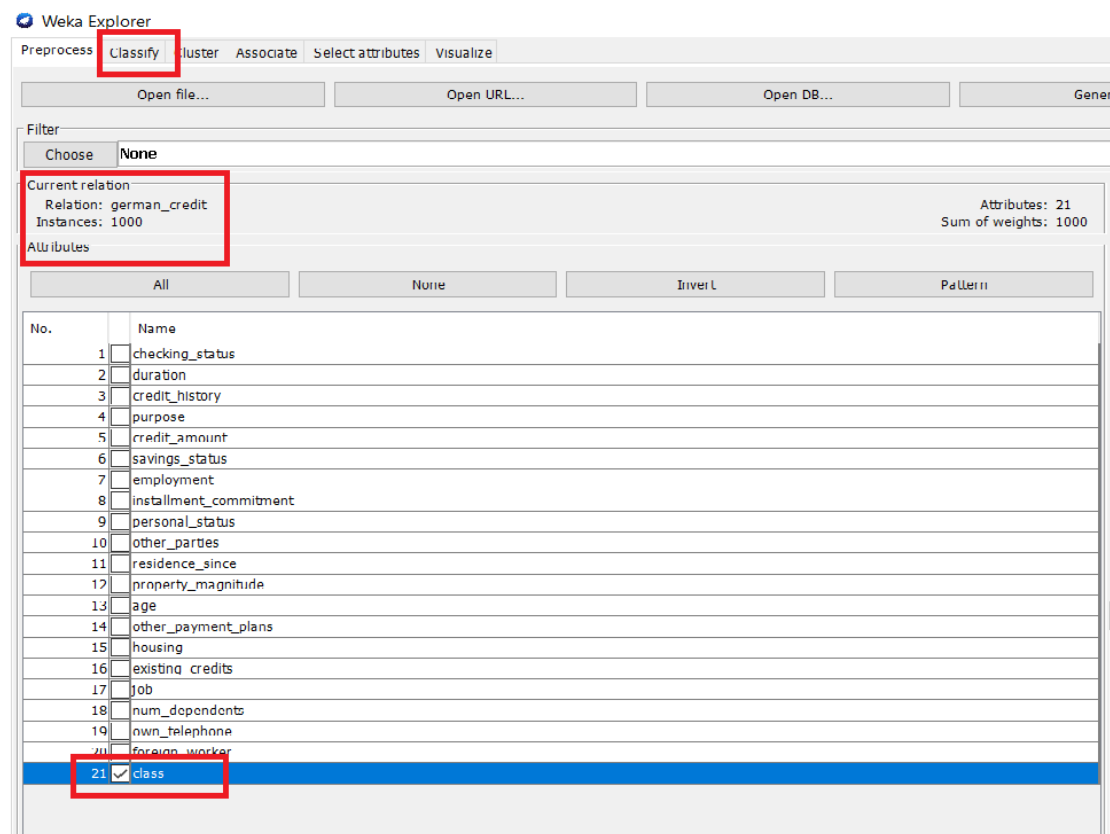


# Data Mining Assignment – II

## German Credit Dataset:

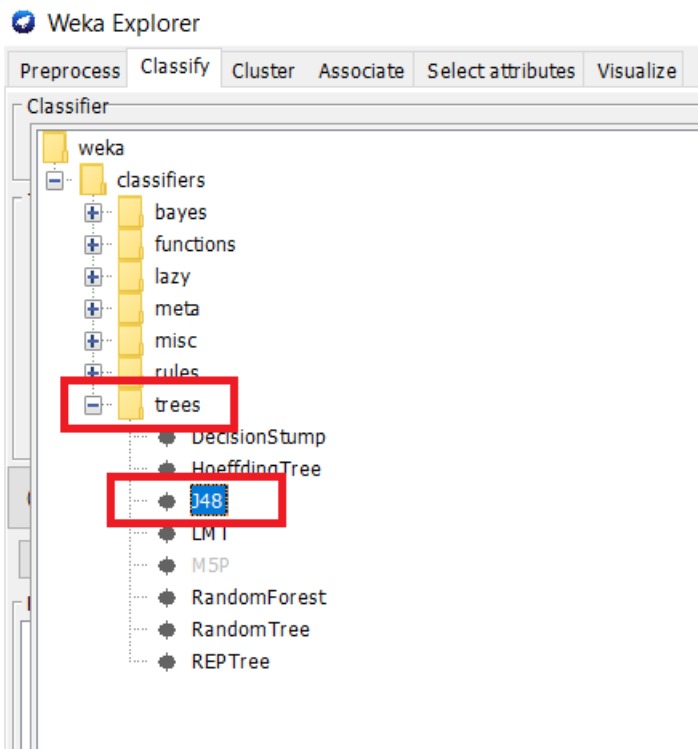
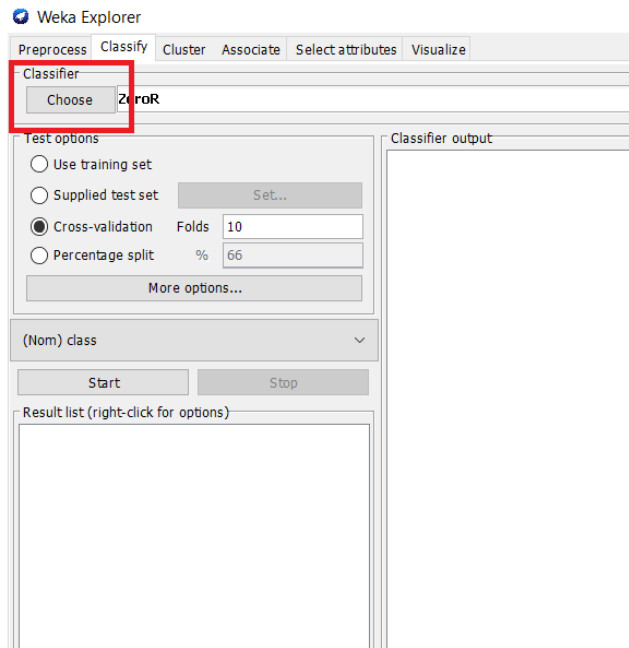
### Steps:

Open german credit dataset in weka. Select class attribute and then click classify.



### Decision Tree:

Click choose option in the classify, select trees and then select J48, as shown below.



In test options select use training set and the start.

The number of leaves in tree are 103 and the size of the tree is 140.

Accuracy is 85.5% when percentage split is 66%.

# 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 % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

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 % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

11:03:51 - trees.J48

Classifier output

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: 0.16 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 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	

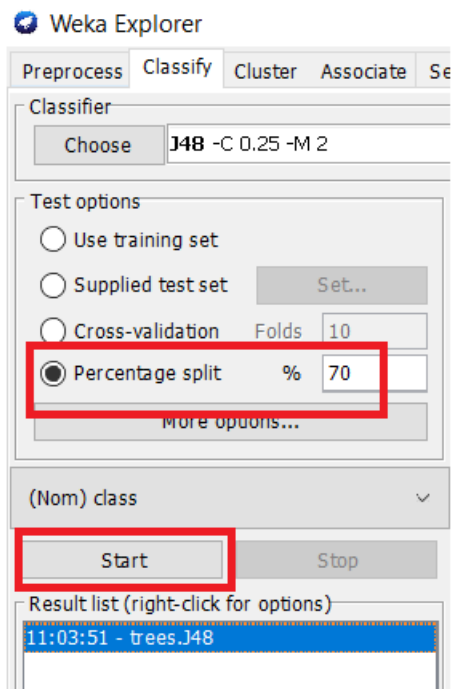
=== Detailed Accuracy By Class ===

	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

Accuracy is 73.66% when percentage split is 70%.



```
Classifier output
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: 0.03 seconds

=== Evaluation on test split ===

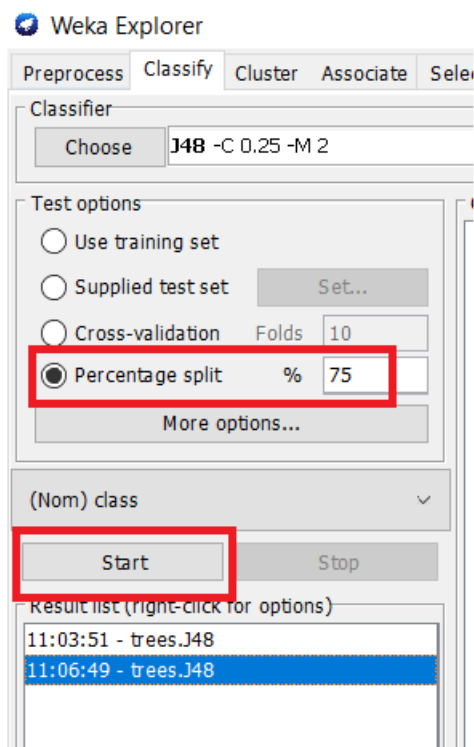
Time taken to test model on test split: 0 seconds

=== 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
               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
Weighted Avg.   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
```

Accuracy is 76% when percentage split is 75%.



Number of Leaves : 103

Size of the tree : 140

Time taken to build model: 0.01 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
	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
```

Accuracy is 77% when percentage split is 80%.

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 % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options)

11:03:51 - trees.J48

11:06:49 - trees.J48

11:19:46 - trees.J48

Classifier output

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: 0.03 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

Accuracy is 72.6% when cross validation is 8.

**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 **8**
- ☐ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options):

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48

Classifier output:

```
| | 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: 0.02 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	

Accuracy is 74.1% when cross validation is 6.

