

Priprema podataka

NIKOLA MILIKIĆ

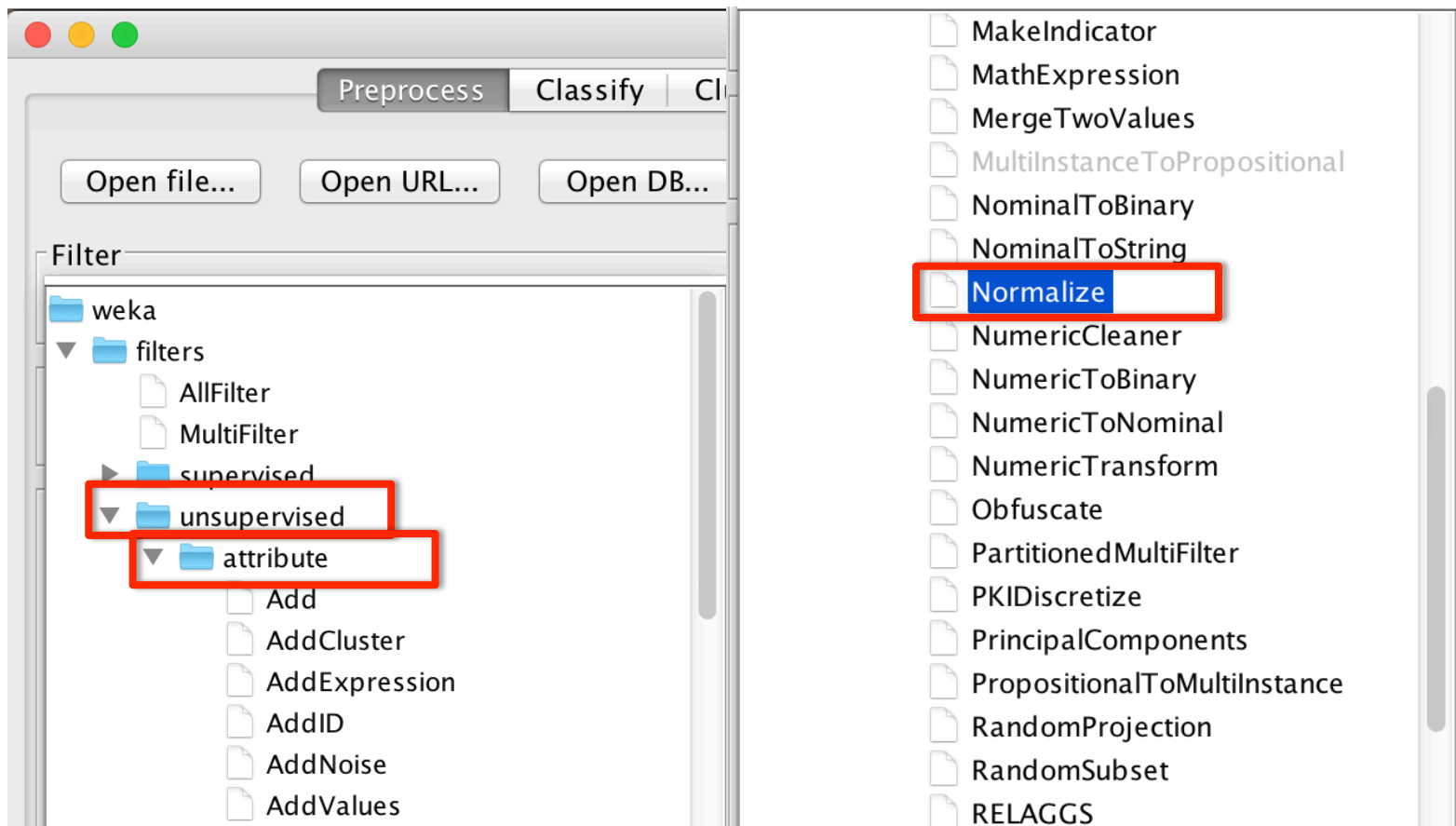
EMAIL: nikola.milikic@fon.bg.ac.rs

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

Normalizacija

Normalizacija je svodenje vrednosti na neki opseg (obično 0 - 1)

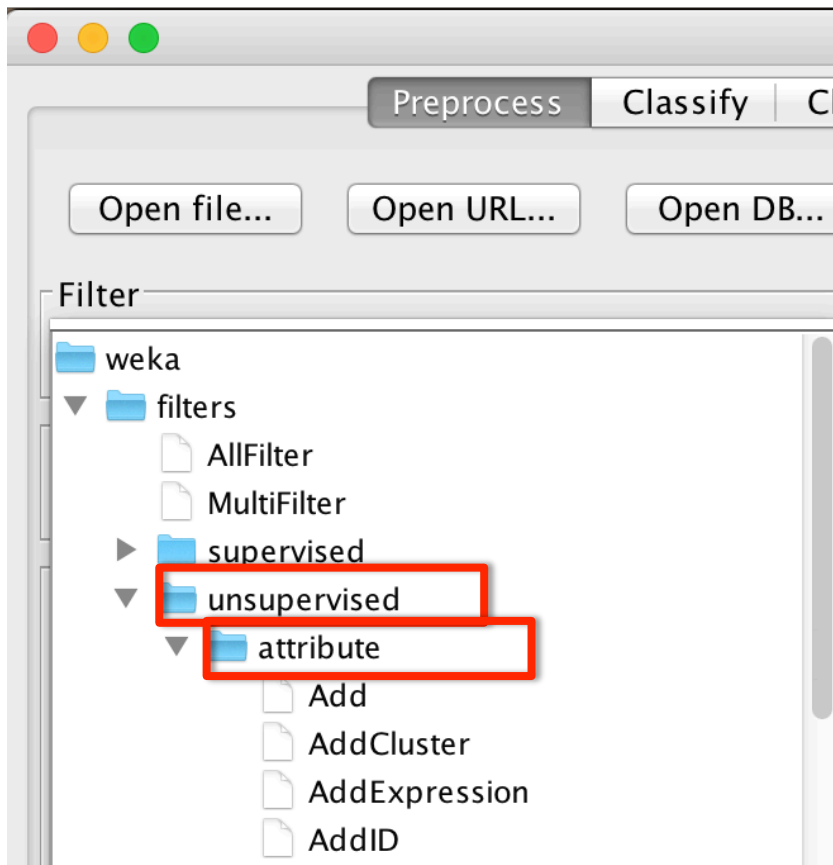
FishersIrisDataset.arff



Standardizacija

Standardizacija je svođenje srednje vrednosti na 0, a standardne devijacije na vrednost 1

