

DATA MINING LAB -3

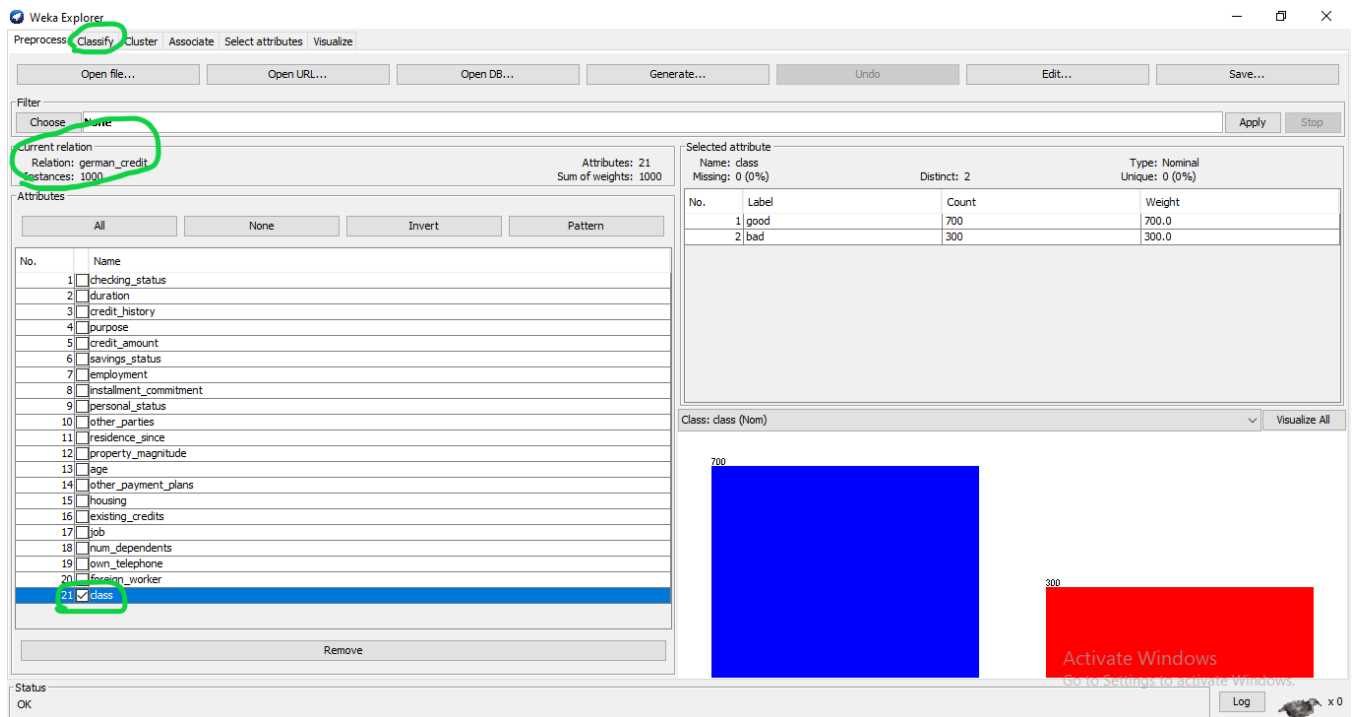
Q.) TO DERIVE A DECISION TREE FROM A GIVEN DATASET:

STEP 1: open Weka.

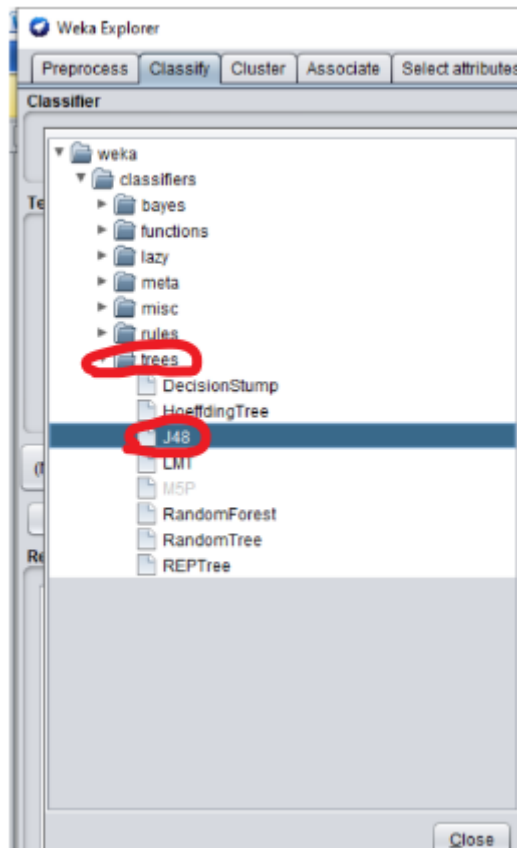
STEP 2: open explorer -> choose required dataset.

