

# Introduction to WEKA

1. Use the following learning schemes to analyze the zoo data (in [zoo.arff](#)):

- Decision stump - `weka.classifiers.DecisionStump`

**Program** Weka Workbench

**Preprocess** **Classify** Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

**Classifier** Choose **DecisionStump**

**Test options**

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %
- 

(Nom) type

Result list (right-click for options)

08:12:36 - trees.DecisionStump

**Classifier output**

Incorrectly Classified Instances 40 39.604 %

Kappa statistic 0.4481

Mean absolute error 0.1337

Root mean squared error 0.259

Relative absolute error 60.9785 %

Root relative squared error 78.5181 %

Total Number of Instances 101

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	mammal
	1.000	0.494	0.333	1.000	0.500	0.411	0.742	0.324	bird
	0.000	0.000	?	0.000	?	?	0.558	0.072	reptile
	0.000	0.000	?	0.000	?	?	0.673	0.186	fish
	0.000	0.000	?	0.000	?	?	0.527	0.056	amphibian
	0.000	0.000	?	0.000	?	?	0.647	0.119	insect
	0.000	0.000	?	0.000	?	?	0.715	0.162	invertebrate
Weighted Avg.	0.604	0.098	?	0.604	?	?	0.810	0.525	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	<-- classified as
41	0	0	0	0	0	0	0	a = mammal
0	20	0	0	0	0	0	0	b = bird
0	5	0	0	0	0	0	0	c = reptile
0	13	0	0	0	0	0	0	d = fish
0	4	0	0	0	0	0	0	e = amphibian
0	8	0	0	0	0	0	0	f = insect
0	10	0	0	0	0	0	0	g = invertebrate

- OneR - `weka.classifiers.OneR`

Program Weka Workbench

Preprocess **Classify** Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

Classifier Choose **OneR -B 6**

Test options

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds **10**
- ☐ Percentage split % **66**

More options...

(Nom) type

Start Stop

Result list (right-click for options)

- 08:12:36 - trees.DecisionStump
- 08:13:50 - rules.OneR

Classifier output

```

Correctly Classified Instances 40 42.0743 %
Incorrectly Classified Instances 56 57.4257 %
Kappa statistic 0.045
Mean absolute error 0.1641
Root mean squared error 0.4051
Relative absolute error 74.8424 %
Root relative squared error 122.7774 %
Total Number of Instances 101
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.967	0.414	1.000	0.586	0.117	0.517	0.414	mammal	
0.000	0.000	?	0.000	?	?	0.500	0.198	bird	
0.000	0.000	?	0.000	?	?	0.500	0.050	reptile	
0.000	0.000	?	0.000	?	?	0.500	0.129	fish	
0.500	0.000	1.000	0.500	0.667	0.700	0.750	0.520	amphibian	
0.000	0.000	?	0.000	?	?	0.500	0.079	insect	
0.000	0.000	?	0.000	?	?	0.500	0.099	invertebrate	
Weighted Avg.	0.426	0.392	?	0.426	?	?	0.517	0.263	

=== Confusion Matrix ===

```

a b c d e f g <-- classified as
41 0 0 0 0 0 0 | a = mammal
20 0 0 0 0 0 0 | b = bird
5 0 0 0 0 0 0 | c = reptile
13 0 0 0 0 0 0 | d = fish
2 0 0 0 2 0 0 | e = amphibian
8 0 0 0 0 0 0 | f = insect
10 0 0 0 0 0 0 | g = invertebrate
  
```

Status OK Log x 0

- Decision table - weka.classifiers.DecisionTable -R

Program Weka Workbench

Preprocess **Classify** Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

Classifier Choose **DecisionTable -X 1 -S "weka.attributeSelection.Bestfirst-D 1 -N 5"**

Test options

- ☐ Use training set
- ☐ Supplied test set Set...
- ☒ Cross-validation Folds **10**
- ☐ Percentage split % **66**

More options...

(Nom) type

Start Stop

Result list (right-click for options)

- 08:12:36 - trees.DecisionStump
- 08:13:50 - rules.OneR
- 08:14:28 - rules.DecisionTable

Classifier output