FishersIrisDataset.arff



- ☐ Obfuscate
- ☐ PartitionedMultiFilter
- ☐ PKIDiscretize
- ☐ PrincipalComponents
- ☐ PropositionalToMultiInstance
- ☐ RandomProjection
- ☐ RandomSubset
- ☐ RELAGGS
- ☐ Remove
- ☐ RemoveType
- ☐ RemoveUseless
- ☐ Reorder
- ☐ ReplaceMissingValues
- ☒ Standardize
- ☐ StringToNominal

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 pristup – uzima 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

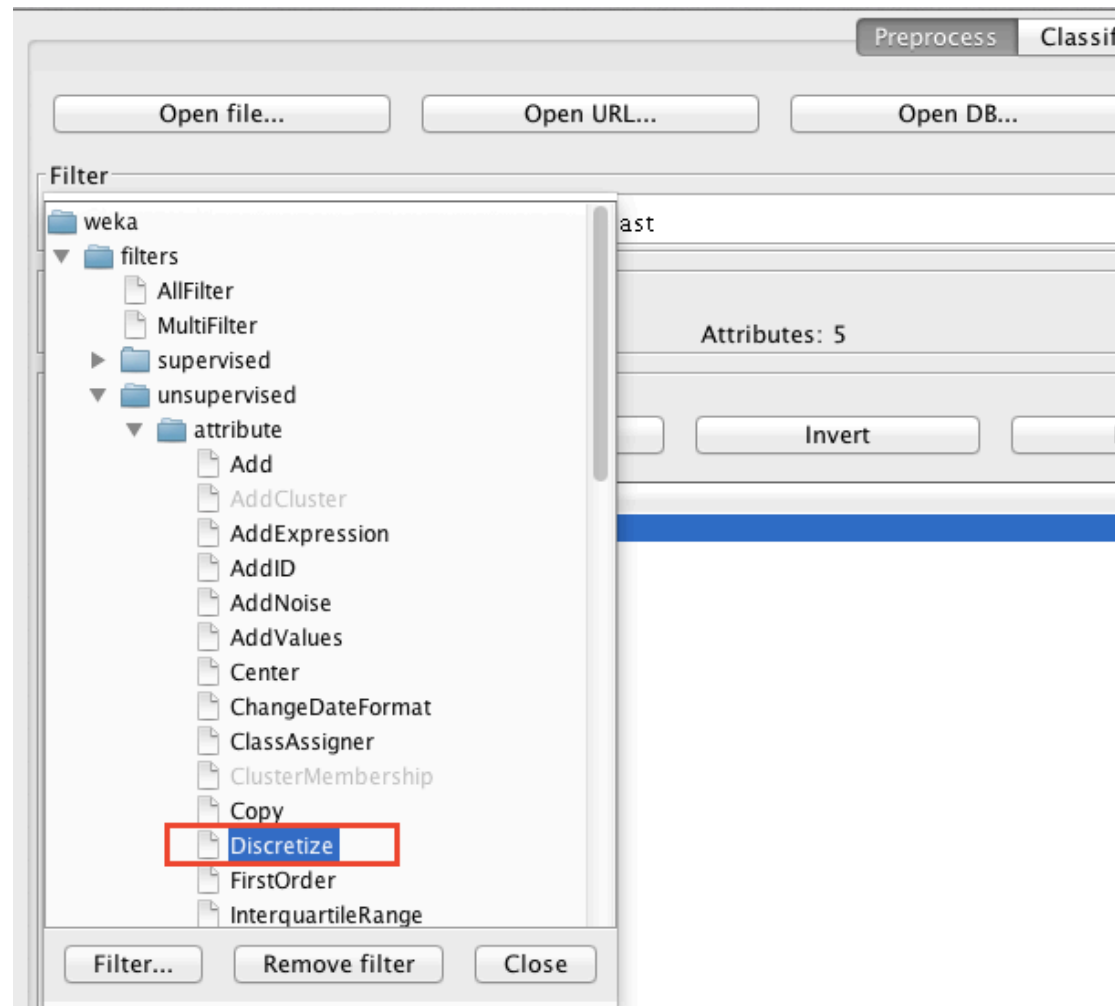
Diskretizacija u Weka-i

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***.

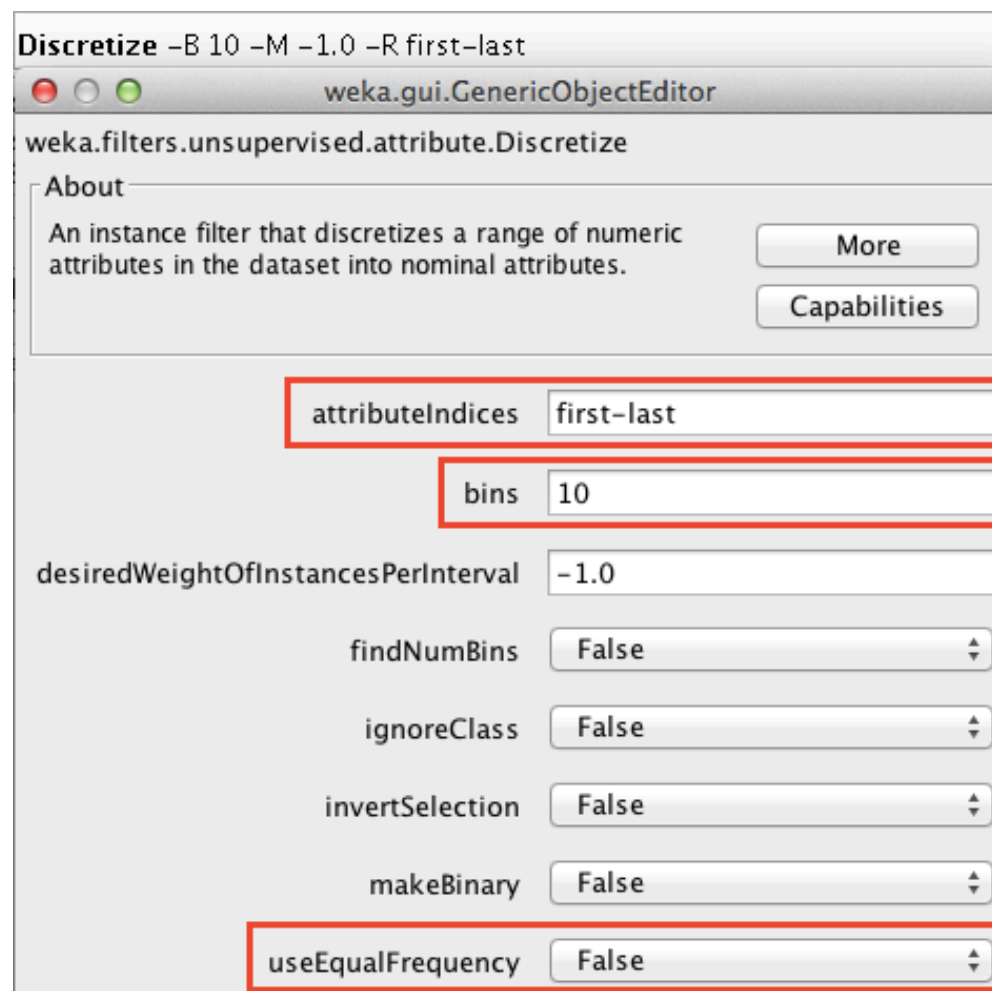


FishersIrisDataset.arff

Diskretizacija u Weka-i

Po defaultu se primenjuje Diskretizacija sa jednakim širinama opsega.

