

Diskretizacija i selekcija atributa. Klasterovanje

NIKOLA MILIKIĆ

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

URL: <http://nikola.milicic.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.

$$\text{širina} = (\text{maks. vrednost} - \text{min. vrednost}) / N$$

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

$$\text{Širina} = (100 - 0) / 5 = 20$$

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

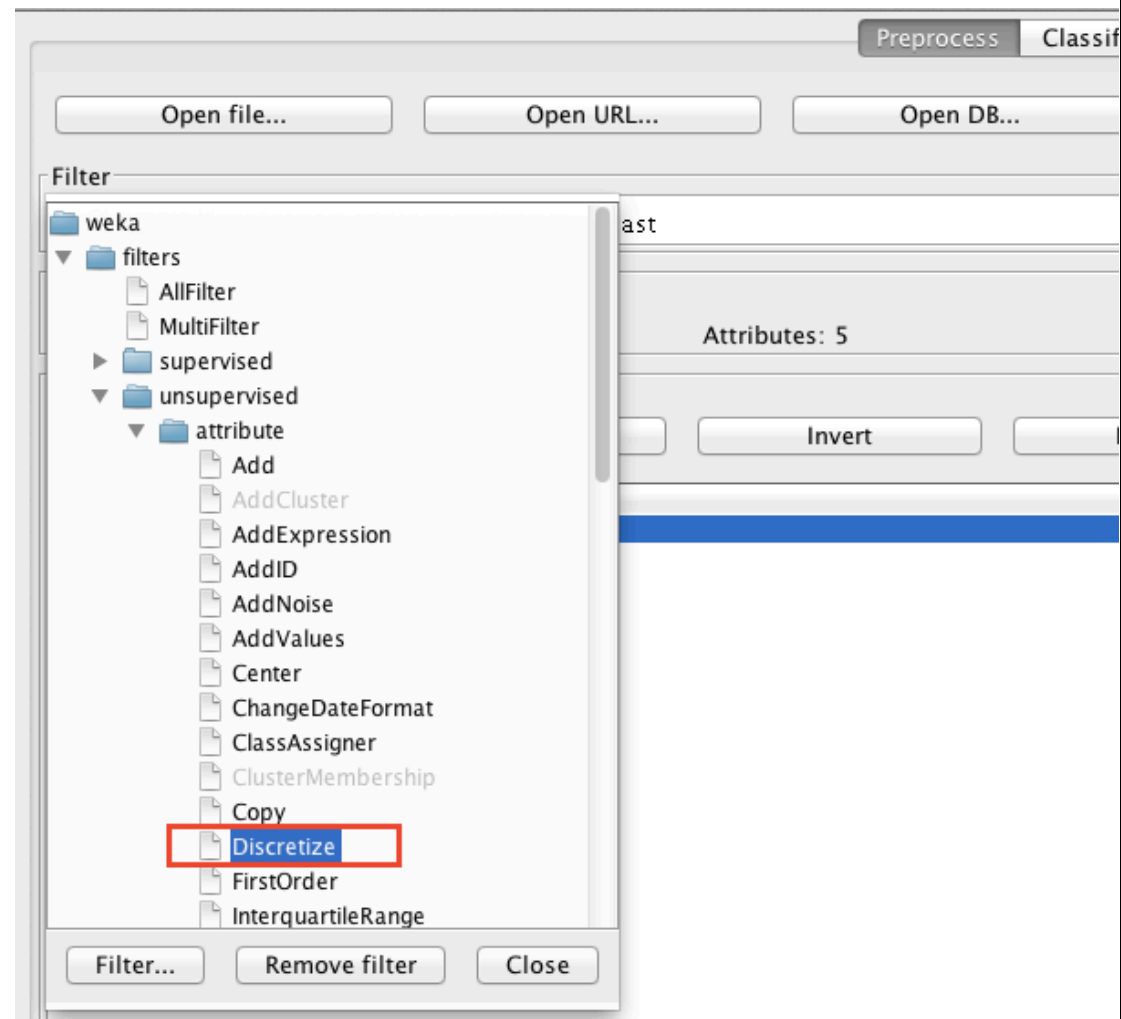
Diskretizacija u Weka-i

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*.

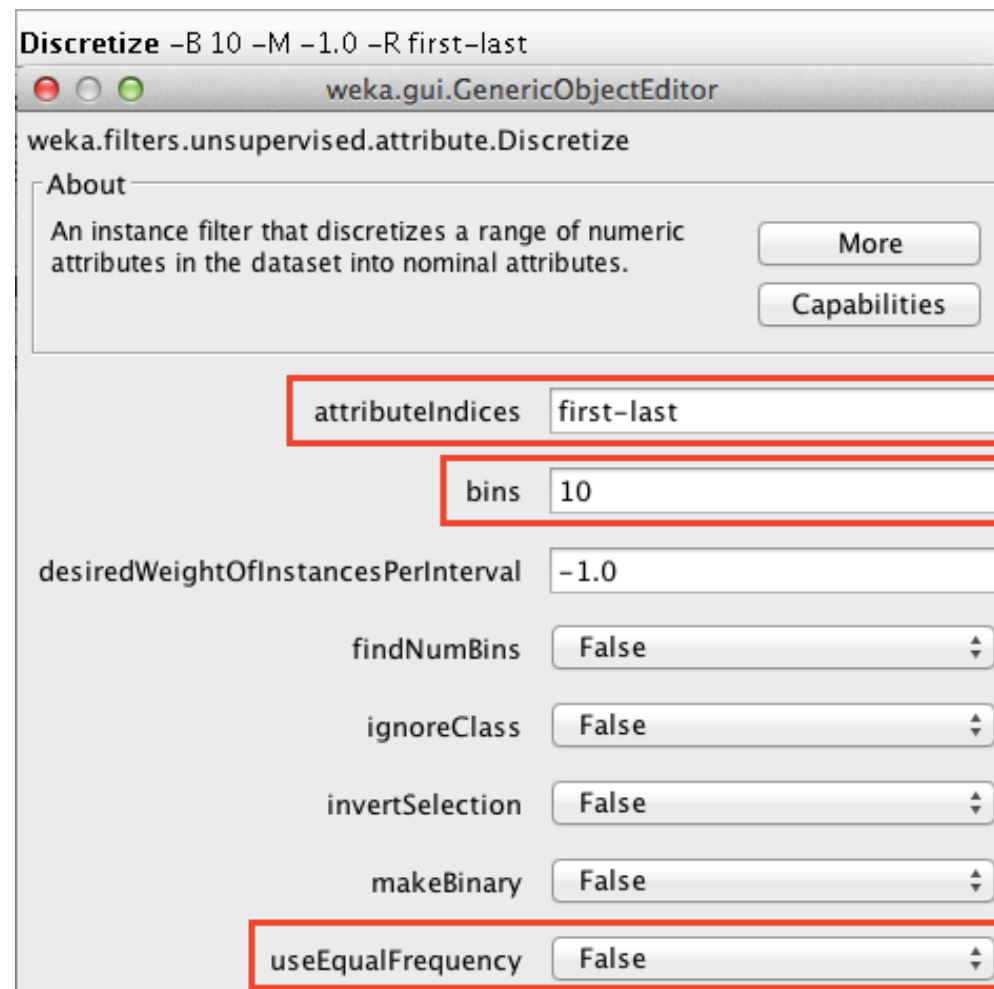


FishersIrisDataset.arff

Diskretizacija u Weka-i

Po defaultu se primenjuje Diskretizacija sa jednakim širinama opsega.

