PROJECT 1 – DATA ANALYSIS (KR-VS-KP)

Contents

1 INTRODUCTION

The purpose of this coursework is to analyse KRKPA7 dataset by using 4 different data-mining software tools, namely Weka, RapidMiner, SAS Enterprise Miner and IBM SPSS Modeler.

Given dataset describes the Chess End-game King + Rook versus King + Pawn on A7 (abbreviated as KRKPA7). The pawn on A7 means – being 1 square away from queening. The white player begins. (1)

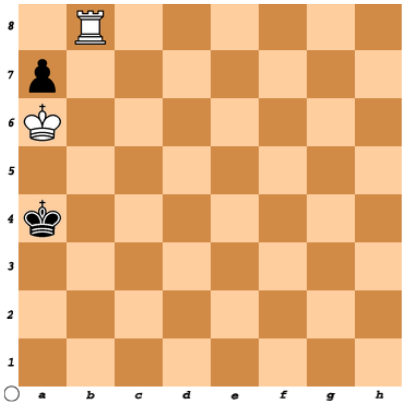
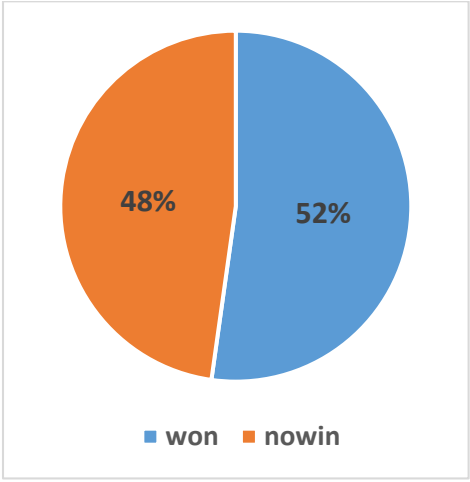


Figure 1 - Example game setup (2)

1.1 DATA DESCRIPTION

- Number of instances: 3196
- Number of attributes: 36
- Classes: White-can-win("won") and White-cannot-win("nowin")
- Missing attributes: none
- Class distribution: 52% – White-can-win, 48% White-cannot-win



Each instance in dataset is a sequence of 37 attributes – first 36 nominal attributes show the board-description, the last 37th attribute is the classification "won" or "nowin".

1.2 ATTRIBUTES

Table 1 - Attributes representing board position (3)

#	Abbreviation	Description
1	bkblk	the BK is not in the way
2	bknwy	the BK is not in the BR's way
3	bkon8	the BK is on rank 8 in a position to aid the BR
4	bkona	the BK is on file A in a position to aid the BR
5	bkspr	the BK can support the BR
6	bkxbq	the BK is not attacked in some way by the promoted WP
7	bkxcr	the BK can attack the critical square (b7)
8	bkxwp	the BK can attack the WP
9	blxwp	B attacks the WP (BR in direction x = -1 only)
10	bxqsq	one or more Black pieces control the queening square
11	cntxt	the WK is on an edge and not on a8
12	dsopp	the kings are in normal opposition
13	dwipd	the WK distance to intersect point is too great
14	hdchk	there is a good delay because there is a hidden check
15	katri	the BK controls the intersect point
16	mulch	B can renew the check to good advantage
17	qxmsq	the mating square is attacked in some way by the promoted WP
18	r2ar8	the BR does not have safe access to file A or rank 8
19	reskd	the WK can be reskewered via a delayed skewer
20	reskr	the BR alone can renew the skewer threat
21	rimmx	the BR can be captured safely
22	rkxwp	the BR bears on the WP (direction x = -1 only)
23	rxmsq	the BR attacks a mating square safely
24	simpl	a very simple pattern applies
25	skach	the WK can be skewered after one or more checks
26	skewr	there is a potential skewer as opposed to fork
27	skrxp	the BR can achieve a skewer or the BK attacks the WP
28	spcop	there is a special opposition pattern present
29	stlmt	the WK is in stalemate
30	thrsk	there is a skewer threat lurking
31	wkcti	the WK cannot control the intersect point
32	wkna8	the WK is on square a8
33	wknck	the WK is in check
34	wkovl	the WK is overloaded
35	wkpos	the WK is in a potential skewer position
36	wtoeg	the WK is one away from the relevant edge

1.3 DATA PRE-PROCESSING & PREPARATION

As a result of fact that all attributes are either binominal or polynomial, no missing values are present and class distribution is almost equal – no data pre-processing is needed.

Unfortunately the given file format (.arff) is not supported by IBM SPSS Modeler and SAS Enterprise Miner. The solution is to convert file format by using Weka.

