

Analýza dat pomocí systému Weka, Rapid miner a Enterprise miner

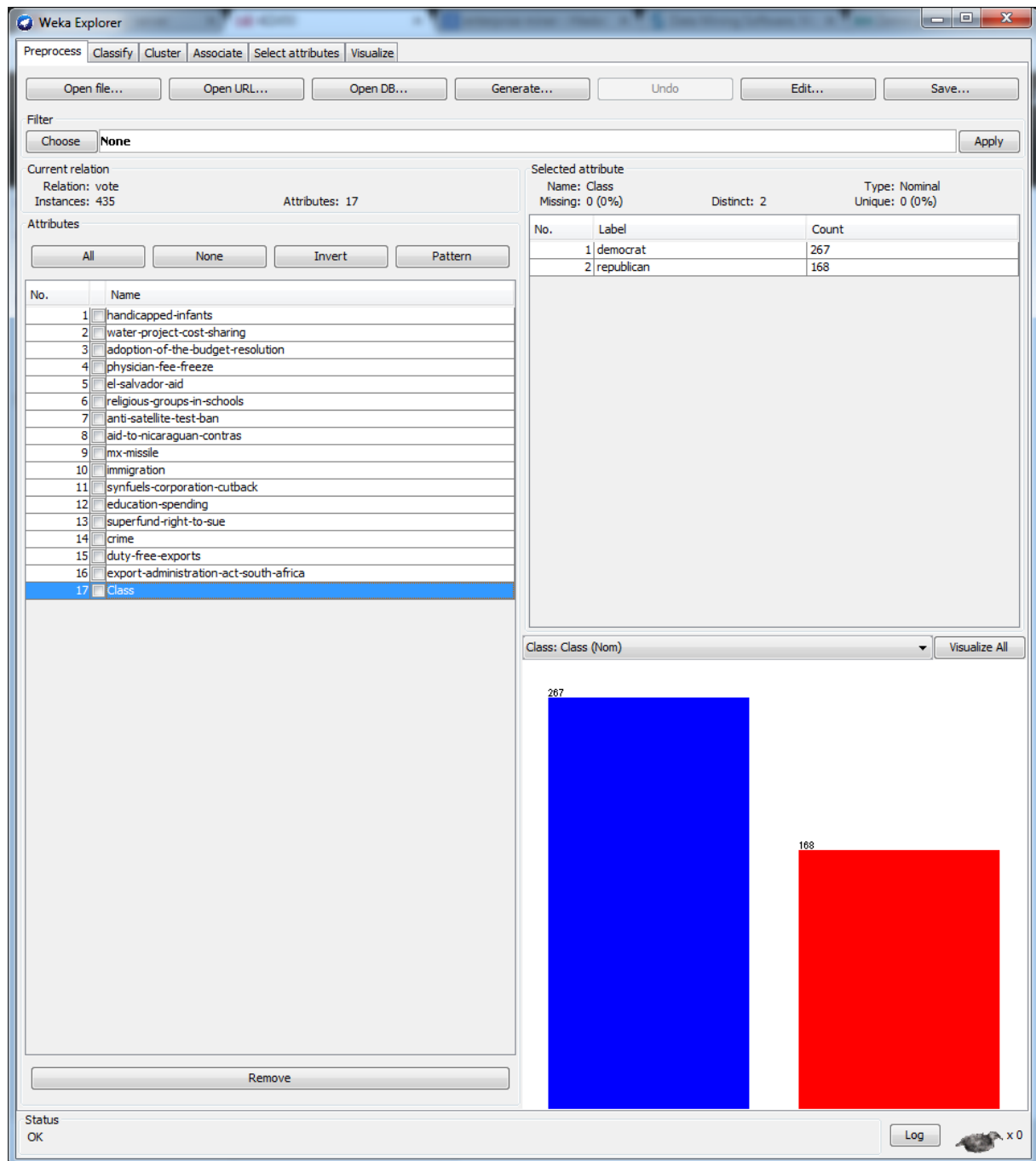
Dobývání znalostí z databází

Přidělená data a jejich popis

Data určená pro zpracování jsou označena jako „Vote“. Jedná se o hlasování kongresu USA z roku 1984. K dispozici je celkem 17 atributů, z čehož je jeden hlavní - class (rozložení sil v kongresu – 267 demokratů a 168 republikánů). Ostatní atributy jsou kategoriálního typu (pro / proti).

Weka

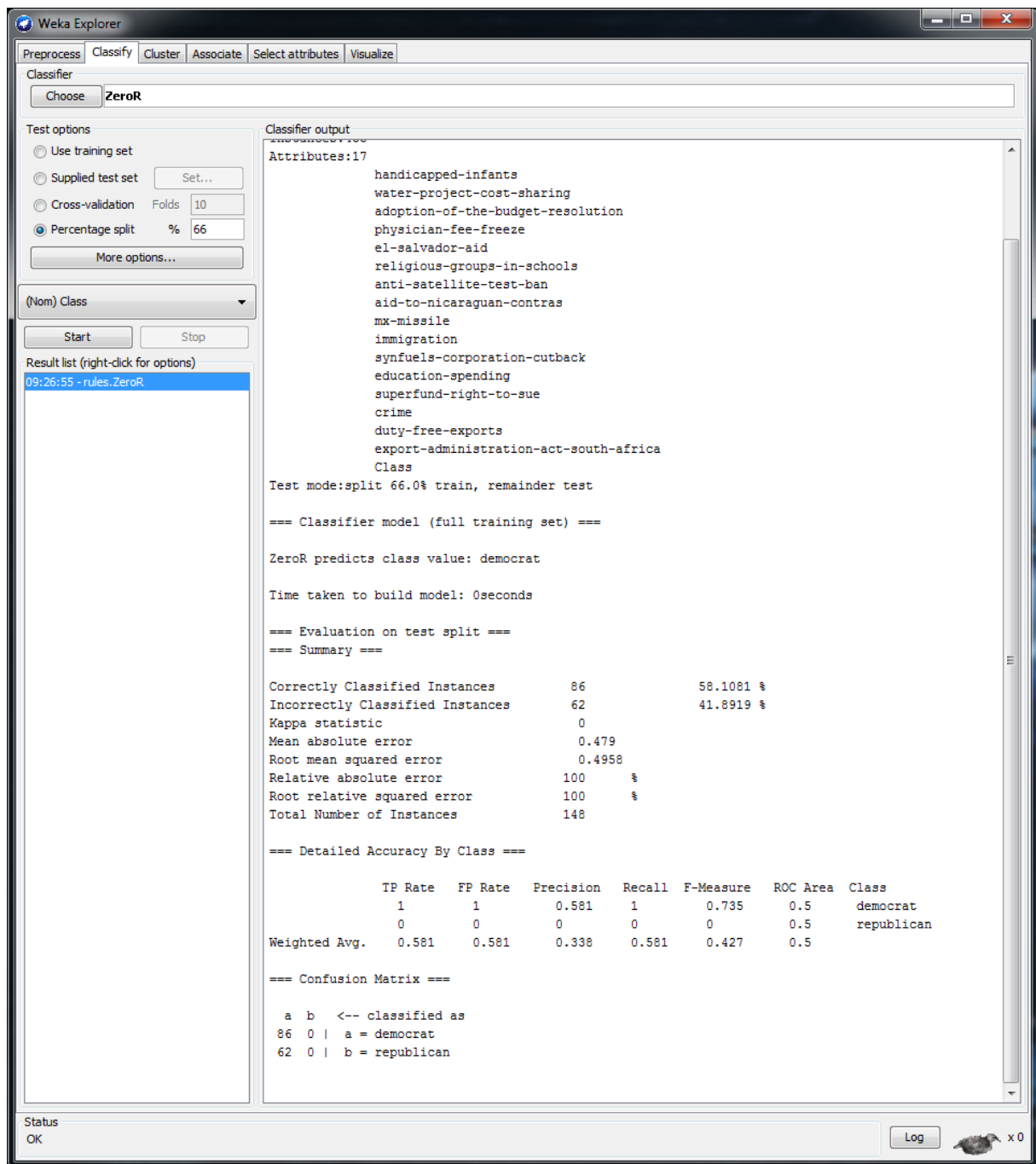
Na prvním snímku je vidět stav po načtení dat do systému Weka. Cílem je najít takový algoritmus, který data zanalyzuje s nejlepším vypovídajícím hodnocením.



Celý soubor dat bude pro aplikaci jednotlivých algoritmů rozdělen na 2/3 dat trénovacích a 1/3 testovacích.