- ***attributeIndices*** - vrednost *first-last* označava da diskretizujemo 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

Pritiskom na **Apply** se primenjuje odabrani filter

Current relation
Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove...
Instances: 150 Attributes: 5

Attributes

- Sepal Length
- Sepal Width
- Petal Length
- Petal Width
- Species

Selected attribute
Name: Sepal Length Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count
1	'(-inf-4.66]'	9
2	'(4.66-5.02]'	23
3	'(5.02-5.38]'	14
4	'(5.38-5.74]'	27
5	'(5.74-6.1]'	22
6	'(6.1-6.46]'	20
7	'(6.46-6.82]'	18
8	'(6.82-7.18]'	6
9	'(7.18-7.54]'	5
10	'(7.54-inf]'	6

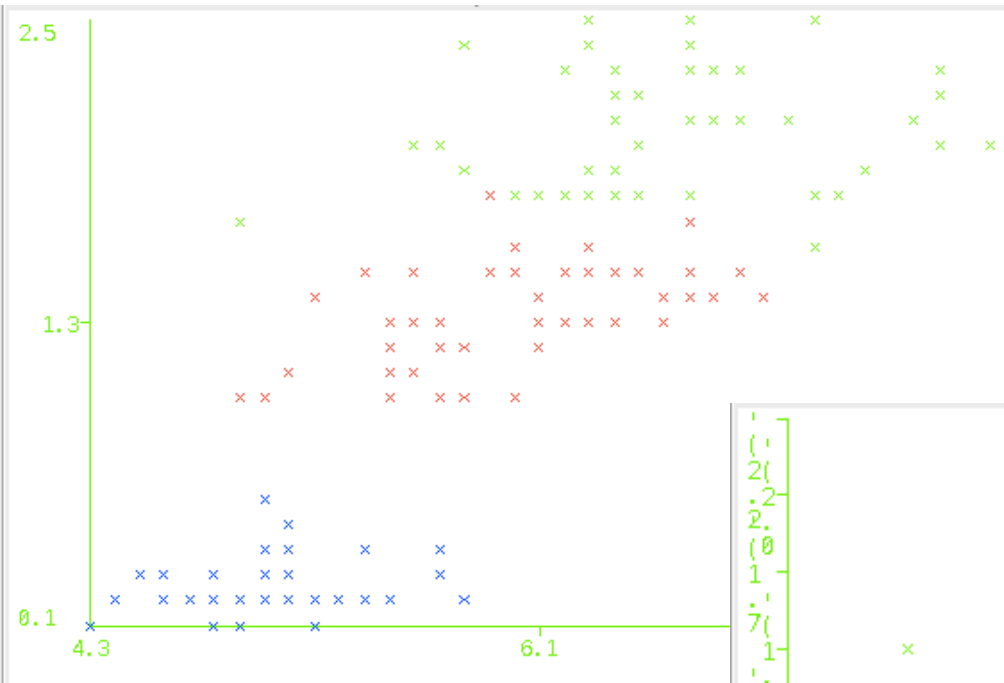
Class: Species (Nom) Visualize All

Remove

Status: OK Log x 0

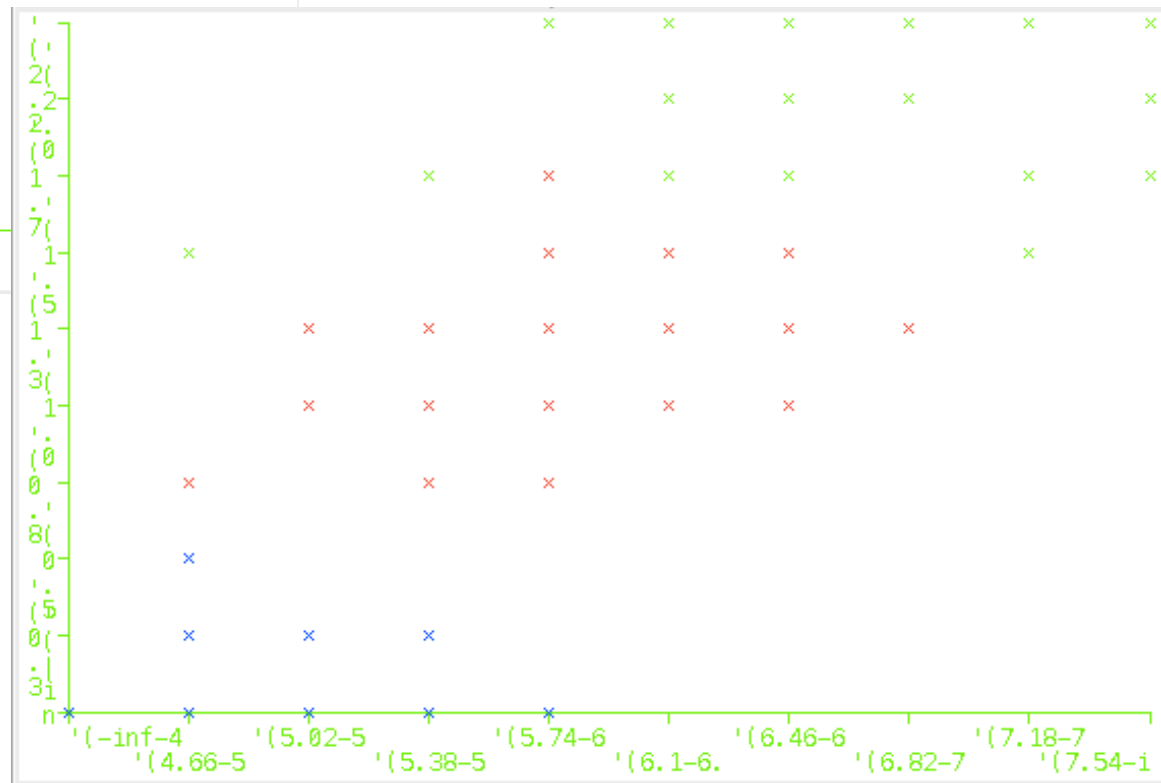
Dobijeni podopsezi vrednosti

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 attribute.
- 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

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

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

Filter: Choose None Apply

Current relation
Relation: 1990census
Instances: 32561
Attributes: 15

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> income

Remove

Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 73
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom) Visualize All

Status: OK Log x 0

Primer selekcije atributa

The screenshot shows the Weka software interface with the 'Preprocess' tab selected. The 'Filter' panel on the left contains a tree view of filters. Under 'supervised' > 'attribute', the 'AttributeSelection' filter is highlighted with a red box. A yellow callout bubble points to it with the text: 'Želimo da primenimo selekciju atributa' (We want to apply attribute selection).