- ***attributeIndices*** - vrednost *first-last* označava da diskretizujemo sve attribute. 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

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter: Choose **Discretize -B 10 -M -1.0 -R first-last** **Apply**

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

Attributes
All None Invert Pattern

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

Remove

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

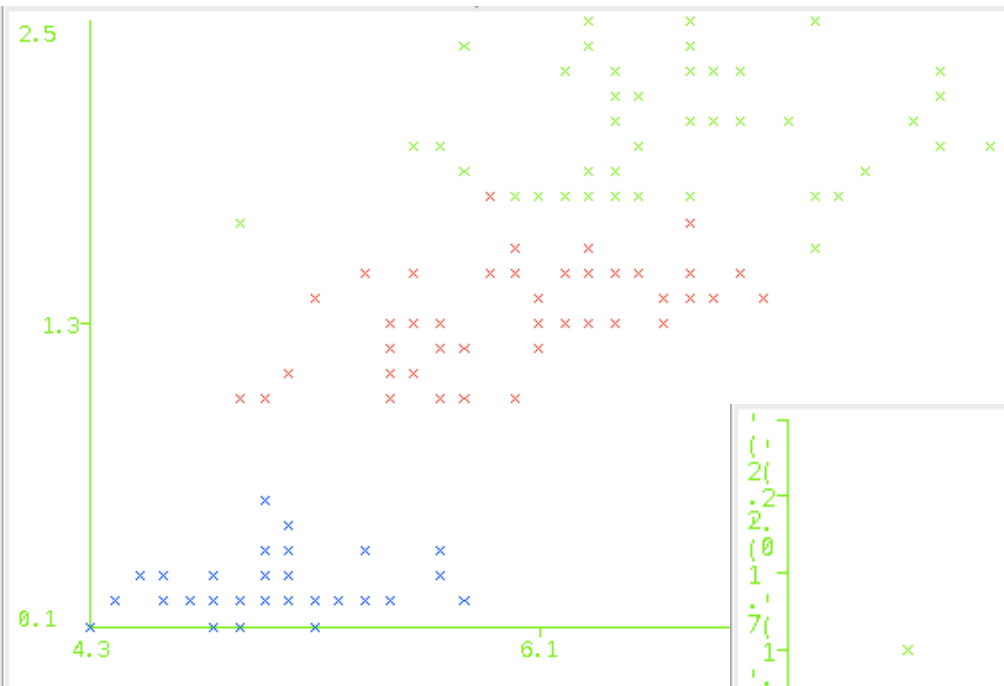
Dobijeni podopsezi vrednosti

Class: Species (Nom) Visualize All

Sepal Length Bin	Setosa (Blue)	Versicolour (Red)	Virginica (Cyan)	Total
1	9	0	0	9
2	23	0	0	23
3	14	0	0	14
4	27	0	0	27
5	22	0	0	22
6	20	0	0	20
7	18	0	0	18
8	6	0	0	6
9	5	0	0	5
10	6	0	0	6