ZeroR

Jako první algoritmus bylo aplikováno pravidlo zero rule, které hledá dominantní třídu (skupinu objektu se stejnými charakteristikami).



The screenshot shows the Weka Explorer interface with the ZeroR classifier selected. The 'Test options' section shows 'Percentage split' at 66%. The 'Classifier output' pane displays the following results:

```
Attributes:17
handicapped-infants
water-project-cost-sharing
adoption-of-the-budget-resolution
physician-fee-freeze
el-salvador-aid
religious-groups-in-schools
anti-satellite-test-ban
aid-to-nicaraguan-contras
mx-missile
immigration
synfuels-corporation-cutback
education-spending
superfund-right-to-sue
crime
duty-free-exports
export-administration-act-south-africa
Class

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

ZeroR predicts class value: democrat

Time taken to build model: 0seconds

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

Correctly Classified Instances      86          58.1081 %
Incorrectly Classified Instances    62          41.8919 %
Kappa statistic                     0
Mean absolute error                 0.479
Root mean squared error             0.4958
Relative absolute error             100 %
Root relative squared error         100 %
Total Number of Instances          148

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      1       1       0.581      1       0.735     0.5      democrat
      0       0       0         0       0         0.5      republican
Weighted Avg.   0.581   0.581   0.338    0.581   0.427     0.5

=== Confusion Matrix ===

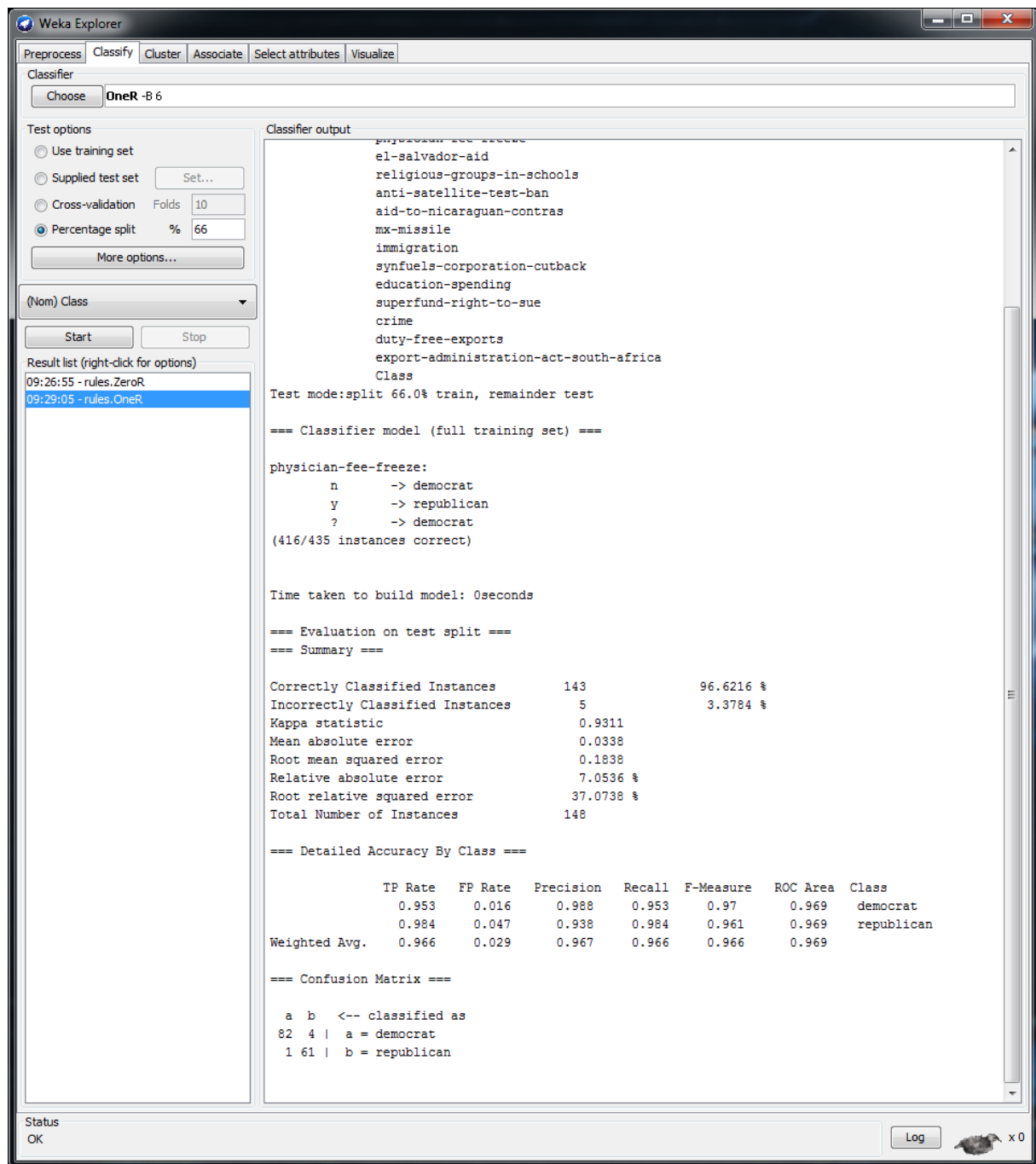
  a  b  <-- classified as
86  0  | a = democrat
62  0  | b = republican
```

The 'Result list' on the left shows '09:26:55 - rules.ZeroR' selected. The 'Status' bar at the bottom indicates 'OK'.

Zde je vidět, že tento algoritmus neposkytuje dostatečně silný výsledek, neboť má chybovost přes 40%

OneR

Dalším algoritmem je one rule, algoritmus určí pravidlo pro každý ukazatel a ten který má nejmenší počet chyb je stanoven jako „one rule“ a všechny ostatní indikátory jsou s ním poměřovány.



The screenshot shows the Weka Explorer interface with the OneR classifier selected. The classifier output pane displays the following results:

```
physician-fee-freeze
el-salvador-aid
religious-groups-in-schools
anti-satellite-test-ban
aid-to-nicaraguan-contras
mx-missile
immigration
synfuels-corporation-cutback
education-spending
superfund-right-to-sue
crime
duty-free-exports
export-administration-act-south-africa
Class
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

physician-fee-freeze:
  n    -> democrat
  y    -> republican
  ?    -> democrat
(416/435 instances correct)

Time taken to build model: 0seconds

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

Correctly Classified Instances      143      96.6216 %
Incorrectly Classified Instances     5       3.3784 %
Kappa statistic                     0.9311
Mean absolute error                  0.0338
Root mean squared error              0.1838
Relative absolute error              7.0536 %
Root relative squared error          37.0738 %
Total Number of Instances           148

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.953    0.016    0.988     0.953    0.97       0.969    democrat
      0.984    0.047    0.938     0.984    0.961     0.969    republican
Weighted Avg.   0.966    0.029    0.967     0.966    0.966     0.969

=== Confusion Matrix ===

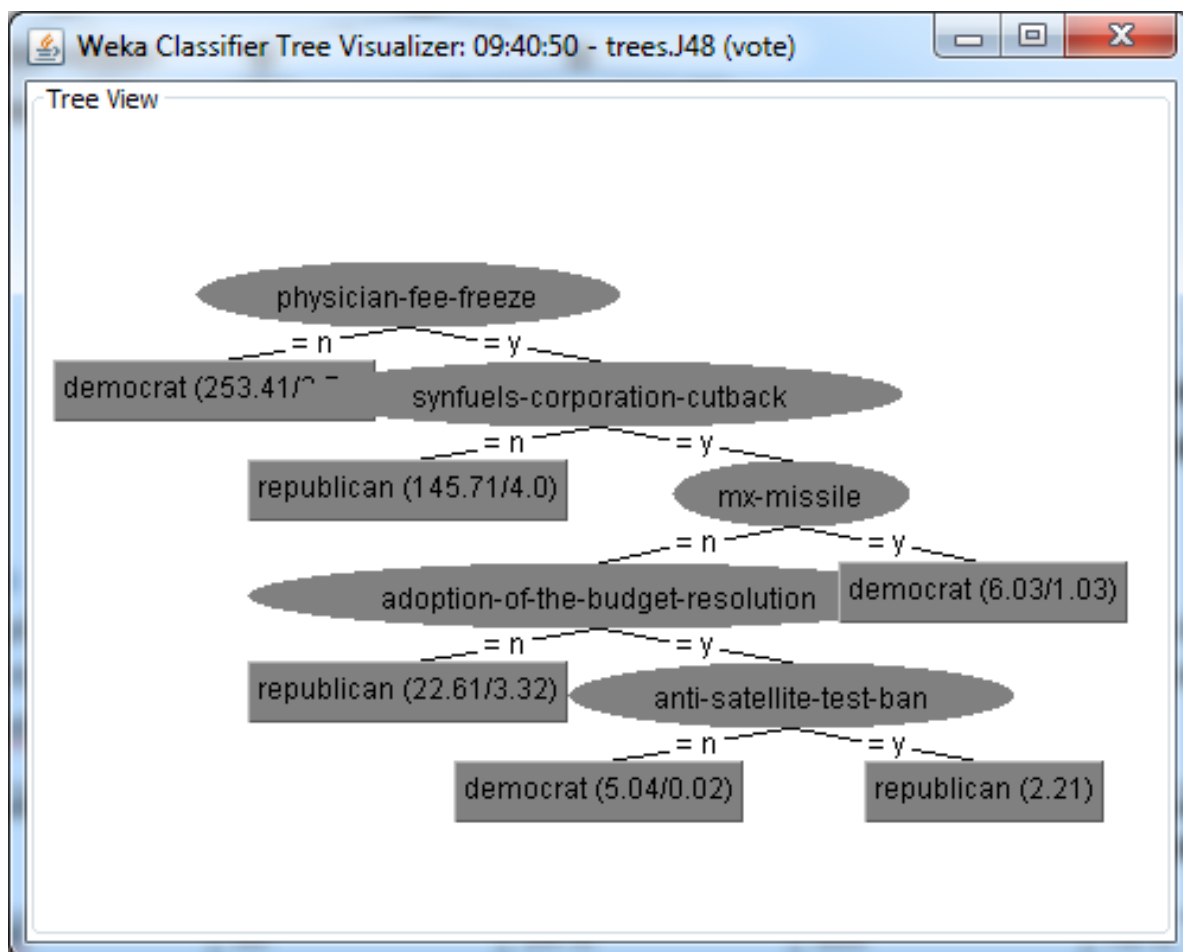
  a  b  <-- classified as
82  4  | a = democrat
 1 61 | b = republican
```

