

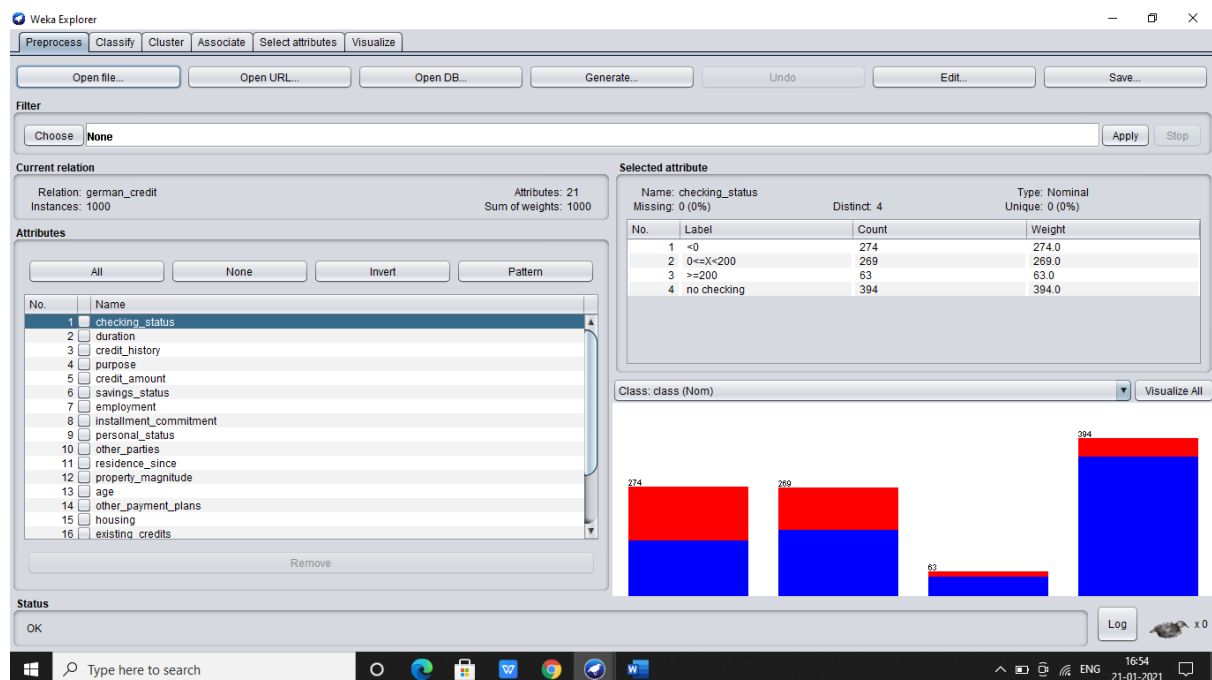
Performing Classification in German Credit Data set

Classification: It is one of the data mining tasks which is used to find a model that distinguishes data classes. There are several techniques to perform classification. Some of them are decision trees, K-Nearest neighbour, neural networks etc.,

Steps to perform classification in weka using decision trees:

Decision Tree algorithm is one of the supervised learning algorithms. It can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules derived from training data.

- Load the german credit dataset into weka platform.



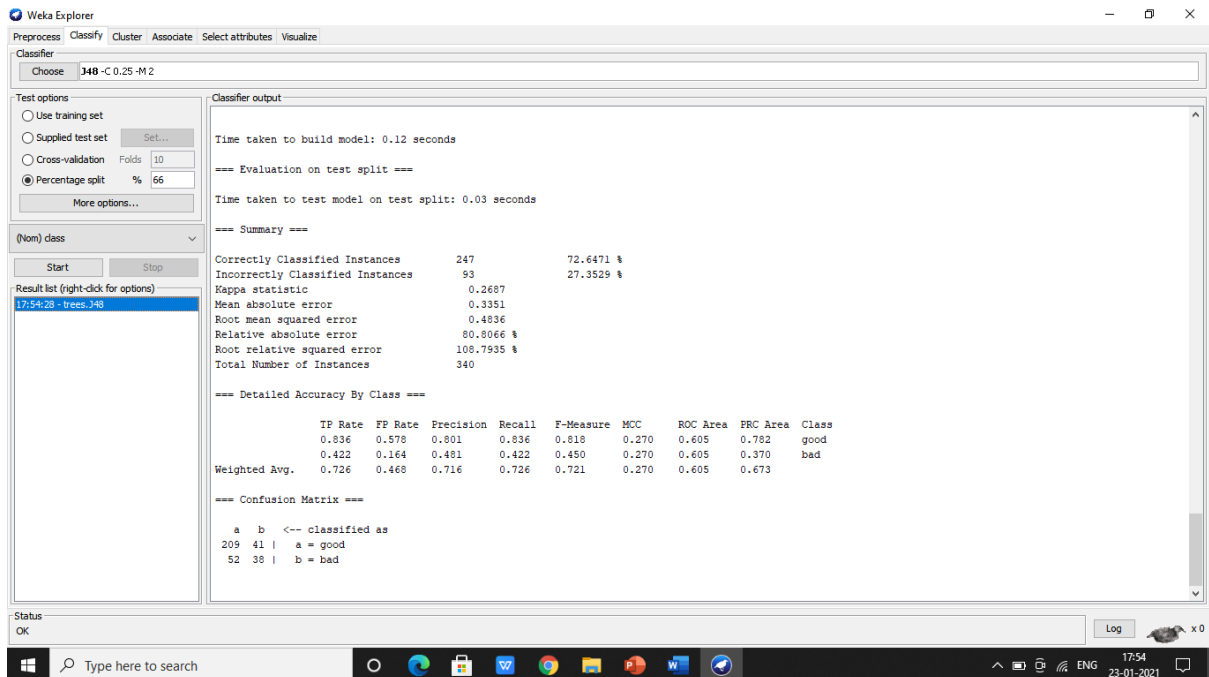
- To classify the dataset opt the classify option and choose the required algorithm. In present case we choose J48 to use decision tree algorithm. Now perform the classification by choose cross validation and percentage split.