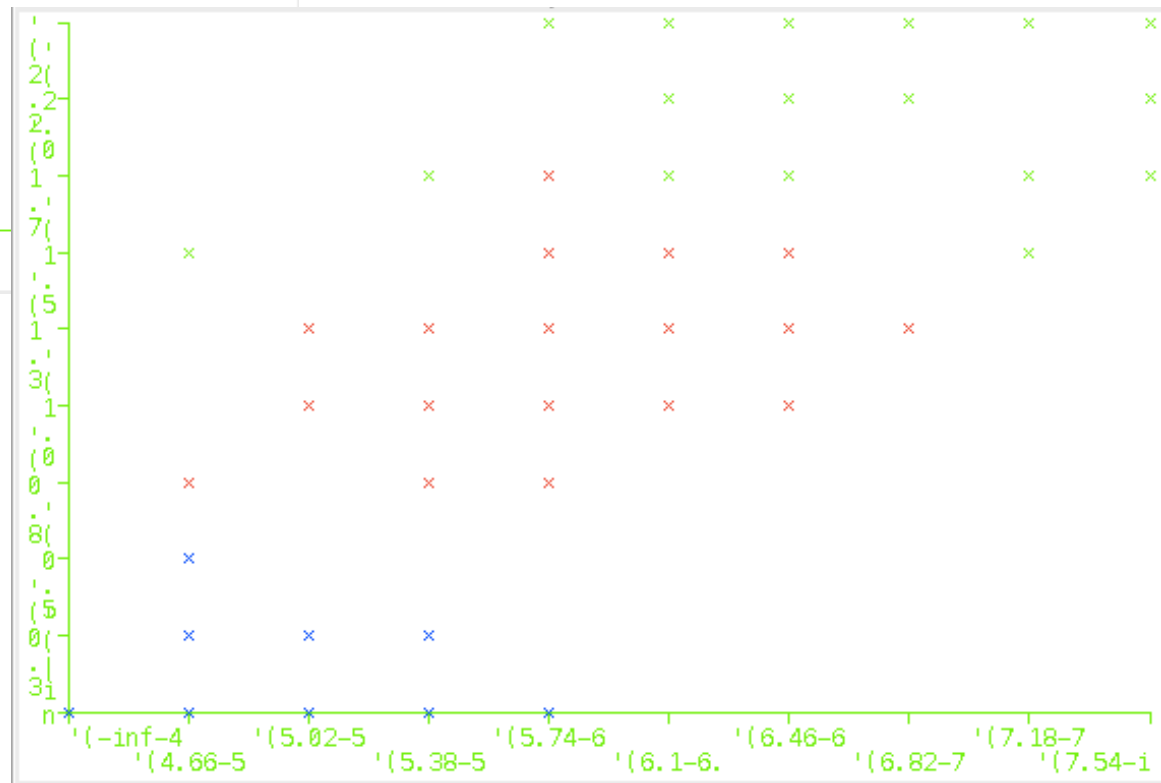
Status: OK Log x 0

Podaci pre i posle diskretizacije



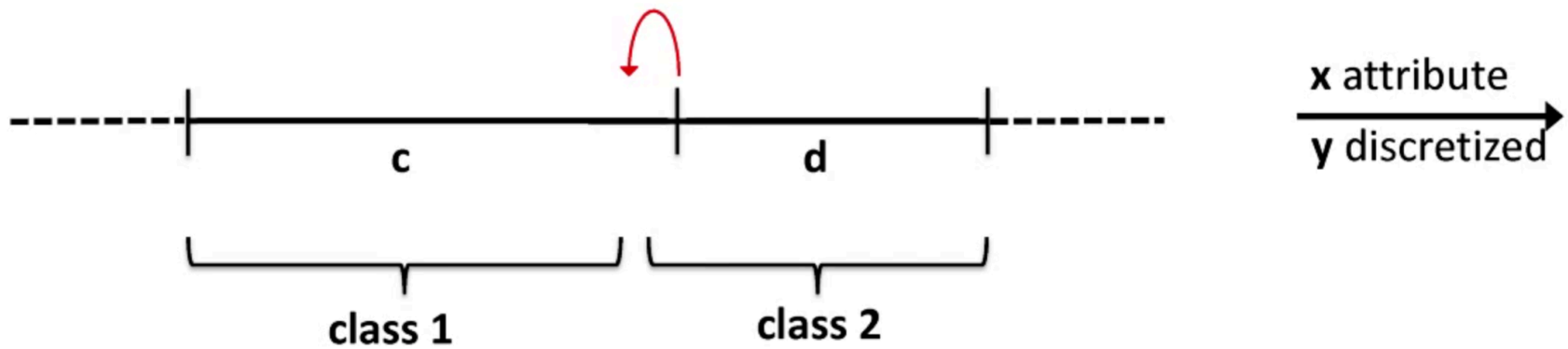
Podaci pre
diskretizacije

Podaci nakon
diskretizacije



Nadgledana diskretizacija

- Šta ako sve instance u jednom binu pripadaju jednoj klasi, a sve instance drugog bina pripadaju drugoj klasi osim prvog koji pripada prvoj klasi?



- Nadgledana diskretizacija uzima u obzir i klasu

Nadgledana diskretizacija

- Jedan od pristupa je koristiti entropiju
- U primeru *weather.numeric.arff*, kod atributa *temperature*

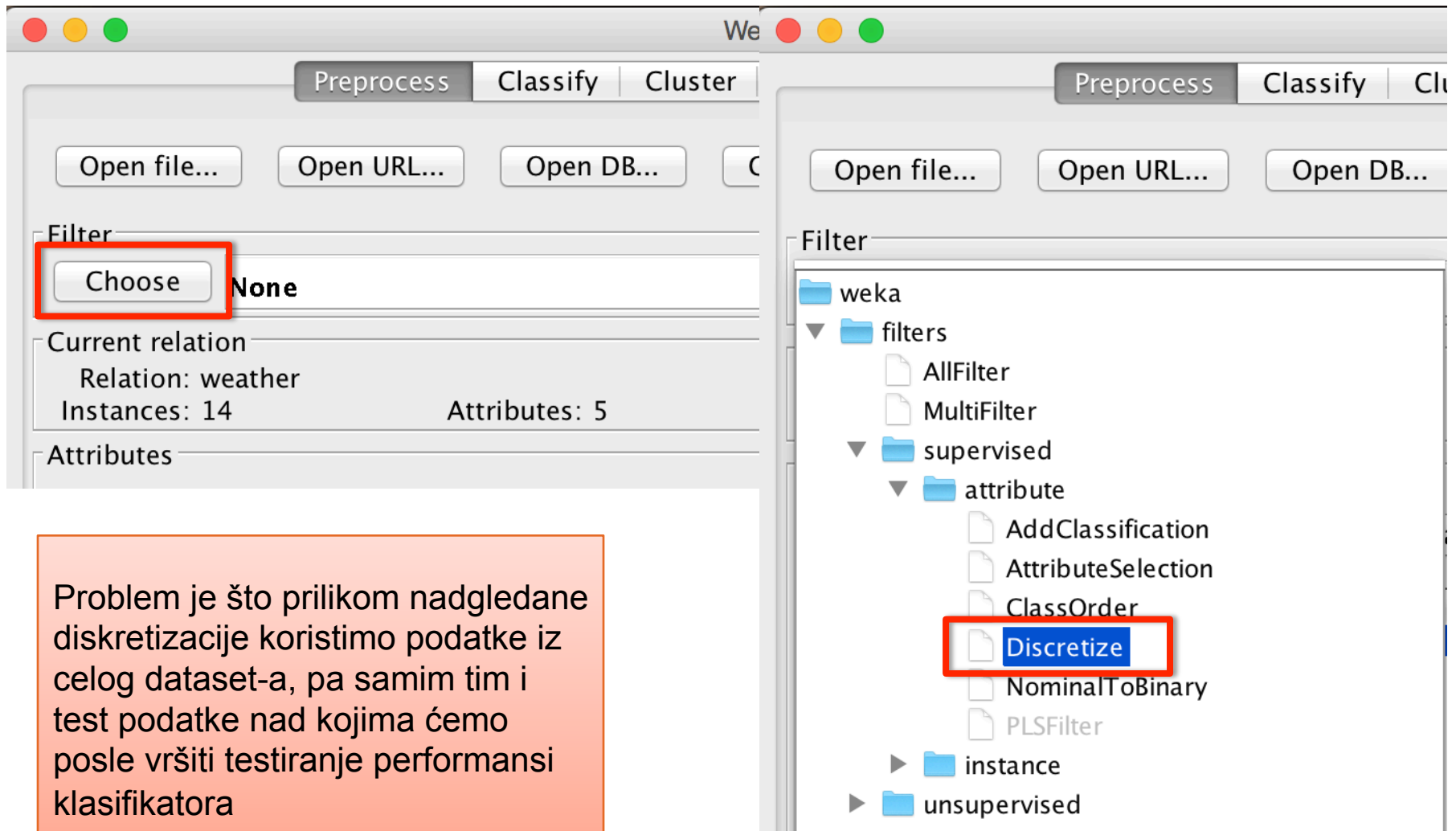
64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no
						yes	yes				
4 yes, 1 no					5 yes, 4 no						
entropy = 0.934 bits											