The status bar at the bottom indicates "Status OK".

Tento algoritmus dosáhl velmi slušného výsledku s přesností přesahující 96%.

J48

Třetím algoritmem je rozhodovací strom, který lze použít především na kategoriální atributy (tedy náš případ). Pokud by uživatel chtěl použít rozhodovací strom na nominální atributy, musel by data patřičným způsobem upravit (např. pomocí intervalů, rozdělení do skupin atp.)



Nejprve jsem zde uvedl obrázek stromu pro představu, kolik větví má. Na dalším obrázku je potom možné vidět velmi solidní výsledek s přesností na 97%, což představuje pouhé 4 chyby.

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)

09:26:55 - rules.ZeroR

09:29:05 - rules.OneR

09:40:50 - trees.J48

Classifier output

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

```

physician-fee-freeze = n: democrat (253.41/3.75)
physician-fee-freeze = y
|   synfuels-corporation-cutback = n: republican (145.71/4.0)
|   synfuels-corporation-cutback = y
|   |   mx-missile = n
|   |   |   adoption-of-the-budget-resolution = n: republican (22.61/3.32)
|   |   |   adoption-of-the-budget-resolution = y
|   |   |   |   anti-satellite-test-ban = n: democrat (5.04/0.02)
|   |   |   |   anti-satellite-test-ban = y: republican (2.21)
|   |   |   mx-missile = y: democrat (6.03/1.03)

```

Number of Leaves : 6

Size of the tree : 11

Time taken to build model: 0.03seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	144	97.2973 %
Incorrectly Classified Instances	4	2.7027 %
Kappa statistic	0.9447	
Mean absolute error	0.0608	
Root mean squared error	0.1539	
Relative absolute error	12.6846 %	
Root relative squared error	31.0328 %	
Total Number of Instances	148	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.965	0.016	0.988	0.965	0.976	0.99	democrat
	0.984	0.035	0.953	0.984	0.968	0.99	republican
Weighted Avg.	0.973	0.024	0.973	0.973	0.973	0.99	

=== Confusion Matrix ===

```

a b  <-- classified as
83 3 | a = democrat
1 61 | b = republican

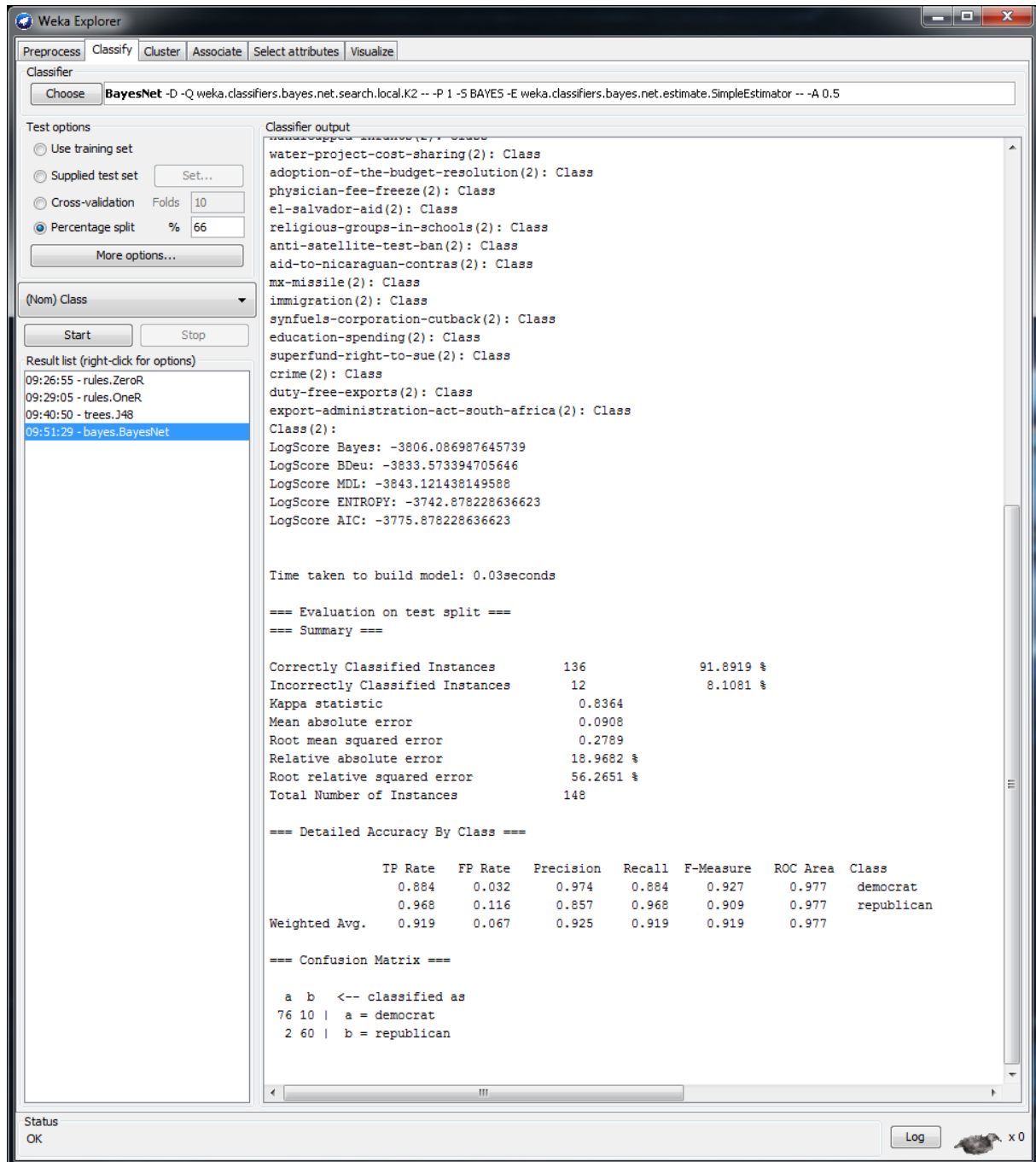
```

Status OK

Log x 0

BayesNet

Bayesovská síť využívá pravděpodobnostních vztahů mezi jevy.



