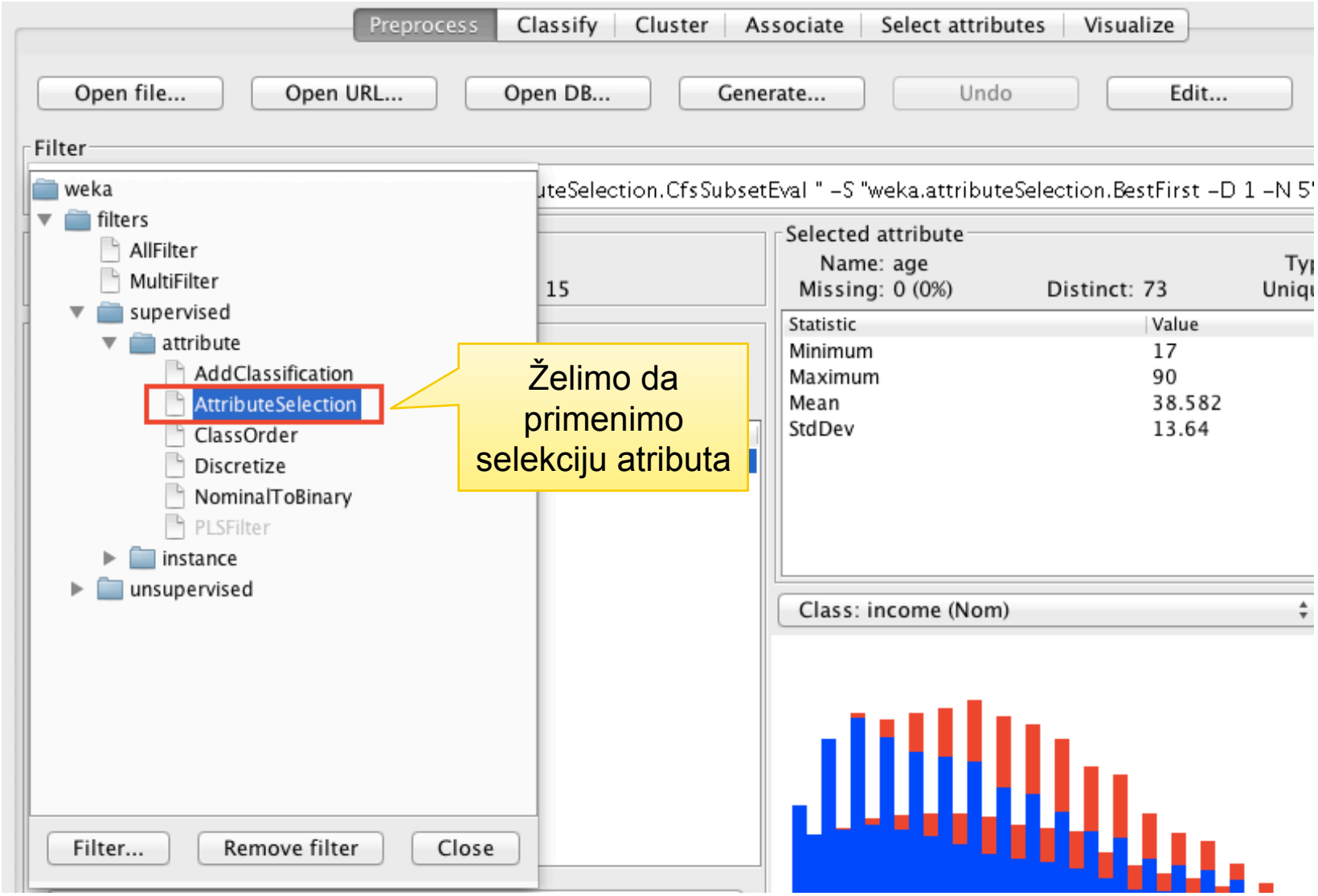
- *Filter* metoda – koriste se procene na osnovu generalnih svojstava podataka
- *Wrapper* metoda – podskupovi atributa se evaluiraju primenom algoritma mašinskog učenja koji će se koristiti nad skupom podataka. Naziv Wrapper se koristi iz razloga što je algoritam učenja "zapakovan" u samom procesu selekcije. Biće odabran onaj podskup atributa za koje dati algoritam učenja daje najbolje rezultate.