- Uzima se razdelnik sa najmanjom entropijom (najvećom informacionom dobiti)

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

Nadgledana diskretizacija u Weka-i

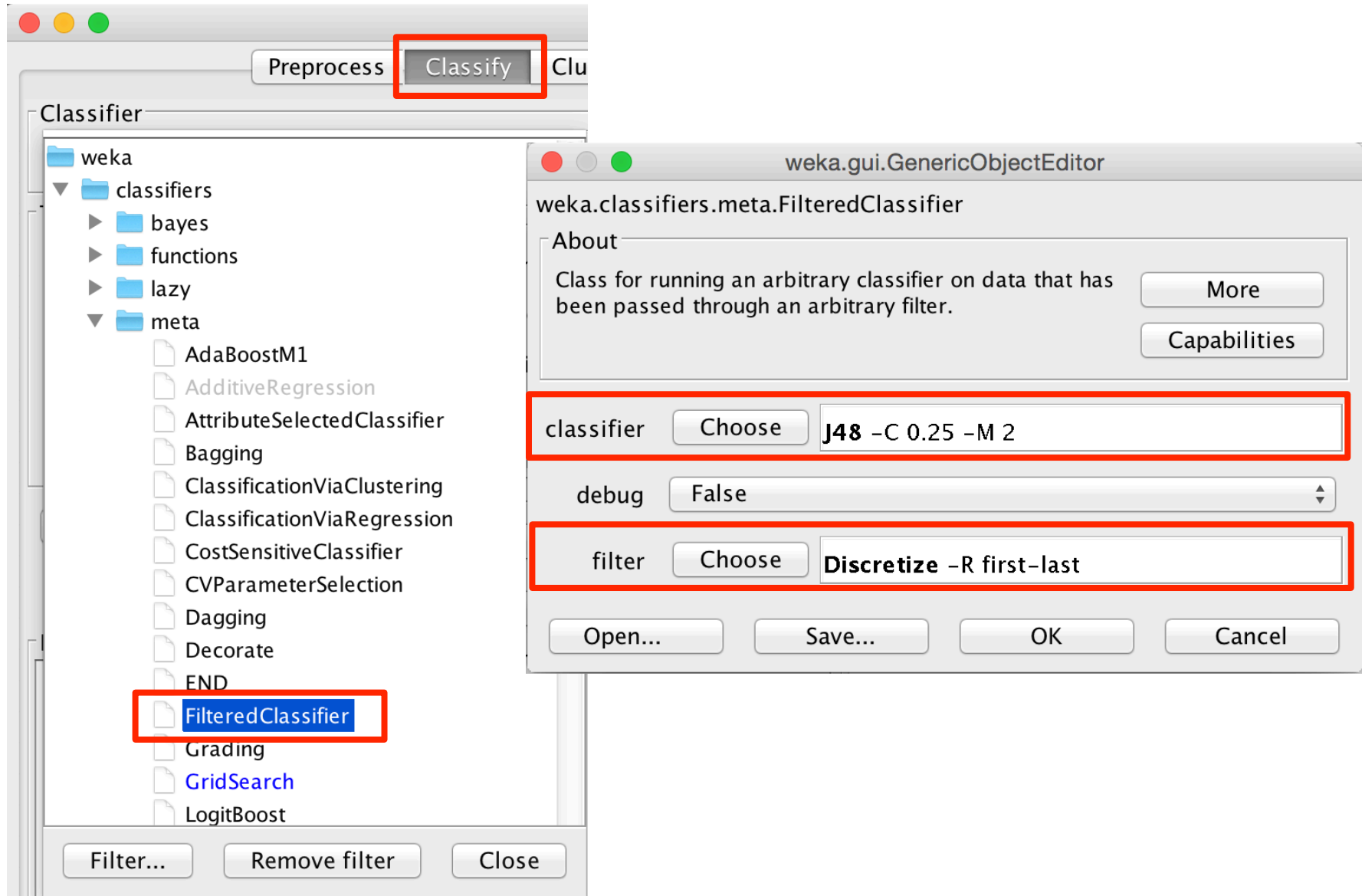
weather.numeric.arff



The image displays two screenshots of the Weka software interface, specifically the 'Preprocess' tab. The left screenshot shows the 'Filter' dropdown menu set to 'None', with a red box highlighting the 'Choose' button. The right screenshot shows the 'Filter' dropdown menu expanded, displaying a tree of filter categories. The 'Discretize' filter is highlighted with a red box.

Problem je što prilikom nadgledane diskretizacije koristimo podatke iz celog dataset-a, pa samim tim i test podatke nad kojima ćemo posle vršiti testiranje performansi klasifikatora

meta>FilteredClassifier



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

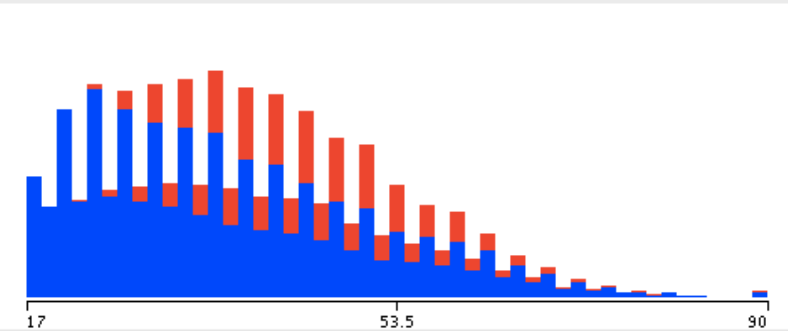
No.	Name
1	☒ age
2	☐ workclass
3	☐ fnlwgt
4	☐ education
5	☐ education-num
6	☐ marital-status
7	☐ occupation
8	☐ relationship
9	☐ race
10	☐ sex
11	☐ capital-gain
12	☐ capital-loss
13	☐ hours-per-week
14	☐ native-country
15	☐ income

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



The histogram shows the distribution of income values. The x-axis represents income values from 17 to 90, with a major tick at 53.5. The y-axis represents frequency. The bars are colored blue and red, showing a right-skewed distribution with a peak around 20-25.

Status
OK

Log x 0

Primer selekcije atributa

The screenshot displays the Weka software interface. At the top, there are tabs for 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below these are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', and 'Edit...'. The 'Filter' list on the left shows a tree structure with 'weka' at the root, followed by 'filters', 'supervised', and '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'. The right panel shows the 'Selected attribute' section with a table of statistics for the 'age' attribute. Below this, a dropdown menu shows 'Class: income (Nom)'. At the bottom right, there is a bar chart with blue and red bars.

