

Exploring German Credit Dataset Using Naive Bayes Classifier

1) The class attribute is classified using naive bayes by the following steps :

- Open weka and load the german credit dataset
- Select class attribute and click on the classify button
- Select training set option and then naive bayes classifier
- Click on the start button to get classifier output

The screenshot shows the Weka Classifier window with the Naive Bayes classifier selected. The 'Test options' section on the left shows 'Use training set' selected, 'Percentage split' at 65%, and '(Nom) class' set to 'none'. The 'Classifier output' section on the right displays the following data:

Attribute	Count	Percentage
skilled	445.0	187.0
high qualif/self emp/mgt	98.0	52.0
[total]	704.0	304.0
num_dependents	1.1557	1.1533
mean	0.3626	0.3603
std. dev.	700	300
weight sum	1	1
precision		
own_telephone		
none	410.0	188.0
yes	292.0	114.0
[total]	702.0	302.0
foreign_worker		
yes	668.0	297.0
no	34.0	5.0
[total]	702.0	302.0

Time taken to build model: 0.03 seconds
Time taken to test model on training data: 0.02 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	772	77.2 %
Incorrectly Classified Instances	228	22.8 %
Kappa statistic	0.43	
Mean absolute error	0.3621	
Root mean squared error	0.4077	
Relative absolute error	67.1408 %	
Root relative squared error	88.8752 %	
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.	0.873	0.463	0.815	0.873	0.843	0.433	0.899	0.906	good
	0.537	0.127	0.644	0.537	0.585	0.433	0.809	0.609	bad

=== Confusion Matrix ===

	a	b	←- classified as
611 69	a = good		
139 161	b = bad		

The 'Result list' on the left shows the execution history, with the last entry being '13:35:13 - bayes.NaiveBayes'.

2) Time taken to build the model : 0.03 sec.

Default percentage split : 66%

Default cross validation folds : 10

Accuracy : 77.2%.

3) Percentage split : 70%

Default cross validation folds : 10

Accuracy : 75.33

The screenshot shows the Orange3 software interface with the NaiveBayes classifier selected. The 'Test options' section on the left has 'Percentage split' set to 70% and 'Cross-validation folds' set to 10. The 'Classifier output' section on the right displays the following data:

Attribute	Value
skilled	445.0
high qualif/self emp/agent	98.0
[total]	704.0

Attribute	Value
num_dependents	1.1557
mean	0.3626
std. dev.	0.3603
weight sum	700
precision	1

Attribute	Value
own_telephone	410.0
no	292.0
yes	118.0
[total]	702.0

Attribute	Value
foreign_worker	668.0
yes	34.0
no	5.0
[total]	702.0

Time taken to build model: 0.01 seconds
Time taken to test model on test split: 0 seconds

==== Summary ====

Metric	Value
Correctly Classified Instances	226
Incorrectly Classified Instances	74
Kappa statistic	0.2537
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	0.864	0.354	0.788	0.815	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
196	25	a = good
99	40	b = bad

4) Percentage split : 75%

Default cross validation folds : 10

Accuracy : 76.8%

The screenshot shows the Orange3 software interface with the NaiveBayes classifier selected. The 'Test options' section on the left has 'Percentage split' set to 75% and 'Cross-validation folds' set to 10. The 'Classifier output' section on the right displays the following data:

Attribute	Value
skilled	445.0
high qualif/self emp/agent	98.0
[total]	704.0

Attribute	Value
num_dependents	1.1557
mean	0.3626
std. dev.	0.3603
weight sum	700
precision	1

Attribute	Value
own_telephone	410.0
no	292.0
yes	118.0
[total]	702.0

Attribute	Value
foreign_worker	668.0
yes	34.0
no	5.0
[total]	702.0

Time taken to build model: 0.01 seconds
Time taken to test model on test split: 0.01 seconds

==== Summary ====

Metric	Value
Correctly Classified Instances	192
Incorrectly Classified Instances	58
Kappa statistic	0.4093
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

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
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	
Weighted Avg.	0.768	0.365	0.768	0.768	0.768	0.403	0.806	0.830	

==== Confusion Matrix ====

a	b	classified as
150	29	a = good
29	37	b = bad

5) Percentage split : 85%

Default cross validation folds : 10

Accuracy : 74.66%

The screenshot shows the Orange3 software interface with the 'Classify' tab selected. The 'NaiveBayes' classifier is chosen. In the 'Test options' section, 'Percentage split' is set to 85% and 'Cross-validation' is set to 10 folds. The '(Nom) class' dropdown is set to 'Start'. The 'Result list' on the left shows a list of recent runs, with '13:22:27 - NaiveBayes' selected. The 'Classifier output' pane on the right displays the following data:

Attribute	Value
skilled	445.0 187.0
high qualif/self emp/mgt	95.0 52.0
[total]	704.0 304.0
num_dependents	1.1557 1.1503
mean	0.3626 0.3603
std. dev	700 300
weight sum	1 1
precision	
own_telephone	410.0 188.0
no	292.0 114.0
yes	702.0 302.0
foreign_worker	668.0 297.0
yes	34.0 5.0
no	702.0 302.0

Time taken to build model: 0.02 seconds
Time taken to test model on test split: 0.02 seconds

==== Evaluation on test split ====

==== Summary ====

Metric	Value
Correctly Classified Instances	112
Incorrectly Classified Instances	38
Kappa statistic	0.3751
Mean absolute error	0.287
Root mean squared error	0.4129
Relative absolute error	71.0916 %
Root relative squared error	94.7676 %
Total Number of Instances	150

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.759	0.376	0.864	0.788	0.824	0.380	0.795	0.504	good
0.622	0.212	0.489	0.622	0.548	0.380	0.796	0.558	bad
Weighted Avg.	0.747	0.337	0.772	0.747	0.756	0.380	0.796	

==== Confusion Matrix ====

a \ b	classified as
89 24	a = good
14 23	b = bad

Status: OK

6) After trying out multiple combinations of percentage split and cross validation it is observed that the best accuracy i.e 77.2%.