2 WEKA

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Weka is open source software issued under the GNU General Public License, developed by Machine Learning Group at the University of Waikato, New Zealand. (4)

2.1 MODELLING

First step in order to be able to process modelling algorithms is the data import and explorer tool. Successful data import and selecting of target attribute can be seen on figure 2.

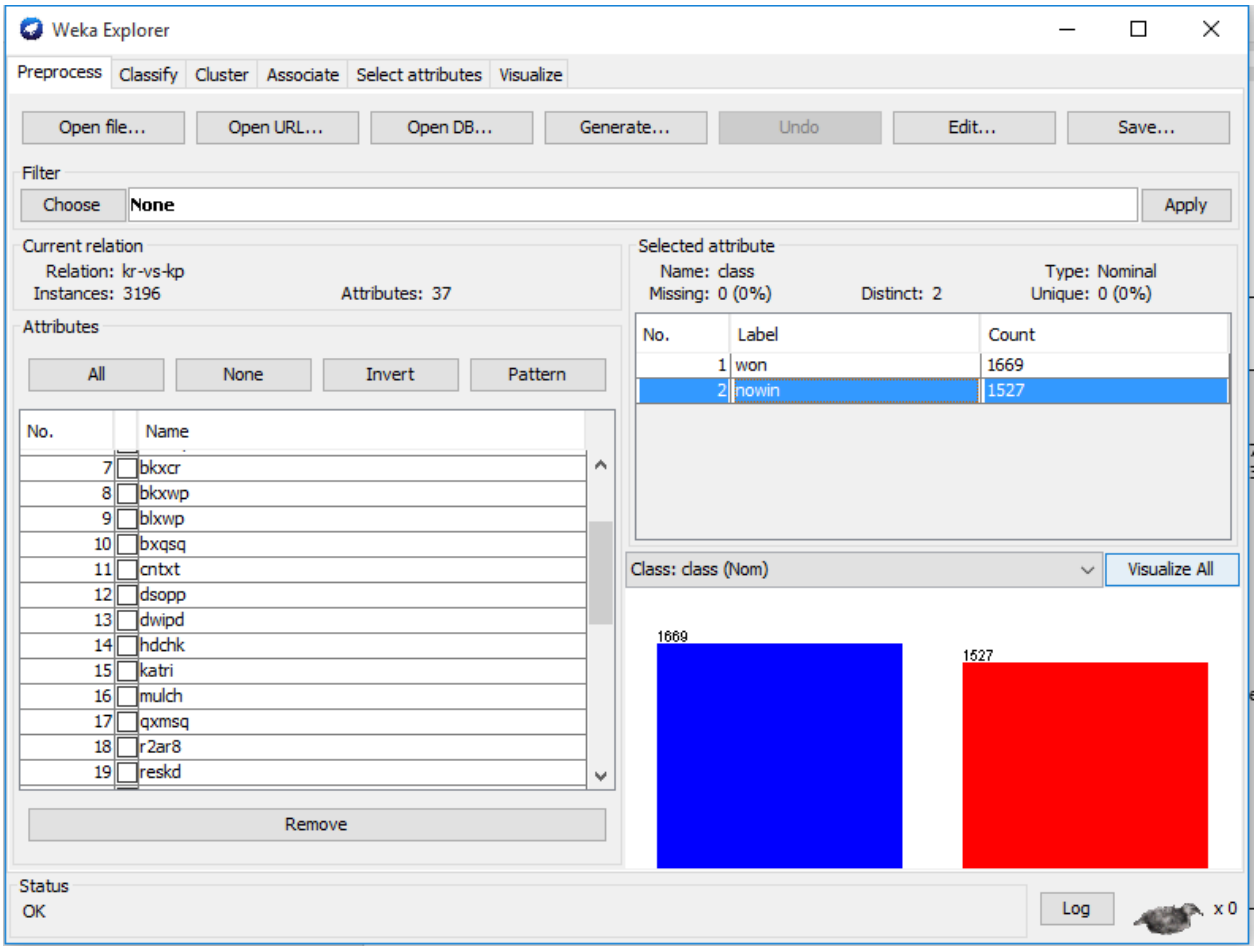


Figure 2 - Weka explorer window

2.1.1 ZeroRule

This simple rule tries to split instances according to the target class attribute – therefore the percentage of correctly classified instances is only 52.43%.