Primer selekcije atributa

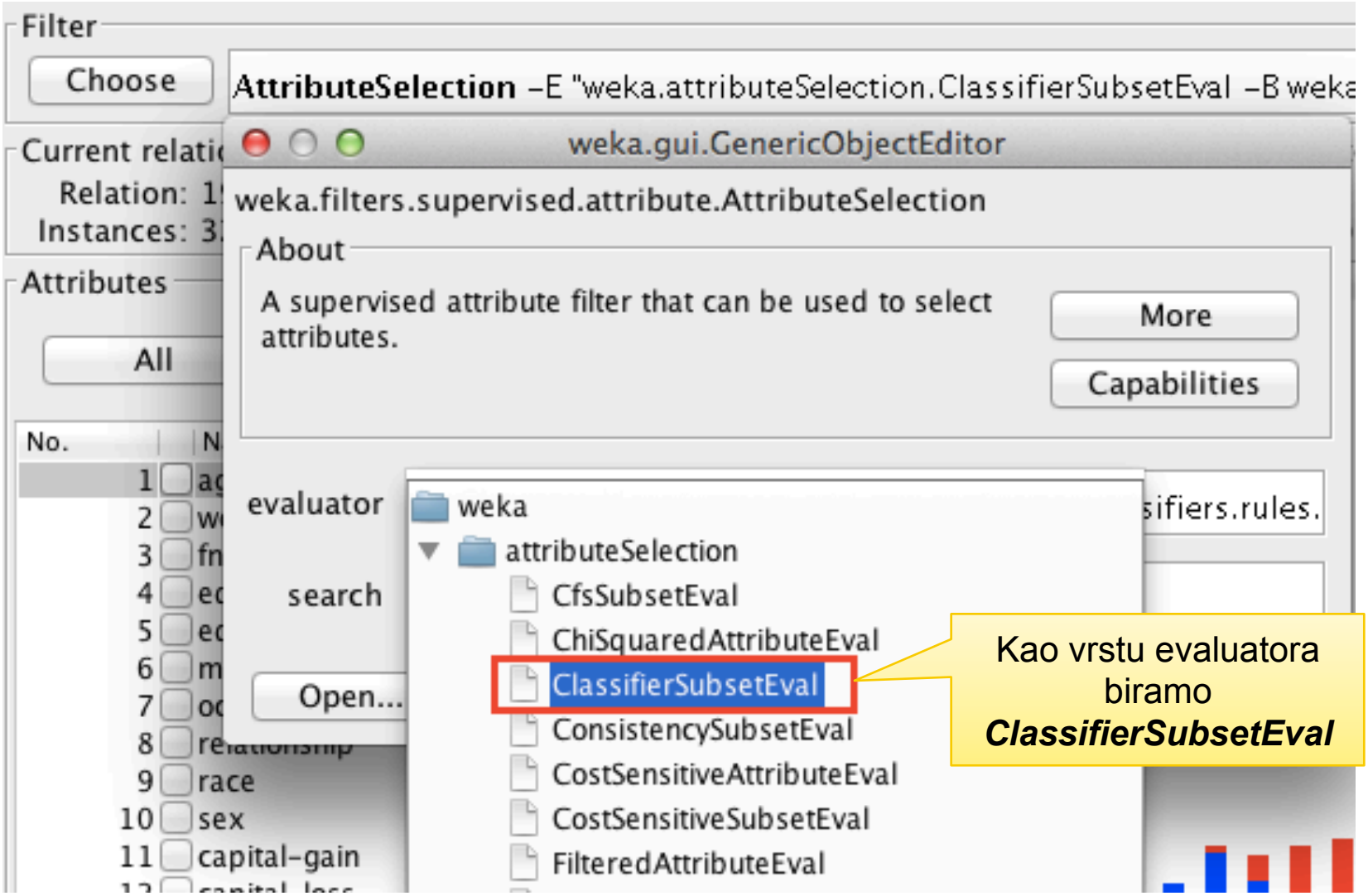
census90-income.arff



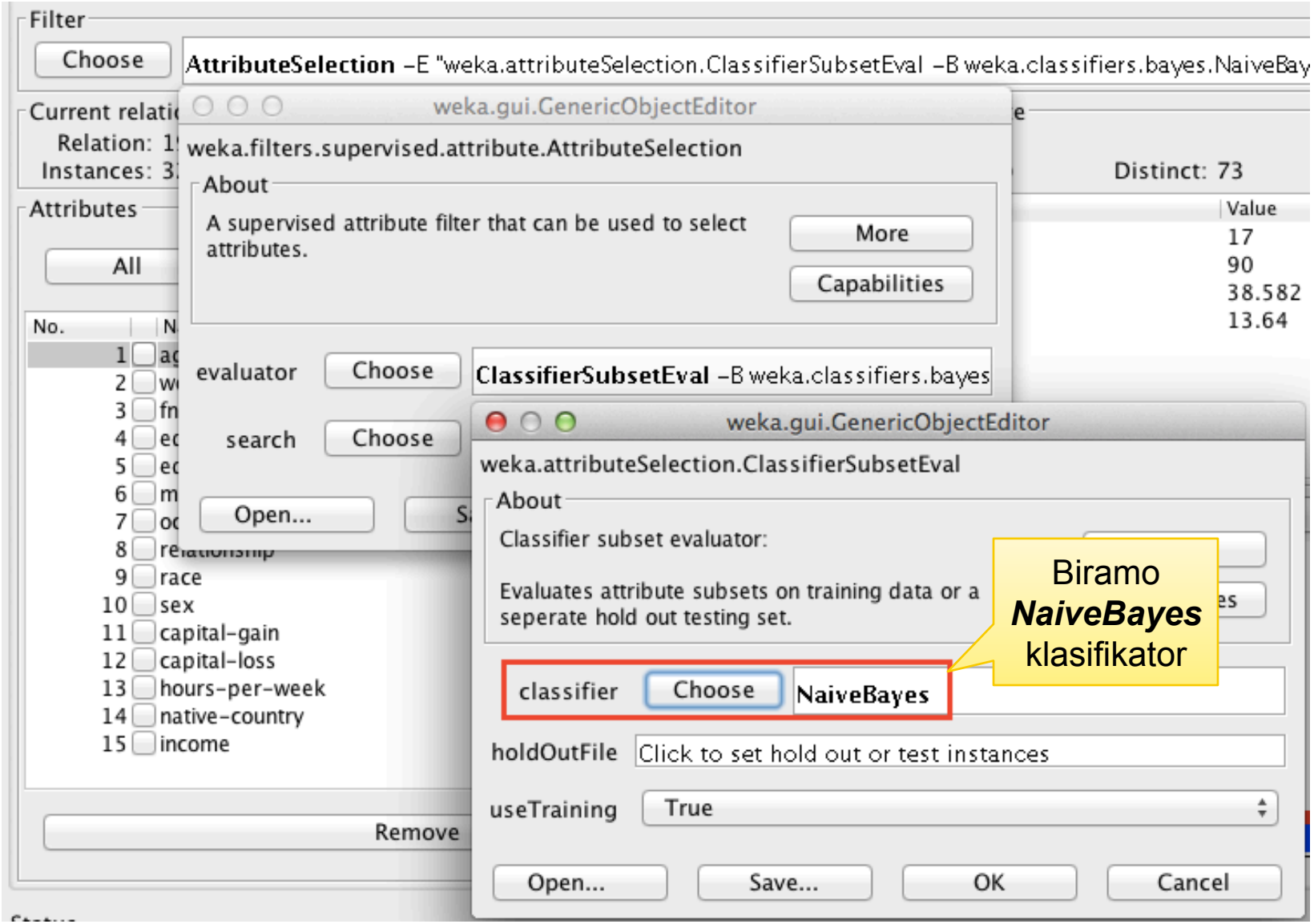
Primer selekcije atributa



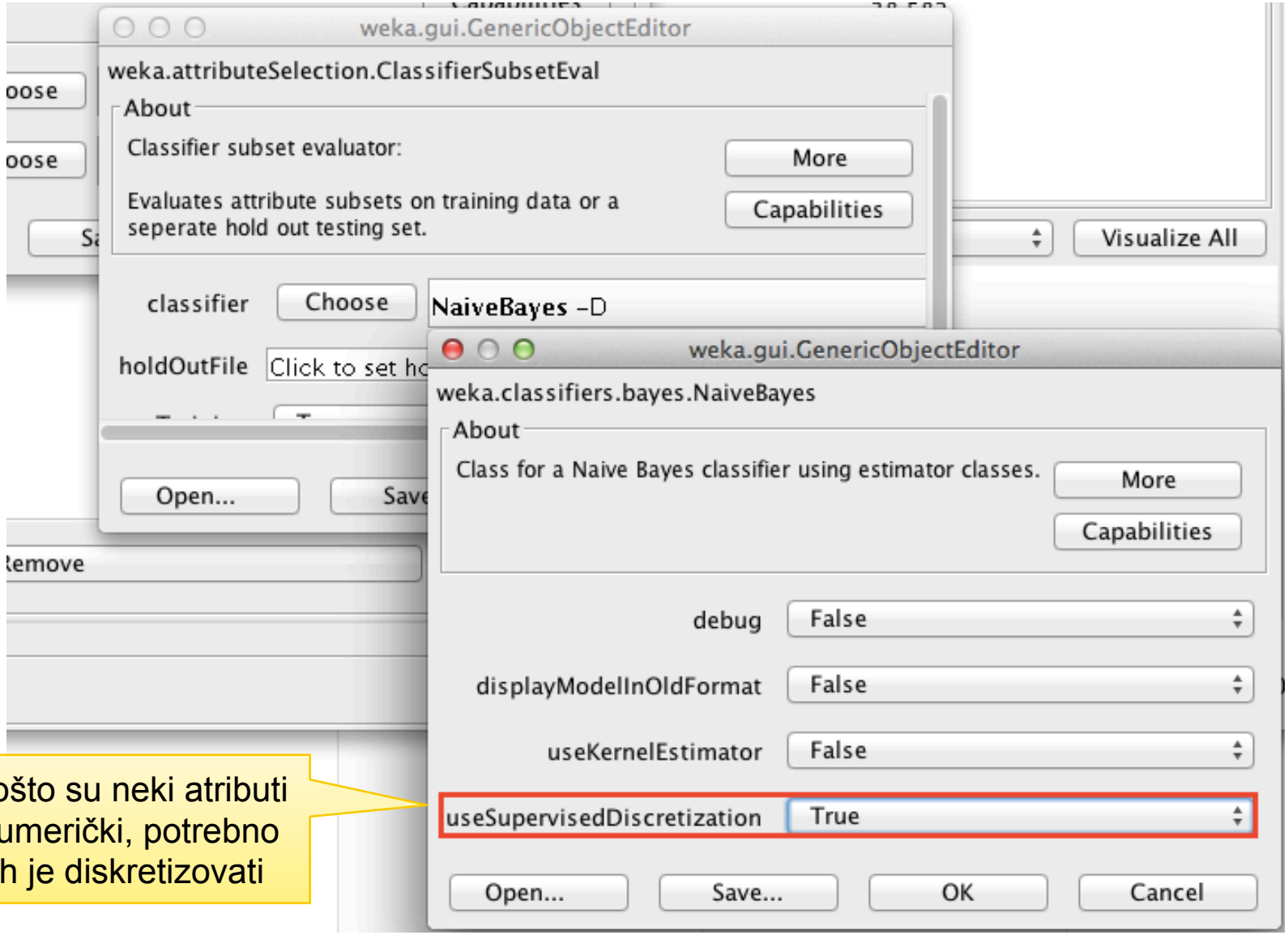
Primer selekcije atributa



Primer selekcije atributa

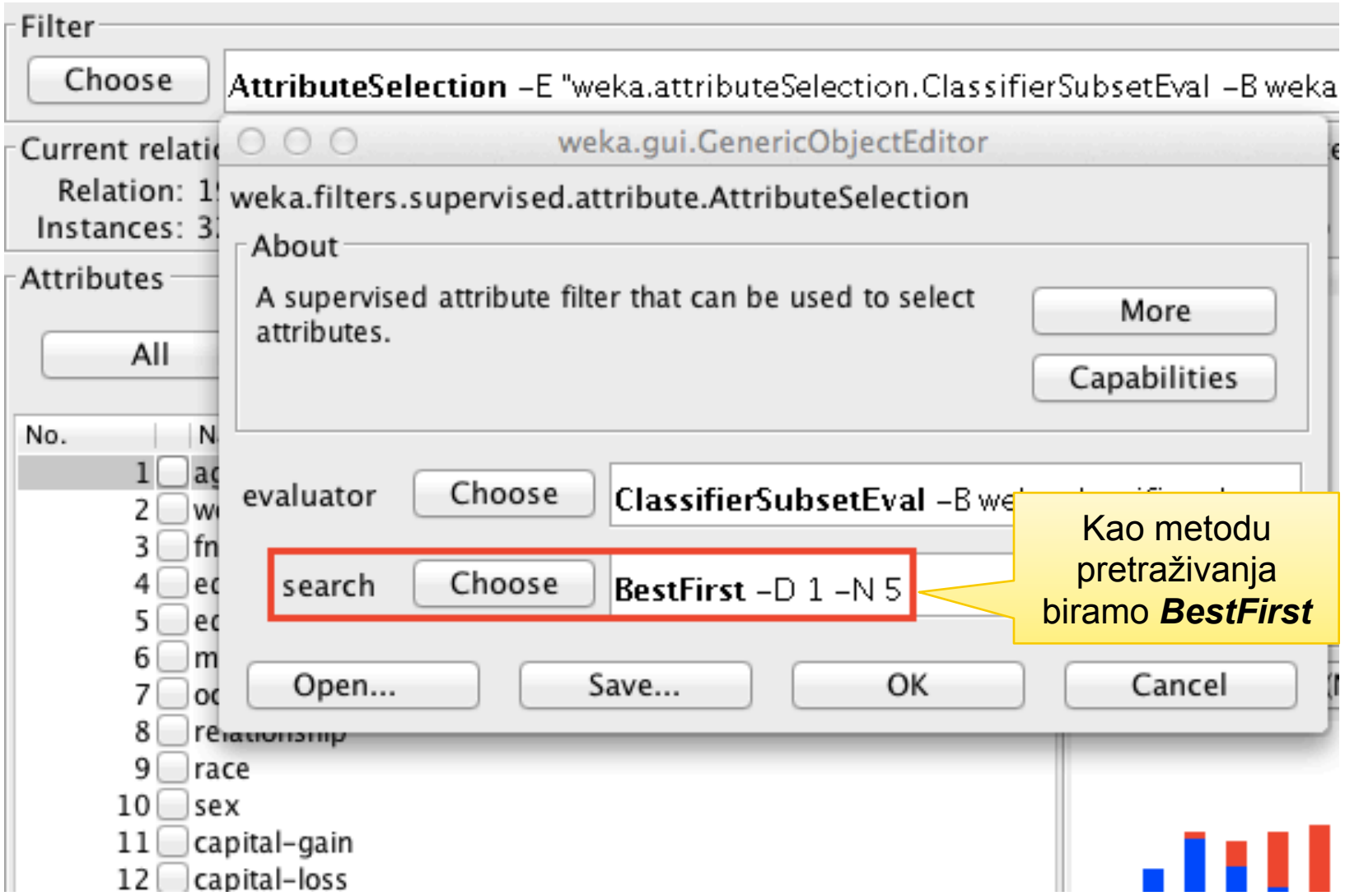


Primer selekcije atributa

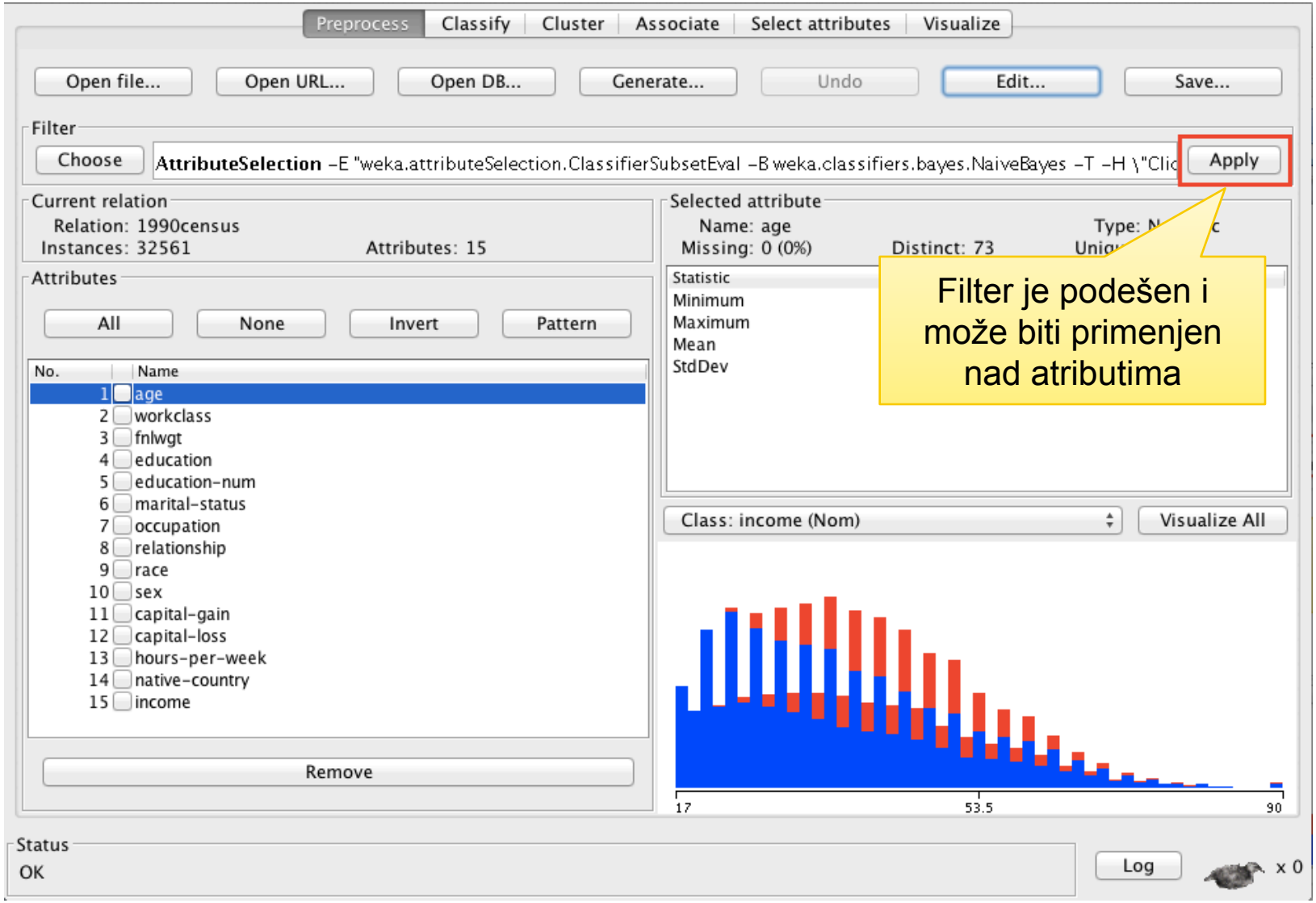