The screenshot shows the Weka Explorer interface with the BayesNet classifier selected. The classifier output pane displays the following results:

```
==== Classifier output =====
water-project-cost-sharing(2): Class
adoption-of-the-budget-resolution(2): Class
physician-fee-freeze(2): Class
el-salvador-aid(2): Class
religious-groups-in-schools(2): Class
anti-satellite-test-ban(2): Class
aid-to-nicaraguan-contras(2): Class
mx-missile(2): Class
immigration(2): Class
synfuels-corporation-cutback(2): Class
education-spending(2): Class
superfund-right-to-sue(2): Class
crime(2): Class
duty-free-exports(2): Class
export-administration-act-south-africa(2): Class
Class(2):
LogScore Bayes: -3806.086987645739
LogScore BDeu: -3833.573394705646
LogScore MDL: -3843.121438149588
LogScore ENTROPY: -3742.878228636623
LogScore AIC: -3775.878228636623

Time taken to build model: 0.03seconds

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

Correctly Classified Instances      136           91.8919 %
Incorrectly Classified Instances     12            8.1081 %
Kappa statistic                     0.8364
Mean absolute error                  0.0908
Root mean squared error              0.2789
Relative absolute error              18.9682 %
Root relative squared error          56.2651 %
Total Number of Instances           148

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.884    0.032    0.974    0.884    0.927    0.977    democrat
          0.968    0.116    0.857    0.968    0.909    0.977    republican
Weighted Avg.   0.919    0.067    0.925    0.919    0.919    0.977

=== Confusion Matrix ===

  a  b  <-- classified as
 76 10 | a = democrat
  2 60 | b = republican
```

The status bar at the bottom shows "Status OK" and a "Log" button.

Tento algoritmus nabízí výsledek s téměř 92% přesností.

MultilayerPerceptron

Předposledním algoritmem je neuronová síť jež zanáší jednotlivé výskyty do mnohvrstevného grafu.

The screenshot shows the Weka Explorer interface with the MultilayerPerceptron classifier selected. The 'Test options' section shows 'Percentage split' at 66%. The 'Result list' on the left shows the MultilayerPerceptron model as the best performer. The 'Classifier output' pane displays the following information:

Classifier output

```
Attrib handicapped-infants 0.35330395613128446
Attrib water-project-cost-sharing 0.2426637469087909
Attrib adoption-of-the-budget-resolution 1.8885404170424036
Attrib physician-fee-freeze -3.202516780814178
Attrib el-salvador-aid -0.44509423052144925
Attrib religious-groups-in-schools 1.2682985503633262
Attrib anti-satellite-test-ban -0.5126567535821926
Attrib aid-to-nicaraguan-contras -0.033646300201523996
Attrib mx-missile 0.46293058210908866
Attrib immigration 0.4295117741977552
Attrib synfuels-corporation-cutback -0.02510660260560449
Attrib education-spending 0.9619990173173805
Attrib superfund-right-to-sue 1.683572262186216
Attrib crime -0.960203035254648
Attrib duty-free-exports 0.41399956486293704
Attrib export-administration-act-south-africa 0.453100855947657
```

Class democrat
Input
Node 0

Class republican
Input
Node 1

Time taken to build model: 0.67seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	146	98.6486 %
Incorrectly Classified Instances	2	1.3514 %
Kappa statistic	0.9724	
Mean absolute error	0.0222	
Root mean squared error	0.1134	
Relative absolute error	4.6257 %	
Root relative squared error	22.8741 %	
Total Number of Instances	148	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.977	0	1	0.977	0.988	0.999	democrat
	1	0.023	0.969	1	0.984	0.999	republican
Weighted Avg.	0.986	0.01	0.987	0.986	0.987	0.999	

=== Confusion Matrix ===

```
a b <-- classified as
84 2 | a = democrat
0 62 | b = republican
```

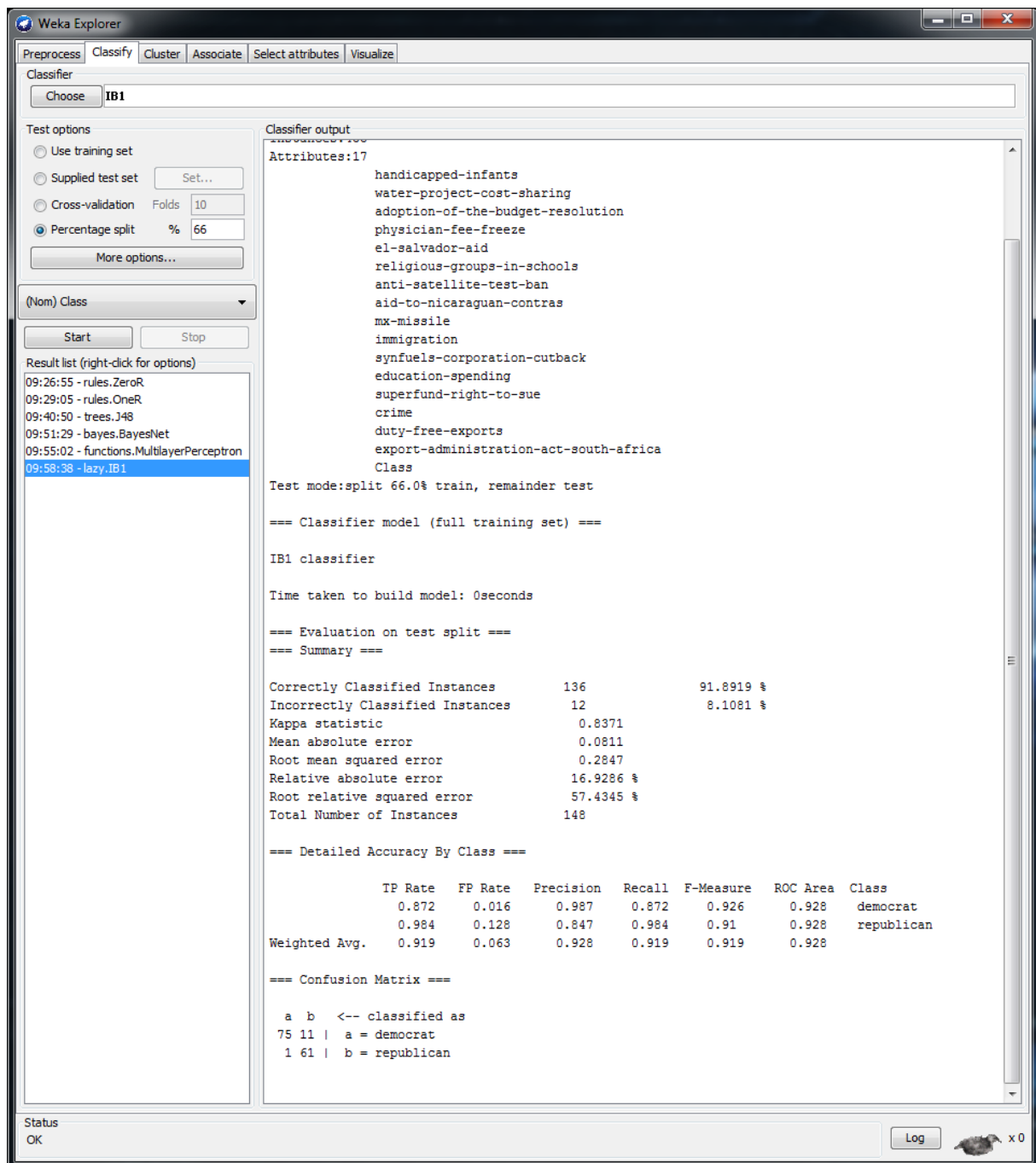
Status: OK

Log x 0

Toto testování nabízí zatím nejlepší výsledek s přesností na 98,6%, tedy s pouze dvěma chybami.

IB1

Jedná se o zkratku „instance based learning algorithm“. Tento algoritmus se snaží najít nejbližší příklad v tréninkové části dat tomu v testovací části a dle toho určit jeho třídu.



The screenshot shows the Weka Explorer window with the IB1 classifier selected. The 'Test options' section shows 'Percentage split' at 66%. The 'Classifier output' pane displays the following results:

```

Attributes:17
handicapped-infants
water-project-cost-sharing
adoption-of-the-budget-resolution
physician-fee-freeze
el-salvador-aid
religious-groups-in-schools
anti-satellite-test-ban
aid-to-nicaraguan-contras
mx-missile
immigration
synfuels-corporation-cutback
education-spending
superfund-right-to-sue
crime
duty-free-exports
export-administration-act-south-africa
Class

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

IB1 classifier

Time taken to build model: 0seconds

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

Correctly Classified Instances      136           91.8919 %
Incorrectly Classified Instances    12            8.1081 %
Kappa statistic                    0.8371
Mean absolute error                 0.0811
Root mean squared error             0.2847
Relative absolute error             16.9286 %
Root relative squared error         57.4345 %
Total Number of Instances          148

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.872    0.016    0.987    0.872    0.926    0.928    democrat
      0.984    0.128    0.847    0.984    0.91    0.928    republican
Weighted Avg.   0.919    0.063    0.928    0.919    0.919    0.928

=== Confusion Matrix ===

  a  b  <-- classified as
75 11 | a = democrat
 1 61 | b = republican

```

The 'Result list' on the left shows the selected model: 09:58:38 - lazy.IB1.

