

Treniranje i testiranje

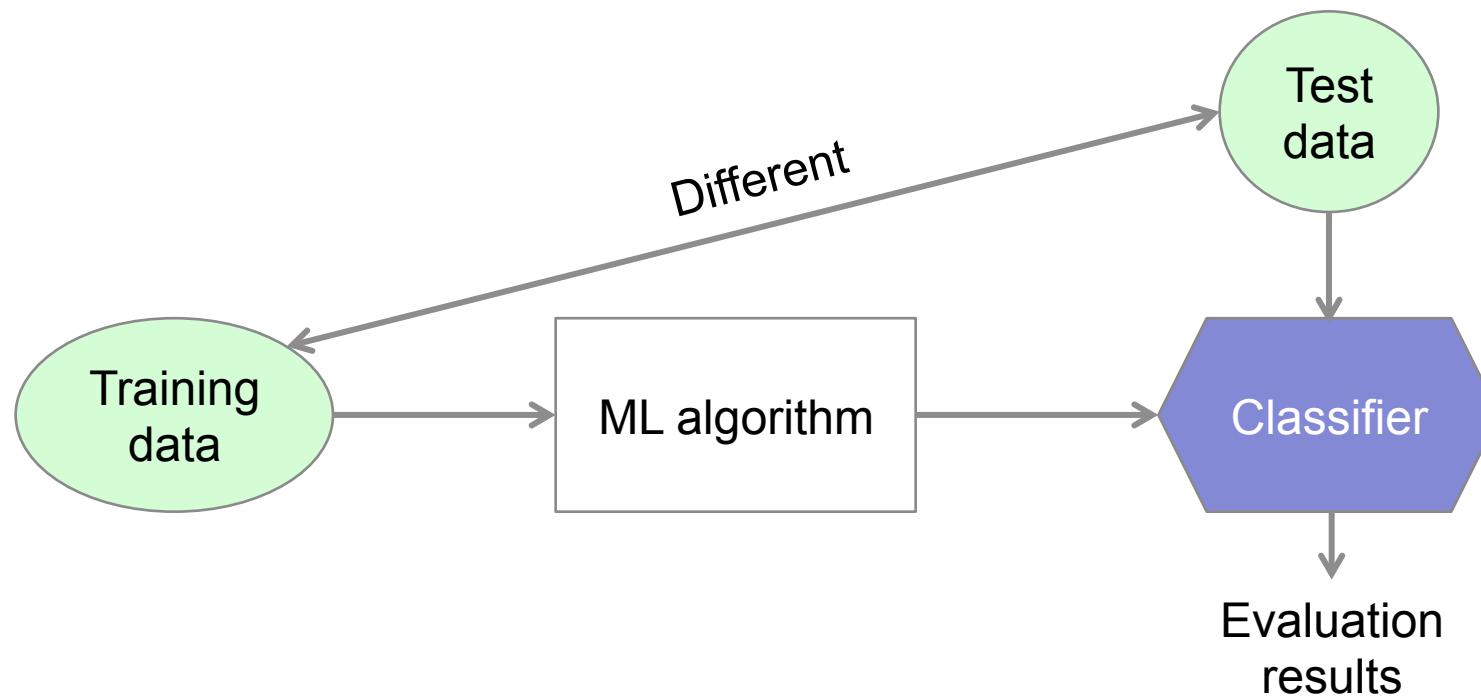
NIKOLA MILIKIĆ

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

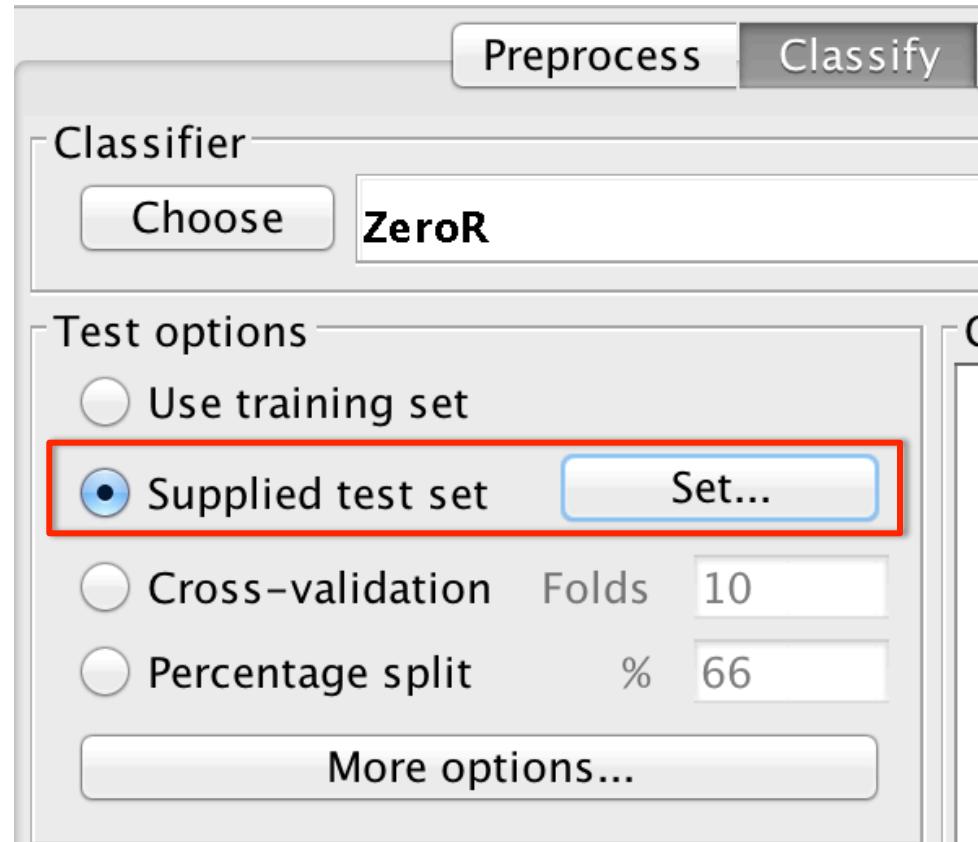
URL: <http://nikola.milikic.info>

Treniranje i testiranje

- Podatke za trening koristimo za kreiranje modela ML
- Podatke za testiranje koristimo za procenu performansi modela ML
- Podaci za trening i testiranje treba da budu različiti, međusobno nezavisni i dobijeni na slučajan način