The main window displays the command: `AttributeSelection.CfsSubsetEval -S "weka.attributeSelection.BestFirst -D 1 -N 5"`

The 'Selected attribute' section shows the following statistics for the 'age' attribute:

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Below this, the class is set to 'income (Nom)'. At the bottom right, a bar chart displays the distribution of the 'age' attribute across the 'income' classes, with blue bars for one class and red bars for another.

Primer selekcije atributa

The image shows a screenshot of the Weka GUI. In the foreground, a 'Filter' dialog box is open, displaying the 'AttributeSelection' filter. The dialog title is 'AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka'. The main content area shows the class name 'weka.filters.supervised.attribute.AttributeSelection' and an 'About' section with the text: 'A supervised attribute filter that can be used to select attributes.' There are 'More' and 'Capabilities' buttons. Below this, a file browser window is open, showing a tree view of the 'weka' package. The 'attributeSelection' sub-package is expanded, and 'ClassifierSubsetEval' is selected and highlighted with a red box. A yellow callout bubble points to this selection with the text: 'Kao vrstu evaluatora biramo **ClassifierSubsetEval**'. In the background, the 'Filter' dialog's 'Attributes' list is visible, showing a table with columns 'No.' and 'Name'. The table contains 12 rows of attribute names, each with a checkbox. The first row is selected.

No.	Name
1	age
2	work
3	fnbr
4	education
5	education-num
6	marital
7	occupation
8	relationship
9	race
10	sex
11	capital-gain
12	capital-loss

Primer selekcije atributa

The screenshot displays the Weka GUI interface for attribute selection. The main window shows a list of attributes with checkboxes, and a 'Filter' section at the top. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for 'weka.attributeSelection.ClassifierSubsetEval'. The 'About' section describes it as a 'Classifier subset evaluator' that evaluates attribute subsets. The 'classifier' field is highlighted with a red box and contains 'NaiveBayes'. A yellow callout box points to this field with the text 'Bramo NaiveBayes klasifikator'. Other fields include 'holdOutFile' (set to 'Click to set hold out or test instances') and 'useTraining' (set to 'True').

Filter

Choose AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes

Current relation: weka.gui.GenericObjectEditor

Relation: 1 weka.filters.supervised.attribute.AttributeSelection

Instances: 3

About

A supervised attribute filter that can be used to select attributes.

More

Capabilities

Attributes

All

No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

1 age

2 work-vol

3 fnlwgt

4 educ

5 educ-yr

6 m-married

7 occ

8 relationship

9 race

10 sex

11 capital-gain

12 capital-loss

13 hours-per-week

14 native-country

15 income

evaluator Choose ClassifierSubsetEval -B weka.classifiers.bayes

search Choose

Open... Save...

weka.gui.GenericObjectEditor

weka.attributeSelection.ClassifierSubsetEval

About

Classifier subset evaluator:

Evaluates attribute subsets on training data or a separate hold out testing set.