➤ Here choose dataset is “German Credit Dataset.”

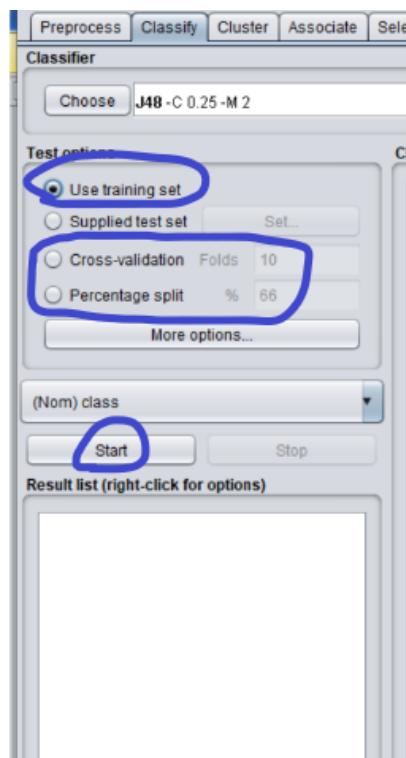
STEP 3: select class attribute from the given dataset and On the top click the classify tab.



STEP 4: from classify select tree in that select J48.



STEP 5: Choose the training set from the test option and then start.



STEP 6:

Classifier

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

Test options

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

(Nom) class

Start Stop

Result list (right-click for options)

22:33:08 - trees.J48

Classifier output

```

| | | purpose = retraining: good (0.0)
| | | purpose = business
| | | | housing = rent
| | | | | existing_credits <= 1: good (2.0)
| | | | | existing_credits > 1: bad (2.0)
| | | | | housing = own: good (6.0)
| | | | | housing = for free: bad (1.0)
| | | | purpose = other: good (1.0)
| | | savings_status = 500<=X<1000: good (11.0/3.0)
| | | savings_status = >=1000: good (13.0/3.0)
| | | savings_status = no known savings: good (41.0/5.0)
| | credit_amount > 9857: bad (20.0/3.0)
| checking_status = >=200: good (63.0/14.0)
| checking_status = no checking: good (394.0/46.0)

```

Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.3 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances	855	85.5 %
--------------------------------	-----	--------

Classifier

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

Test options

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

(Nom) class

Start Stop

Result list (right-click for options)

22:33:08 - trees.J48

Classifier output

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 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
Weighted Avg.	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

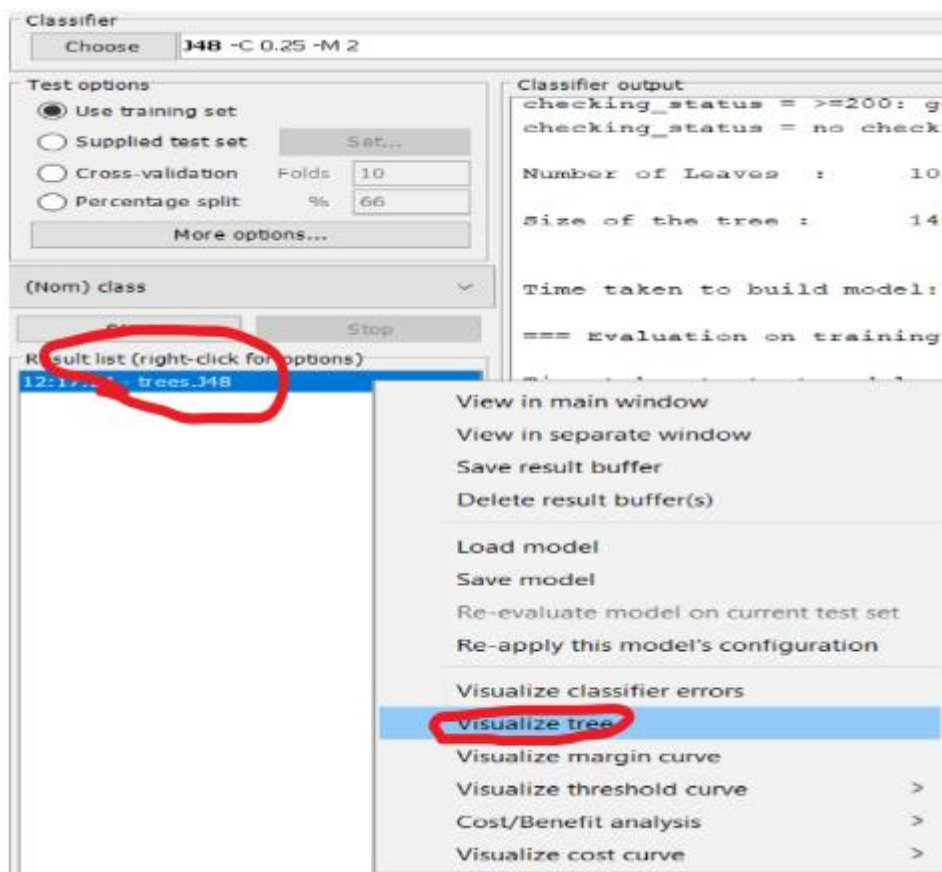
=== Confusion Matrix ===

a	b	<-- classified as
669	31	a = good
114	186	b = bad

➤ Here the number of leaves: 103.