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

Correctly Classified Instances         670               52.4257 %
Incorrectly Classified Instances       608               47.5743 %
Kappa statistic                          0
Mean absolute error                      0.499
Root mean squared error                  0.4994
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.524     1         0.688      0.5      won
                 0         0          0         0         0          0.5      nowin
Weighted Avg.    0.524     0.524      0.275     0.524     0.361      0.5

=== Confusion Matrix ===

   a   b   <-- classified as
 670   0 |   a = won
 608   0 |   b = nowin
```

2.1.2 ONERULE

OneRule algorithm is slightly more complex than the ZeroRule – it chooses the best attribute in dataset to predict the target – in this case attribute bxqsq was chosen.

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

Correctly Classified Instances         869               67.9969 %
Incorrectly Classified Instances       409               32.0031 %
Kappa statistic                          0.3465
Mean absolute error                      0.32
Root mean squared error                  0.5657
Relative absolute error                 64.1359 %
Root relative squared error            113.2733 %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.873     0.533      0.644     0.873     0.741      0.67     won
                 0.467     0.127      0.77      0.467     0.581      0.67     nowin
Weighted Avg.    0.68      0.34       0.704     0.68      0.665      0.67

=== Confusion Matrix ===

   a   b   <-- classified as
 585  85 |   a = won
 324 284 |   b = nowin
```

2.1.3 NAIVE BAYES

The algorithm based on conditional probability model has given better results, but still not absolutely satisfying.

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

Correctly Classified Instances        1120               87.6369 %
Incorrectly Classified Instances       158               12.3631 %
Kappa statistic                          0.7517
Mean absolute error                      0.2098
Root mean squared error                  0.3026
Relative absolute error                 42.0375 %
Root relative squared error             60.5957 %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.9       0.15       0.869     0.9       0.884      0.953    won
                 0.85      0.1        0.885     0.85      0.867      0.953    nowin
Weighted Avg.    0.876     0.126      0.877     0.876     0.876      0.953

=== Confusion Matrix ===

   a   b   <-- classified as
 603  67 |   a = won
  91 517 |   b = nowin
```

2.1.4 ID3 DECISION TREE

Entropy based algorithm produced perfect classification tree with high success rate of 99.77%.

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

Correctly Classified Instances        1275               99.7653 %
Incorrectly Classified Instances         3                0.2347 %
Kappa statistic                          0.9953
Mean absolute error                      0.0023
Root mean squared error                  0.0485
Relative absolute error                  0.4704 %
Root relative squared error              9.7012 %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.005      0.996     1         0.998      0.998    won
                 0.995     0          1         0.995     0.998      0.998    nowin
Weighted Avg.    0.998     0.003      0.998     0.998     0.998      0.998

=== Confusion Matrix ===

   a   b   <-- classified as
 670   0 |   a = won
   3 605 |   b = nowin
```

2.1.5 J48 Decision Tree

J48 is an implementation of C4.5 algorithm in Weka. It should extend the ID3 algorithm, but in this case – it gives worse results.

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

Correctly Classified Instances        1271               99.4523 %
Incorrectly Classified Instances         7                0.5477 %
Kappa statistic                          0.989
Mean absolute error                      0.008
Root mean squared error                  0.0724
Relative absolute error                  1.5936 %
Root relative squared error             14.5027 %
Total Number of Instances             1278

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.008      0.993     0.997     0.995      0.999    won
                 0.992     0.003      0.997     0.992     0.994      0.999    nowin
Weighted Avg.    0.995     0.006      0.995     0.995     0.995      0.999

