

Training and Testing

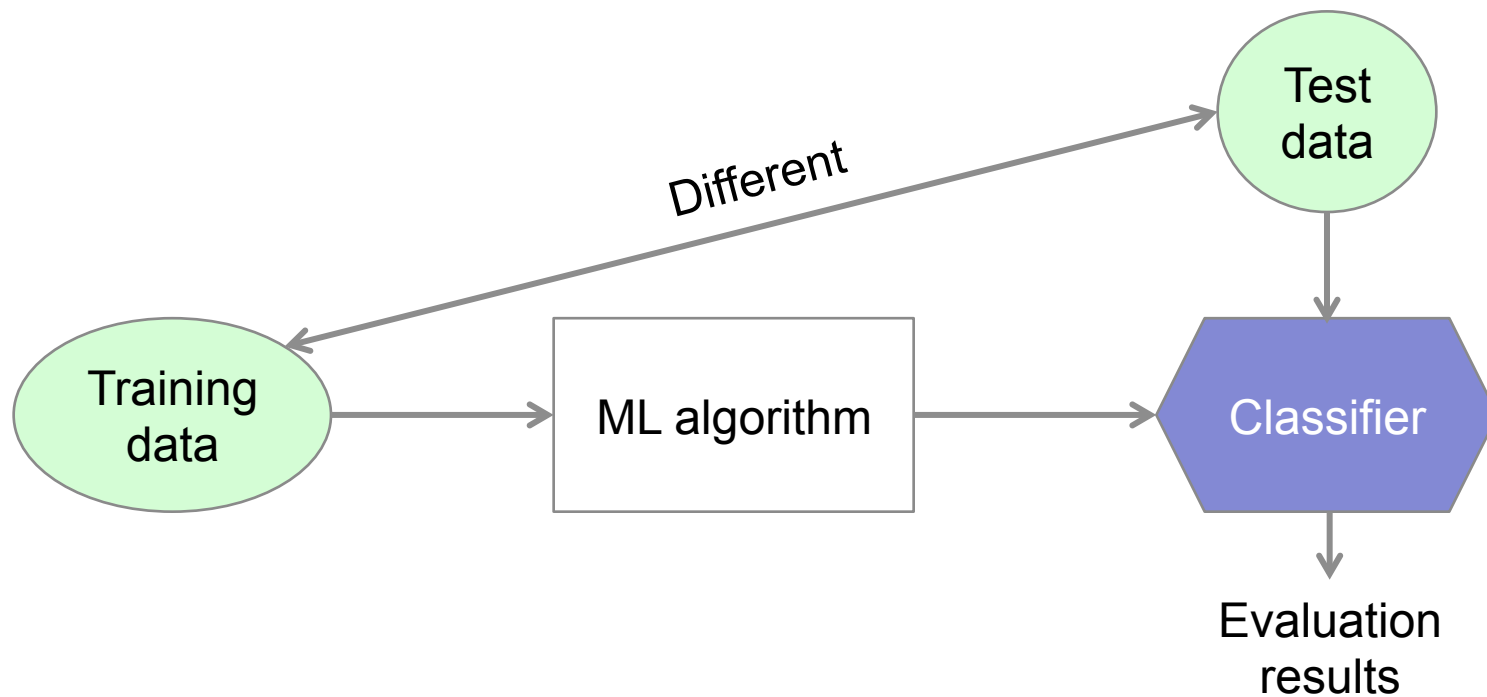
UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