Výsledkem je 12 chybných zařazení.

Shrnutí – Weka

Následující tabulka je seřazena podle spolehlivosti jednotlivých algoritmů. Na základě daných výsledků bych volil mezi neuronovou sítí a rozhodovacím stromem. Respektive, i když má neuronová síť lepší výsledek, vzhledem k plochosti a jednoduchosti stromu, bych dal nejspíš přednost právě tomuto algoritmu.

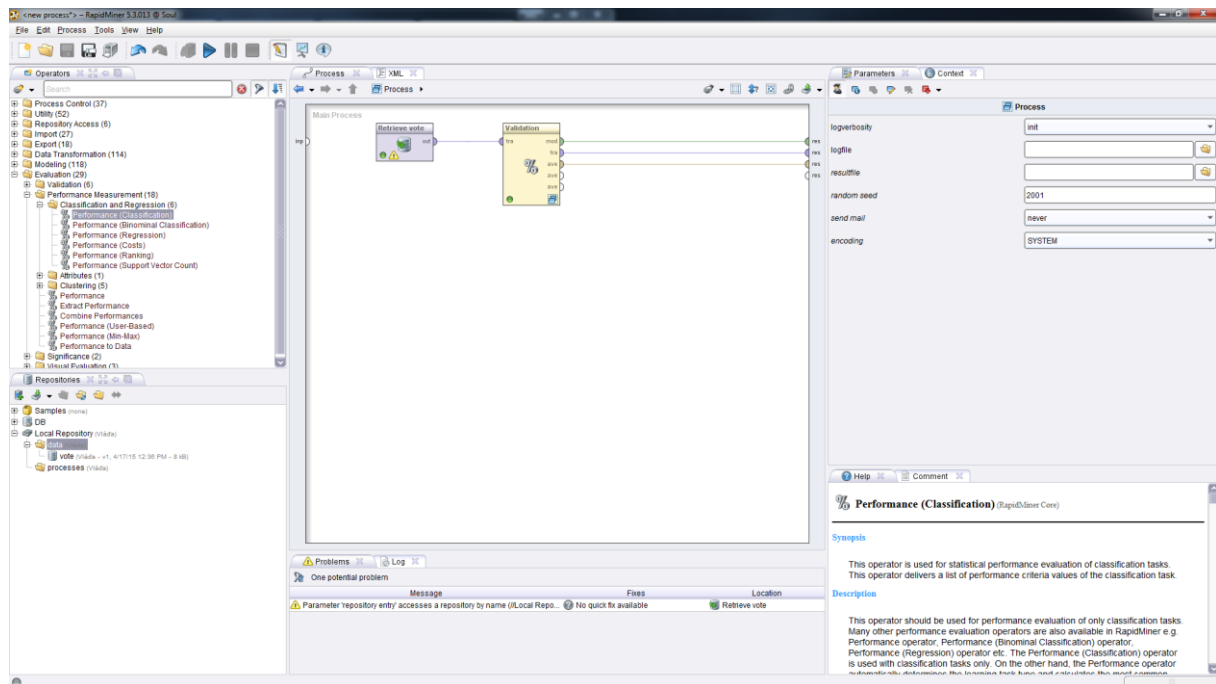
	Počet chyb	Spolehlivost
MultilayerPerceptron	2	98,6%
J48	4	97,3%
OneR	5	96,6%
BayesNet	12	91,9%
IB1	12	91,9%
ZeroR	62	58,1%

Rapid miner

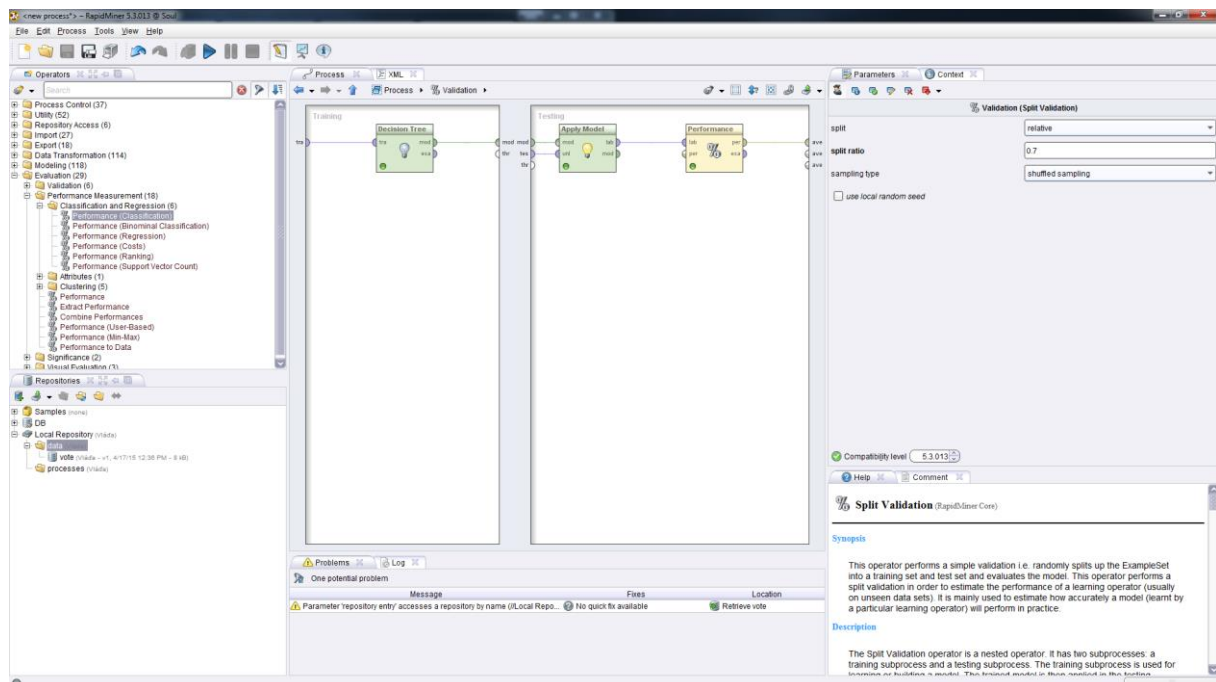
Tento software má podstatně lepší grafické UI, které umožňuje lepší orientaci v práci s daty. Na druhou stranu klade větší požadavky na znalosti a dovednosti uživatele, aby daný výstup fungoval tak, jak má.

Já jsem použil pro všechny níže uvedené modely následující strukturu:

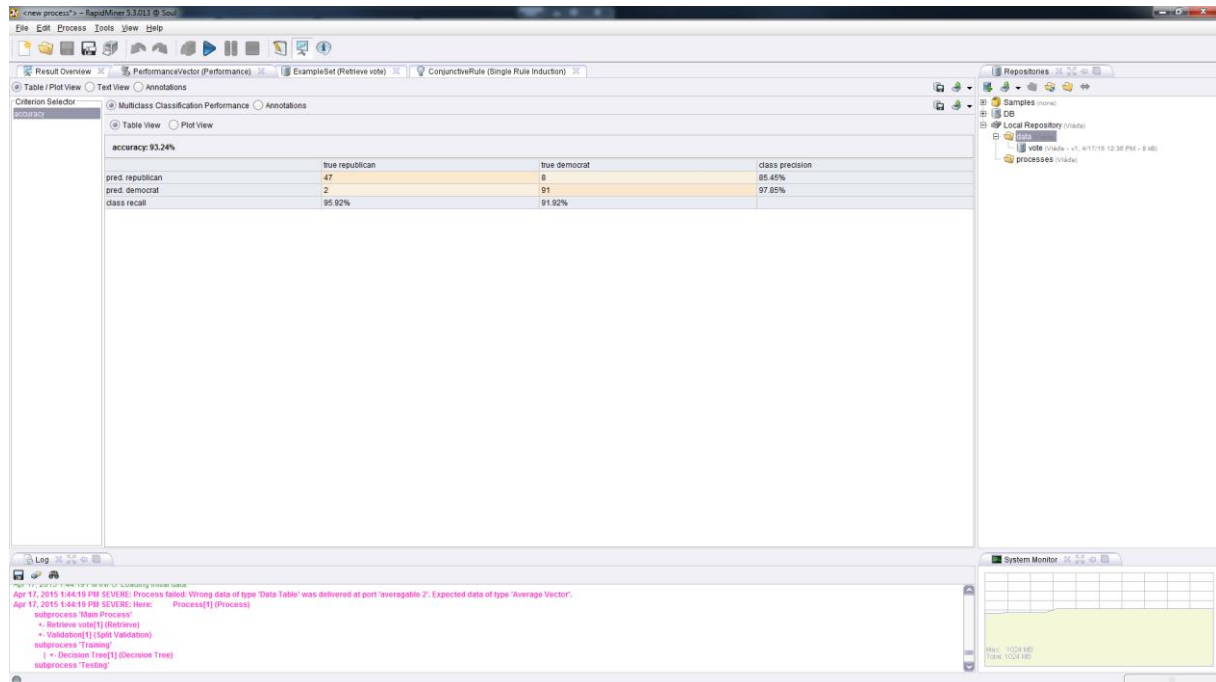
Vstupní data -> vyhodnocení (stejně jako u weky 66% jsou data trénovací a zbytek testovací)