Filter

- weka
 - filters
 - AllFilter
 - MultiFilter
 - supervised
 - attribute
 - AddClassification
 - AttributeSelection**
 - ClassOrder
 - Discretize
 - NominalToBinary
 - PLSFilter
 - instance
 - unsupervised

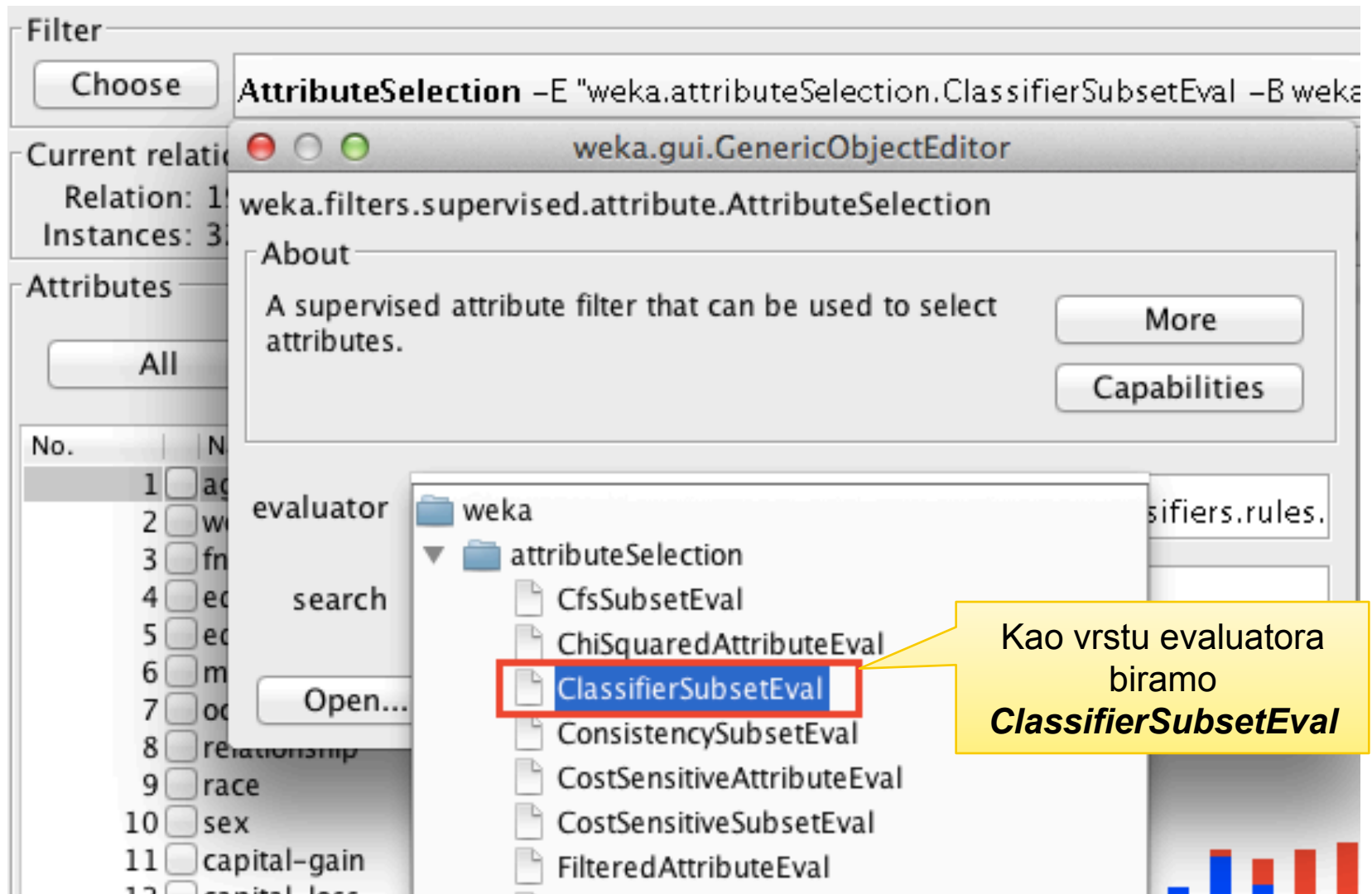
Selected attribute

Name: age	Typ
Missing: 0 (0%)	Uniqu
Distinct: 73	

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom)

Primer selekcije atributa



Primer selekcije atributa

The screenshot displays the Weka GUI with several windows open for attribute selection. The main window shows a list of attributes with checkboxes. Overlaid on this is the 'AttributeSelection' window, which contains an 'About' dialog and a 'ClassifierSubsetEval' window. The 'ClassifierSubsetEval' window has a 'classifier' field set to 'NaiveBayes', highlighted by a red box and a yellow callout bubble. The callout bubble contains the text 'Bramo NaiveBayes klasifikator'. The 'ClassifierSubsetEval' window also shows 'holdOutFile' and 'useTraining' options. In the background, a table of distinct values is visible.

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

weka.gui.GenericObjectEditor
weka.filters.supervised.attribute.AttributeSelection