**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 **6**
- ☐ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options):

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48

Classifier output:

```
| | 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: 0.02 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 ===

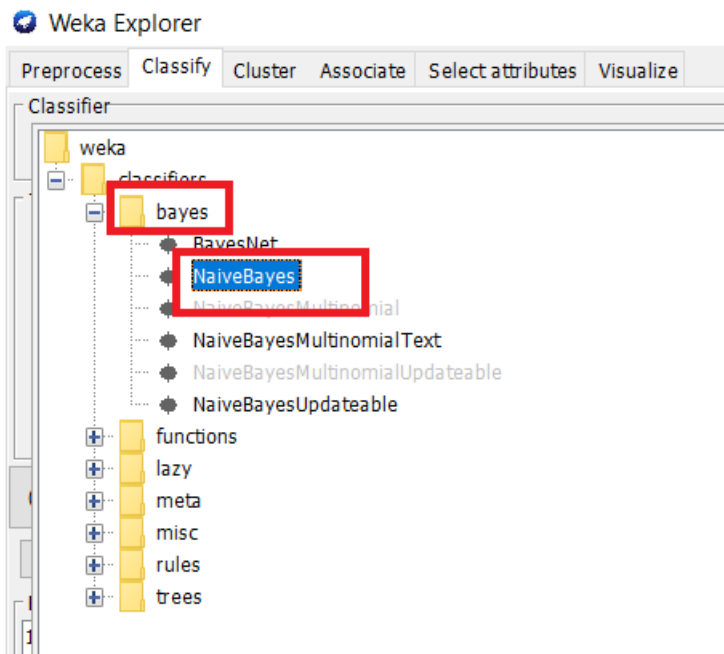
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.859	0.533	0.790	0.859	0.823	0.349	0.680	0.778	good
	0.467	0.141	0.586	0.467	0.519	0.349	0.680	0.497	bad
Weighted Avg.	0.741	0.416	0.729	0.741	0.732	0.349	0.680	0.694	

=== Confusion Matrix ===

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

# Naïve Bayes:

Click choose option in the classify, select bayes and then select naivebayes, as shown below.



Accuracy is 75.4% when cross validation is 10.

Classifier

Choose **NaiveBayes**

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)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes**

Classifier output

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: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	754	75.4	%
Incorrectly Classified Instances	246	24.6	%
Kappa statistic	0.3813		
Mean absolute error	0.2936		
Root mean squared error	0.4201		
Relative absolute error	69.8801	%	
Root relative squared error	91.6718	%	
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	0.787	0.891	good
	0.497	0.136	0.611	0.497	0.548	0.385	0.787	0.577	bad
Weighted Avg.	0.754	0.393	0.743	0.754	0.746	0.385	0.787	0.797	

=== Confusion Matrix ===

a	b	<-- classified as
605	95	a = good
151	149	b = bad



Accuracy is 75.9% when cross validation is 8.

Test options

☐ Use training set

☐ Supplied test set

☒ Cross validation Folds

☐ Percentage split %

(Nom) class

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes

Classifier output

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 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	65.0657 %	
Root relative squared error	91.7659 %	
Total Number of Instances	1000	

=== Detailed Accuracy By Class ===

	TP Data	FP Data	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.066	0.490	0.005	0.066	0.034	0.399	0.709	0.093	good
	0.510	0.134	0.618	0.510	0.558	0.398	0.788	0.575	bad
Weighted Avg.	0.759	0.303	0.749	0.759	0.752	0.399	0.709	0.790	

=== Confusion Matrix ===

a	b	<-- classified as
606	94	a = good
147	153	b = bad

Accuracy is 75.4% when cross validation is 6.

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 6

☐ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes
- 11:37:15 - bayes.NaiveBayes

Classifier output

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 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	0.393	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

Accuracy is 76.47% when percentage split is 66.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 6

☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes
- 11:37:15 - bayes.NaiveBayes
- 11:39:19 - bayes.NaiveBayes

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 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
```

Accuracy is 75.33% when percentage split is 70.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 6

☒ Percentage split % 70

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes
- 11:37:15 - bayes.NaiveBayes
- 11:39:19 - bayes.NaiveBayes
- 11:40:54 - bayes.NaiveBayes

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 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

Accuracy is 76.8% when percentage split is 75.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 6

☒ Percentage split % 75

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes
- 11:37:15 - bayes.NaiveBayes
- 11:39:19 - bayes.NaiveBayes
- 11:40:54 - bayes.NaiveBayes
- 11:42:27 - bayes.NaiveBayes

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 seconds

=== 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	

=== 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
155 29 | a = good
29 37 | b = bad
```

Accuracy is 74.5% when percentage split is 80.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 6

☒ Percentage split % 80

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 11:03:51 - trees.J48
- 11:06:49 - trees.J48
- 11:19:46 - trees.J48
- 11:23:43 - trees.J48
- 11:25:47 - trees.J48
- 11:29:16 - bayes.NaiveBayes
- 11:34:44 - bayes.NaiveBayes
- 11:37:15 - bayes.NaiveBayes
- 11:39:19 - bayes.NaiveBayes
- 11:40:54 - bayes.NaiveBayes
- 11:42:27 - bayes.NaiveBayes
- 11:44:21 - bayes.NaiveBayes

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 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		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	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
Weighted Avg.	0.745	0.358	0.761	0.745	0.751	0.368	0.796	0.825	

=== Confusion Matrix ===

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