- Size of tree: 140
- Time taken to build a model: 0.3sec
- We get “Confusion Matrix” as the output.
- When **percentage split = 66%** the accuracy will be 85.5%.

STEP 7: To get a tree. Right click on the option in the result list and select the visualize tree option.



STEP 8: the output of the tree will be like this:



Activate Windows
Go to Settings to activate Windows.

- When Percentage split = 70% the accuracy will be 73.6%.
- Time to build a model = 0.04 sec.

Weka Explorer

Preprocess Classifier Cluster Associate Select attributes Visualize

Classifier

Choose 348 -C 0.25 -M 2

Test options

- ☐ Use training set
- ☐ Supplier test set
- ☐ Cross-validation Folds 10
- ☒ Percentage split % 70

More options...

(Nom) class

Start Stop

Result list (right-click for options)

15:01:02 - trees.148

15:12:44 - trees.148

Classifier output

Time taken to build model: 0.04 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

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	78.2126 %	
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
Weighted Avg.	0.869	0.633	0.793	0.869	0.829	0.263	0.636	0.794	good
	0.367	0.131	0.500	0.367	0.423	0.263	0.636	0.424	bad

=== Confusion Matrix ===

	a	b	<-- classified as
192 29	a = good		
50 29	b = bad		

Status OK

Activate Windows
Go to Settings to activate Windows.

Log x0

- When **Percentage split = 80%** the accuracy will be 77%
- Time to build a model = 0.02 sec

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. In the 'Test options' section, 'Percentage split' is selected with a value of 80. The 'Start' button is also highlighted. The 'Classifier output' pane displays the following information:

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	154	77 %
Incorrectly Classified Instances	46	23 %
Kappa 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
0.852	0.471	0.841	0.852	0.847	0.387	0.691	0.833	good	
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	<-- classified as
127	22	a = good
24	27	b = bad

When cross-validation folds is increased by 20 and percentage split = 80%.

- Time taken to build a model = 0.02 sec
- Accuracy will be = 69.8%.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. In the 'Test options' section, 'Cross-validation' is selected with 'Folds' set to 20. The 'Start' button is also highlighted. The 'Classifier output' pane displays the following information:

NUMBER OF LEAVES : 103

Size of the tree : 140

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	698	69.8 %
Incorrectly Classified Instances	302	30.2 %
Kappa statistic	0.2264	
Mean absolute error	0.3571	
Root mean squared error	0.4883	
Relative absolute error	85.0006 %	
Root relative squared error	106.5538 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.837	0.627	0.757	0.837	0.795	0.231	0.621	0.732	good	
0.373	0.163	0.496	0.373	0.426	0.231	0.621	0.415	bad	
Weighted Avg.	0.698	0.488	0.679	0.698	0.684	0.231	0.621	0.637	

=== Confusion Matrix ===

a	b	<-- classified as
586	114	a = good
188	112	b = bad

When cross validation folds = 15 and the percentage split = 80%.

- Time to build a model = 0.01 sec.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Test options' section on the left has 'Cross-validation' selected with 'Folds' set to 15 and 'Percentage split' set to 80%. The 'Classifier output' section on the right displays the results for a 'J48 - C 0.25 - M 2' model. The output includes the number of leaves (103), the size of the tree (140), and the time taken to build the model (0.01 seconds). A summary of stratified cross-validation results is shown, including the number of correctly and incorrectly classified instances, kappa statistic, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. A detailed accuracy by class table is also provided, showing TP Rate, FP Rate, Precision, Recall, F-Measure, MOC, ROC Area, and FRC Area for both 'good' and 'bad' classes. Finally, a confusion matrix is displayed, showing the relationship between actual and predicted classes.