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

ClassifierSubsetEval -B weka.classifiers.bayes

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**

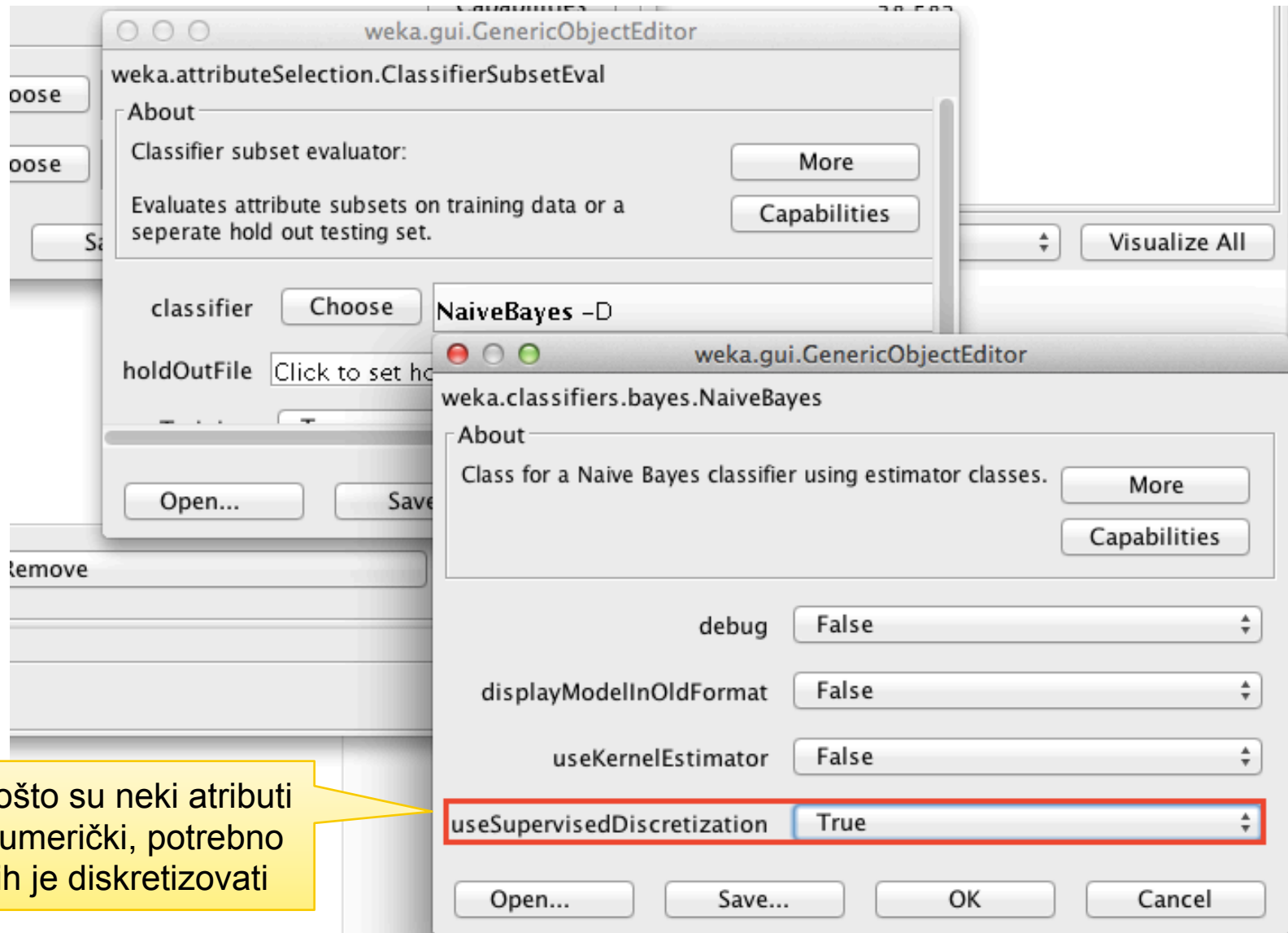
holdOutFile Click to set hold out or test instances

useTraining True

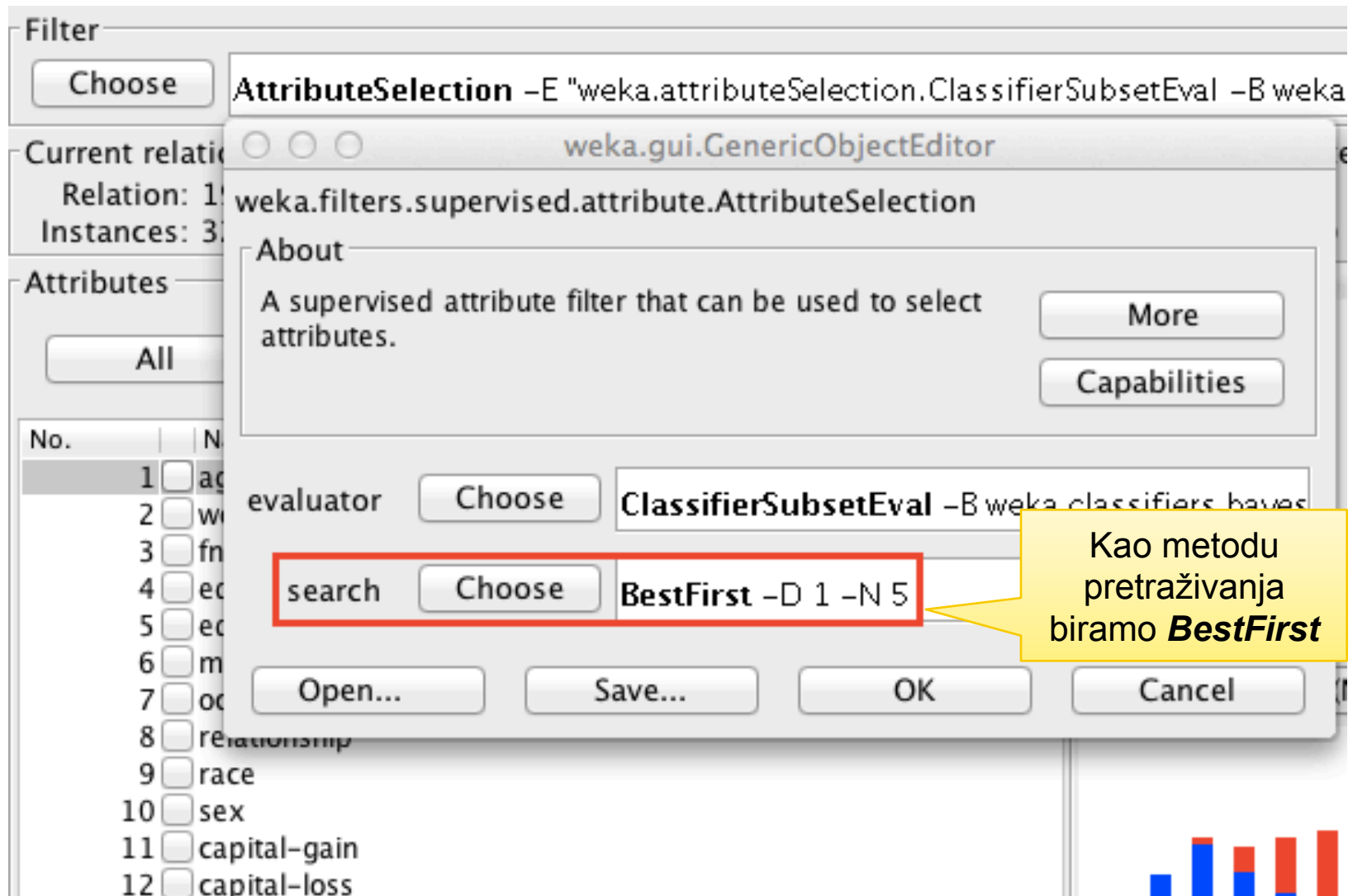
Distinct: 73

Value
17
90
38.582
13.64

Primer selekcije atributa



Primer selekcije atributa



Primer selekcije atributa

The screenshot shows the Weka software interface with the 'Attribute Selection' tab active. The 'Filter' section at the top shows 'AttributeSelection' with a command line: `-E "weka.attributeSelection.ClassifierSubsetEval" -B weka.classifiers.bayes.NaiveBayes -T -H \{Classifier\}`. The 'Apply' button is highlighted with a red box. A yellow callout box points to the 'Apply' button with the text: 'Filter je podešen i može biti primenjen nad atributima'.

