

# DATA MINING ASSIGNMENT 2

## Decision tree classification

**TASK 1:** One type of model that you can create is a decision tree. Train a decision tree using the complete dataset as the training data. Report the model obtained after training.

### PROCEDURE:

- 1) Open Weka GUI Chooser.
- 2) Select EXPLORER present in Applications.
- 3) Select Preprocess Tab.
- 4) Go to OPEN file and browse the file that is already stored in the system "credit-g.arff".
- 5) Go to Classify tab.
- 6) Here the c4.5 algorithm has been chosen which is entitled as j48 in Java and can be selected by clicking the button choose and select tree j48.
- 7) Select Test options "Use training set".
- 8) Select class attribute.
- 9) Click Start.
- 10) Now we can see the output details in the Classifier output.
- 11) Right click on the result list and select "visualize tree" option.

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) class

Start Stop

Result list (right-click for options)

15:35:34 - trees.J48

Classifier output

Time taken to test model on training data: 0.08 seconds

=== Summary ===

Correctly Classified Instances	855	85.5	%
Incorrectly Classified Instances	145	14.5	%
Kappa statistic	0.6251		
Mean absolute error	0.2312		
Root mean squared error	0.34		
Relative absolute error	55.0377	%	
Root relative squared error	74.2015	%	
Total Number of Instances	1000		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.956	0.380	0.854	0.956	0.902	0.640	0.857	0.905	good
	0.620	0.044	0.857	0.620	0.720	0.640	0.857	0.783	bad
Weighted Avg.	0.855	0.279	0.855	0.855	0.847	0.640	0.857	0.869	

=== Confusion Matrix ===

a b <-- classified as

669	31	a = good
114	186	b = bad

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

8

Percentage split

%

80

More options...

(Nom) class

Start

Stop

Result list (right-click for options)

15:35:34 - trees.J48

15:37:50 - trees.J48

15:38:05 - trees.J48

15:38:16 - trees.J48

15:38:26 - trees.J48

15:41:37 - trees.J48

15:41:49 - trees.J48

Classifier output

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	726	72.6	%
Incorrectly Classified Instances	274	27.4	%
Kappa statistic	0.2996		
Mean absolute error	0.3319		
Root mean squared error	0.4692		
Relative absolute error	78.9988	%	
Root relative squared error	102.3972	%	
Total Number of Instances	1000		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.856	0.577	0.776	0.856	0.814	0.305	0.663	0.765	good
	0.423	0.144	0.557	0.423	0.481	0.305	0.663	0.469	bad
Weighted Avg.	0.726	0.447	0.710	0.726	0.714	0.305	0.663	0.676	