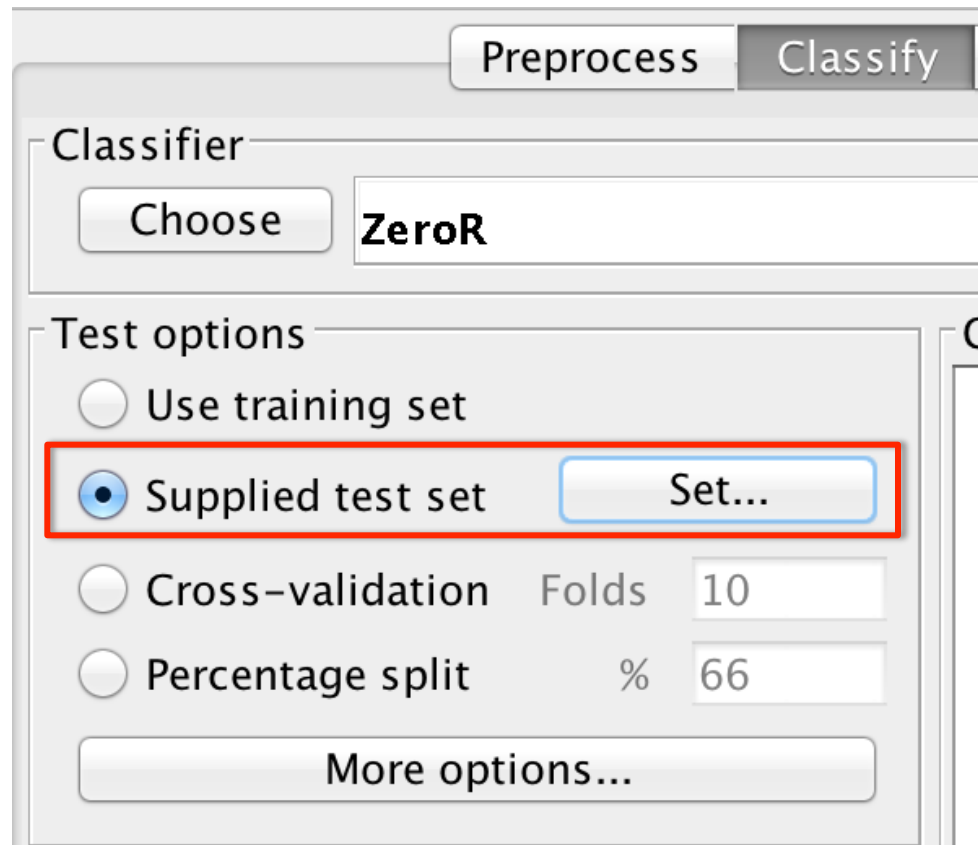
URL: <http://krcadinac.com>

Training and Testing

- Training data is used for building a ML model
- Testing data is used for measuring performance of a ML model
- Training and testing data should be different, mutually independent and created by random sampling