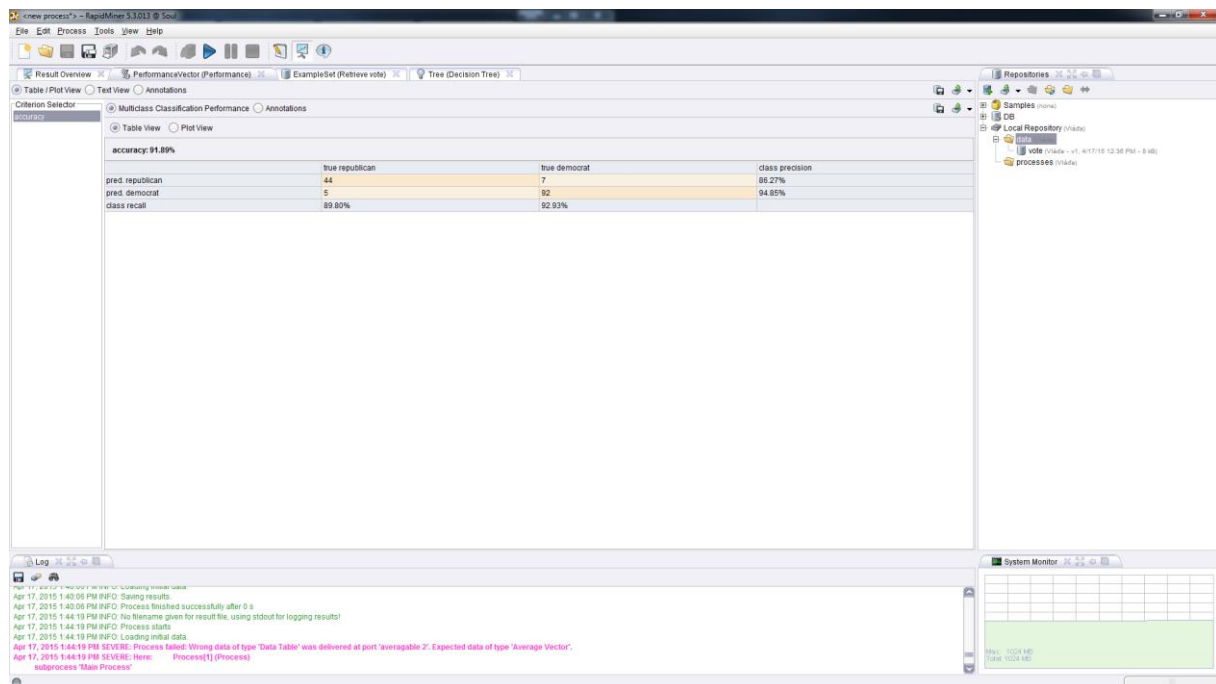
Ve vyhodnocení se bude měnit testovací model – dle jednotlivých testů (na obrázku „decision tree“).



Single rule



Decision tree



new process" - RapidMiner 3.3.0.11 - [Tool]

File Edit Process Tools View Help

Result Overview PerformanceVector (Performance) ExampleSet (Retrieve vote) SimpleDistribution (Naive Bayes)

Table / Plot View Test View Annotations

Criterion Selector

Table View Plot View

accuracy: 85.14%

	true republican	true democrat	class precision
pred republican	41	14	74.55%
pred democrat	8	85	91.40%
class recall	83.67%	85.88%	

Log

```

initprocess: Training
  | -> Decision Tree(1) (Decision Tree)
initprocess: Test
  | -> Apply Model(1) (Apply Model)
  | -> Performance(1) (Performance (Classification))
  | ->
Apr 17, 2015 1:40:00 PM INFO: No filename given for result file, using stdout for logging results!
Apr 17, 2015 1:40:00 PM INFO: Process starts
Apr 17, 2015 1:40:00 PM INFO: Loading initial data
  
```

System Monitor

RAM: 1024 MB
CPU: 100.00%

The screenshot displays the RapidMiner 3.3.0.11 interface. The main window shows a workflow execution log with the following entries:

```

Apr 17, 2015 2:10:57 PM INFO Process starts
Apr 17, 2015 2:10:57 PM INFO Loading initial data
Apr 17, 2015 2:10:57 PM SEVERE: Process failed: The operator Perceptron does not have sufficient capabilities for the given data set: binomical attributes not supported
Apr 17, 2015 2:10:57 PM SEVERE: Here: Process[1] (Process)
    subprocess: Main Process
        + Retrieve vote[1] (Retrieve)
        + Validation[1] (Split Validation)
        subprocess: Training
            + + + Process output[1] (Process output)
  
```

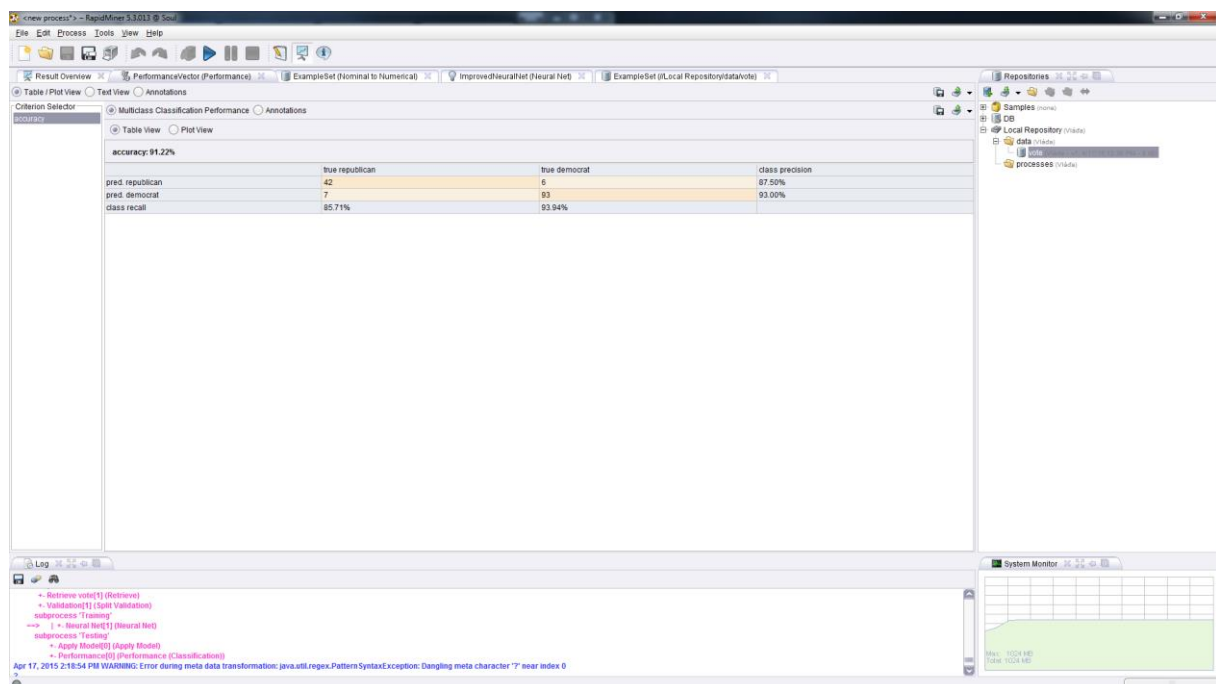
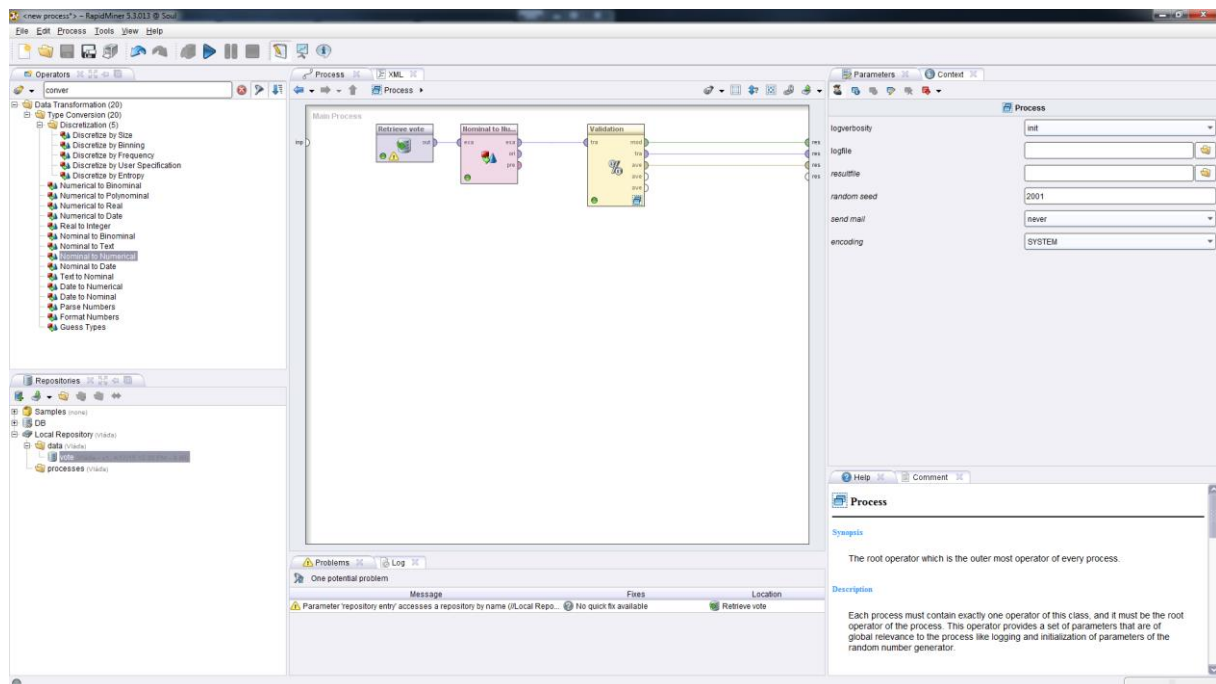
Below the log, a table displays classification results for a dataset with two classes: 'true republican' and 'true democrat'. The table includes columns for predicted class, true class, and class precision.