Observations By changing percentage splits and cross validation folds:

Percentage split: It is a percentage which describe the percent of data used for training and testing purpose.

Cross Validation: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. The procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

- By classifying with default percentage split i.e 66% and with cross validation folds 10 the accuracy observed is 72.64% .Time taken to build model is 0.12 sec.



The screenshot shows the Weka Explorer interface with the Classifier tab selected. The classifier chosen is J48 -C 0.25 -M 2. The test options are set to Percentage split (66%) and Cross-validation (10 folds). The classifier output displays the following information:

Time taken to build model: 0.12 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances	247	72.6471 %
Incorrectly Classified Instances	93	27.3529 %
Kappa statistic	0.2687	
Mean absolute error	0.3351	
Root mean squared error	0.4836	
Relative absolute error	80.8066 %	
Root relative squared error	108.7935 %	
Total Number of Instances	340	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.836	0.578	0.801	0.836	0.818	0.270	0.605	0.782	good
	0.422	0.164	0.481	0.422	0.450	0.270	0.605	0.370	bad
Weighted Avg.	0.726	0.468	0.716	0.726	0.721	0.270	0.605	0.673	

=== Confusion Matrix ===

```

a  b  <-- classified as
209 41 | a = good
 52 38 | b = bad

```

- When the percentage split is 70% and with cross validation folds 10 the accuracy observed is 73.67% .Time taken to build model is 0.06 sec.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Test options

- ☐ Use training set
- ☐ Supplied test set **Set...**
- ☐ Cross-validation Folds **10**
- ☒ Percentage split % **70**

(Nom) class

Start Stop

Result list (right-click for options)

- 17:54:28 - trees.J48
- 17:57:00 - trees.J48
- 17:57:11 - trees.J48
- 17:57:24 - trees.J48
- 18:27:52 - trees.J48**

Classifier output

Time taken to build model: 0.06 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 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
	0.737	0.501	0.716	0.737	0.722	0.263	0.636	0.696	

=== Confusion Matrix ===

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

Status: OK

Log x 0

- When the percentage split is 75% and with cross validation folds 10 the accuracy observed is 76% .Time taken to build model is 0.04 sec.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Test options

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

(Nom) class

Start Stop

Result list (right-click for options)

- 17:54:28 - trees.J48
- 17:57:00 - trees.J48
- 17:57:11 - trees.J48
- 17:57:24 - trees.J48
- 18:27:52 - trees.J48
- 18:36:24 - trees.J48**

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	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
Weighted Avg.	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
	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	

Status: OK

Log x 0

- When the percentage split is 80% and with cross validation folds 10 the accuracy observed is 77% .Time taken to build model is 0.02 sec.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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)

- 17:54:28 - trees.J48
- 17:57:00 - trees.J48
- 17:57:11 - trees.J48
- 17:57:24 - trees.J48
- 18:27:52 - trees.J48
- 18:36:24 - trees.J48
- 18:40:42 - trees.J48**

Classifier output

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	

Status: OK

Log x 0

18:42 23-01-2021

- Number of leaves : 103
- Size of the tree: 140

- When the percentage split is 66% and with cross validation folds 8 the accuracy observed is 72.6% .

The screenshot shows the Weka Explorer interface. The 'Classifier' dropdown is set to 'J48 -C 0.25 -M 2'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 8. The 'Classifier output' pane displays the following results:

```

Classifier output
Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.07 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 0.4662
Relative absolute error 78.9968 %
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   0.461   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

```

- When the percentage split is 66% and with cross validation folds 6 the accuracy observed is 74.1% .