=== Confusion Matrix ===

   a   b   <-- classified as
 668   2 |   a = won
   5 603 |   b = nowin
```

2.2 Results

Algorithm	Correctly classified	Success rate
ZeroRule	670	52,43%
OneRule	869	68,00%
NaiveBayes	1120	87,64%
ID3	1275	99,77%
J48	1271	99,45%

The best algorithms were decision trees – namely ID3 and J48, NaïveBayes provided also quite high success rate, but on the other hand simple classification algorithms such as ZeroRule and OneRule are unusable for this dataset.

3 RapidMiner

RapidMiner is a software platform developed for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization. (5)

RapidMiner company releases the previous version of the software under an OSI-certified open source license on Sourceforge, the current version is commercial with licences starting from US$999.

3.1 Modelling

RapidMiner provides users with highly intuitive modelling process window that covers all data-mining steps. Firstly, users have to specify the class attribute – in RapidMiner called label, secondly multiply the data stream for validation boxes and in the end – very intuitively specify what will be included in final report by connecting "boxes" with the output column.

The project setup can be seen on figure 3.

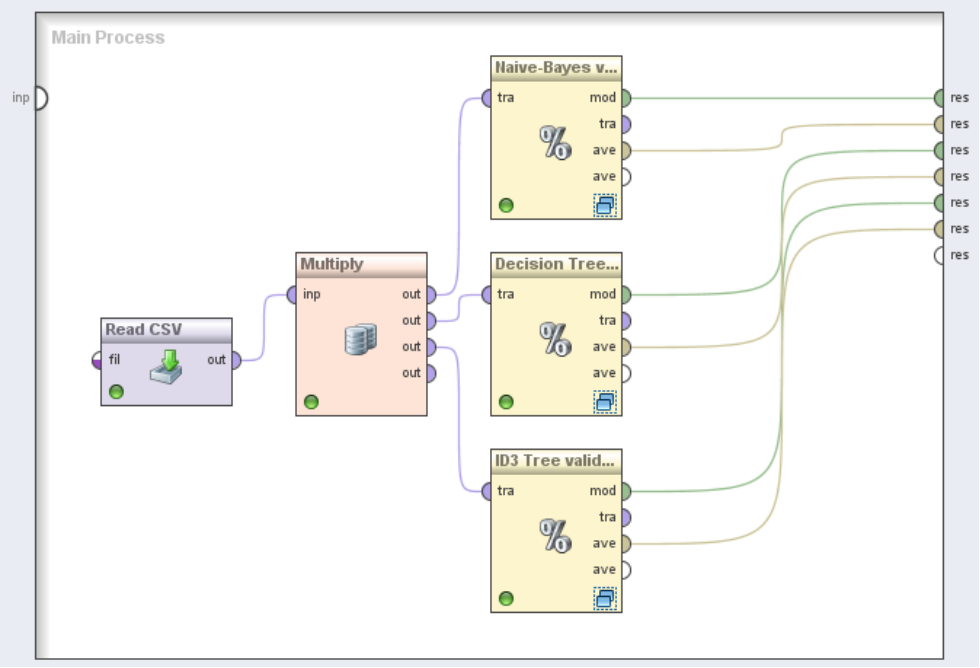


Figure 3 - RapidMiner project setup

3.1.1 NAIVE BAYES

The design of NaiveBayes validation box can be seen on figure 4.

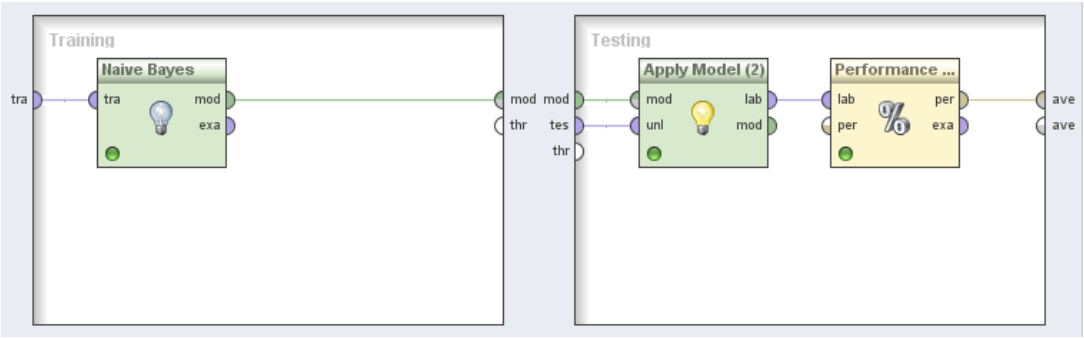


Figure 4 - RapidMiner Naive Bayes validation box in detail

```
PerformanceVector:
accuracy: 87.92% +/- 2.58% (mikro: 87.92%)
ConfusionMatrix:
True:   won     nowin
won:    1491    208
nowin:  178     1319
precision: 88.09% +/- 2.23% (mikro: 88.11%) (positive class: nowin)
```

3.1.2 DECISION TREE

The modelling process of the decision tree is the same as in case of Naive Bayes algorithm and can be seen on figure 5.

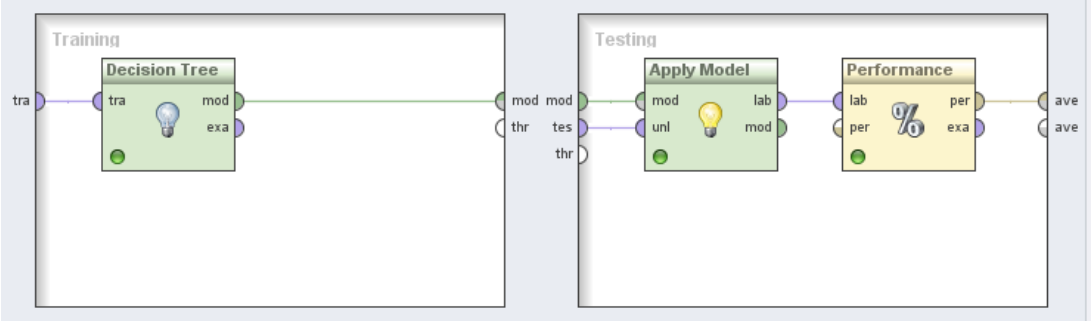


Figure 5 - RapidMiner Decision tree validation box in detail