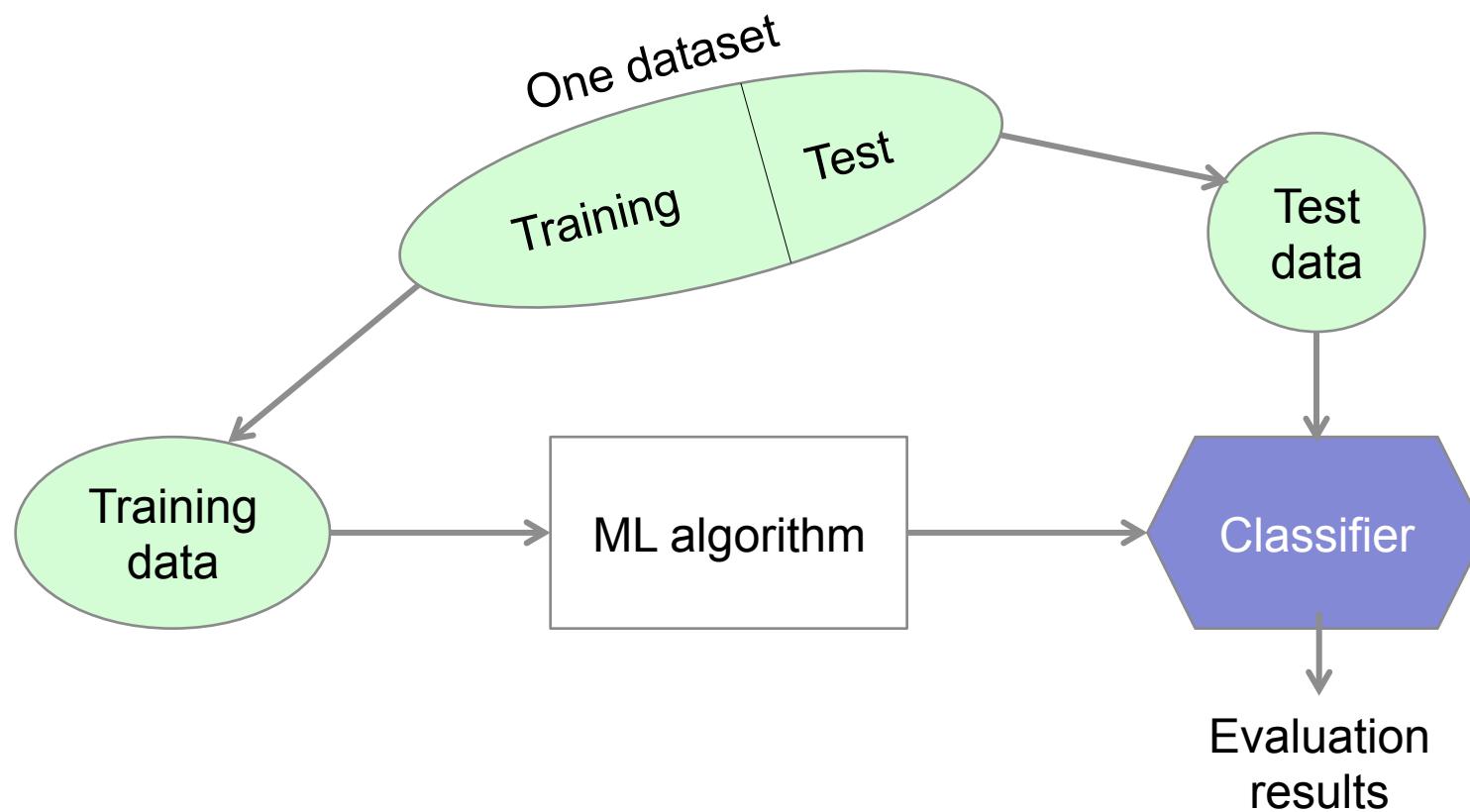


Odabir podataka za testiranje u Weka-i

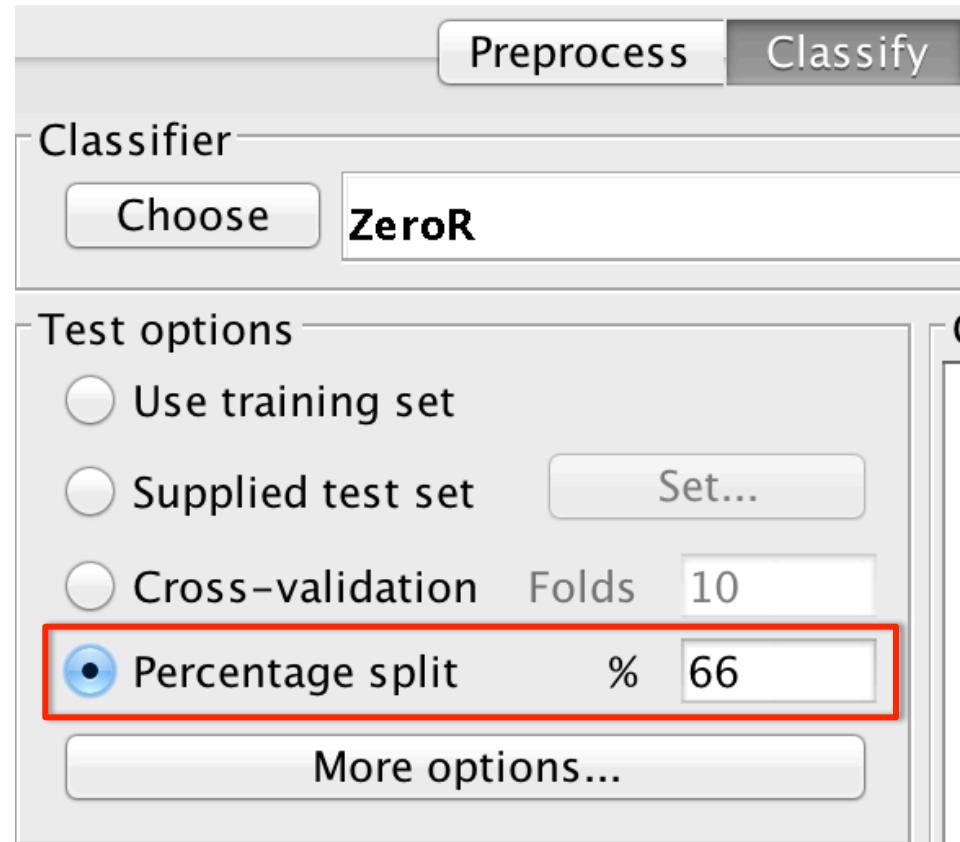


Treniranje i testiranje

- Ukoliko imamo samo jedan dataset (npr. u jednom fajlu), onda je potrebno podeliti taj dataset na slučajan način na podatke za trening i za testiranje