```

Correctly Classified Instances 67 66.1961 %
Incorrectly Classified Instances 34 33.8039 %
Kappa statistic 0.8127
Mean absolute error 0.1302
Root mean squared error 0.2142
Relative absolute error 59.3758 %
Root relative squared error 64.9211 %
Total Number of Instances 101
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.100	0.872	1.000	0.932	0.886	0.996	0.991	mammal	
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	bird	
0.000	0.010	0.000	0.000	0.000	-0.023	0.916	0.258	reptile	
1.000	0.011	0.929	1.000	0.963	0.958	0.996	0.957	fish	
0.750	0.021	0.600	0.750	0.667	0.656	0.977	0.508	amphibian	
1.000	0.032	0.727	1.000	0.842	0.839	0.978	0.658	insect	
0.200	0.011	0.667	0.200	0.308	0.333	0.848	0.406	invertebrate	
Weighted Avg.	0.861	0.047	0.819	0.861	0.824	0.805	0.976	0.849	

=== Confusion Matrix ===

```

a b c d e f g <-- classified as
41 0 0 0 0 0 0 | a = mammal
0 20 0 0 0 0 0 | b = bird
3 0 0 1 0 0 1 | c = reptile
0 0 0 13 0 0 0 | d = fish
1 0 0 0 3 0 0 | e = amphibian
0 0 0 0 0 8 0 | f = insect
2 0 1 0 2 3 2 | g = invertebrate
  
```

Status OK Log x 0

- C4.5 - the J48 classifier

Weka Workbench

Program

Preprocess **Classify** Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

Classifier

Choose **J48 -C 0.25 -M 2**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) type

Start Stop

Result list (right-click for options)

- 08:12:36 - trees.DecisionStump
- 08:13:50 - rules.OneR
- 08:14:28 - rules.DecisionTable
- 08:15:08 - trees.J48**

Classifier output

Correctly Classified Instances 93 92.0792 %

Incorrectly Classified Instances 8 7.9208 %

Kappa statistic 0.8955

Mean absolute error 0.0225

Root mean squared error 0.14

Relative absolute error 10.2478 %

Root relative squared error 42.4398 %

Total Number of Instances 101

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	mammal
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	bird
0.600	0.010	0.750	0.600	0.667	0.656	0.793	0.420	0.420	reptile
1.000	0.011	0.929	1.000	0.963	0.958	0.994	0.929	0.929	fish
0.750	0.000	1.000	0.750	0.857	0.862	0.872	0.760	0.760	amphibian
0.625	0.032	0.625	0.625	0.625	0.593	0.920	0.677	0.677	insect
0.800	0.033	0.727	0.800	0.762	0.735	0.986	0.812	0.812	invertebrate
Weighted Avg.	0.921	0.008	0.922	0.921	0.920	0.914	0.976	0.908	

=== Confusion Matrix ===

```

a b c d e f g <-- Classified as
41 0 0 0 0 0 0 | a = mammal
0 20 0 0 0 0 0 | b = bird
0 0 3 1 0 1 0 | c = reptile
0 0 0 13 0 0 0 | d = fish
0 0 1 0 3 0 0 | e = amphibian
0 0 0 0 0 5 3 | f = insect
0 0 0 0 0 2 8 | g = invertebrate

```

Status OK

Log x0

- PART - under "rules"

Weka Workbench

Program

Preprocess **Classify** Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI

Classifier

Choose **PART -C 0.25 -M 2**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) type

Start Stop

Result list (right-click for options)

- 08:12:36 - trees.DecisionStump
- 08:13:50 - rules.OneR
- 08:14:28 - rules.DecisionTable
- 08:15:08 - trees.J48
- 08:16:21 - rules.PART**

Classifier output

Correctly Classified Instances 93 92.0792 %

Incorrectly Classified Instances 8 7.9208 %

Kappa statistic 0.8955

Mean absolute error 0.0231

Root mean squared error 0.1435

Relative absolute error 10.5346 %

Root relative squared error 43.4854 %

Total Number of Instances 101

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	mammal
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	bird
0.600	0.010	0.750	0.600	0.667	0.656	0.793	0.420	0.420	reptile
1.000	0.011	0.929	1.000	0.963	0.958	0.994	0.929	0.929	fish
0.750	0.000	1.000	0.750	0.857	0.862	0.872	0.760	0.760	amphibian
0.625	0.032	0.625	0.625	0.625	0.593	0.920	0.677	0.677	insect
0.800	0.033	0.727	0.800	0.762	0.735	0.986	0.812	0.812	invertebrate
Weighted Avg.	0.921	0.008	0.922	0.921	0.920	0.914	0.976	0.908	