Pošto su neki atributi numerički, potrebno ih je diskretizovati

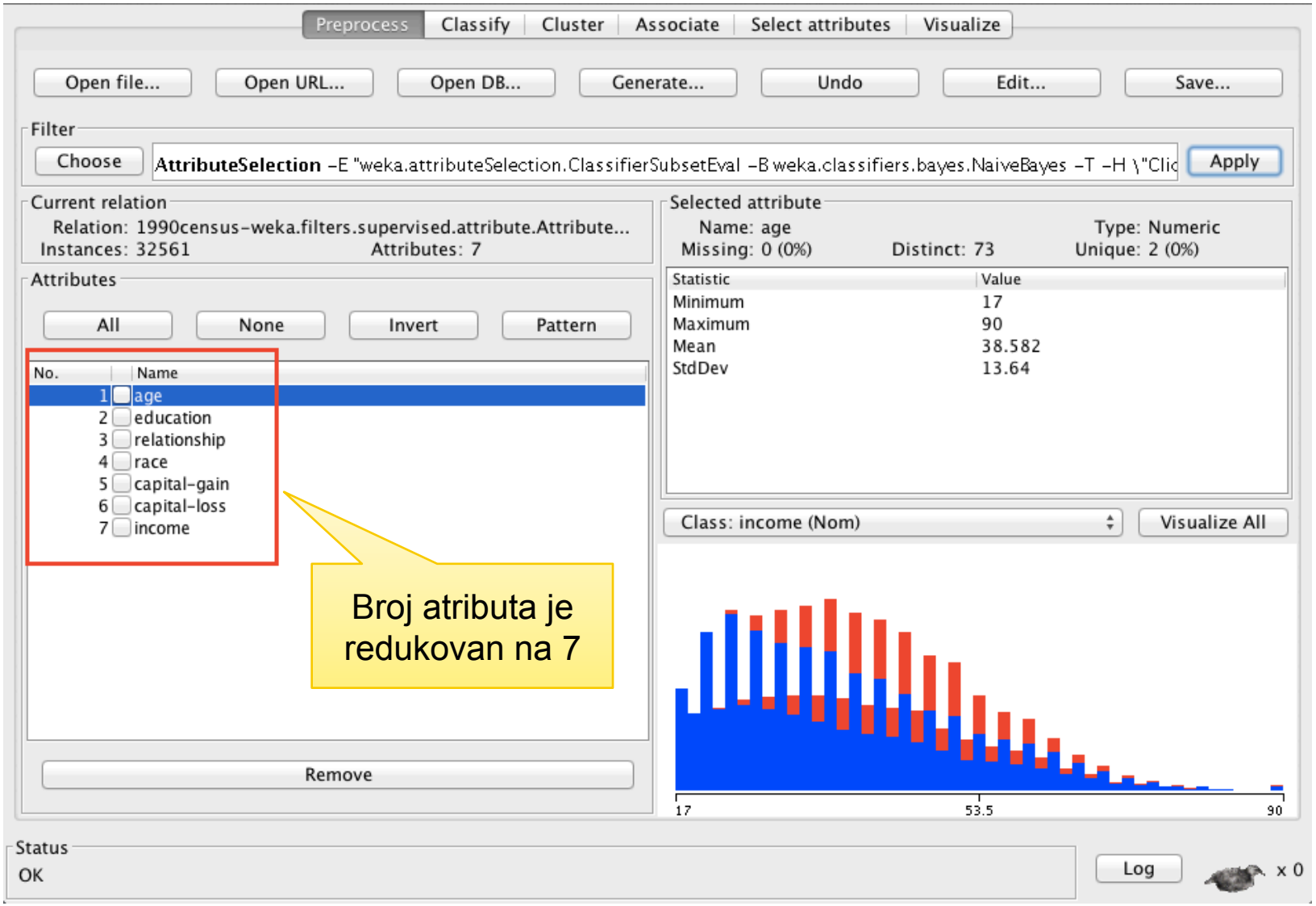
Primer selekcije atributa



Primer selekcije atributa



Primer selekcije atributa



Klasterizacija

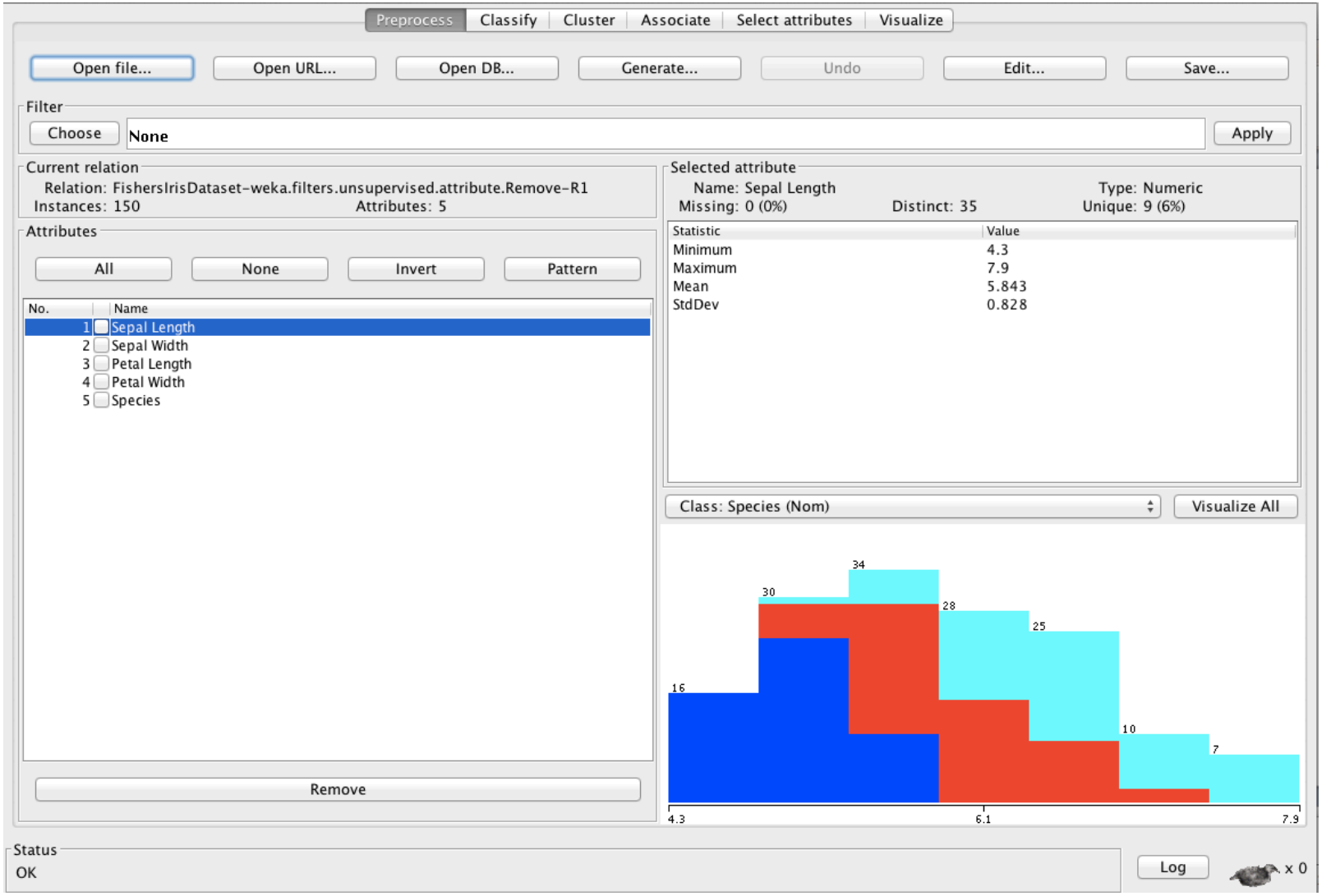
Klasterizacija (eng. Clustering) spada u grupu tehnika nenadgledanog učenja i omogućava grupisanje instanci u grupe, gde unapred ne znamo koje su sve grupe moguće.