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

```

Classifier output
Number of Leaves : 103
Size of the tree : 140
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

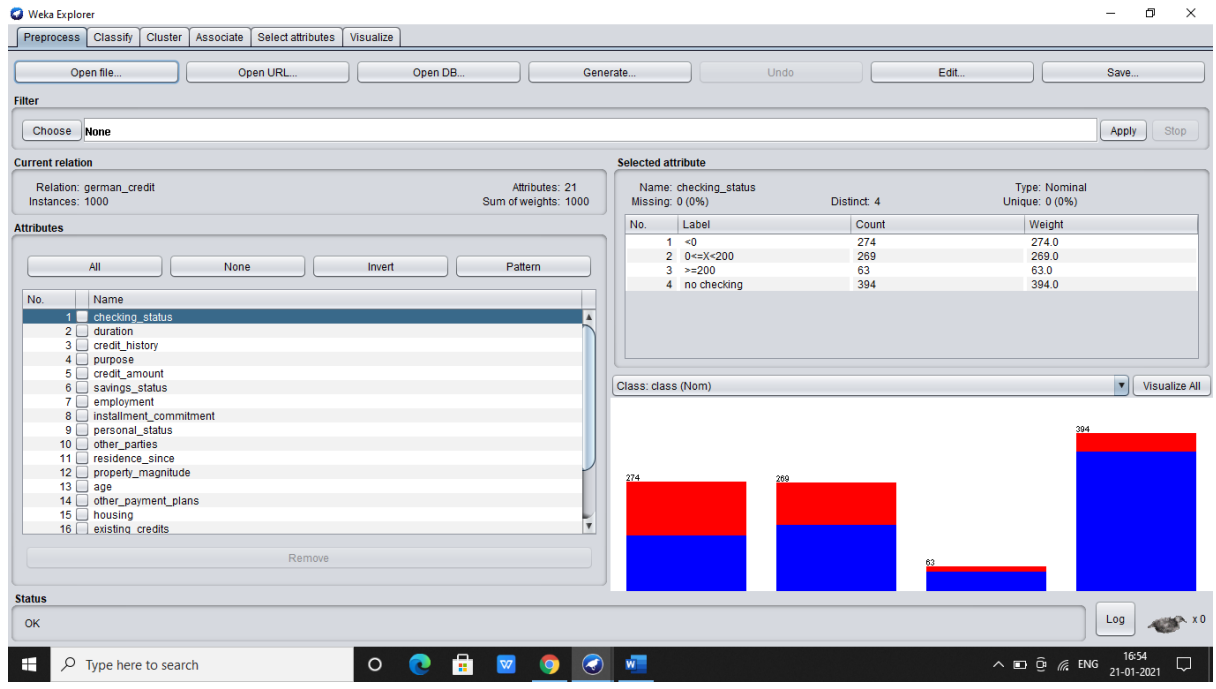
=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.859   0.533   0.790   0.859   0.823   0.349   0.680   0.778   good
0.467   0.141   0.586   0.467   0.519   0.349   0.680   0.497   bad
Weighted Avg.   0.741   0.416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

```

Classification using Naivebayes algorithm:

- Load the german credit dataset into weka platform



- To classify the dataset opt the classify option and choose the required algorithm. In present case we choose Naivebayes. Now perform the classification by choose cross validation and percentage split.
- When the percentage split is 66% and with cross validation folds 10 the accuracy observed is 76.47% .Time taken to build model is 0.01 sec.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

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):

- 17:54:28 - trees.J48
- 17:57:00 - trees.J48
- 17:57:11 - trees.J48
- 17:57:24 - trees.J48
- 18:27:52 - trees.J48
- 18:36:24 - trees.J48
- 18:40:42 - trees.J48
- 18:52:02 - bayes.NaiveBayes
- 19:15:51 - bayes.NaiveBayes**

Classifier output:

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.13 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	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.852	0.478	0.832	0.852	0.842	0.383	0.804	0.921	good	
0.522	0.148	0.560	0.522	0.540	0.383	0.804	0.592	bad	
Weighted Avg.	0.765	0.390	0.760	0.765	0.762	0.383	0.804	0.834	

=== Confusion Matrix ===

a	b	<-- classified as	
213	37	a = good	
43	47	b = bad	

Status: OK

Log

- When the percentage split is 70% and with cross validation folds 10 the accuracy observed is 75.33% .Time taken to build model is 0.01 sec.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options:

- ☐ Use training set
- ☐ Supplied test set **Set...**
- ☐ Cross-validation Folds **10**
- ☒ Percentage split % **70**

More options...