=== Confusion Matrix ===

```

a b c d e f g <-- classified as
41 0 0 0 0 0 0 | a = mammal
0 20 0 0 0 0 0 | b = bird
0 0 3 1 0 1 0 | c = reptile
0 0 0 13 0 0 0 | d = fish
0 0 1 0 3 0 0 | e = amphibian
0 0 0 0 0 5 3 | f = insect
0 0 0 0 0 2 8 | g = invertebrate

```

- How do the classifiers determine whether an animal is a mammal, bird, reptile, fish, amphibian, insect, or invertebrate?
  - Decision Stump: Nothing was identified correctly except the bird and mammal groups. There are plenty of groups that were identified as birds that shouldn't

have been as well. Based on the viewer it appears as though it is not checking past the milk attribute so a lot of these variables are not being correctly identified.

- Decision Table: It looks like amphibian, invertebrate, and reptile were misclassified. This is because they were only measured by a few attributes such as milk, legs, and tail.
- OneR: Looks like only mammal was identified correctly, because only one variable was used to complete the matrix.
- PART: Only fish, mammal, and bird were correctly identified, but since many decision rules were used it appears to be more accurate.
- J48: Same as PART, even down to the CCI.
- Do the decisions made by the classifiers make sense to you?
  - It makes sense but I also do not understand why you would use OneR, because it only analyzes by one parameter compared to the others that are more accurate.
- What can you say about the accuracy of these classifiers when classifying an animal that has not been used for training?
  - J48 and the Decision Stump were the most accurate with a CCI over 90%.
- Why does OneR perform so badly?
  - It only analyzes by a single parameter since it is meant to be simple.