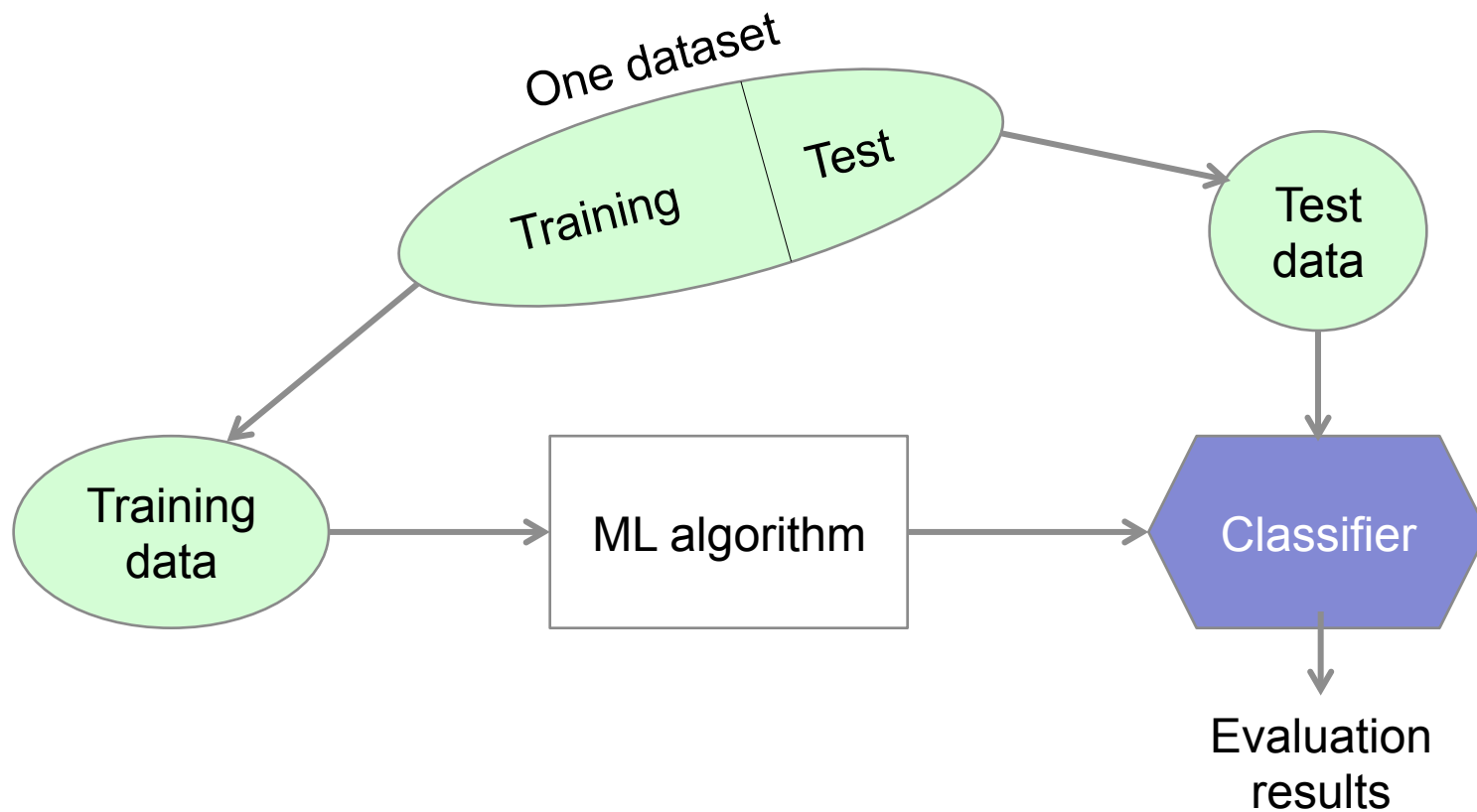


Supplying testing set in Weka

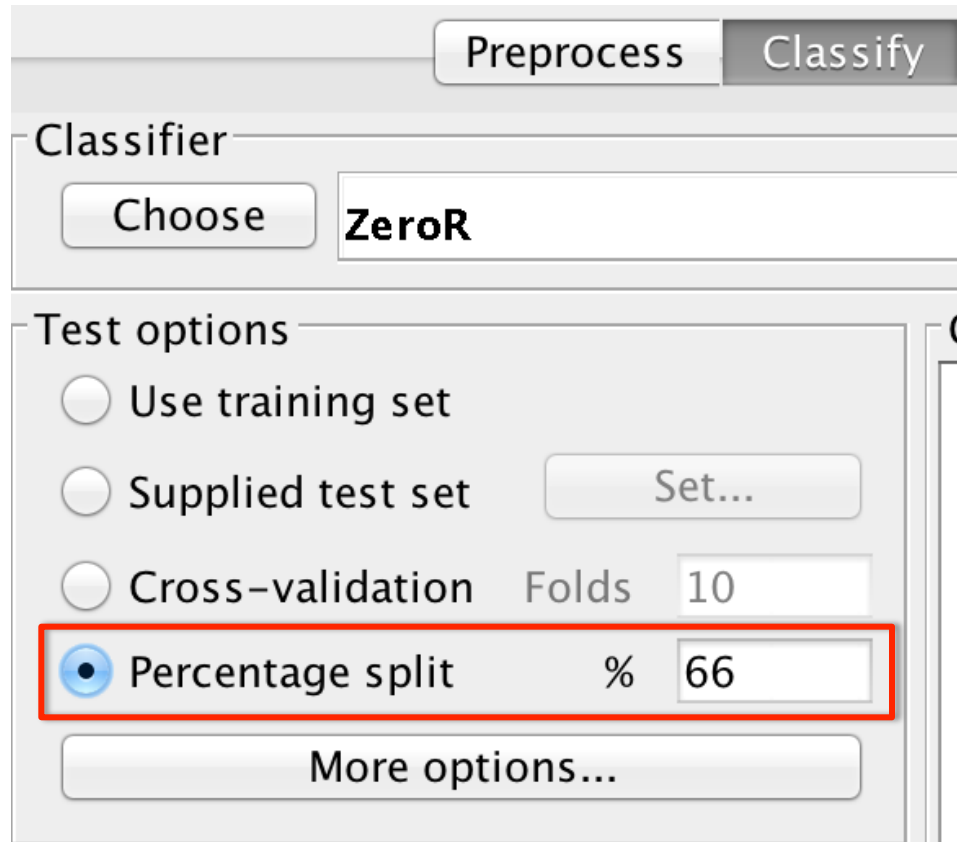


Training and Testing

- In case we have one dataset (for instance, in one file), we need to split the original dataset into subsets for training and testing