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

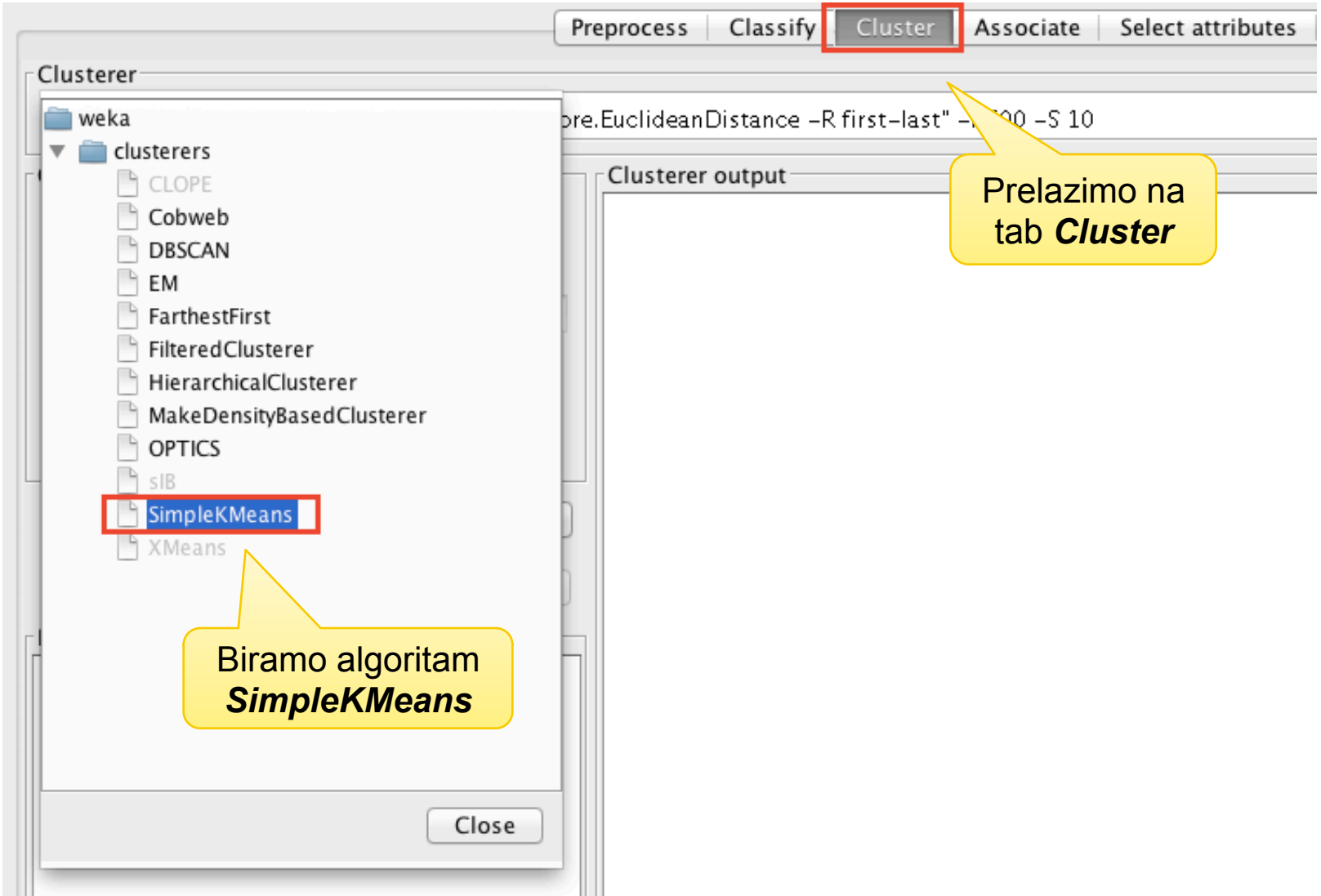
Kao rezultat klasterizacije svakoj instanci je dodeljen novi atribut koji predstavlja klaster kojoj pripada. Može se reći da je klasterovanje uspešno ukoliko su dobijeni klasteri smisleni i ukoliko se mogu imenovati.