classifier Choose NaiveBayes

holdOutFile Click to set hold out or test instances

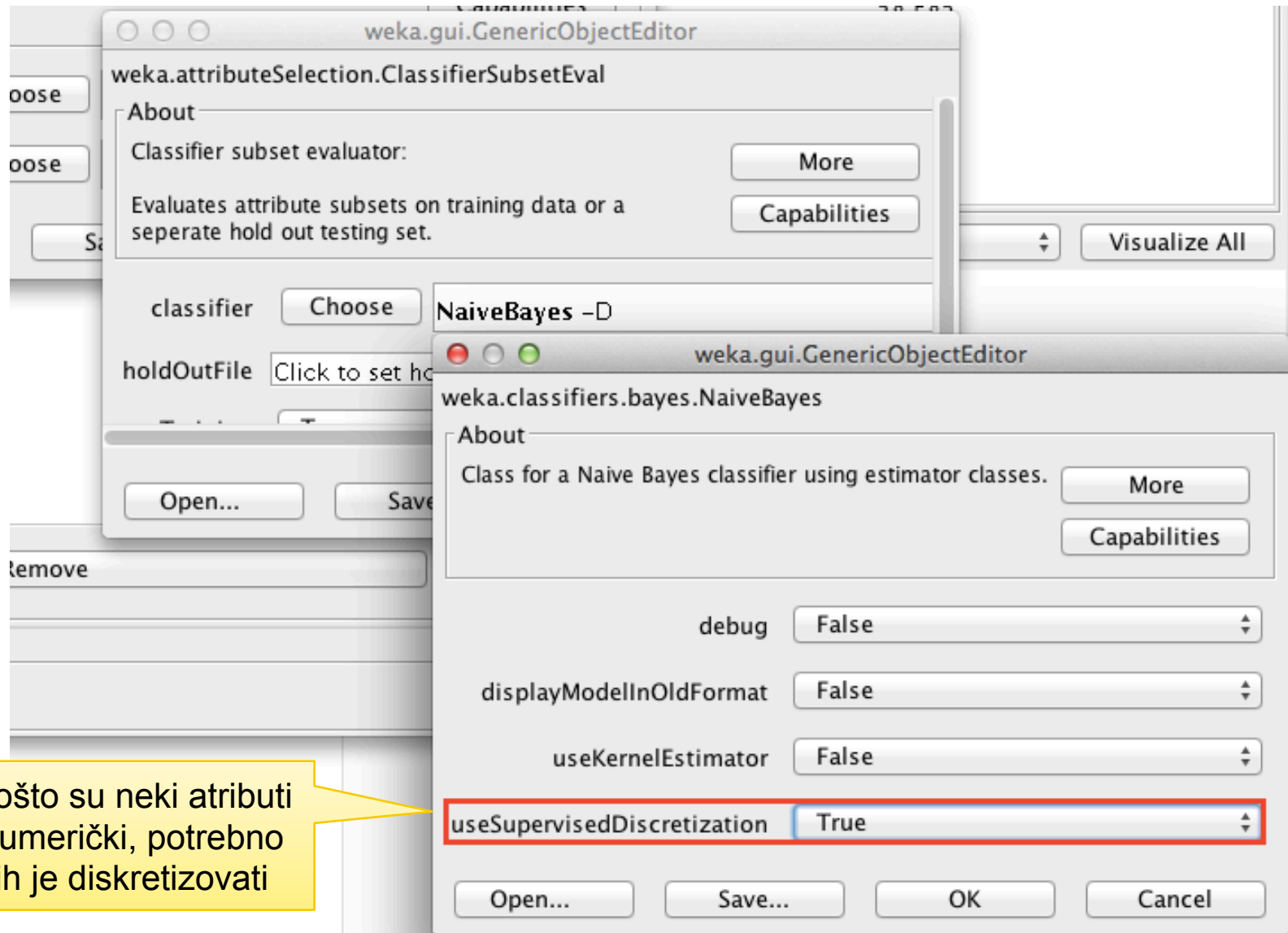
useTraining True

Open... Save... OK Cancel

Distinct: 73

Value
17
90
38.582
13.64

Primer selekcije atributa



Primer selekcije atributa

The screenshot shows the Weka Attribute Selection interface. The 'Filter' section is set to 'AttributeSelection' with the following command: `-E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \{Click`. The 'Apply' button is highlighted with a red box. The 'Current relation' section shows 'Relation: 1990census' and 'Instances: 32561'. The 'Attributes' section lists 15 attributes, with 'age' selected. The 'Selected attribute' section shows 'Name: age', 'Missing: 0 (0%)', and 'Distinct: 73'. The 'Class: income (Nom)' dropdown is set to 'income (Nom)'. A histogram of income is displayed at the bottom right, showing a distribution of income values from 17 to 90. A yellow callout box points to the 'Apply' button with the text: 'Filter je podešen i može biti primenjen nad atributima'.

Filter: `AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \{Click` **Apply**

Current relation
Relation: 1990census
Instances: 32561
Attributes: 15

Attributes

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> income

Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 73
Type: Numeric
Uniformity: 0.9999999999999999

Statistic
Minimum
Maximum
Mean
StdDev

Class: income (Nom) **Visualize All**

17 53.5 90

Status
OK **Log** x 0

Filter je podešen i može biti primenjen nad atributima

Primer selekcije atributa

The screenshot shows the Weka software interface with the 'Preprocess' tab selected. The 'AttributeSelection' filter is applied, and the 'age' attribute is selected. A yellow callout box points to the 'age' attribute in the list, stating 'Broj atributa je redukovan na 7' (The number of attributes is reduced to 7). The 'Selected attribute' panel shows statistics for 'age' (Minimum: 17, Maximum: 90, Mean: 38.582, StdDev: 13.64). The 'Class: income (Nom)' is selected, and a histogram is displayed at the bottom right.

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

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

Filter
Choose `AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \\"Click` Apply

Current relation
Relation: 1990census-weka.filters.supervised.attribute.Attribute...
Instances: 32561 | Attributes: 7

Attributes
All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> education
3	<input type="checkbox"/> relationship
4	<input type="checkbox"/> race
5	<input type="checkbox"/> capital-gain
6	<input type="checkbox"/> capital-loss
7	<input type="checkbox"/> income

Remove

Selected attribute
Name: age | Type: Numeric
Missing: 0 (0%) | Distinct: 73 | Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom) | Visualize All

17 | 53.5 | 90

Status
OK | Log | x 0

Klasterizacija

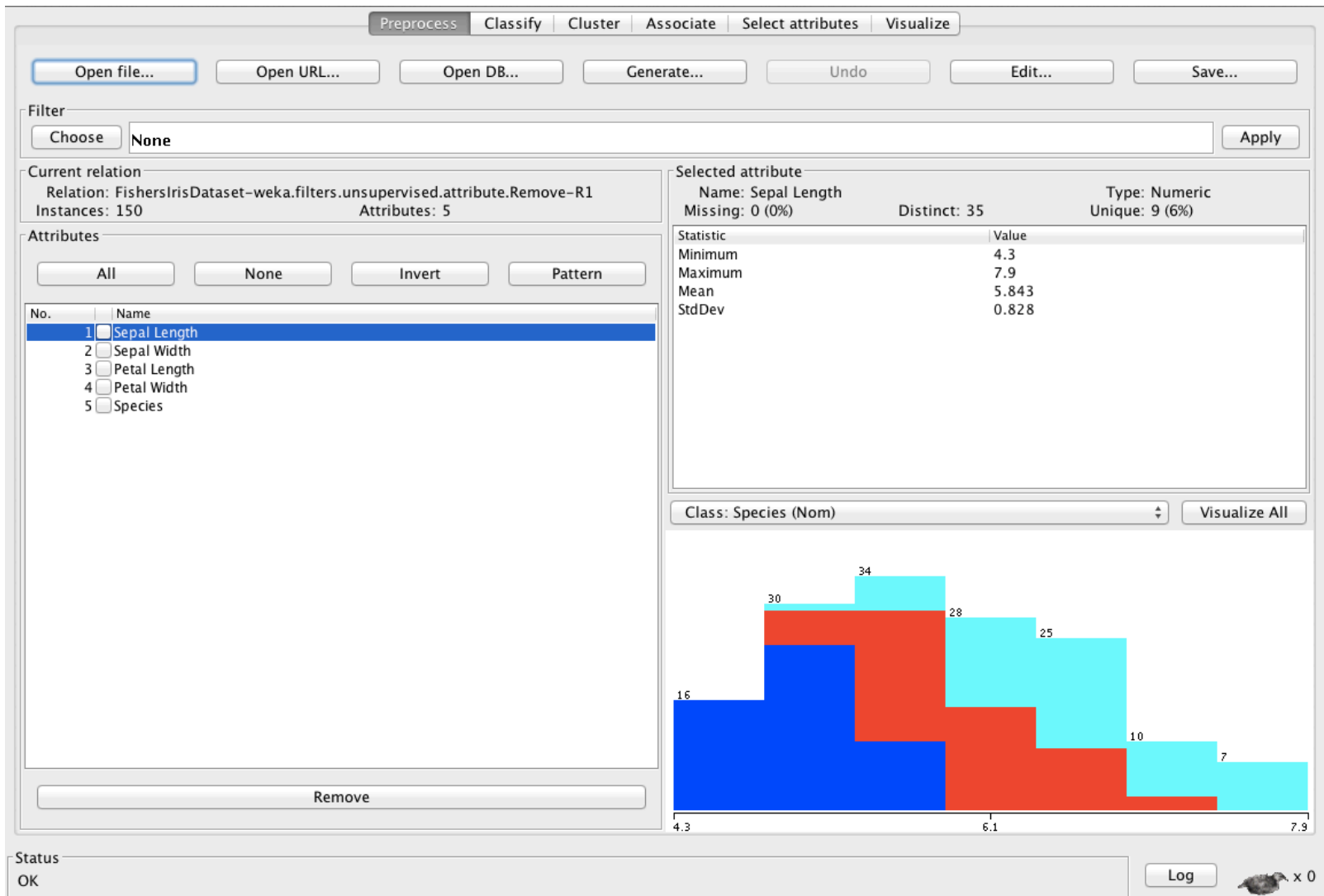
Klasterizacija (eng. Clustering) spada u grupu tehnika nenadgledanog učenja i omogućava grupisanje instanci u grupe, gde unapred ne znamo koje su sve grupe moguće.