```
PerformanceVector:
accuracy: 98.19% +/- 0.54% (mikro: 98.19%)
ConfusionMatrix:
True:   won     nowin
won:    1617    6
nowin:  52      1521
precision: 96.71% +/- 1.12% (mikro: 96.69%) (positive class: nowin)
```

3.1.3 ID3 TREE

ID3 algorithm is implemented as well as in Weka – the inner part of the validation box can be seen on figure 6.

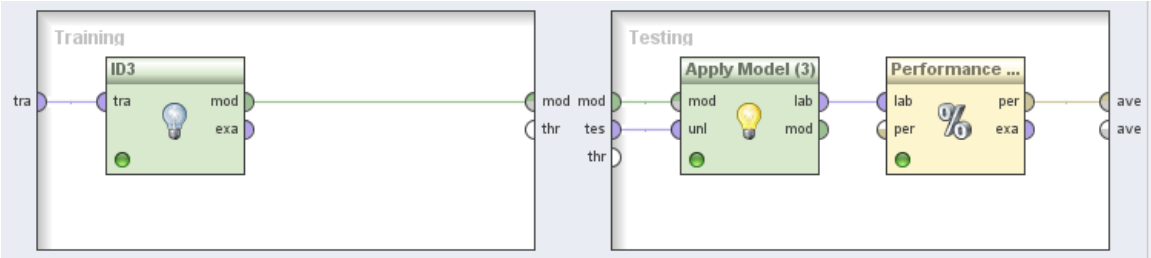


Figure 6 - RapidMiner ID3 tree validation box in detail

```
PerformanceVector:
accuracy: 99.59% +/- 0.49% (mikro: 99.59%)
ConfusionMatrix:
True:   won     nowin
won:    1663    7
nowin:  6       1520
precision: 99.61% +/- 0.67% (mikro: 99.61%) (positive class: nowin)
```

3.2 RESULTS

Algorithm	Correctly classified	Success rate
Naive Bayes	2810	87,92%
Decision Tree	3138	98,19%
ID3 Tree	3183	99,59%

ID3 Tree algorithm resulted again in the best result.

4 IBM SPSS MODELER

IBM SPSS Modeler is an extensive predictive analytics platform that is designed to bring predictive intelligence to decisions made by individuals, groups, systems and the enterprise. By providing a range of advanced algorithms and techniques that include text analytics, entity analytics, decision management and optimization. (6)

IBM SPSS Modeler is commercially licensed product.

4.1 MODELLING

Data mining flow has several components, some of them – such as data preparation can be processed automatically. Partition component enables to split data to training and testing sets afterwards models are made and result can be seen in bi-matrixes.

The project setup can be seen on figure 7.

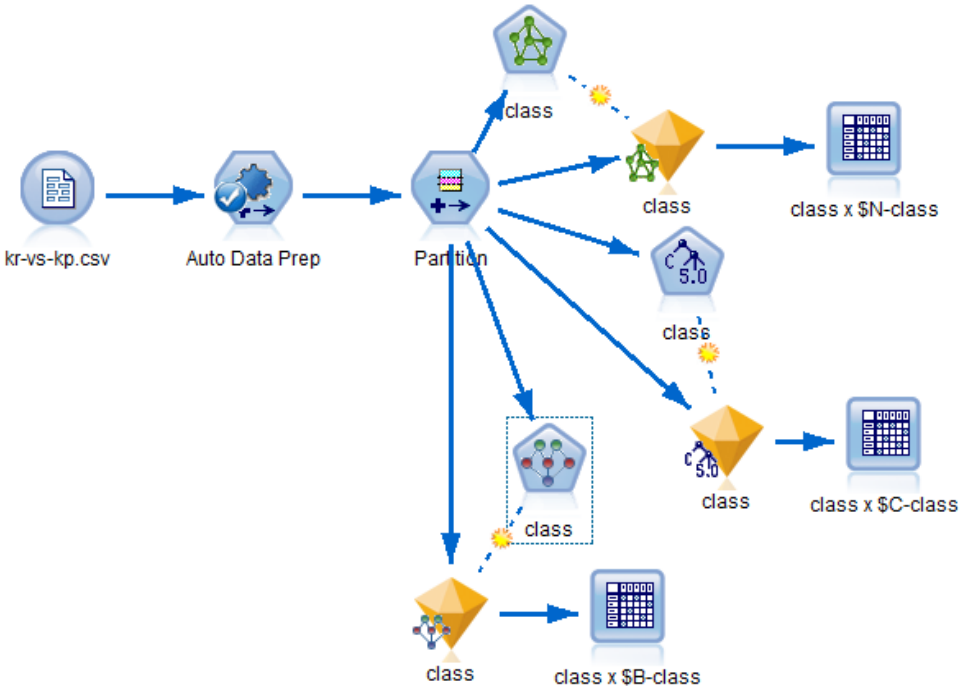


Figure 7 - IBM SPSS Modeler project setup

4.1.1 NEURAL NET

IBM SPSS Modeler enables users to build neural network model from nominal data, results can be seen in the result window and on the figure 8.