K-Means algoritam u Weka-i

FishersIrisDataset.arff



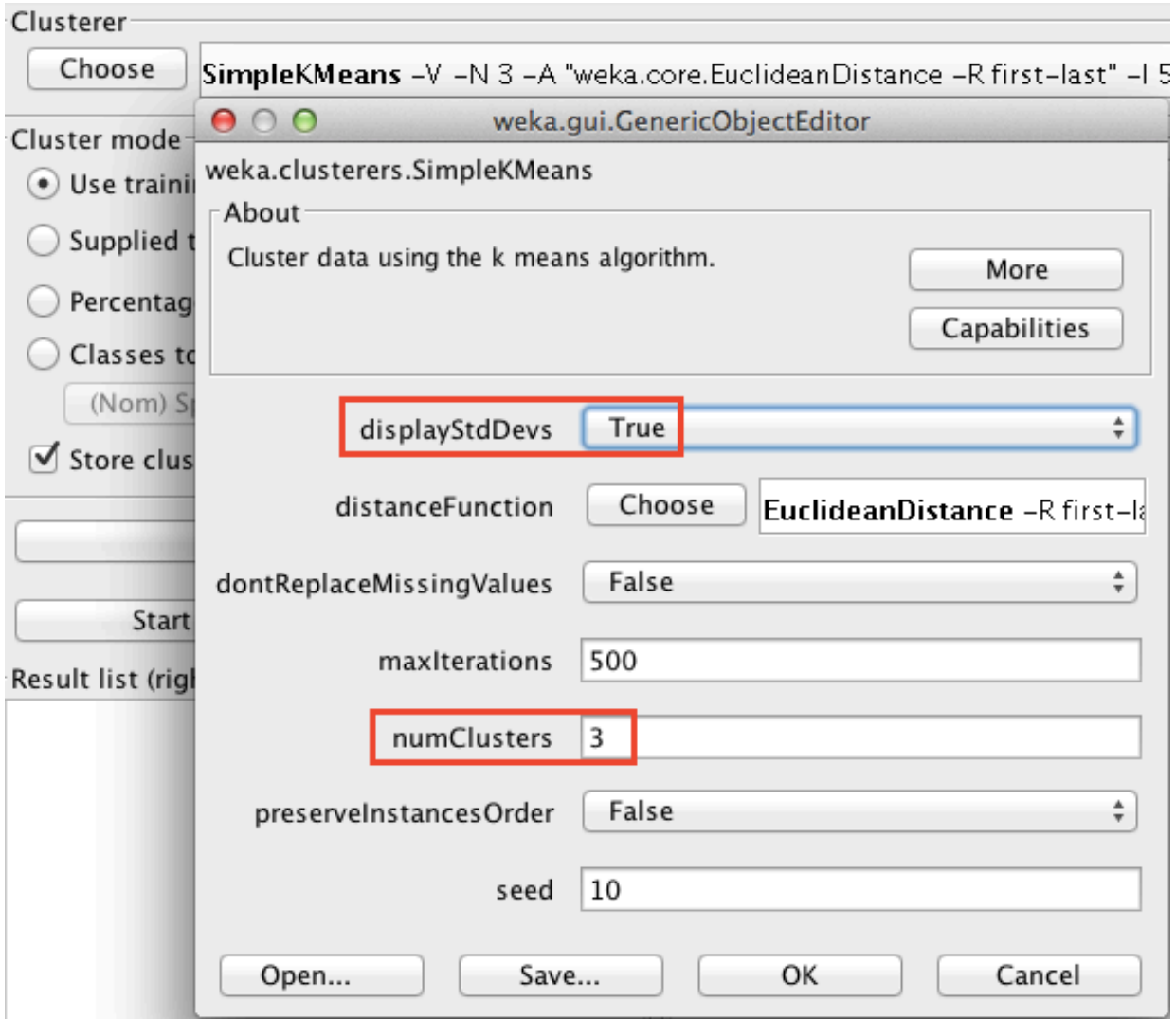
Odabir algoritma za klasterovanje



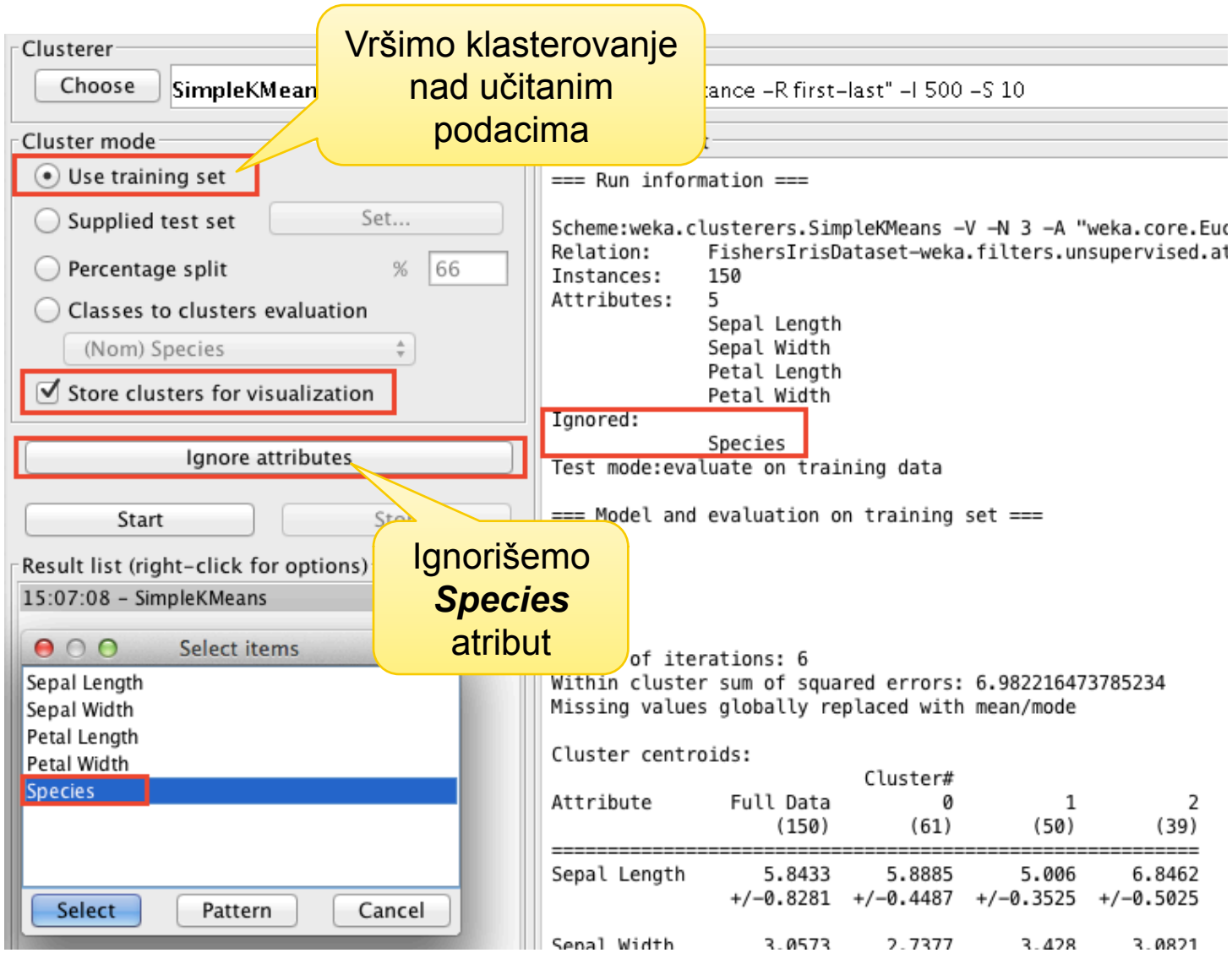
Podešavanje parametara

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

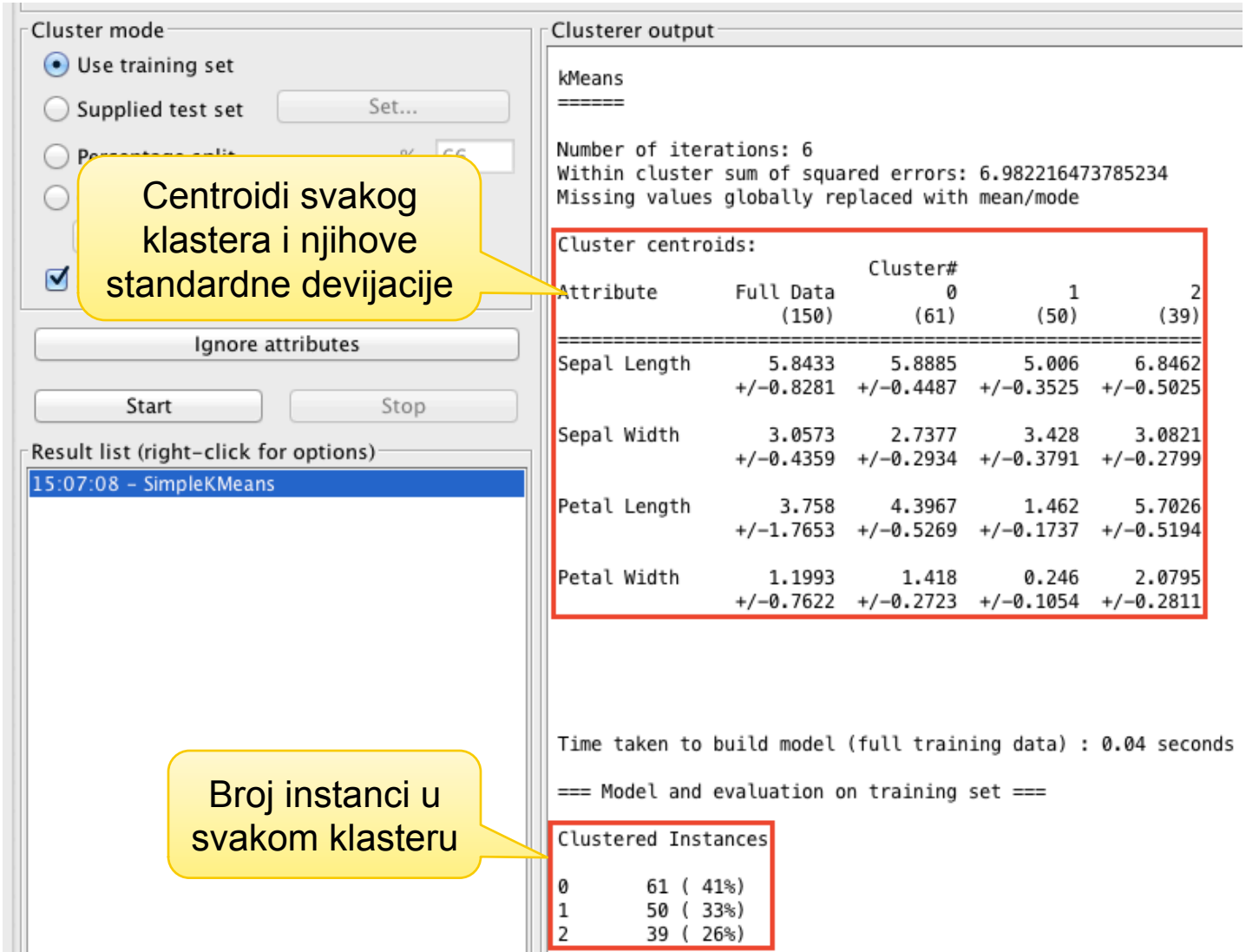
displayStdDevs– ako je *true*, onda će se ispisati vrednosti standardne devijacije