Dividing original dataset into testing and training in Weka



Holdout method – different *random seed* values

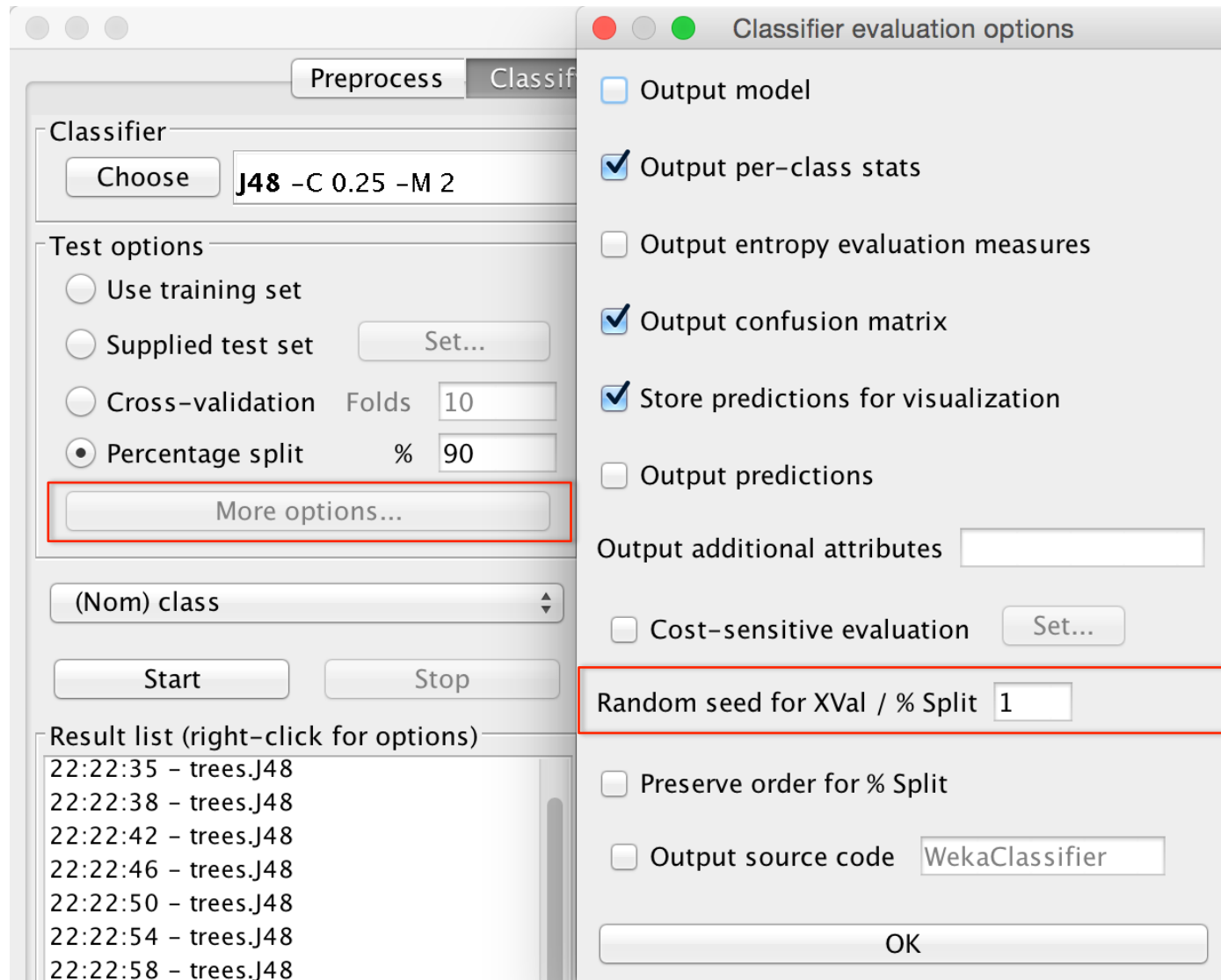
- Random seed is a number (or a vector) used to initialize a pseudorandom number generator
- Testing J48 classifier results over the dataset *diabetes.arff*
- With *Percentage split* set to 90% for different *random seed* values we get different results:

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

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

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

Changing random seed number in Weka

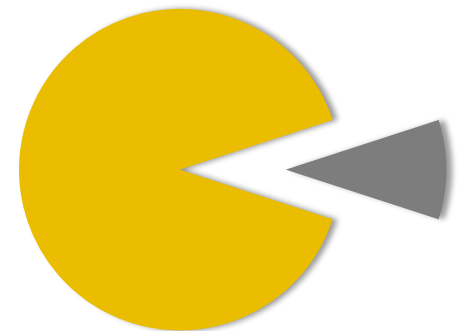


Cross-validation

- Goal of cross-validation is:
 - to overcome the problem of overfitting
 - makes the predictions more general
- Improving the holdout method by reducing the variance among data

Cross-validation

- Includes:
 - Splitting the original dataset into k equal parts (folds)
 - Takes out one fold aside, and performs training over the rest $k-1$ folds and measures the performance
 - Repeats the process k times by taking different fold each time