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



The screenshot displays the Weka software interface with the 'Preprocess' tab selected. The 'Current relation' is 'FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1' with 150 instances and 5 attributes. The 'Attributes' list includes 'Sepal Length', 'Sepal Width', 'Petal Length', 'Petal Width', and 'Species'. The 'Selected attribute' is 'Sepal Length', which is numeric with 35 distinct values and 9 unique values (6%).

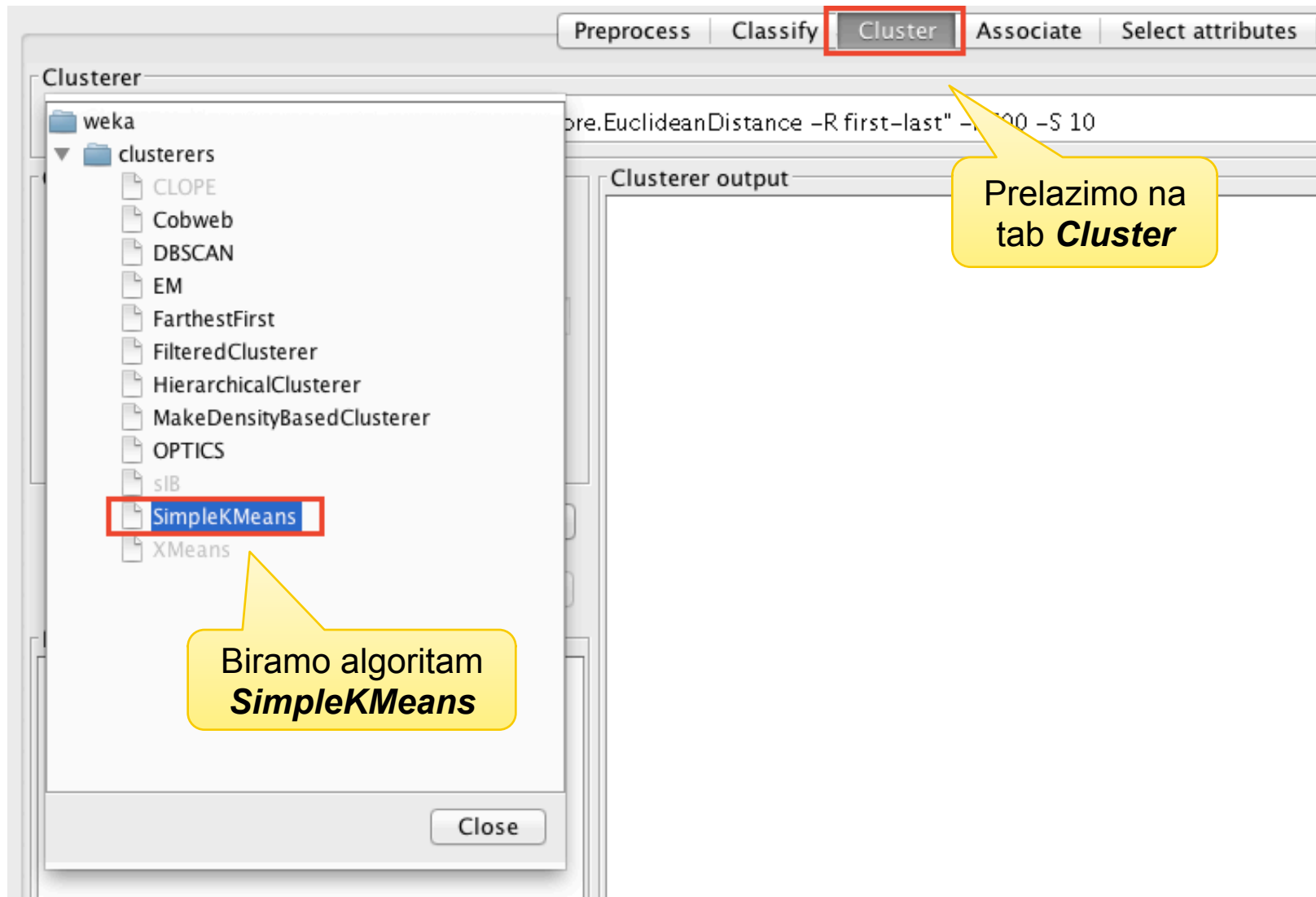
Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

The histogram shows the distribution of 'Sepal Length' values across three clusters (blue, red, cyan). The x-axis represents 'Sepal Length' values from 4.3 to 7.9, and the y-axis represents the number of instances in each cluster.

Cluster	Count
Blue	16
Red	30
Cyan	34

The status bar at the bottom shows 'Status OK' and a 'Log' button.