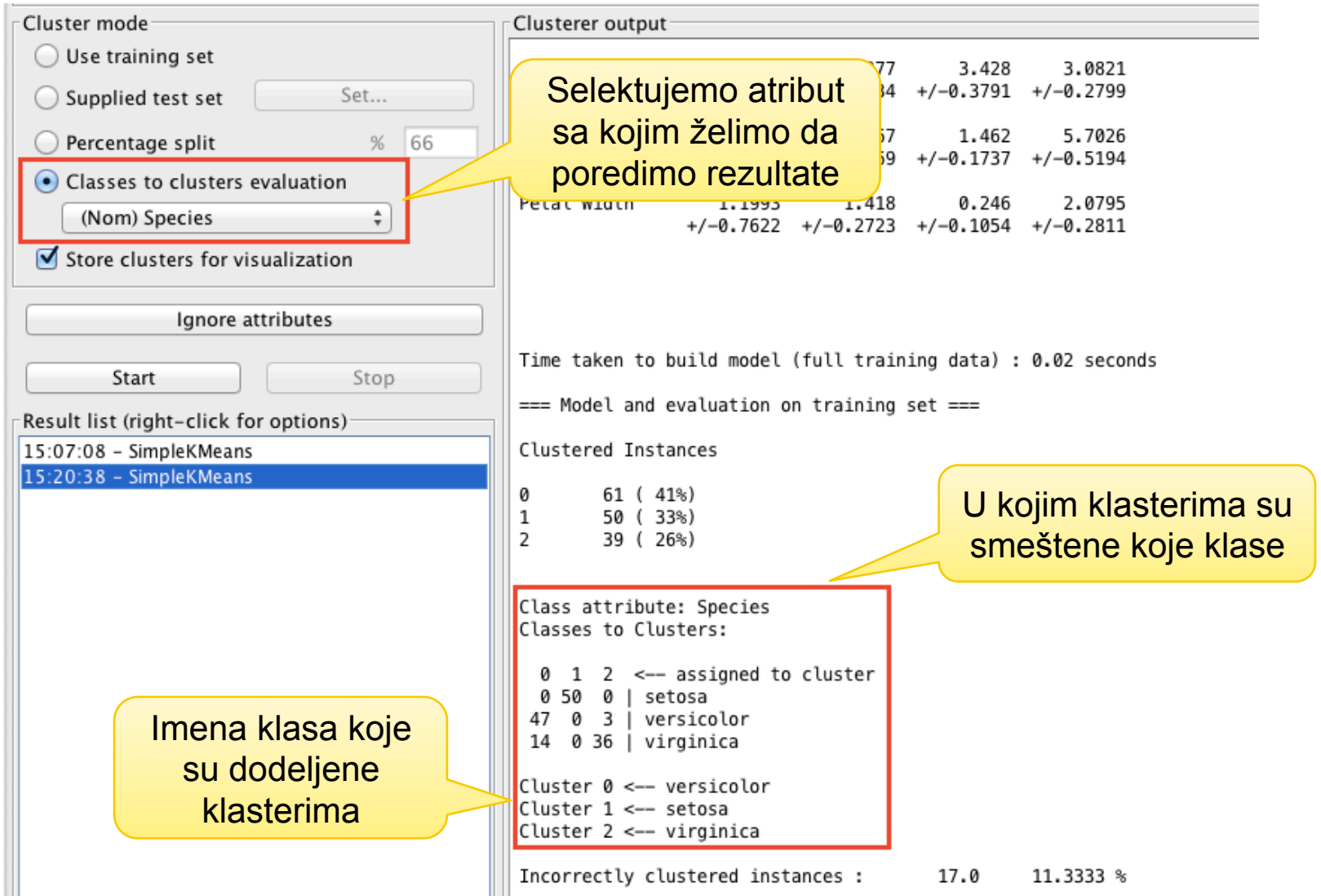
Pokretanje procesa klasterovanja



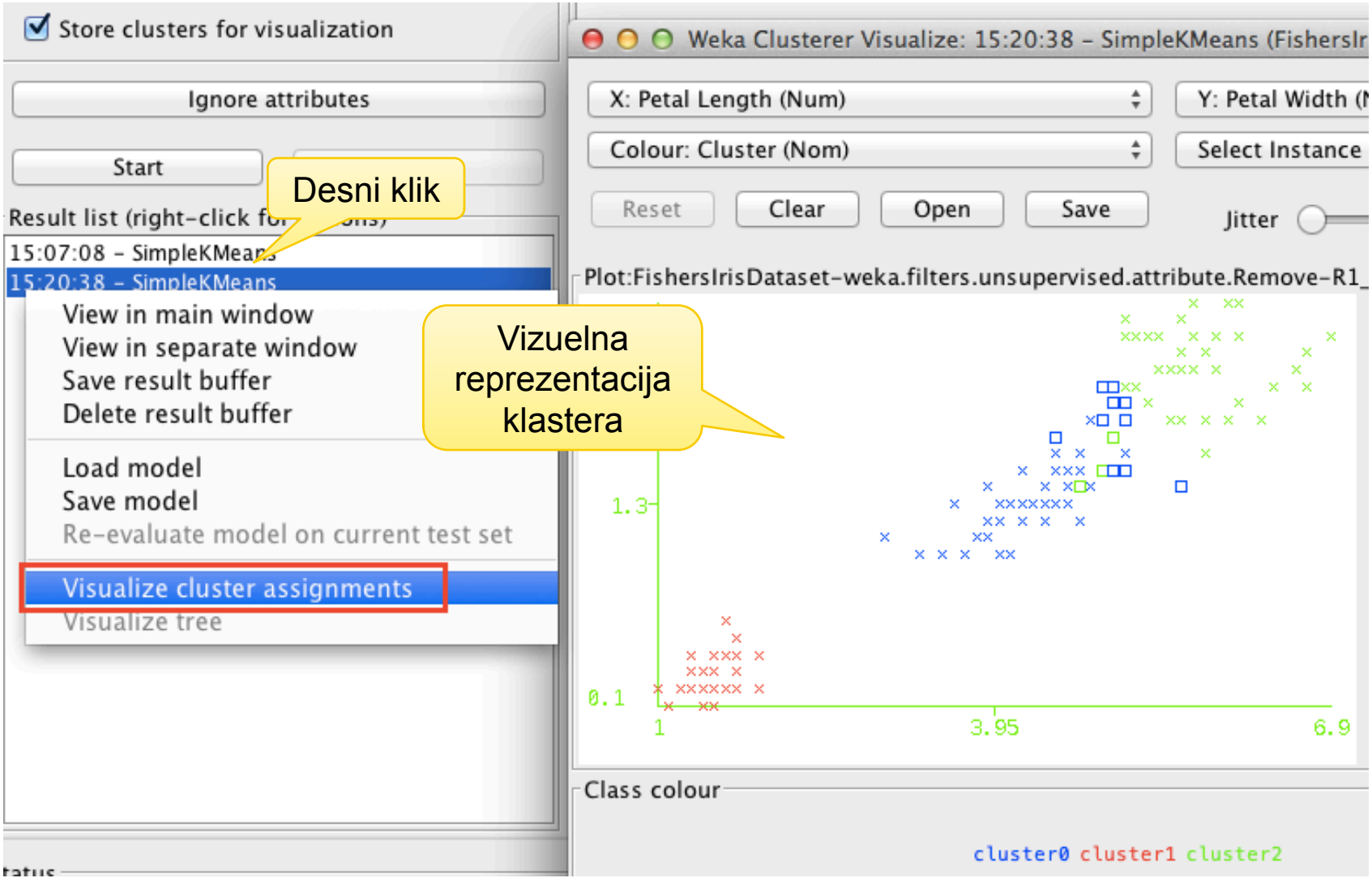
Rezultat klasterovanja



Evaluacija rezultata

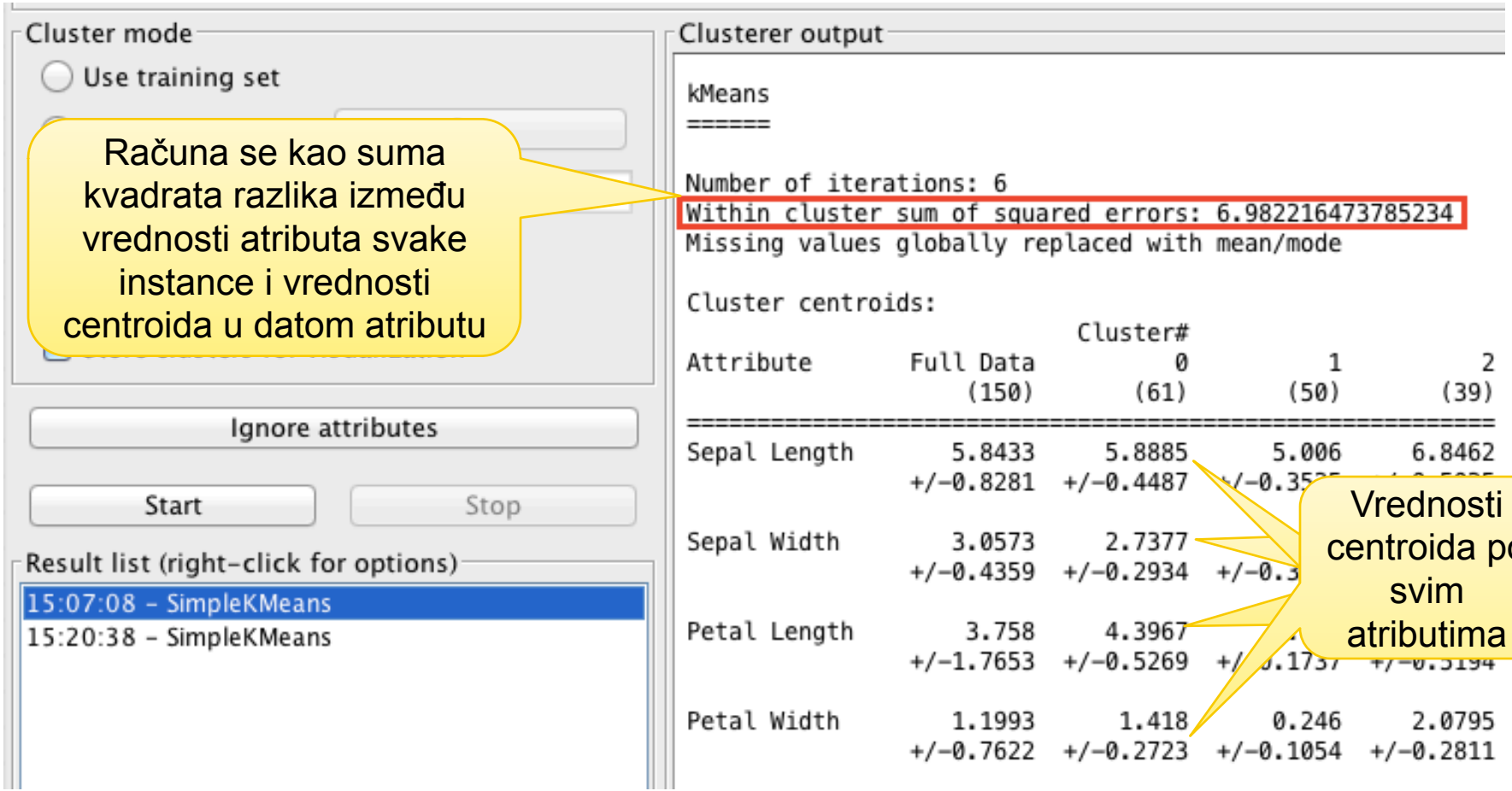


Vizuelizacija klastera



Procena uspešnosti klasterovanja

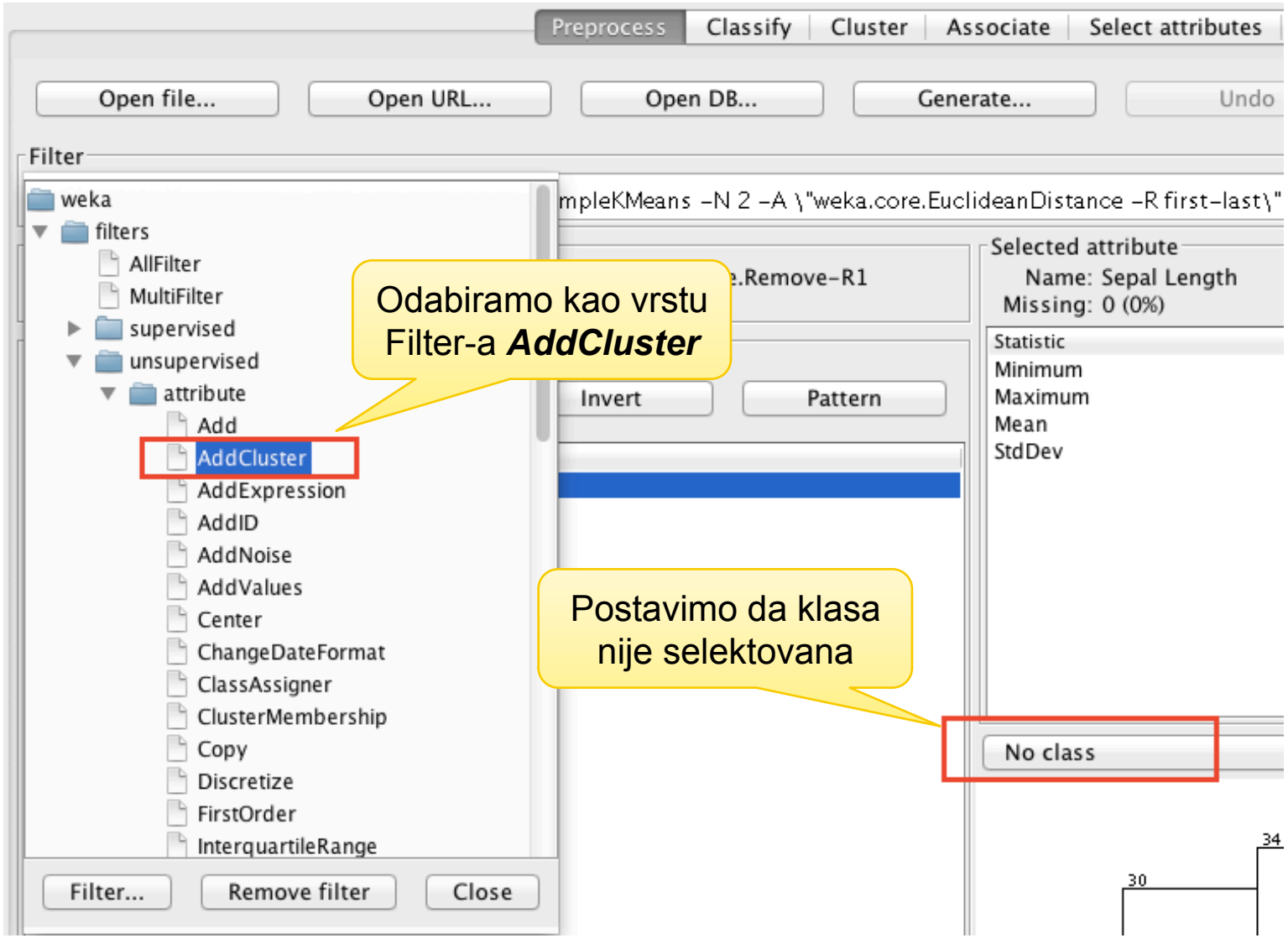
Within cluster sum of squared error (suma kvadrata greške unutar klastera) daje procenu kvaliteta dobijenih klastera



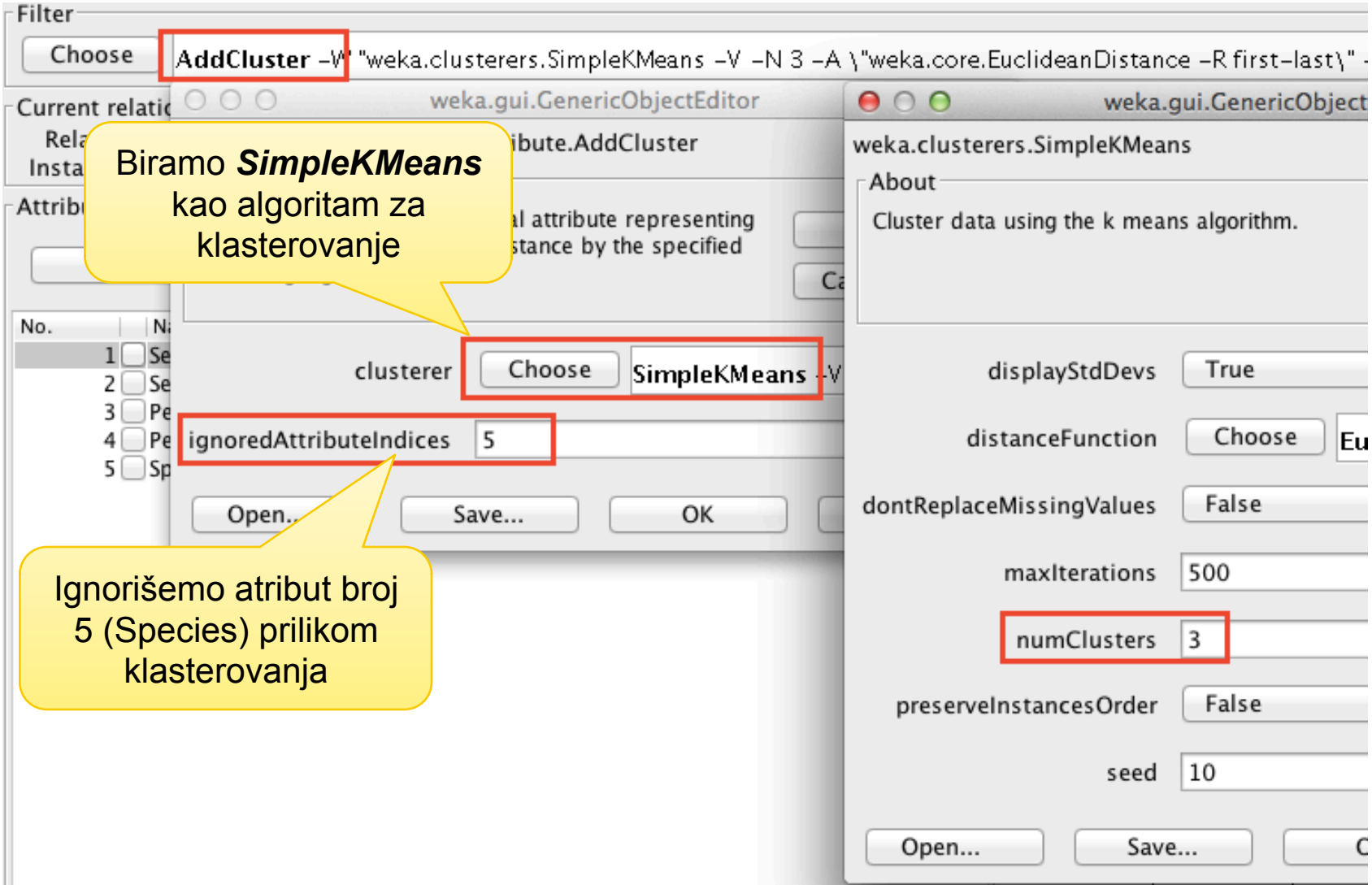
Kako proceniti dobar broj klastera?