=== Confusion Matrix ===

a

b

<-- classified as

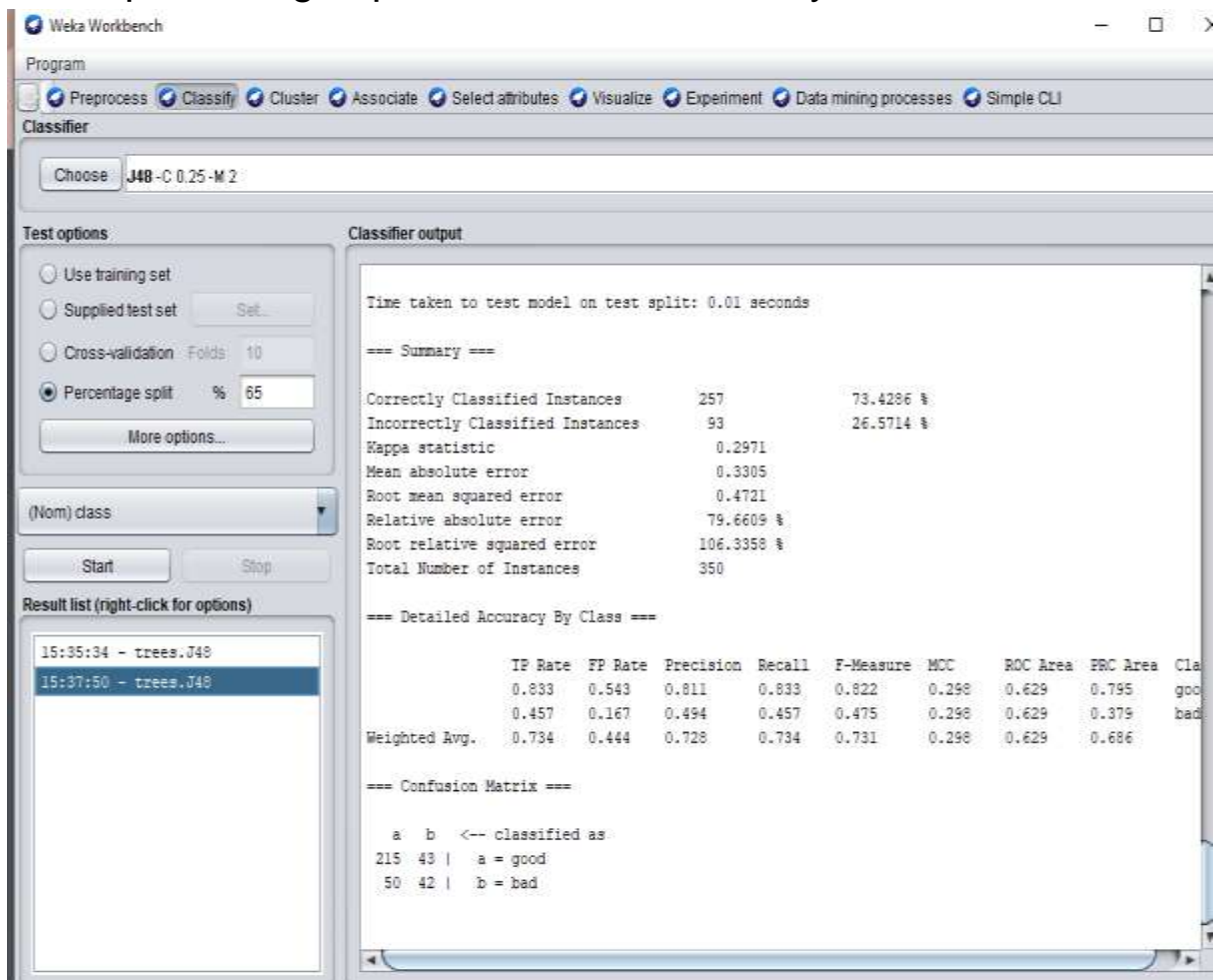
599 101 | a = good

173 127 | b = bad

[illegible]

**TASK 2:** Train a Decision Tree using percentage split and report your results. Increase percentage split by 5% upto 80% starting from 65% and check at which percentage split we are getting the best accuracy.

1. When percentage split is 65%, the accuracy is 73.4286%



The screenshot shows the Weka Workbench interface. The 'Classifier' tab is active, and the 'J48-C 0.25-M 2' classifier is selected. The 'Test options' section is configured with 'Percentage split' at 65%. The 'Classifier output' pane displays the following results:

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	257	73.4286 %
Incorrectly Classified Instances	93	26.5714 %
Kappa statistic	0.2971	
Mean absolute error	0.3305	
Root mean squared error	0.4721	
Relative absolute error	79.6609 %	
Root relative squared error	106.3358 %	
Total Number of Instances	350	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
good	0.833	0.543	0.811	0.833	0.822	0.298	0.629	0.795	good
bad	0.457	0.167	0.494	0.457	0.475	0.298	0.629	0.379	bad
Weighted Avg.	0.734	0.444	0.728	0.734	0.731	0.298	0.629	0.686	

=== Confusion Matrix ===

a \ b	Classified as	
	a = good	b = bad
215 43   a = good	215	43
50 42   b = bad	50	42

2. When percentage split is 70%, the accuracy is 73.6667%

The screenshot shows the Weka Workbench interface. The 'Classifier' tab is active, and the 'J48 - C 0.25 - M 2' classifier is selected. The 'Test options' section on the left has 'Percentage split' set to 70%. The 'Classifier output' section on the right displays the results of the test.

**Test options:**

- ☐ Use training set
- ☐ Supplied test set
- ☐ Cross-validation / folds: 10
- ☒ Percentage split %: 70

**Classifier output:**

Time taken to test model on test split: 0 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	221	73.6667 %
Incorrectly Classified Instances	79	26.3333 %
Kappa statistic	0.2579	
Mean absolute error	0.323	
Root mean squared error	0.47	
Relative absolute error	70.2126 %	
Root relative squared error	105.9524 %	
Total Number of Instances	300	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
good	0.869	0.433	0.793	0.869	0.829	0.263	0.636	0.794	good
bad	0.367	0.131	0.500	0.367	0.423	0.263	0.636	0.424	bad
Weighted Avg.	0.737	0.501	0.716	0.737	0.722	0.263	0.636	0.696	

=== Confusion Matrix ===

a \ b	good	bad
good	192	29
bad	50	29

3. When percentage split is 75%, the accuracy is 76%

The screenshot shows the Weka Workbench interface. The 'Classifier' tab is selected, and the 'J48 -C 0.25 -M 2' classifier is chosen. The 'Test options' section on the left shows 'Percentage split' set to 75%. The 'Classifier output' pane on the right displays the results of the test.