The 'Current relation' section shows 'Relation: 1990census' and 'Instances: 32561'. The 'Attributes' section lists 15 attributes, with 'age' selected (indicated by a blue highlight and a checked checkbox). The 'Selected attribute' section shows 'Name: age', 'Missing: 0 (0%)', 'Distinct: 73', and 'Type: Numeric'. The 'Statistic' section lists 'Minimum', 'Maximum', 'Mean', and 'StdDev'. The 'Class: income (Nom)' is selected, and the 'Visualize All' button is visible. A histogram of the 'income' class is displayed at the bottom right, showing the distribution of income values across the dataset.

Status: OK

Primer selekcije atributa

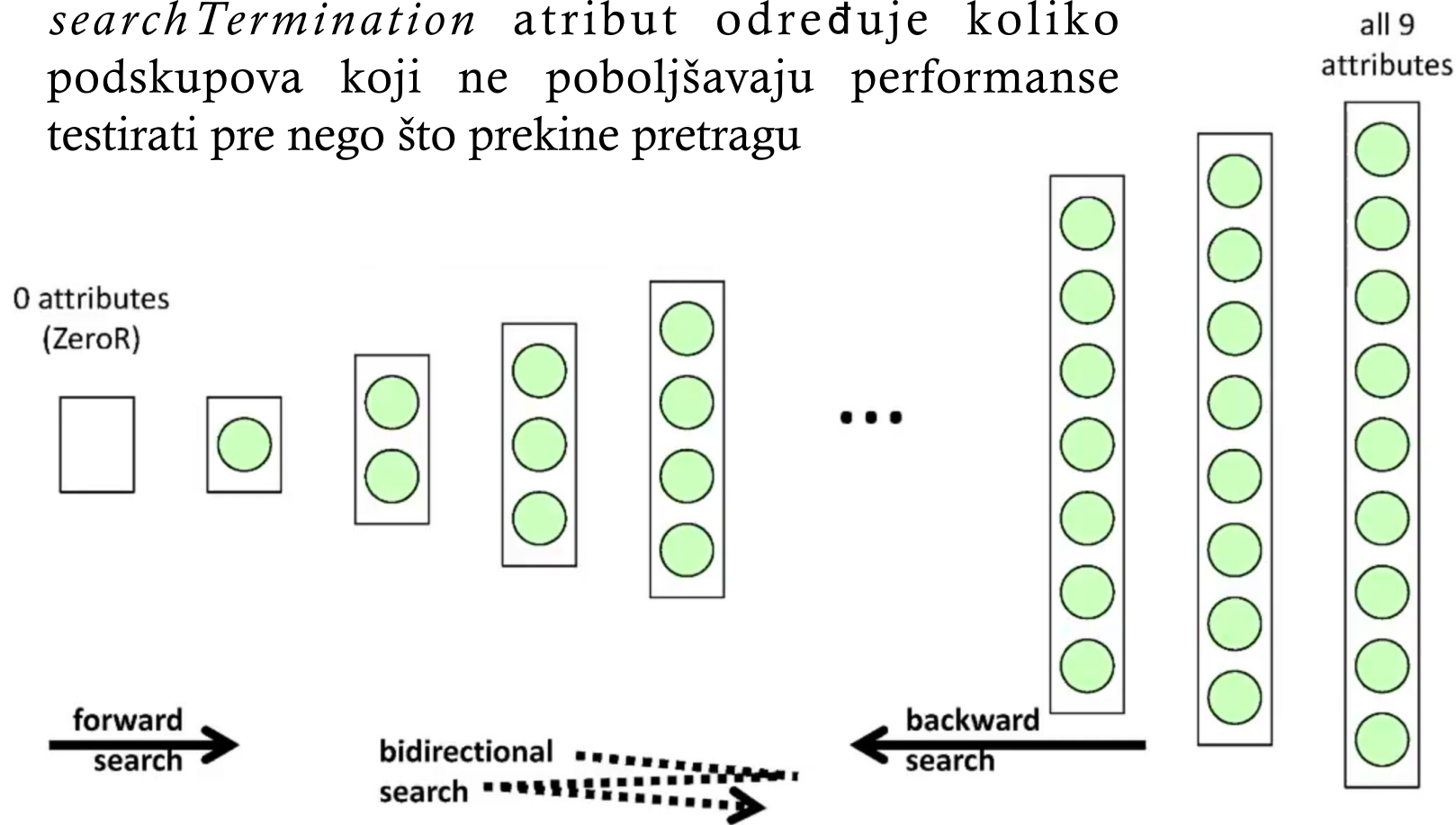
The screenshot shows the Weka software interface with the 'Attribute Selection' tab active. The 'Filter' dropdown is set to 'AttributeSelection' with the command: `-E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \\"Click"`. The 'Current relation' is '1990census-weka.filters.supervised.attribute.Attribute...' with 32561 instances and 7 attributes. The 'Attributes' list on the left shows 7 attributes: age, education, relationship, race, capital-gain, capital-loss, and income. The 'age' attribute is selected and highlighted in blue. A red box highlights the 'age' attribute in the list, and a yellow callout bubble points to it with the text 'Broj atributa je redukovan na 7'. The 'Selected attribute' panel on the right shows details for 'age': Name: age, Missing: 0 (0%), Distinct: 73, Type: Numeric, Unique: 2 (0%). Below this, a table shows statistics for 'age':

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

The 'Class' is set to 'income (Nom)'. A histogram at the bottom right shows the distribution of 'age' values, with the x-axis ranging from 17 to 90. The histogram bars are blue and red. The status bar at the bottom left shows 'Status OK'.

Metod pretrage kod selekcije atributa

- Exhaustive search (512 podskupova atributa)
- Best First: *Forward*, *Backward*, *Bi-directional*
 - *searchTermination* atribut određuje koliko podskupova koji ne poboljšavaju performanse testirati pre nego što prekine pretragu



Preporuke i zahvalnice

Weka Tutorials and Assignments @ The Technology Forge

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

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- Link: <https://www.youtube.com/user/WekaMOOC/>

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

<http://goo.gl/cqdp3l>

Pitanja?

NIKOLA MILIKIĆ

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

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