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

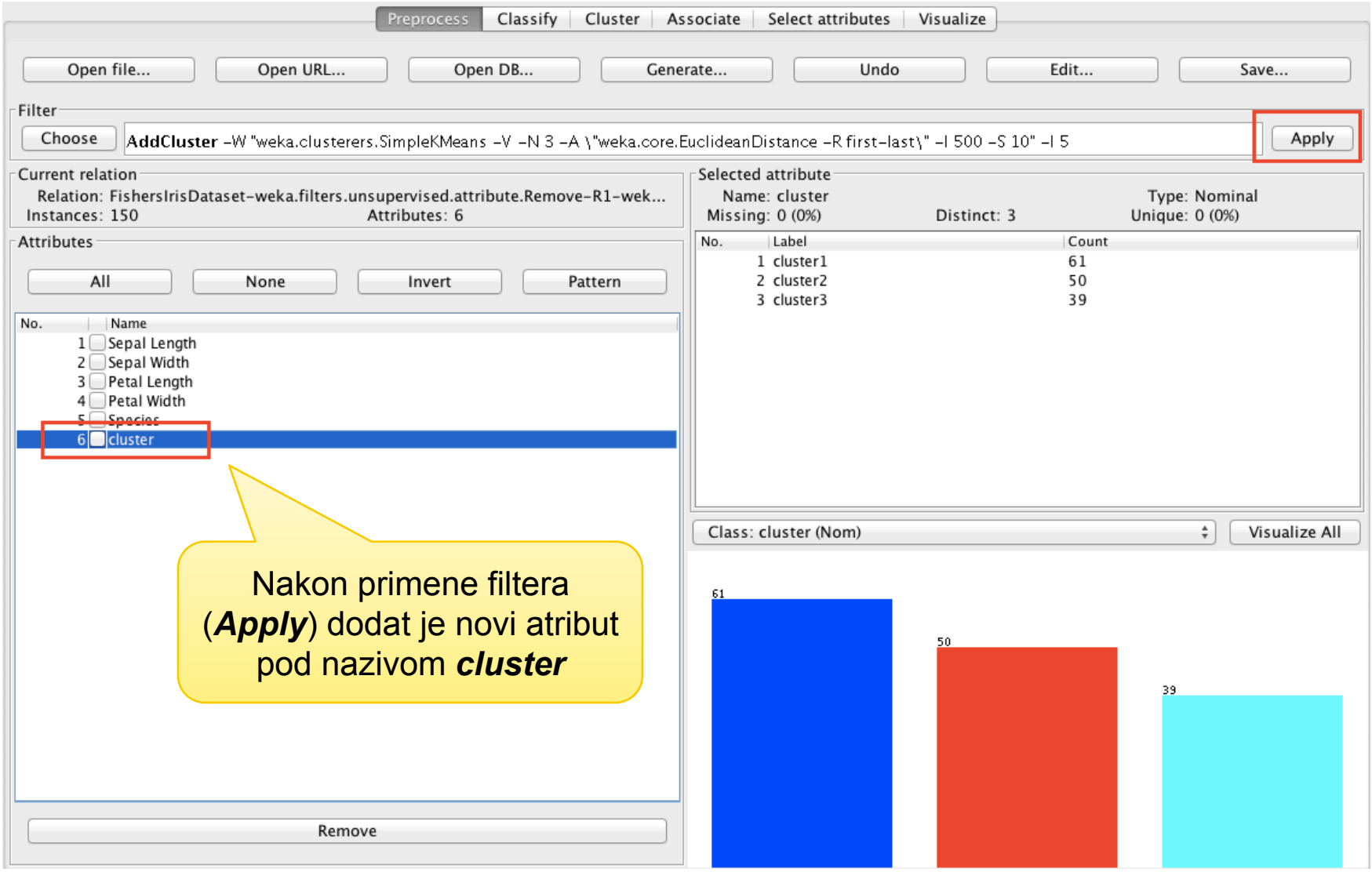
Korišćenje klastera za klasifikaciju



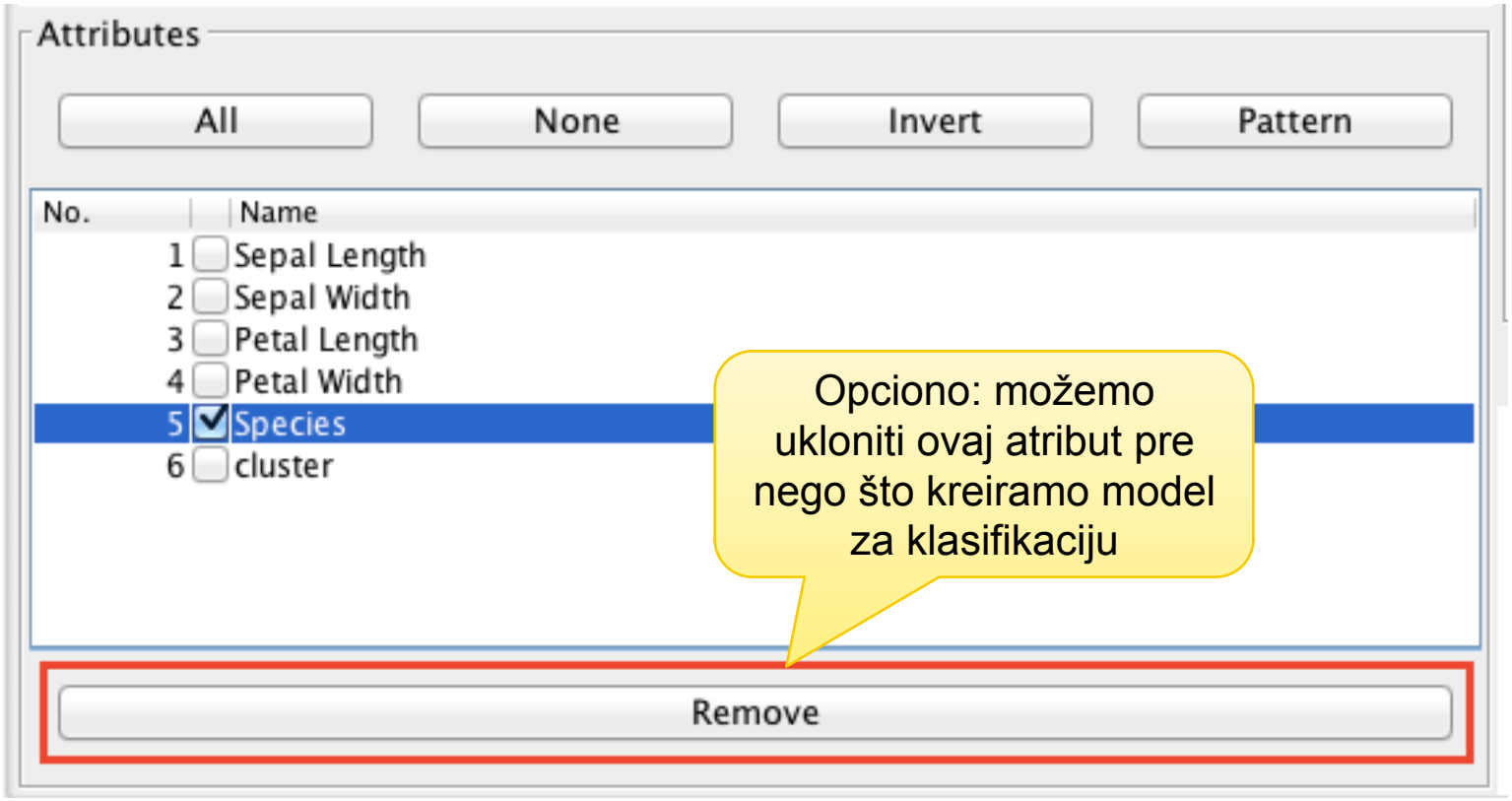
Korišćenje klastera za klasifikaciju



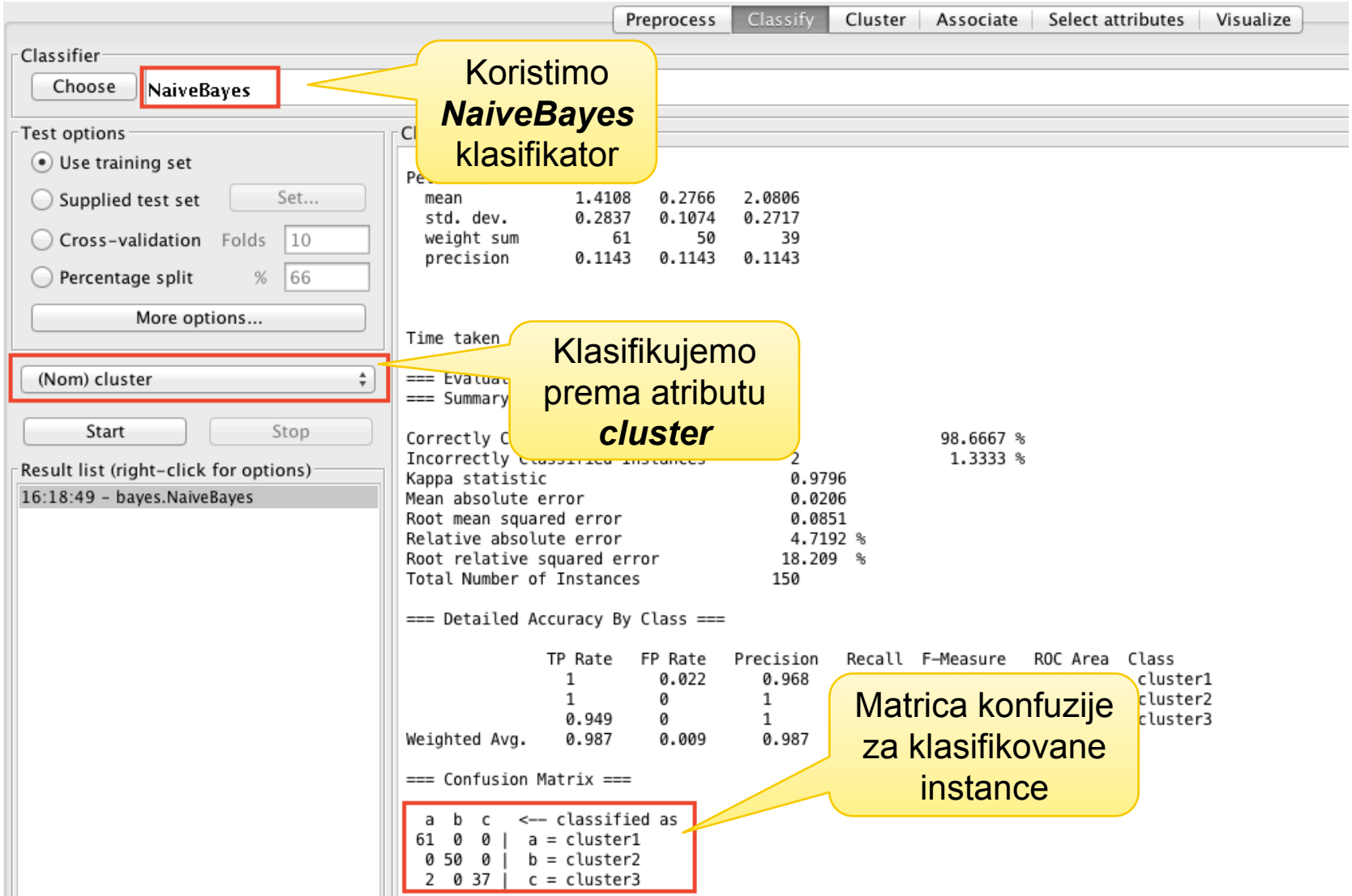
Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju



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:

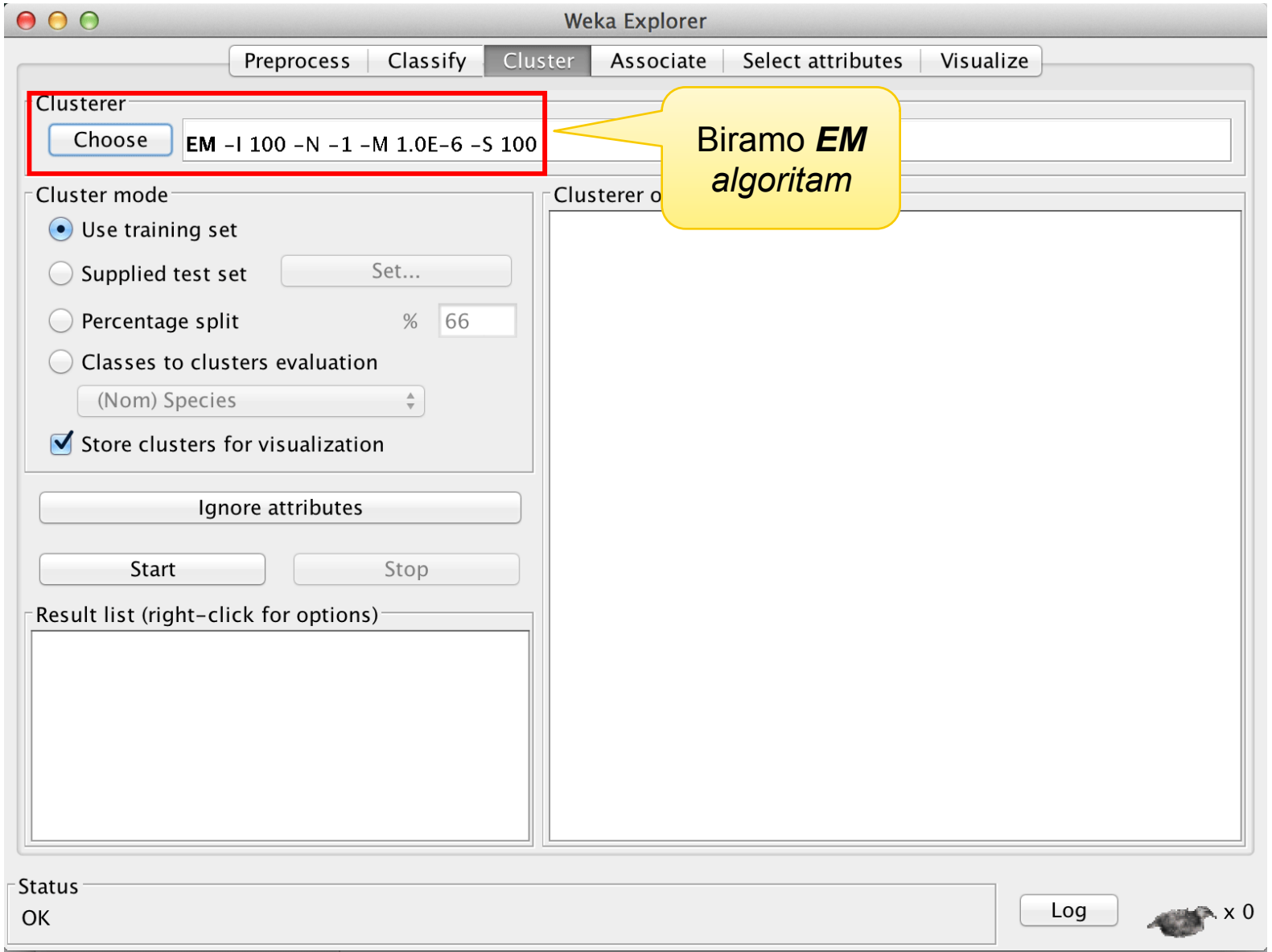
1) Inicijalno, definisati broj klastera (k) I nasumice izabrati vrednosti parametara modela ($\mu_i, \sigma_i, p_i, i=1,k$)