	true republican	true democrat	class precision
pred: republican	47	9	85.45%
pred: democrat	2	91	97.85%
class recall	95.92%	91.92%	

The interface also shows a 'Criterion Selector' on the left with 'Table View' selected, and a 'System Monitor' on the bottom right showing memory usage (1024 MB).

Neural net

Zde bylo potřeba upravit vstupní data na numerické, jelikož tento model neuměl pracovat s binomickými atributy



Shrnutí – Rapid miner

Stejně jako u předchozího SW jsem dané výstupy zaznamenal do tabulky a seřadil dle spolehlivosti předpovědi jednotlivých výskytů. Zde se nejlépe jeví použití single rule (obdoba OneR z weky) nebo random tree.

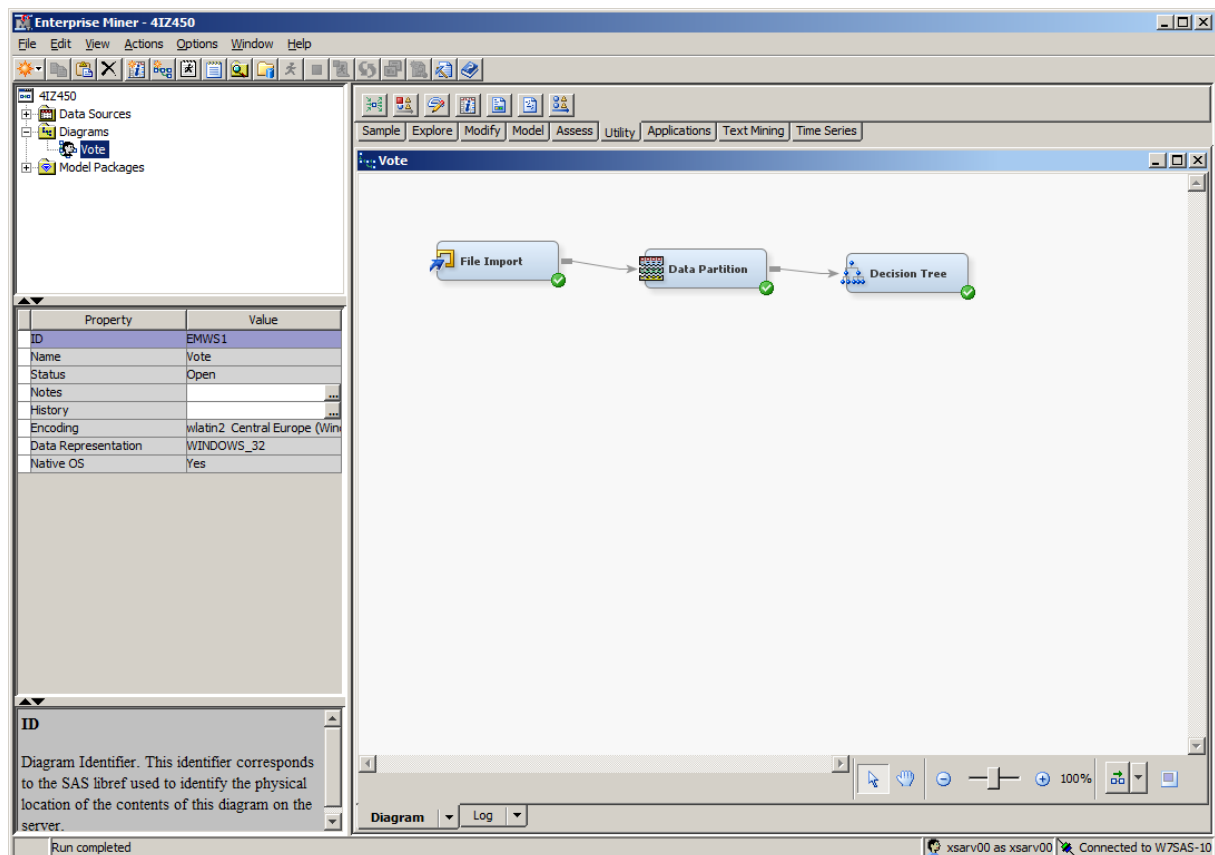
	Počet chyb	Spolehlivost
Single rule	10	93,2%
Random tree	10	93,2%
Decision tree	12	91,9%
Neural net	13	91,2%
BayesNet	22	85,1%

Enterprise miner

Hlavním nedostatkem tohoto systému spatřuji v tom, že neexistuje možnost si daný software vyzkoušet bez zakoupení (ať už formou LITE verze – kdy jsou uživateli zpřístupněny jen některé základní funkce; nebo třeba TRIAL verze, která bývá většinou omezena časovým oknem). S tímto omezením se vážou i moje osobní problémy s daným softwarem, kdy ke zpracování zadané úlohy je potřeba využívat školní PC, které svojí hardwarovou kapacitou nedostačují k plynulé práci s daným softwarem.

Decision tree

Na prvním obrázku je vidět aplikované schéma, které je výchozím pro následnou analýzu pomocí různých metod. Zde je zvolenou metodou decision tree.