```
Matrix of class by $N-class
============================
        nowin   won
nowin   1510    17
won     95      1574

Correct         96.50%
```

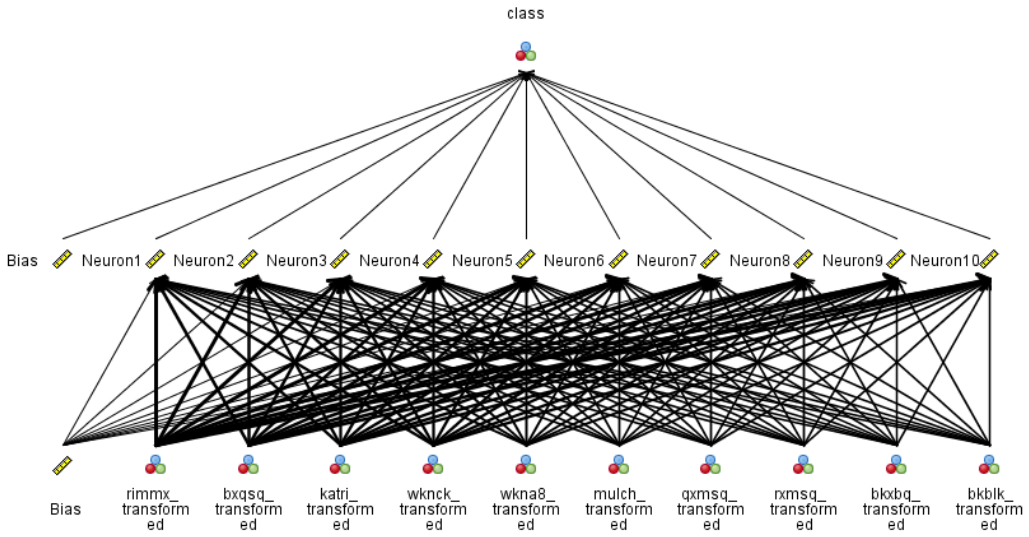


Figure 8 - IBM SPSS Neural network visualization

4.1.2 BAYES NET

Second model, that can be used for modelling this dataset in SPSS is Bayes Net. The software tool gives interesting visualizations that can be further enhanced – as it is shown on figure 9.

```
Matrix of class by $B-class
============================
        nowin   won
nowin   1412    115
won     117     1552

Correct         92.74%
```

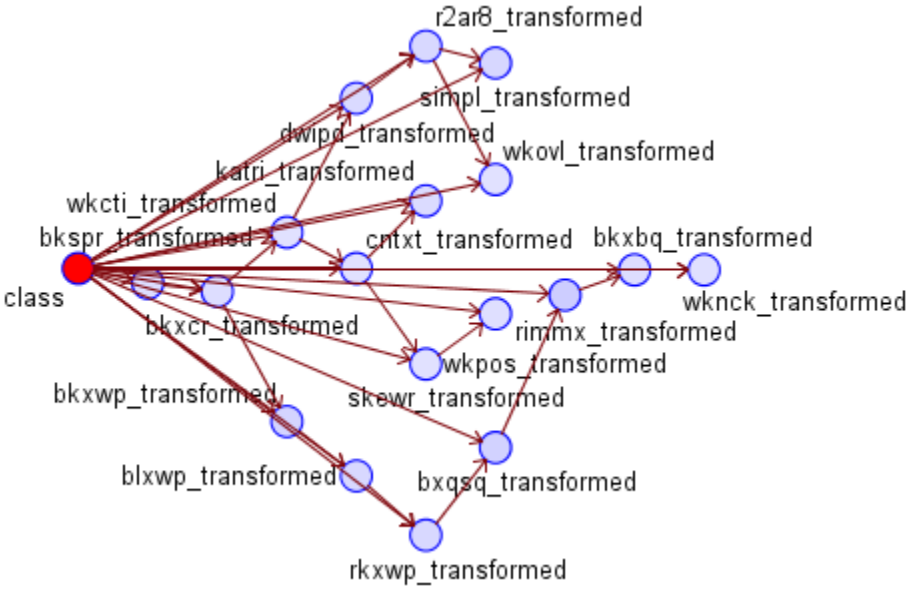


Figure 9 - IBM SPSS Bayes Net visualization

4.1.3 C5.0

C5.0 is a decision tree algorithm that is significantly faster than C4.5 with comparable results.