Podela dataseta na podatke za trening i test u Weka-i



Holdout metoda - različite vrednosti *random seed*

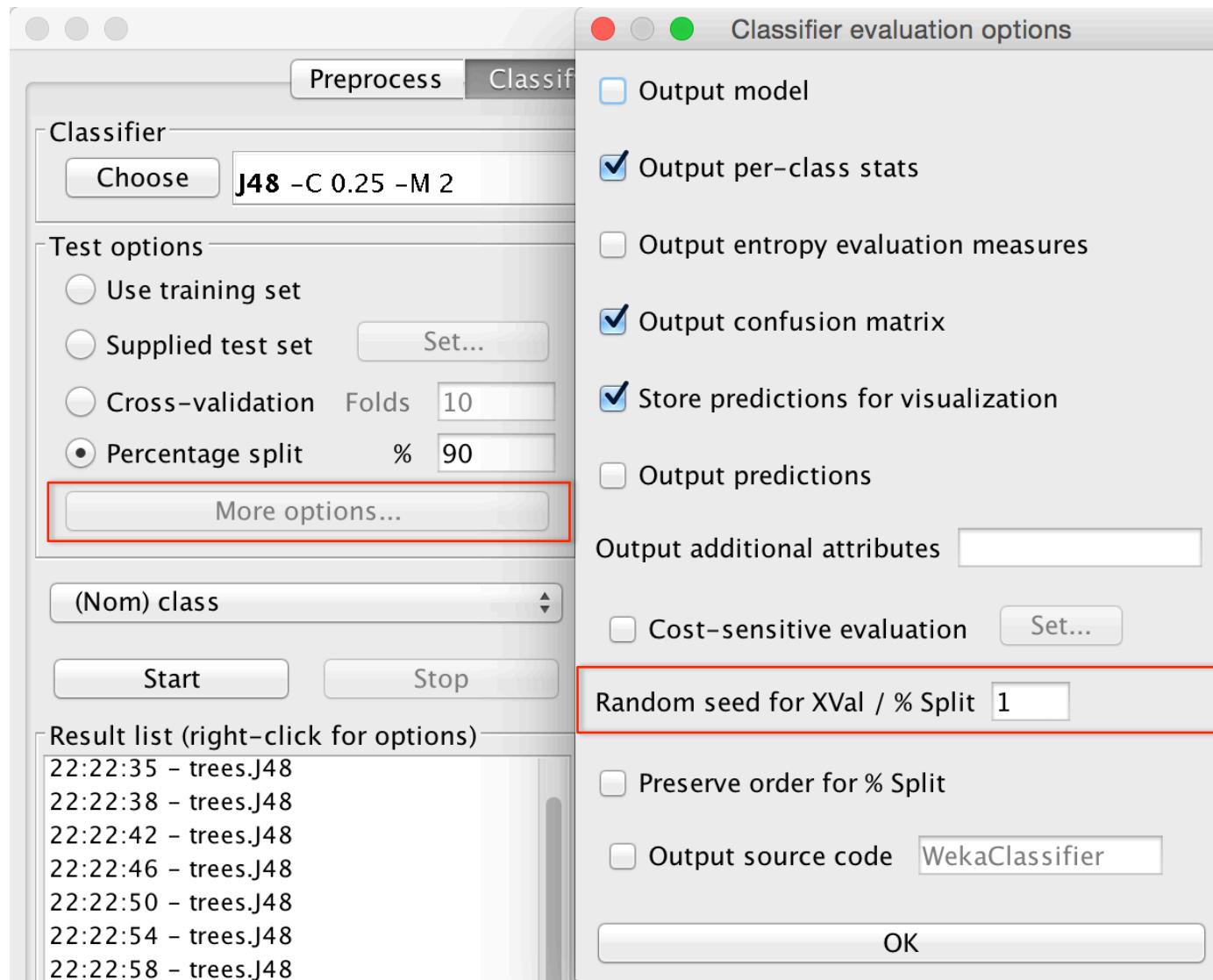
- Random seed je broj (ili vektor) koji se koristi prilikom inicijalizacije generatora pseudo-slučajnih brojeva
- Testiramo rezultate J48 klasifikatora nad datasetom *diabetes.arff*
- Sa *Percentage split* na 90% za različite vrednosti *random seed* dobijamo različite rezultate:

Random seed	1	2	3	4	5	6	7	8	9	10
Accuracy	0.753	0.779	0.805	0.74	0.714	0.701	0.792	0.714	0.805	0.675

$$\bar{x} = \frac{\sum x_i}{n} = 0.7478$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \sigma = 0.046$$

Promena random seed u Weka-i

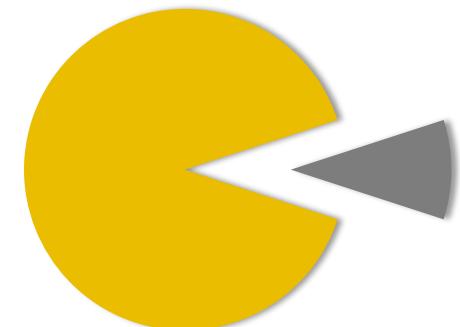
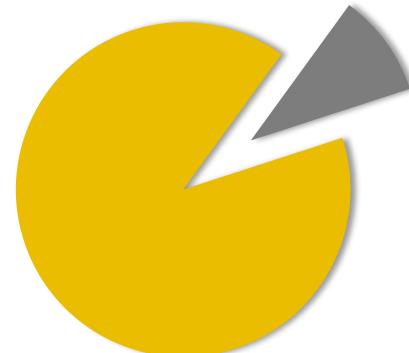


Kros validacija

- Cilj kros validacije je da:
 - spreči problem overfitting
 - učini predikcije generalnijim
- Unapređuje metodu deljenja dataseta procentualno time što smanjuje varijansu među podacima

Kros validacija

- Uključuje:
 - deljenje originalnog dataseta na k jednakih delova (folds)
 - Odvaja jedan deo za validaciju, a trening vrši nad preostalih $k-1$ delova i meri performanse
 - Ponavlja proces k puta tako što svaki put uzima drugi deo za validaciju



10-fold cross-validation

- $k = 10$
- Dataset se podeli na 10 jednakih delova (folds)
- Jedan deo se zadži u svakoj iteraciji
- Svaki deo se jednom koristi za testiranje rezultata, 9 puta za treniranje
- Izračuna se prosek rezultata



Stratifikovana kros-validacija

- Obezbeđuje da se u svakom delu (fold) nalazi odgovarajuća proporcija podataka

Kros-validacija u Weka-i

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. In the 'Classifier' section, 'J48 -C 0.25 -M 2' is chosen. Under 'Test options', 'Cross-validation' is selected with 10 folds, highlighted by a red box. A yellow arrow points from a callout box to the 'Stratified cross-validation' output section. The 'Classifier output' pane displays the stratified cross-validation results.

Weka Explorer

Preprocess Classify Cluster Associate Select attribut

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 90

More options...

Classifier output

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified
Incorrectly Classified
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error

Weka po defaultu radi stratifikovanu kros-validaciju

Kros validacija - različite vrednosti *random seed*

- Testiramo rezultate J48 klasifikatora nad datasetom *diabetes.arff*
- Sa *Cross-validation* na *10 folds* za različite vrednosti *random seed* dobijamo različite rezultate:

Random seed	1	2	3	4	5	6	7	8	9	10
Accuracy	0.738	0.75	0.755	0.755	0.743	0.756	0.736	0.74	0.745	0.73

$$\bar{x} = 0.7448$$

$$\sigma = 0.0008$$

Manja devijacija i varijansa kod kros-validacije

- Prethodni rezultati za holdout metodu:

$$\bar{x} = 0.7478$$

$$\sigma = 0.046$$

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>