2. Use the following learning schemes to analyze the bolts data ( [bolts.arff](#) without the TIME attribute):

- Decision stump - `weka.classifiers.DecisionStump`

```

TOTAL
SPEED2
NUMBER2
SENS
TIME
T20BOLT
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Stump

Classifications

TIME <= 32.19 : 18.378275862068968
TIME > 32.19 : 74.94454545454549
TIME is missing : 33.934000000000001

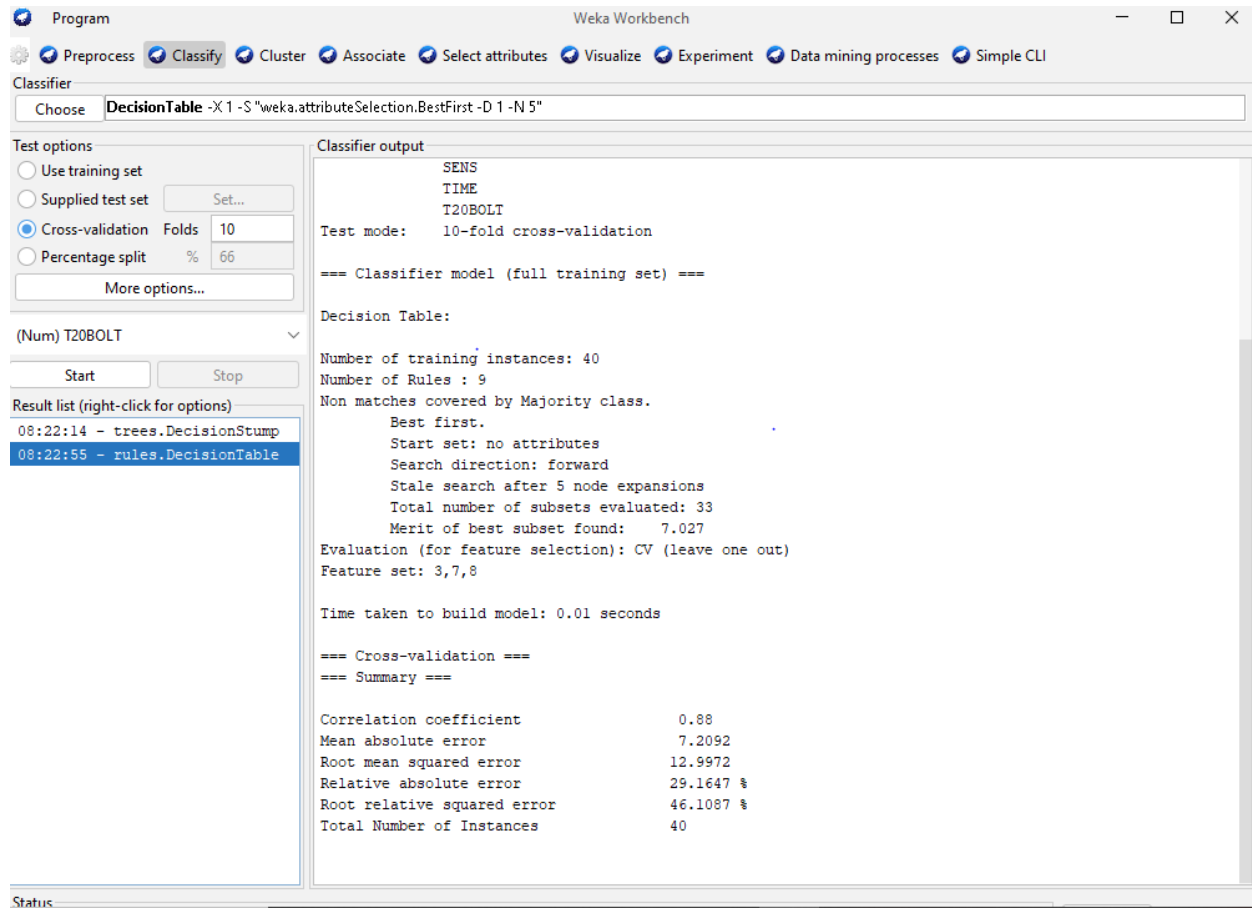
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

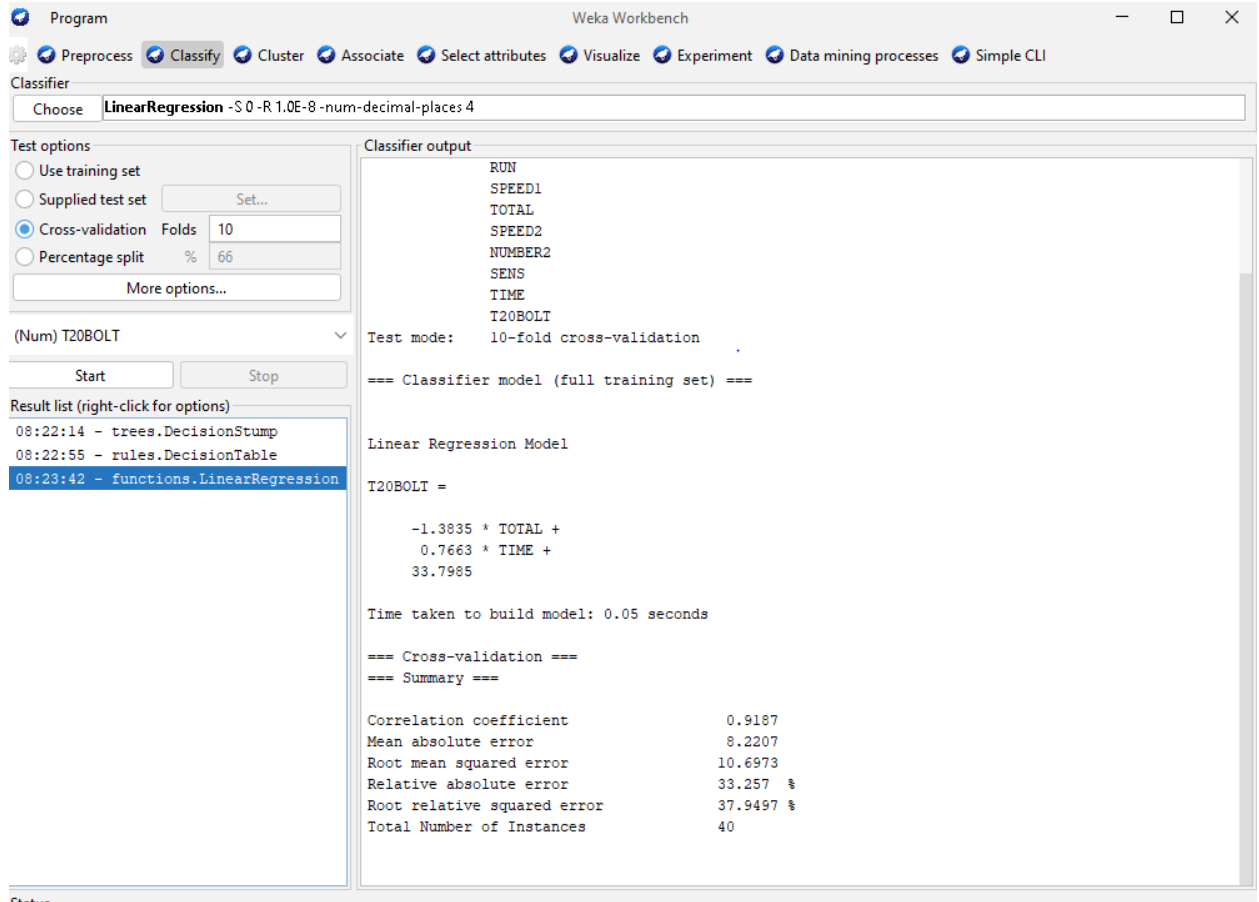
Correlation coefficient      0.8677
Kendall's tau               0.0657
Spearman's rho              0.2955
Mean absolute error         9.1809
Root mean squared error     13.6626
Relative absolute error     37.1415 %
Root relative squared error 48.4693 %
Total Number of Instances   40