**Test options**

- ☐ Use training set
- ☐ Supplied test set
- ☐ Cross-validation Folds: 10
- ☒ Percentage split %: 75

**Classifier output**

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	190	76	%
Incorrectly Classified Instances	60	24	%
Kappa statistic	0.3232		
Mean absolute error	0.3073		
Root mean squared error	0.4365		
Relative absolute error	74.6884 %		
Root relative squared error	98.4212 %		
Total Number of Instances	250		

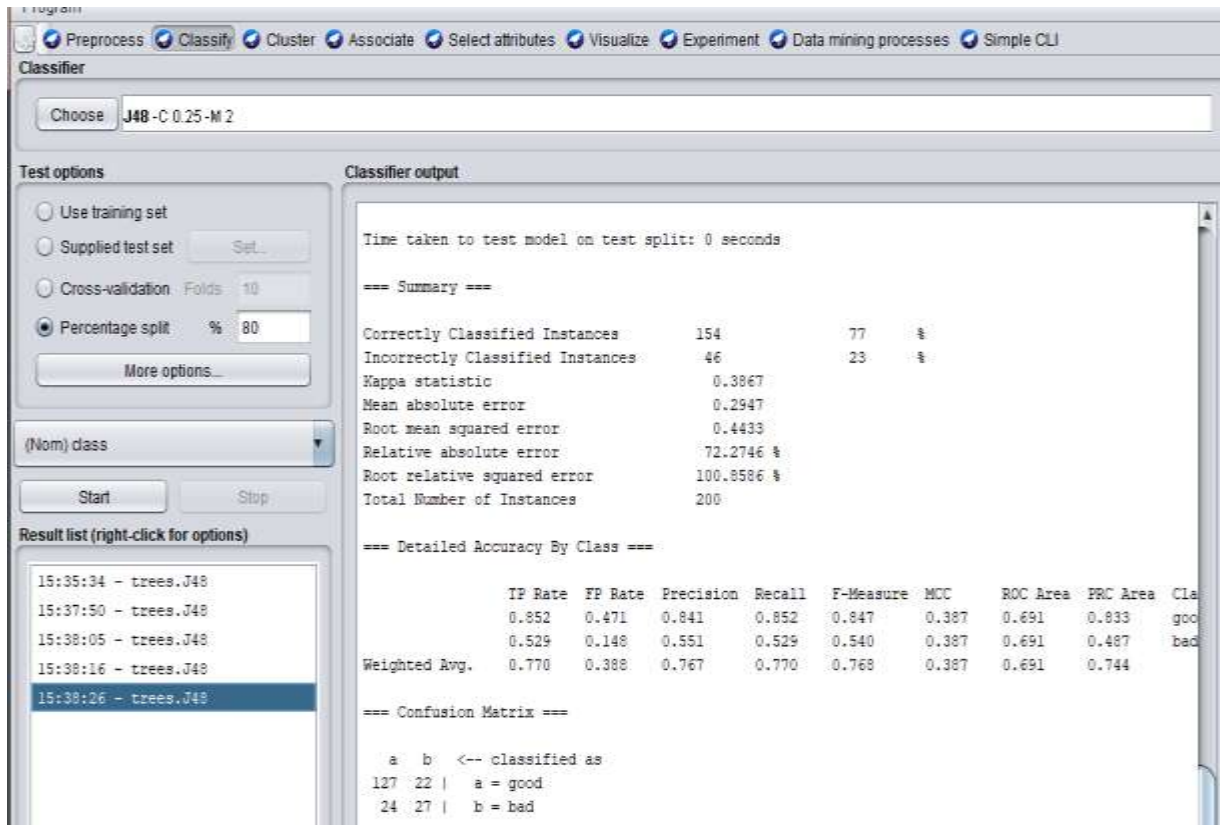
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.886	0.591	0.807	0.886	0.845	0.330	0.673	0.820	good
	0.409	0.114	0.563	0.409	0.474	0.330	0.673	0.478	bad
Weighted Avg.	0.760	0.465	0.742	0.760	0.747	0.330	0.673	0.730	

=== Confusion Matrix ===

a	b	<-- classified as	
163	21	a	= good
39	27	b	= bad

4. When percentage split is 80%, the accuracy is 77%



The screenshot shows the Weka Classifier window. The 'Classifier' dropdown is set to 'J48 -C 0.25-M 2'. Under 'Test options', 'Percentage split' is selected with a value of 80%. The 'Classifier output' pane displays the following results:

Time taken to test model on test split: 0 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	154	77 %
Incorrectly Classified Instances	46	23 %
Nappa statistic	0.3867	
Mean absolute error	0.2947	
Root mean squared error	0.4433	
Relative absolute error	72.2746 %	
Root relative squared error	100.8586 %	
Total Number of Instances	200	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
good	0.852	0.471	0.841	0.852	0.847	0.387	0.691	0.833	good
bad	0.529	0.148	0.551	0.529	0.540	0.387	0.691	0.487	bad
Weighted Avg.	0.770	0.388	0.767	0.770	0.768	0.387	0.691	0.744	

=== Confusion Matrix ===

a \ b	good	bad
good	127	22
bad	24	27

CONCLUSION: When the percentage split is 80%, the accuracy is high(77%).



## TASK 3: Train a Decision Tree using cross validation and report your results.

1. When cross validation folds : 10, accuracy is 70.5%

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

☒ Cross-validation: Folds 10

☐ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options)

15:35:34 - trees.J48

15:37:50 - trees.J48

15:38:05 - trees.J48

15:38:16 - trees.J48

15:38:26 - trees.J48

15:41:57 - trees.J48

Classifier output

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	705	70.5 %
Incorrectly Classified Instances	295	29.5 %
Kappa statistic	0.2467	
Mean absolute error	0.3467	
Root mean squared error	0.4796	
Relative absolute error	82.5233 %	
Root relative squared error	104.6565 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PBC Area	Class
	0.840	0.610	0.763	0.840	0.799	0.251	0.639	0.746	good
	0.390	0.160	0.511	0.390	0.442	0.251	0.635	0.449	bad
Weighted Avg.	0.705	0.475	0.687	0.705	0.652	0.251	0.639	0.657	

=== Confusion Matrix ===

a b <-- classified as

565 112 | a = good

183 117 | b = bad

## 2. When cross validation folds : 8, accuracy is 72.6%

The screenshot shows the Weka Workbench interface. The 'Classifier' tab is selected, and the 'J48 -C 0.25 -M 2' classifier is chosen. The 'Test options' section shows 'Cross-validation' selected with 'Folds' set to 8. The 'Classifier output' pane displays the results of the stratified cross-validation.

**Test options**

- ☐ Use training set
- ☐ Supplied test set (Set...)
- ☒ Cross-validation Folds: 8
- ☐ Percentage split %: 80
- More options...

(Nom) class

Start Stop

**Result list (right-click for options)**

- 15:35:34 - trees.J48
- 15:37:50 - trees.J48
- 15:38:05 - trees.J48
- 15:38:16 - trees.J48
- 15:38:26 - trees.J48
- 15:41:37 - trees.J48
- 15:41:49 - trees.J48

**Classifier output**

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	726	72.6 %
Incorrectly Classified Instances	274	27.4 %
Kappa statistic	0.2996	
Mean absolute error	0.3319	
Root mean squared error	0.4692	
Relative absolute error	78.9988 %	
Root relative squared error	102.3972 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
good	0.856	0.577	0.776	0.856	0.814	0.305	0.663	0.765	good
bad	0.423	0.144	0.557	0.423	0.481	0.305	0.663	0.469	bad
Weighted Avg.	0.726	0.447	0.710	0.726	0.714	0.305	0.663	0.676	

=== Confusion Matrix ===

a	b	-- classified as
599	101	a = good
173	127	b = bad

3. When cross validation folds : 6, accuracy is 74.1%

The screenshot shows the Weka Workbench interface. The 'Classify' tab is selected. Under 'Classifier', 'J48 - C 0.25-M 2' is chosen. In the 'Test options' section, 'Cross-validation' is selected with 'Folds' set to 6. The 'Classifier output' pane displays the following results:

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	741	74.1 %
Incorrectly Classified Instances	259	25.9 %
Kappa statistic	0.3453	
Mean absolute error	0.3239	
Root mean squared error	0.4479	
Relative absolute error	77.0782 %	
Root relative squared error	97.737 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
good	0.859	0.533	0.790	0.859	0.823	0.349	0.680	0.778	good
bad	0.467	0.141	0.586	0.467	0.519	0.349	0.680	0.497	bad
Weighted Avg.	0.741	0.416	0.729	0.741	0.732	0.349	0.680	0.694	

=== Confusion Matrix ===

a	b	<-- classified as
601	99	a = good
160	140	b = bad

The 'Result list' on the left shows a series of entries for 'trees.J48' with timestamps, with the entry '15:43:29 - trees.J48' selected.

**CONCLUSION:** The accuracy is high(74.1%) when cross validation folds: 6