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

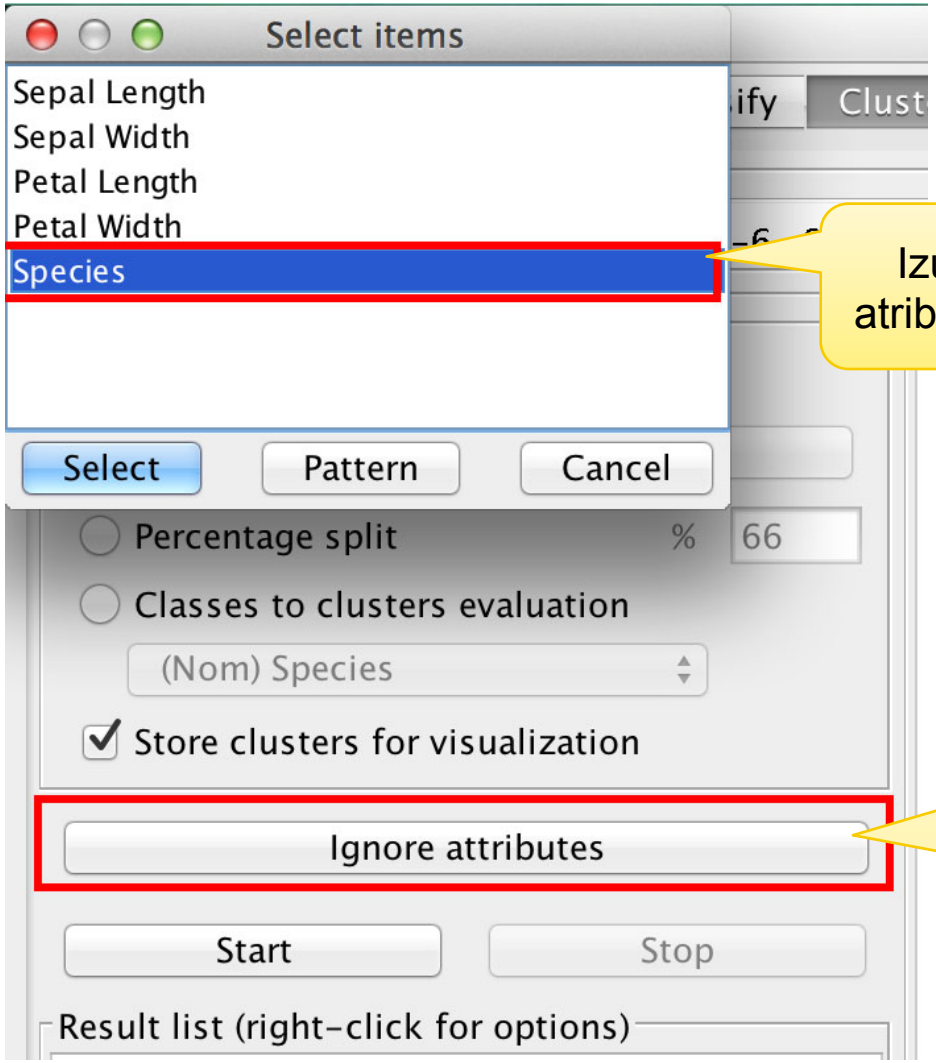
3) Na osnovu verovatnoća pripadnosti klasterima (instanci iz dataset-a), odrediti nove vrednosti parametara modela

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

Korišćenje EM algoritma



Ne uzimanje u obzir klase



Izuzimamo atribut **Species**

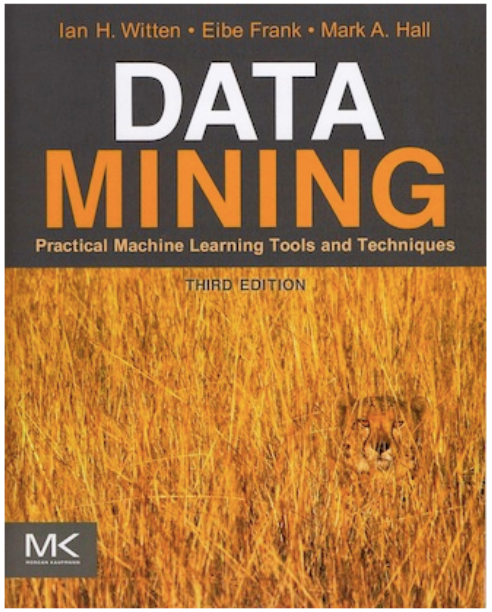
Selektovanje atributa koji neće biti korišćeni prilikom klasterovanja

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