Výsledkem je následující matice správných a chybných přiřazení:

Results - Node: Decision Tree Diagram: Vote						
Output						
Statistics	Statistics Label	Train	Validation	Test		
84						
85						
86	_NOBS_	Sum of Frequencies	173.000	130.000	132.000	
87	_MISC_	Misclassification Rate	0.046	0.046	0.038	
88	_MAX_	Maximum Absolute Error	0.981	0.981	0.981	
89	_SSE_	Sum of Squared Errors	14.894	11.296	9.417	
90	_ASE_	Average Squared Error	0.043	0.043	0.036	
91	_RASE_	Root Average Squared Error	0.207	0.208	0.189	
92	_DIV_	Divisor for ASE	346.000	260.000	264.000	
93	_DFT_	Total Degrees of Freedom	173.000	.	.	
94						
95						
96						
97						
98	Classification Table					
99						
100	Data Role=TRAIN Target Variable=Class					
101						
102						
103	Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
104						
105	DEMOCRAT	DEMOCRAT	98.0583	94.3925	101	58.3815
106	REPUBLICAN	DEMOCRAT	1.9417	3.0303	2	1.1561
107	DEMOCRAT	REPUBLICAN	8.5714	5.6075	6	3.4682
108	REPUBLICAN	REPUBLICAN	91.4286	96.9697	64	36.9942
109						
110						
111	Data Role=VALIDATE Target Variable=Class					
112						
113						
114	Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
115						
116	DEMOCRAT	DEMOCRAT	97.4359	95	76	58.4615
117	REPUBLICAN	DEMOCRAT	2.5641	4	2	1.5385
118	DEMOCRAT	REPUBLICAN	7.6923	5	4	3.0769
119	REPUBLICAN	REPUBLICAN	92.3077	96	48	36.9231
120						
121						
122						
123						
124	Event Classification Table					
125						
126	Data Role=TRAIN Target=Class					
127						
128	False	True	False	True		
129	Negative	Negative	Positive	Positive		
130						
131	2	101	6	64		
132						
133						
134	Data Role=VALIDATE Target=Class					
135						

Vzhledem k tomu, že Enterprise miner rozděluje danou matici na všechny nastalé jevy, je pro srovnání potřeba sečíst ty správné výskyty – neboli první a poslední řádek (Target: democrat, Outcome: democrat; Target: republican, Outcome: republican). Zde se dostaneme na číslo 95,37%. Tento výsledek byl dosažen na testovacích datech, které představovali 40% z celkového souboru.

Autoneural

Results - Node: AutoNeural Diagram: Vote						
Output						
518	_AVER_	Average Error Function	0.615	0.728	0.805	
519	_ERR_	Error Function	212.699	189.395	212.642	
520	_MISC_	Misclassification Rate	0.150	0.154	0.182	
521	_WRONG_	Number of Wrong Classifications	26.000	20.000	24.000	
522						
523						
524						
525						
526	Classification Table					
527						
528	Data Role=TRAIN Target Variable=Class					
529						
530			Target	Outcome	Frequency	Total
531	Target	Outcome	Percentage	Percentage	Count	Percentage
532						
533	DEMOCRAT	DEMOCRAT	96.5517	78.5047	84	48.5549
534	REPUBLICAN	DEMOCRAT	3.4483	4.5455	3	1.7341
535	DEMOCRAT	REPUBLICAN	26.7442	21.4953	23	13.2948
536	REPUBLICAN	REPUBLICAN	73.2558	95.4545	63	36.4162
537						
538						
539	Data Role=VALIDATE Target Variable=Class					
540						
541			Target	Outcome	Frequency	Total
542	Target	Outcome	Percentage	Percentage	Count	Percentage
543						
544	DEMOCRAT	DEMOCRAT	92.8571	81.25	65	50.0000
545	REPUBLICAN	DEMOCRAT	7.1429	10.00	5	3.8462
546	DEMOCRAT	REPUBLICAN	25.0000	18.75	15	11.5385
547	REPUBLICAN	REPUBLICAN	75.0000	90.00	45	34.6154
548						
549						
550						
551						
552	Event Classification Table					
553						
554	Data Role=TRAIN Target=Class					
555						
556	False	True	False	True		
557	Negative	Negative	Positive	Positive		
558						
559	3	84	23	63		
560						
561						
562	Data Role=VALIDATE Target=Class					
563						
564	False	True	False	True		
565	Negative	Negative	Positive	Positive		
566						
567	5	65	15	45		
568						
569						

S touto funkcí bylo dosaženo přesnosti pouhých 84,96% na stejných datech se stejným rozdělením.

Neural network

Results - Node: Neural Network Diagram: Vote					
File Edit View Window					
Output					
400	_AVER_	Average Error Function	0.003	0.149	0.331
401	_ERR_	Error Function	0.983	38.696	87.479
402	_MISC_	Misclassification Rate	0.000	0.038	0.053
403	_WRONG_	Number of Wrong Classifications	0.000	5.000	7.000
404					
405					
406					
407	Classification Table				
408					
409	Data Role=TRAIN Target Variable=Class				
410					
411					
412					
413	Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count
414					Total Percentage
415	DEMOCRAT	DEMOCRAT	100	100	107
416	REPUBLICAN	REPUBLICAN	100	100	66
417					
418	Data Role=VALIDATE Target Variable=Class				
419					
420					
421					
422	Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count
423					Total Percentage
424	DEMOCRAT	DEMOCRAT	100.000	93.75	75
425	DEMOCRAT	REPUBLICAN	9.091	6.25	5
426	REPUBLICAN	REPUBLICAN	90.909	100.00	50
427					
428	Event Classification Table				
429					
430	Data Role=TRAIN Target=Class				
431					
432					
433	False	True	False	True	
434	Negative	Negative	Positive	Positive	
435	0	107	.	66	
436					
437					
438					
439					
440	Data Role=VALIDATE Target=Class				
441					
442					
443	False	True	False	True	
444	Negative	Negative	Positive	Positive	
445	0	75	5	50	
446					
447					
448					
449					
450	Assessment Score Rankings				
451					

Zajímavé je, že u nerálních sítí bylo na testovacích datech dosaženo 100% přesnosti, ale na validační části dat (30% z celku) pouze 96,15%.

Regression

Results - Node: Regression Diagram: Vote					
File Edit View Window					
Output					
265	_RMSE_	Root Mean Squared Error	0.002	0.232	0.282
266	_SBC_	Schwarz's Bayesian Criterion	170.259	.	.
267	_SSE_	Sum of Squared Errors	0.002	13.955	20.953
268	_SUNW_	Sum of Case Weights Times Freq	346.000	260.000	264.000
269	_MISC_	Misclassification Rate	0.000	0.054	0.083
270					
271					
272					
273					
274	Classification Table				
275					
276	Data Role=TRAIN Target Variable=Class				
277					
278			Target	Outcome	Frequency
279	Target	Outcome	Percentage	Percentage	Count
280					Total
281	DEMOCRAT	DEMOCRAT	100	100	107
282	REPUBLICAN	REPUBLICAN	100	100	66
283					61.8497
284					38.1503
285	Data Role=VALIDATE Target Variable=Class				
286					
287			Target	Outcome	Frequency
288	Target	Outcome	Percentage	Percentage	Count
289					Total
290	DEMOCRAT	DEMOCRAT	100.000	91.25	73
291	DEMOCRAT	REPUBLICAN	12.281	8.75	7
292	REPUBLICAN	REPUBLICAN	87.719	100.00	50
293					56.1538
294					5.3846
295					38.4615
296					
297	Event Classification Table				
298					
299	Data Role=TRAIN Target=Class				
300					
301	False	True	False	True	
302	Negative	Negative	Positive	Positive	
303					
304	0	107	.	66	
305					
306	Data Role=VALIDATE Target=Class				
307					
308	False	True	False	True	
309	Negative	Negative	Positive	Positive	
310					
311	0	73	7	50	
312					
313					
314					
315					
316					

Stejně je tomu tak i u regresní funkce kdy je rozdíl mezi přesností trénovacích a validačních dat ještě větší neboli 100% ku 94,61%.

Shrnutí Enterprise miner

Tento software ve srovnání s předchozími dvěma je rozhodně nejsložitější na znalosti uživatele. Vše lze parametricky upravovat, a tudíž umožňuje uživateli absolutní kontrolu nad prováděnou analýzou. Taktéž nabízí poměrně širokou škálu formy výstupu (grafické i numerické). Osobně bych tento program označil za velmi dobrý pro profesionální práci ovšem absolutně nevhodný pro začátečníky, či uživatele bez velmi hlubokých znalostí veškerých použitých algoritmů (právě z důvodu rozličného množství potřebných parametrů při jejich aplikaci).

A nakonec výsledná tabulka:

	Počet chyb	Spolehlivost - trénovací data	Spolehlivost - validační data
Decision tree	8 / 6	95,37%	95,38%
Autoneural	26 / 20	84,96%	84,61%
Neural network	0 / 5	100,00%	96,15%
Regression	0 / 7	100,00%	94,61%

Na základě těchto výsledků bych nejspíše volil rozhodovací strom, neboť jeho spolehlivost nemá takové výkyvy jako regrese a neurální síť a přitom dosahuje velmi vysokých hodnot.

Shrnutí práce

Po vyzkoušení všech tří navrhovaných softwarů se při zpracování následujícího úkolu budu rozhodovat mezi Rapid minerem a Wekou. Výstupy jsou pro mě čitelné ve všech testovaných případech stejně. S Wekou je jednodušší práce, ovšem Rapid miner má přívětivější uživatelské rozhraní.