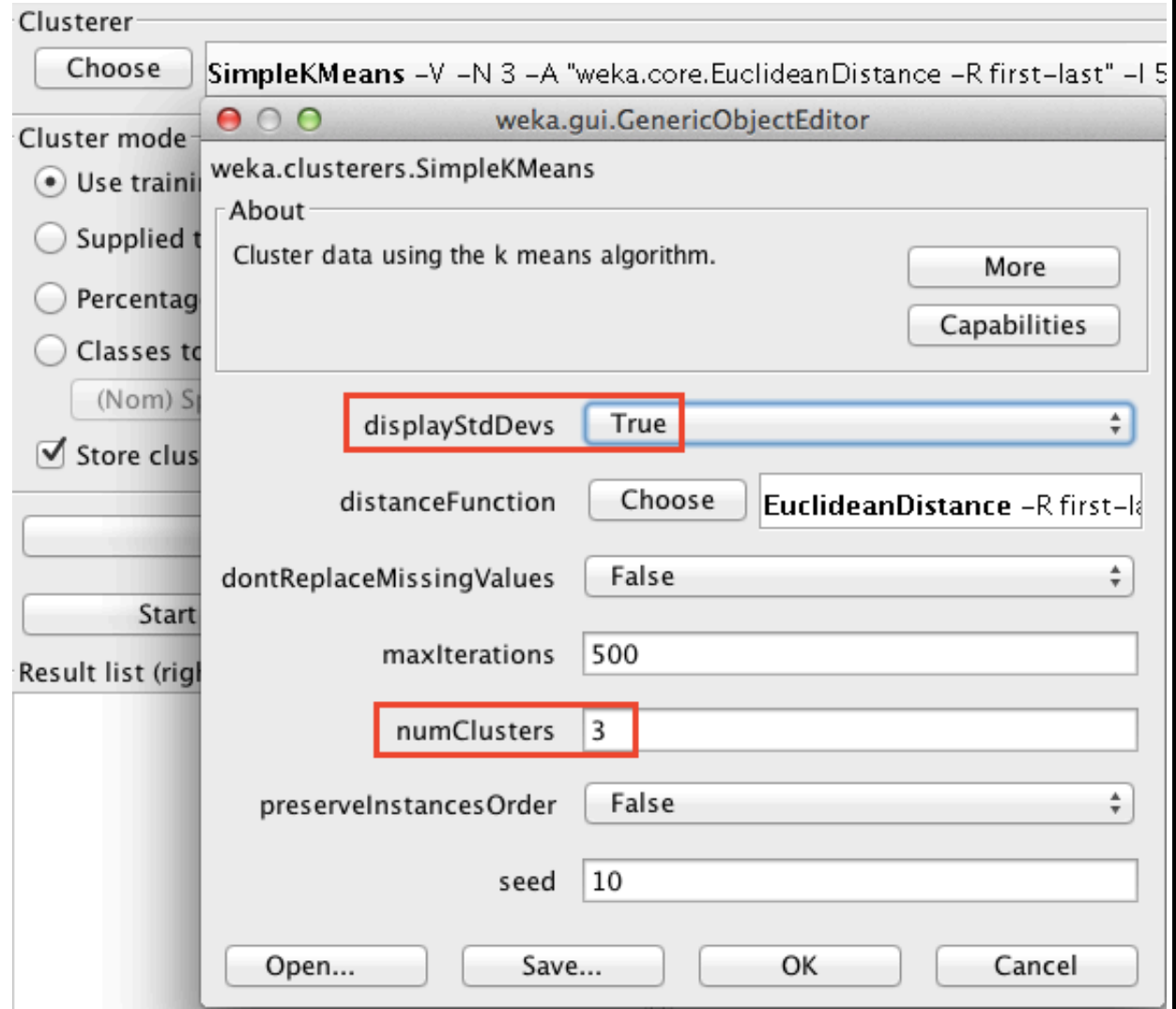
Odabir algoritma za klasterovanje



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

The screenshot displays the Weka GUI for the SimpleKMeans clustering algorithm. The 'Clusterer' is set to 'SimpleKMeans'. The 'Cluster mode' is 'Use training set', and 'Store clusters for visualization' is checked. The 'Ignore attributes' field contains 'Species'. The 'Start' button is visible. A yellow callout bubble points to the 'Use training set' radio button with the text 'Vršimo klasterovanje nad učitanim podacima'. Another yellow callout bubble points to the 'Ignore attributes' field with the text 'Ignorišemo Species atribut'. The 'Result list' shows a window titled 'Select items' with 'Species' selected. The 'Run information' panel shows the command line and the following output:

```
=== Run information ===
Scheme:weka.clusterers.SimpleKMeans -V -N 3 -A "weka.core.Euc
Relation: FishersIrisDataset-weka.filters.unsupervised.at
Instances: 150
Attributes: 5
Sepal Length
Sepal Width
Petal Length
Petal Width
Ignored: Species
Test mode:evaluate on training data
=== Model and evaluation on training set ===
of iterations: 6
Within cluster sum of squared errors: 6.982216473785234
Missing values globally replaced with mean/mode
Cluster centroids:
Attribute      Full Data      Cluster#
                (150)          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
```

Rezultat klasterovanja

Cluster mode

Use training set
 Supplied test set Set...
 Percentage split % 66
 Centroidi svakog klastera i njihove standardne devijacije

Ignore attributes

Start Stop

Result list (right-click for options)

15:07:08 - SimpleKMeans

Clusterer output

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

Cluster centroids:

Attribute	Full Data (150)	Cluster#		
		0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	3.428 +/-0.3791	3.0821 +/-0.2799
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	1.462 +/-0.1737	5.7026 +/-0.5194
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-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%)

Broj instanci u svakom klasteru

Evaluacija rezultata

The screenshot shows the Orange3 SimpleKMeans widget interface. On the left, the 'Cluster mode' section has three radio buttons: 'Use training set', 'Supplied test set', and 'Percentage split'. The 'Classes to clusters evaluation' radio button is selected and highlighted with a red box. Below it, a dropdown menu shows '(Nom) Species'. A yellow callout bubble points to this dropdown with the text 'Selektujemo atribut sa kojim želimo da poredimo rezultate'. Below the dropdown is a checked checkbox 'Store clusters for visualization'. There are 'Ignore attributes', 'Start', and 'Stop' buttons. The 'Result list' shows two entries: '15:07:08 - SimpleKMeans' and '15:20:38 - SimpleKMeans', with the latter selected. A yellow callout bubble points to this list with the text 'Imena klasa koje su dodeljene klasterima'. On the right, the 'Clusterer output' pane shows a table of statistics for 'Petal width' and 'Time taken to build model'. A yellow callout bubble points to the 'Clustering' section of the output with the text 'U kojim klasterima su smeštene koje klase'. Below this, a red box highlights the 'Class attribute: Species' and 'Classes to Clusters:' section, which shows a mapping of clusters to species names. At the bottom, the 'Incorrectly clustered instances' section shows 17.0 instances, which is 11.3333% of the total.

Cluster mode

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation
(Nom) Species
- Store clusters for visualization

Result list (right-click for options)

- 15:07:08 - SimpleKMeans
- 15:20:38 - SimpleKMeans

Clusterer output

	77	3.428	3.0821
	34	+/-0.3791	+/-0.2799
	67	1.462	5.7026
	69	+/-0.1737	+/-0.5194
Petal width	1.1995	1.418	0.246
	+/-0.7622	+/-0.2723	+/-0.1054
			2.0795

Time taken to build model (full training data) : 0.02 seconds

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

Clustering

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

Class attribute: Species
Classes to Clusters:

0	1	2	<-- assigned to cluster
0	50	0	setosa
47	0	3	versicolor
14	0	36	virginica

Cluster 0 <-- versicolor
Cluster 1 <-- setosa
Cluster 2 <-- virginica

Incorrectly clustered instances : 17.0 11.3333 %

Selektujemo atribut sa kojim želimo da poredimo rezultate

U kojim klasterima su smeštene koje klase

Imena klasa koje su dodeljene klasterima

Vizuelizacija klastera

The image shows the Weka Clusterer Visualize interface. On the left, a panel contains a checkbox for "Store clusters for visualization" (checked), an "Ignore attributes" button, and a "Start" button. Below these is a "Result list (right-click for...)" with two entries: "15:07:08 - SimpleKMeans" and "15:20:38 - SimpleKMeans". A yellow callout bubble labeled "Desni klik" points to the second entry. A context menu is open over the second entry, listing options: "View in main window", "View in separate window", "Save result buffer", "Delete result buffer", "Load model", "Save model", "Re-evaluate model on current test set", "Visualize cluster assignments" (highlighted with a red box), and "Visualize tree". A yellow callout bubble labeled "Vizuelna reprezentacija klastera" points to the "Visualize cluster assignments" option.