(Nom) class: **Start** **Stop**

Result list (right-click for options):

- 17:54:28 - trees.J48
- 17:57:00 - trees.J48
- 17:57:11 - trees.J48
- 17:57:24 - trees.J48
- 18:27:52 - trees.J48
- 18:36:24 - trees.J48
- 18:40:42 - trees.J48
- 18:52:02 - bayes.NaiveBayes
- 19:15:51 - bayes.NaiveBayes
- 19:21:39 - bayes.NaiveBayes**

Classifier output:

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances	226	75.3333 %
Incorrectly Classified Instances	74	24.6667 %
Kappa statistic	0.3537	
Mean absolute error	0.2851	
Root mean squared error	0.4116	
Relative absolute error	69.0347 %	
Root relative squared error	92.7794 %	
Total Number of Instances	300	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.842	0.494	0.827	0.842	0.834	0.354	0.788	0.916	good	
0.506	0.158	0.533	0.506	0.519	0.354	0.788	0.547	bad	
Weighted Avg.	0.753	0.405	0.749	0.753	0.751	0.354	0.788	0.819	

=== Confusion Matrix ===

a	b	<-- classified as	
186	35	a = good	
39	40	b = bad	

Status: OK

Log

- When the percentage split is 75% and with cross validation folds 10 the accuracy observed is 76.8% .Time taken to build model is 0.02 sec.

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The test options are set to 'Percentage split' at 75% and 'Cross-validation' with 10 folds. The classifier output displays the following summary:

```

=== Summary ===
Correctly Classified Instances      192      76.8 %
Incorrectly Classified Instances    58      23.2 %
Kappa statistic                    0.403
Mean absolute error                 0.2778
Root mean squared error             0.4029
Relative absolute error             67.5042 %
Root relative squared error        90.8443 %
Total Number of Instances         250
  
```

The detailed accuracy by class is as follows:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.842	0.439	0.842	0.842	0.842	0.403	0.806	0.924	good
	0.561	0.158	0.561	0.561	0.561	0.403	0.806	0.567	bad

The confusion matrix is:

```

=== Confusion Matrix ===
 a b  <-- classified as
155 29 | a = good
 29 37 | b = bad
  
```

- When the percentage split is 80% and with cross validation folds 10 the accuracy observed is 74.5% .Time taken to build model is 0.01 sec.

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The test options are set to 'Percentage split' at 80% and 'Cross-validation' with 10 folds. The classifier output displays the following summary:

```

=== 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
  
```

The detailed accuracy by class is as follows:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.799	0.412	0.850	0.799	0.824	0.368	0.796	0.923	good
	0.588	0.201	0.500	0.588	0.541	0.368	0.796	0.539	bad

The confusion matrix is:

```

=== Confusion Matrix ===
 a b  <-- classified as
119 30 | a = good
 21 30 | b = bad
  
```

- When the percentage split is 66% and with cross validation folds 8 the accuracy observed is 75.9% .

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose NaiveBayes

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds 8
- ☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options):

- 11:00:28 - bayes.NaiveBayes
- 11:03:09 - bayes.NaiveBayes

Classifier output:

```
no 34.0 5.0
[total] 702.0 302.0

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 759 75.9 %
Incorrectly Classified Instances 241 24.1 %
Kappa statistic 0.3957
Mean absolute error 0.2936
Root mean squared error 0.4205
Relative absolute error 69.8657 %
Root relative squared error 91.7659 %
Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.866 0.490 0.805 0.866 0.834 0.399 0.789 0.893 good
0.510 0.134 0.619 0.510 0.559 0.399 0.789 0.575 bad
Weighted Avg. 0.759 0.383 0.749 0.759 0.752 0.399 0.789 0.798

=== Confusion Matrix ===
a b <-- classified as
606 94 | a = good
147 153 | b = bad
```

Status: OK

Log

- When the percentage split is 66% and with cross validation folds 6 the accuracy observed is 75.4%.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose NaiveBayes

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds 6
- ☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options):

- 11:00:28 - bayes.NaiveBayes
- 11:03:09 - bayes.NaiveBayes

Classifier output:

```
no 34.0 5.0
[total] 702.0 302.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 754 75.4 %
Incorrectly Classified Instances 246 24.6 %
Kappa statistic 0.3813
Mean absolute error 0.2955
Root mean squared error 0.4222
Relative absolute error 70.3237 %
Root relative squared error 92.1377 %
Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.864 0.503 0.800 0.864 0.831 0.385 0.785 0.890 good
0.497 0.136 0.611 0.497 0.548 0.385 0.785 0.573 bad
Weighted Avg. 0.754 0.383 0.743 0.754 0.746 0.385 0.785 0.795

=== Confusion Matrix ===
a b <-- classified as
605 95 | a = good
151 149 | b = bad
```

Status: OK

Log