10-fold cross-validation

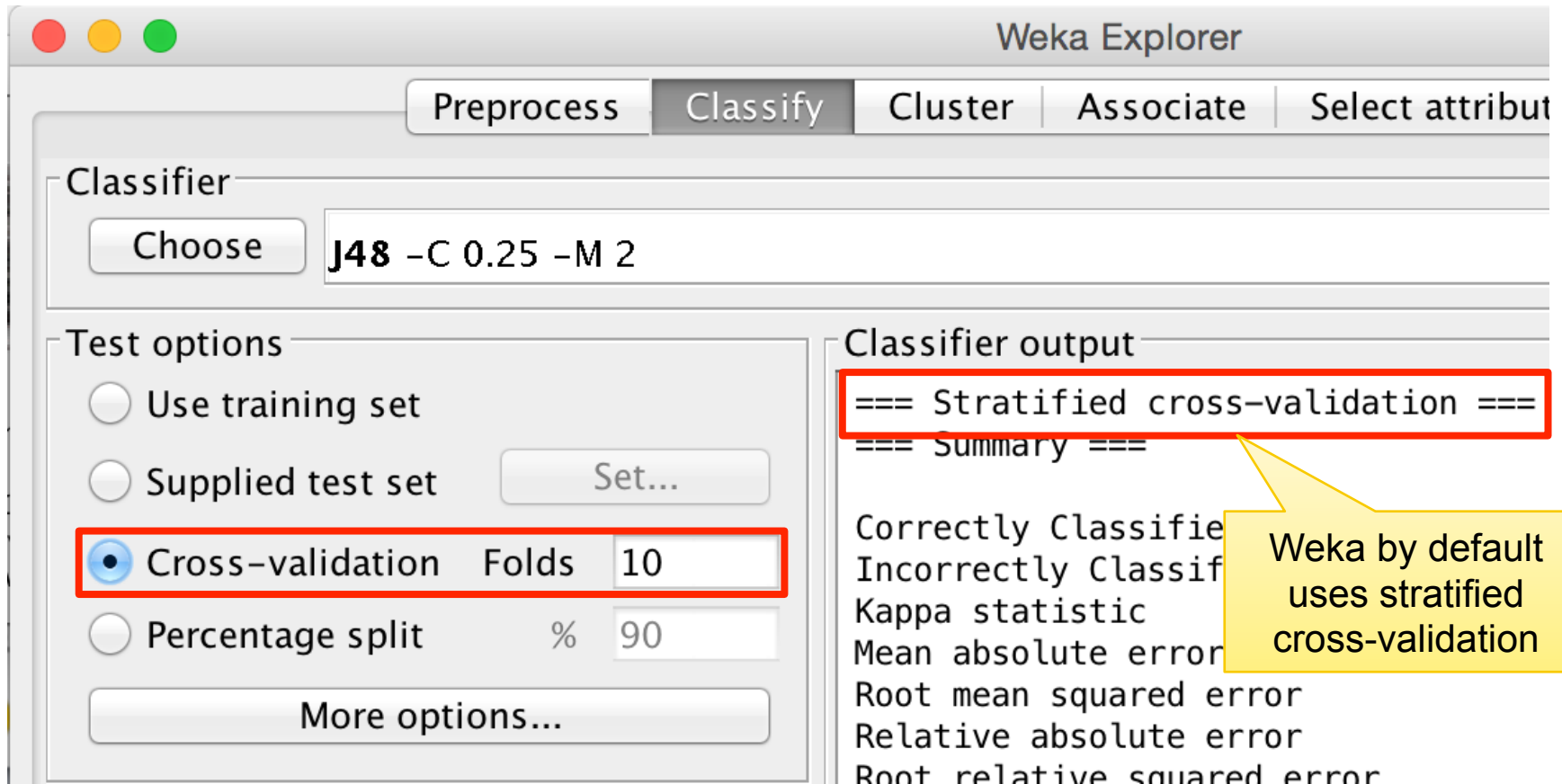
- $k = 10$
- Dataset is divided into 10 equal parts (folds)
- One fold is set aside in each iteration
- Each fold is used once for testing, nine times for training
- Average the scores



Stratified cross-validation

- Ensures that each fold has the right proportion of each class value

Cross-validation in Weka



Cross validation- different *random seed*

- Testing results of J48 classifier over dataset *diabetes.arff*
- With *Cross-validation* set to *10 folds* for different *random seed* values we get different results:

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

$$\bar{x} = 0.7448$$

$$\sigma = 0.0008$$

Smaller deviation
and variance with
cross-validation

- Previous results for holdout method:

$$\bar{x} = 0.7478$$

$$\sigma = 0.046$$

Recommendations and credits

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/>

(Anonymous) survey for your
comments and suggestions:

<http://goo.gl/cqdp3l>

ANY QUESTIONS?

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: <http://krcadinac.com>