The main window, titled "Weka Clusterer Visualize: 15:20:38 - SimpleKMeans (FishersIris)", displays a scatter plot. The X-axis is labeled "X: Petal Length (Num)" with values 1, 3.95, and 6.9. The Y-axis is labeled "Y: Petal Width (Num)" with values 0.1 and 1.3. The plot shows three clusters of data points: cluster0 (red 'x' marks), cluster1 (blue squares), and cluster2 (green 'x' marks). A legend at the bottom right shows "cluster0 cluster1 cluster2" with corresponding colors. The plot title is "Plot:FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1".

Procena uspešnosti klasterovanja

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

Računa se kao suma kvadrata razlika između vrednosti atributa svake instance i vrednosti centroida u datom atributu

Cluster mode
 Use training set

Ignore attributes

Start Stop

Result list (right-click for options)

- 15:07:08 - SimpleKMeans
- 15:20:38 - SimpleKMeans

Clusterer output

kMeans
=====

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

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	+/-0.3	
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	+/-0.1737	+/-0.5194
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Vrednosti centroida po svim atributima

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

The screenshot shows the Weka GUI with the 'Filter' list on the left and the 'Selected attribute' dropdown on the right. The 'AddCluster' filter is highlighted in the list, and the 'No class' option is selected in the dropdown. Two yellow callout boxes provide instructions: 'Odabiramo kao vrstu Filter-a **AddCluster**' and 'Postavimo da klasa nije selektovana'.

Preprocess | Classify | Cluster | Associate | Select attributes

Open file... Open URL... Open DB... Generate... Undo

Filter

- weka
 - filters
 - AllFilter
 - MultiFilter
 - supervised
 - unsupervised
 - attribute
 - Add
 - AddCluster**
 - AddExpression
 - AddID
 - AddNoise
 - AddValues
 - Center
 - ChangeDateFormat
 - ClassAssigner
 - ClusterMembership
 - Copy
 - Discretize
 - FirstOrder
 - InterquartileRange

SimpleKMeans -N 2 -A {"weka.core.EuclideanDistance -R first-last}"

Selected attribute

Name: Sepal Length
Missing: 0 (0%)

Statistic

- Minimum
- Maximum
- Mean
- StdDev

Invert Pattern

No class

Filter... Remove filter Close

30 34

Korišćenje klastera za klasifikaciju

The image shows a screenshot of the Weka GUI with two windows. The main window is titled 'weka.gui.GenericObjectEditor' and shows the 'AddCluster' dialog. The 'clusterer' dropdown is set to 'SimpleKMeans'. The 'ignoredAttributeIndices' field is set to '5'. A yellow callout bubble points to the 'SimpleKMeans' dropdown with the text: 'Biramo *SimpleKMeans* kao algoritam za klasterovanje'. Another yellow callout bubble points to the 'ignoredAttributeIndices' field with the text: 'Ignorišemo atribut broj 5 (Species) prilikom klasterovanja'. The second window is titled 'weka.gui.GenericObjectEditor' and shows the 'SimpleKMeans' configuration dialog. The 'numClusters' field is set to '3'. A red box highlights the 'numClusters' field.

Filter

Choose **AddCluster** -V "weka.clusterers.SimpleKMeans -V -N 3 -A {"weka.core.EuclideanDistance -R first-last}" -

Current relation: weka.gui.GenericObjectEditor

Relation: weka.gui.GenericObjectEditor

Instance: weka.gui.GenericObjectEditor

Attribute: weka.gui.GenericObjectEditor

clusterer: **SimpleKMeans**

ignoredAttributeIndices: 5

Open... Save... OK

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs: True

distanceFunction: Choose Eu

dontReplaceMissingValues: False

maxIterations: 500

numClusters: 3

preserveInstancesOrder: False

seed: 10

Open... Save... C

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	<input type="checkbox"/> Sepal Length
2	<input type="checkbox"/> Sepal Width
3	<input type="checkbox"/> Petal Length
4	<input type="checkbox"/> Petal Width
5	<input type="checkbox"/> Species
6	<input checked="" type="checkbox"/> cluster

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

Selected attribute
Name: cluster
Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

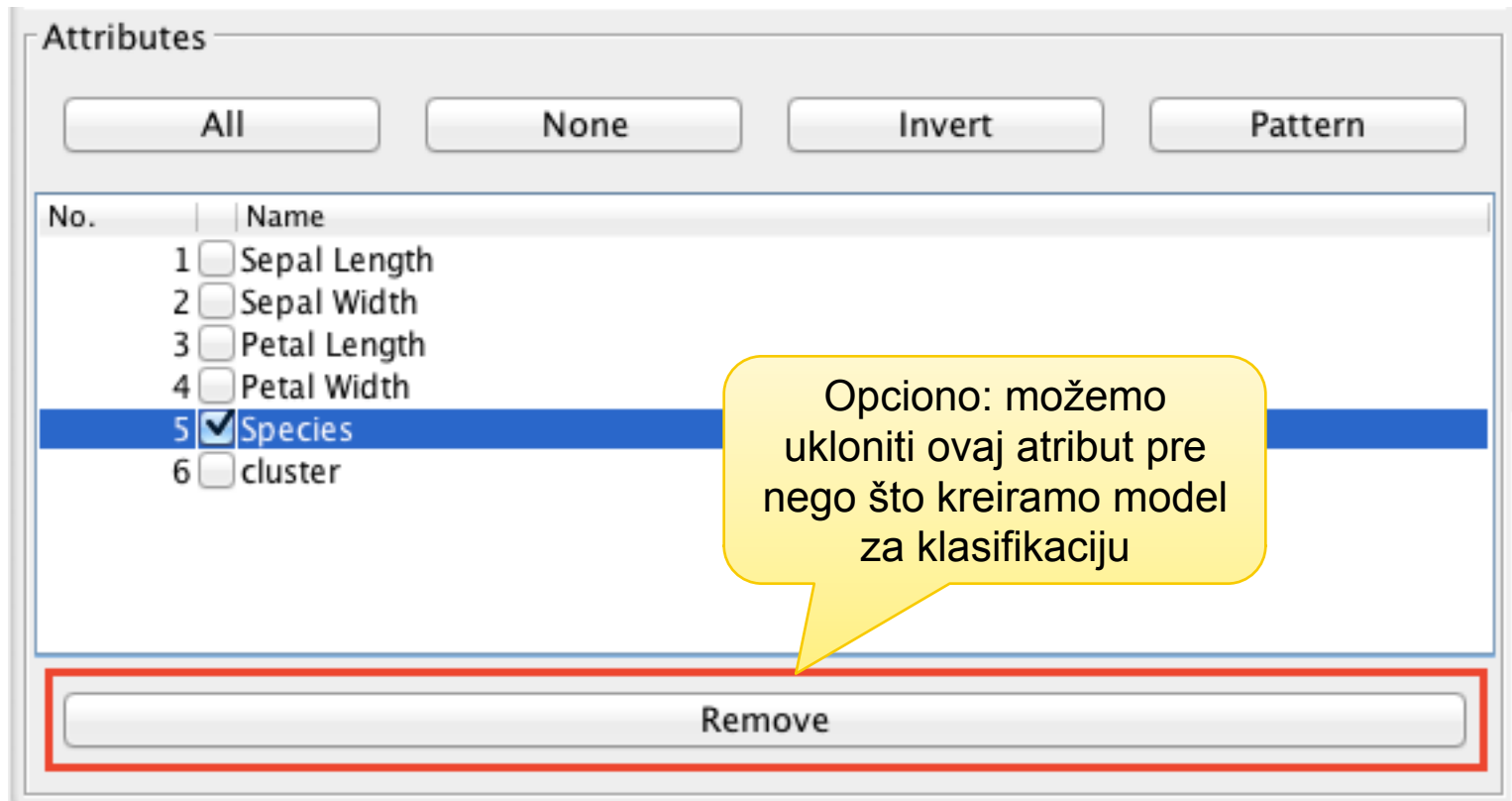
No.	Label	Count
1	cluster1	61
2	cluster2	50
3	cluster3	39