```
Matrix of class by $C-class
===========================
        nowin   won
nowin   1484    43
won     71      1598

Correct         96.43%
```

4.2 RESULTS

Algorithm	Correctly classified	Success rate
Neural Net	3084	96,50%
Bayes Net	2964	92,74%
C5.0	3082	96,43%

This time neural network provided the best result in classification.

5 SAS Enterprise Miner

SAS Enterprise Miner is an advanced analytics data mining tool intended to help users quickly develop descriptive and predictive models through a streamlined data mining process. Enterprise Miner's graphical interface enables users to logically move through the data mining process using the five-step SAS SEMMA approach: sampling, exploration, modification, modelling and assessment. (7)

SAS Enterprise Miner is commercially licensed product.

5.1 Modelling

SAS process pipeline is very similar to IBM SPSS includinfg boxes that are connected together. Process consist of file import – data partition – modelling – model comparison and finally reporter.

Import of the generated CSV file causes no problems, the only thing the users must be aware of is changing the role of the attribute in the Variables windows – as it can be seen on figure 10.

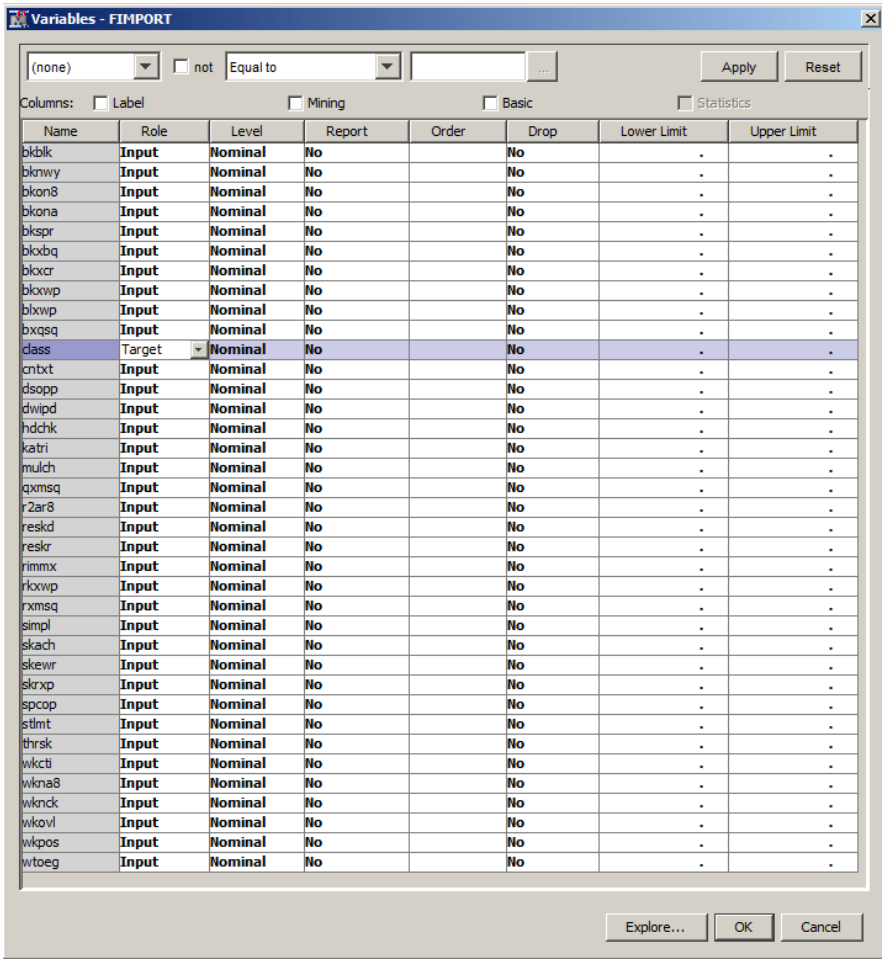


Figure 10 - SAS EM changing role of variable

The split between training and testing data can be set in Data Partition (in this case 60:40). Very appreciated feature of the SAS EM is the reporter box – all necessary statements are exported to a single PDF file.

Process pipeline can be seen on figure 11.

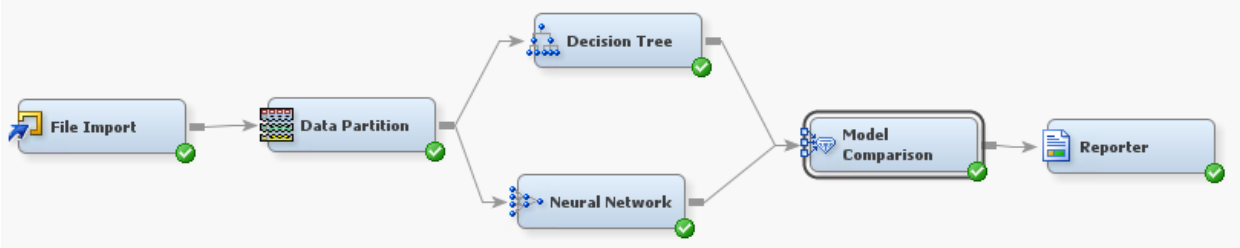


Figure 11 - SAS EM project setup

5.2 RESULTS

The comparison of neural network and decision tree algorithms can be seen in table or chart on figures 12 and 13.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Train: Average Squared Error	Train: Misclassification Rate	Train: Roc Index	Train: Gini Coefficient
Y	Neural	Neural	Neural Network	class		0.003597	0.002610	1.000	1.000
	Tree	Tree	Decision Tree	class		0.051232	0.056889	0.962	0.923

Figure 12 - SAS EM model comparison table