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds 15
- ☐ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 15:01:02 - trees.J48
- 15:12:44 - trees.J48
- 15:13:04 - trees.J48
- 15:14:06 - trees.J48
- 15:15:14 - trees.J48
- 15:16:03 - trees.J48

Classifier output

Number of Leaves : 103

Size of the tree : 140

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	719	71.9 %
Incorrectly Classified Instances	281	28.1 %
Kappa statistic	0.2659	
Mean absolute error	0.3348	
Root mean squared error	0.4689	
Relative absolute error	79.6731 %	
Root relative squared error	102.3148 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

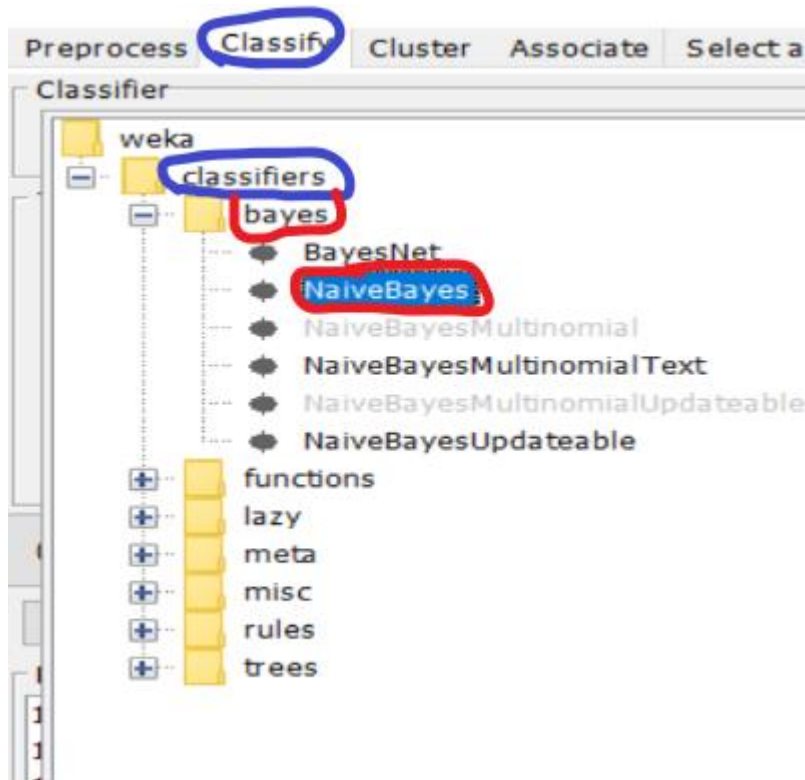
	TP Rate	FP Rate	Precision	Recall	F-Measure	MOC	ROC Area	FRC Area	Class
0.866	0.623	0.764	0.866	0.812	0.274	0.647	0.752	good	
0.377	0.134	0.546	0.377	0.446	0.274	0.647	0.476	bad	
Weighted Avg.	0.719	0.477	0.699	0.719	0.702	0.274	0.647	0.669	

=== Confusion Matrix ===

a	b	<-- classified as
606	94	a = good
187	113	b = bad

NAÏVE BAYES CLASSIFIER:

- To select naïve baye click choose option in the classify tab, select bayes in that select naïve bayes.



When **percentage split = 66%**, cross-validation folds = 10 then

- Time to build a model = 0 sec
- Accuracy = 76.4%.

The screenshot shows the Weka Classifier window with the 'Test options' and 'Classifier output' sections. The 'Percentage split' option is selected with a value of 66%. The 'Classifier output' section shows the results of the test, including a summary of classification performance.

Test options:

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

Classifier output:

```
foreign_worker
yes          668.0    297.0
no           34.0     5.0
[total]      702.0    302.0
```

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances	260	76.4706 %
Incorrectly Classified Instances	80	23.5294 %
Kappa statistic	0.3824	
Mean absolute error	0.2819	
Root mean squared error	0.4005	
Relative absolute error	67.9798 %	
Root relative squared error	90.114 %	
Total Number of Instances	340	

When **percentage split= 80%**, cross-validation = 10.

- Time taken to build a model = 0sec
- Accuracy = 74.5%.

The screenshot shows the Weka Classifier window with the NaiveBayes classifier selected. The 'Test options' section on the left has 'Percentage split' set to 80% and 'Cross-validation' set to 10 folds. The 'Classifier output' pane on the right displays the results of the classification.

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☐ Cross-validation Folds: 10
- ☒ Percentage split %: 80

Classifier output:

foreign_worker		
yes	668.0	297.0
no	34.0	5.0
[total]	702.0	302.0

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances	149	74.5 %
Incorrectly Classified Instances	51	25.5 %
Kappa statistic	0.3657	
Mean absolute error	0.2879	
Root mean squared error	0.4129	
Relative absolute error	70.6169 %	
Root relative squared error	93.9316 %	
Total Number of Instances	200	

Result list (right-click for options):

- 12:17:24 - trees.J48
- 12:41:21 - trees.J48
- 12:42:28 - trees.J48
- 12:44:14 - trees.J48
- 12:47:47 - bayes.NaiveBayes
- 12:49:46 - bayes.NaiveBayes
- 12:50:58 - bayes.NaiveBayes

Activate Windows
Go to Settings to activate