Class: cluster (Nom) Visualize All

Cluster	Count
cluster1	61
cluster2	50
cluster3	39

Remove

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

The screenshot shows the Weka Explorer application window with the 'Cluster' tab selected. The 'Clusterer' dropdown menu is highlighted with a red box, and a yellow callout bubble points to it with the text 'Biram EM algoritam'. The 'Clusterer' field contains the command 'EM -I 100 -N -1 -M 1.0E-6 -S 100'. Below this, the 'Cluster mode' section has 'Use training set' selected. Other options include 'Supplied test set', 'Percentage split' (66%), 'Classes to clusters evaluation' (set to '(Nom) Species'), and 'Store clusters for visualization' (checked). There are buttons for 'Ignore attributes', 'Start', and 'Stop'. The 'Result list' is empty. The status bar at the bottom shows 'Status OK' and a 'Log' button.

Weka Explorer

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

Clusterer


Choose **EM -I 100 -N -1 -M 1.0E-6 -S 100**

Cluster mode

- Use training set
- Supplied test set
- Percentage split %
- Classes to clusters evaluation
- Store clusters for visualization

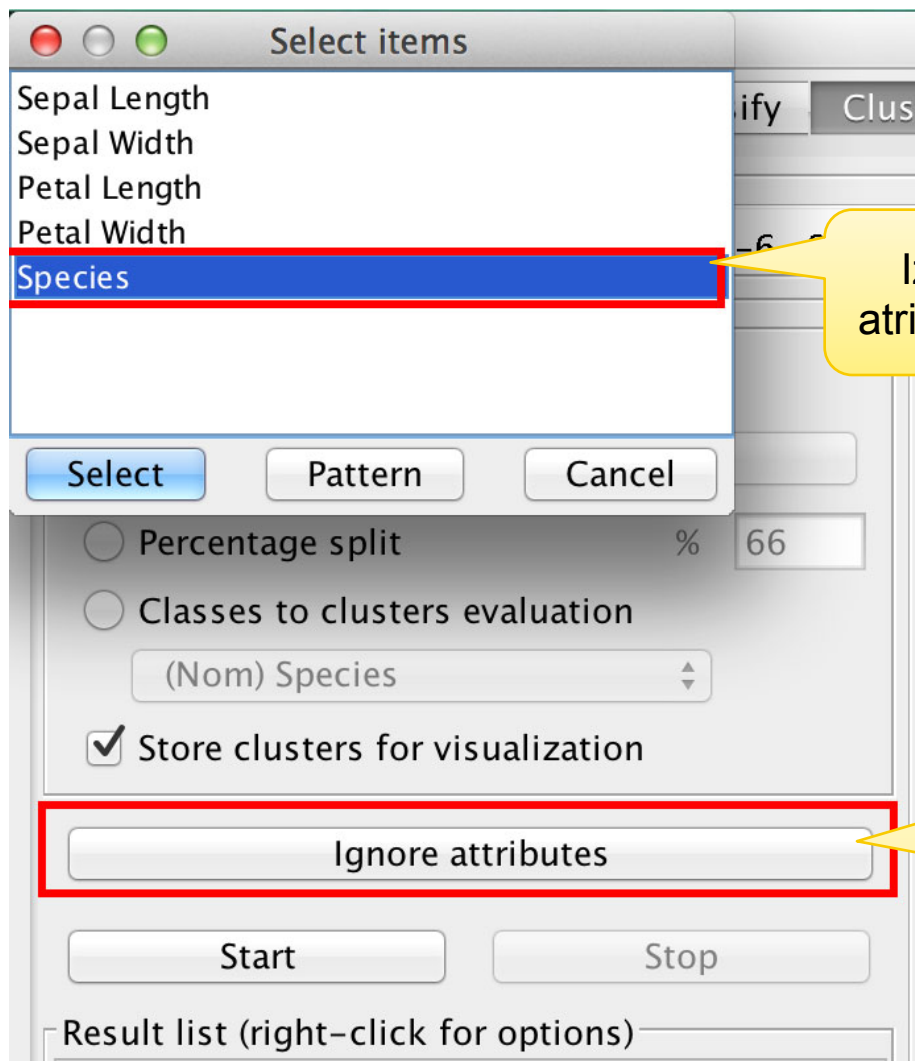
Result list (right-click for options)

Status
OK

 x 0

Biram *EM* algoritam

Ne uzimanje u obzir klase



Izuzimamo atribut **Species**

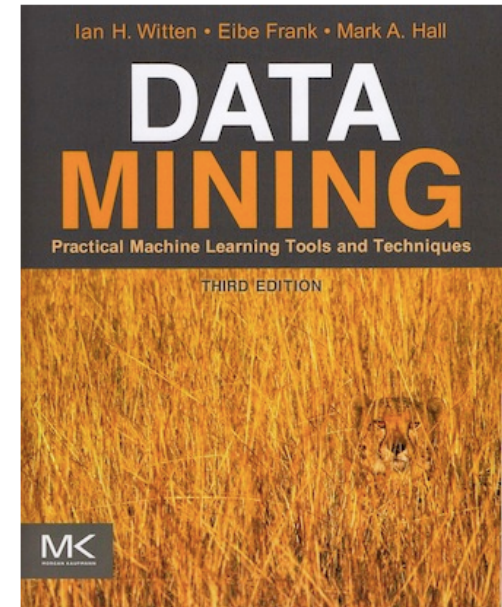
Selektovanje atributa koji neće biti korišćeni prilikom klasterovanja

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.milicic@fon.bg.ac.rs

URL: <http://nikola.milicic.info>