```

- [Decision table - weka.classifiers.DecisionTable -R](#)



- Linear regression - `weka.classifiers.LinearRegression`



- M5' - weka.classifiers.M5'
- === Run information ===
- 
- Scheme: weka.classifiers.rules.M5Rules -M 4.0 -num-decimal-places 4
- Relation: bolts
- Instances: 40
- Attributes: 8
- RUN
- SPEED1
- TOTAL
- SPEED2
- NUMBER2
- SENS
- TIME
- T20BOLT
- Test mode: 10-fold cross-validation
- 
- === Classifier model (full training set) ===
-

- M5 pruned model rules
- (using smoothed linear models) :
- Number of Rules : 4
- 
- Rule: 1
- IF
- TIME <= 32.19
- TOTAL > 15
- THEN
- 
- T20BOLT =
- -0.8868 \* TOTAL
- - 0.4179 \* NUMBER2
- + 0.914 \* TIME
- + 19.7865 [16/4.785%]
- 
- Rule: 2
- IF
- TIME <= 23.225
- THEN
- 
- T20BOLT =
- 1.162 \* TOTAL
- + 0.8889 \* TIME
- + 7.1663 [12/0%]
- 
- Rule: 3
- IF
- RUN <= 28.5
- THEN
- 
- T20BOLT =
- 0.3162 \* RUN
- + 0.2518 \* TOTAL
- - 0.0169 \* TIME
- + 63.1843 [8/23.535%]
- 
- Rule: 4
- 
- T20BOLT =
- 1.2335 \* TOTAL
- + 50.815 [4/30.472%]
- 
-



- 
- Time taken to build model: 0.03 seconds
- 
- === Cross-validation ===
- === Summary ===
- 
- Correlation coefficient                      0.9104
- Mean absolute error                        5.9185
- Root mean squared error                    11.3655
- Relative absolute error                    23.9433 %
- Root relative squared error                40.3201 %
- Total Number of Instances                40
- 
- The dataset describes the time needed by a machine to produce and count 20 bolts. (More details can be found in the file containing the dataset.) Analyze the data.
- What adjustments have the greatest effect on the time to count 20 bolts?
  - I would say according to the above that the decision stump with 86.77 correctly classified instances and the linear regression with 91.87% correlation coefficient because they had the highest values of the analyzed set. I am not comparing PART to the LR because these are different measures.
- According to each classifier, how would you adjust the machine to get the shortest time to count 20 bolts?
  - Decision Stump: Adjust the speed 1 attribute to less than five.
  - Decision Table: Change Speed 1 to a value of 4 and the attribute total to 20.
  - Linear Regression: Make speed 1 the low value and adjust sens to be the high value
  - M5: Based on the two decision rules for speed I would increase the value of speed 1 and lower the values of both sens and total for both those greater and lower than five.

Rating: 9/10

- I believe this rating is fair because some of the questions were hard to grasp with only the reading and the videos, however, the videos were extremely helpful.