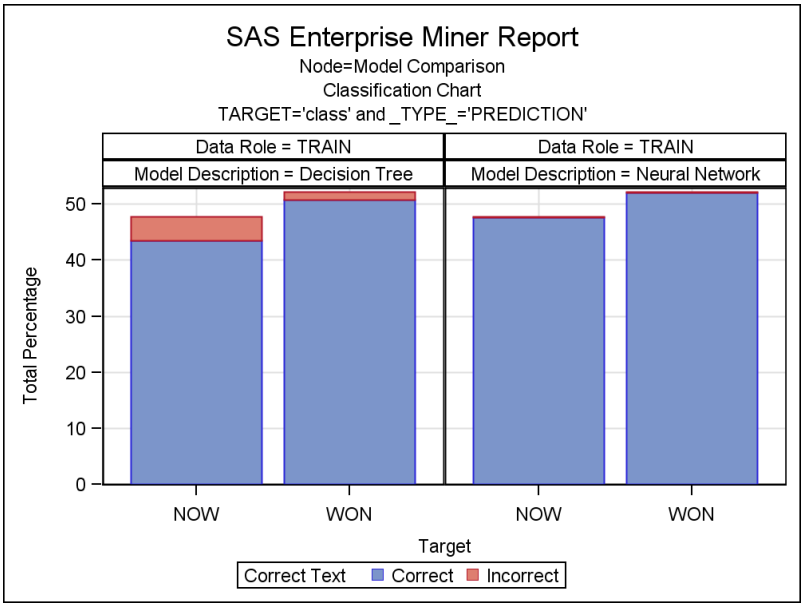


Figure 13 - SAS EM model comparison chart

Algorithm	Correctly classified	Success rate
Neural Network	3188	99,74%
Decission Tree	3014	94,31%

Neural network algorithm provided an excellent result with success rate 99,74%.

6 CONCLUSION

The aim of this project was fulfilled – the dataset KRKPA7 has been analysed by using 4 different data mining software tools as described in previous chapters. Decision trees showed great classification performance throughout all 4 tools, the success rate of neural networks was slightly lower in case of IBM SPSS but neural network model generated by SAS EM outperformed all other models.

Interesting fact was that commercially available tools were not better than open source. What is more I found RapidMiner the most intuitive and full of various features. That is why I would recommend RapidMiner to other data analysts.

REFERENCES

1. University of Perugia. krkpa7 - An educational decision-tree classifier. *Google Project Hosting.* [Online] 2015. https://code.google.com/p/sprinkler/source/browse/progetti/DataMining/data/krkpa7/krkpa7.names?r=48.

2. Rook and pawn versus rook endgame - Wikipedia, the free encyclopedia. *Wikipedia.* [Online] 2015. https://en.wikipedia.org/wiki/Rook_and_pawn_versus_rook_endgame.

3. Michie, D. Consciousness as an Engineering Issue. *Journal of Consciousness Studies.* 1995.

4. Machine Learning Group at the University of Waikat. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. *The University of Waikato.* [Online] 2015. http://www.cs.waikato.ac.nz/ml/weka/.

5. Wikipedia. RapidMiner. *Wikipedia, the free encyclopedia.* [Online] 2015. https://en.wikipedia.org/wiki/RapidMiner.

6. IBM. IBM SPSS Modeler. *IBM - United States.* [Online] 2015. http://www-01.ibm.com/software/analytics/spss/products/modeler/index.html.

7. Reifer, Abie. How SAS Enterprise Miner simplifies the data mining process. *TechTarget.com.* [Online] 2015. http://searchbusinessanalytics.techtarget.com/feature/How-SAS-Enterprise-Miner-simplifies-the-data-mining-process.

TABLE OF FIGURES