

Week 08

Decision Tree algorithms in Rapidminer

- **Dataset:** playtennis and risk
week 08 – datasets.xls
- **Algorithms:** ID3, CHAID, CART, C4.5

play tennis nominal example

parameters

The screenshot displays the RapidMiner interface with a workflow in the 'Main Process' area. The workflow consists of a 'Read Excel' operator connected to an 'ID3' operator. The 'ID3' operator's parameters are highlighted in a red box:

- criteria: gain_ratio
- minimal size f...: 2
- minimal leaf si...: 2
- minimal gain: 0.1

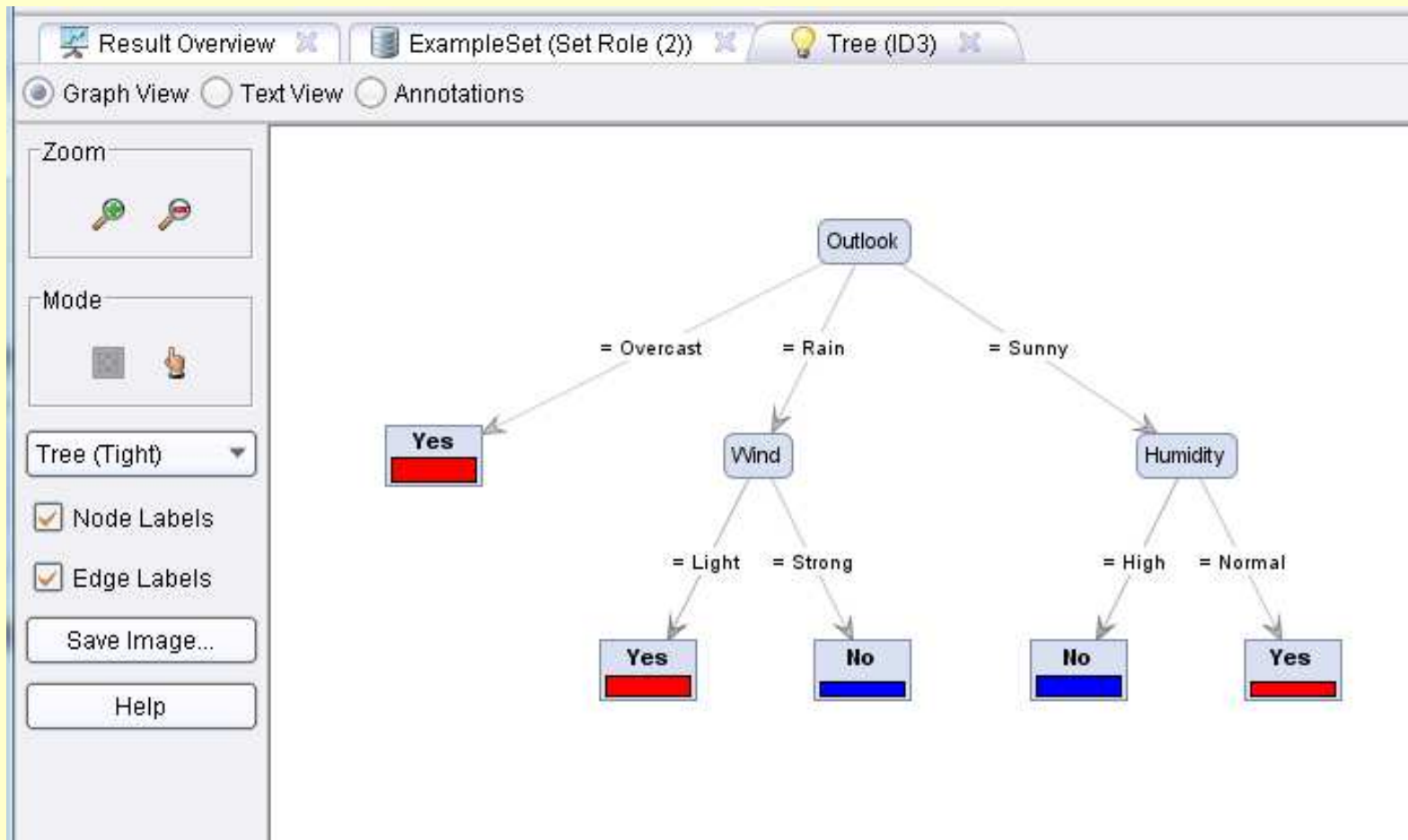
Below the parameters, a 'Help' window is open, providing a description of the ID3 algorithm:

nominal attributes only. Decision trees are powerful classification methods which often can also easily be understood. This decision tree learner works similar to Quinlan's ID3.

Input

- training set: expects: ExampleSet

Output



Result Overview × ExampleSet (Set Role (2)) × Tre

Graph View Text View Annotations

Tree

```
Outlook = Overcast: Yes {No=0, Yes=4}
Outlook = Rain
|   Wind = Light: Yes {No=0, Yes=3}
|   Wind = Strong: No {No=2, Yes=0}
Outlook = Sunny
|   Humidity = High: No {No=3, Yes=0}
|   Humidity = Normal: Yes {No=0, Yes=2}
```

play tennis **numeric** example

The screenshot shows the RapidMiner interface. The main workspace contains a workflow with a 'Read Excel' operator followed by a 'Decision Tree' operator. A red arrow points to the 'Decision Tree' operator. The 'Parameters' panel on the right is open, showing settings for the 'Decision Tree' operator. The 'criterion' is set to 'gain_ra...', 'minimal size' is 4, 'minimal leaf' is 2, 'minimal gain' is 0.1, 'maximal dep.' is 20, 'confidence' is 0.1, and 'number of pr...' is 3. There are also checkboxes for 'no pre pruning' and 'no pruning', both of which are unchecked.

Parameters

Decision Tree

criterion gain_ra...

minimal size... 4

minimal leaf... 2

minimal gain 0.1

maximal dep... 20

confidence 0.1

number of pr... 3

no pre pruning

no pruning

Comment

Help

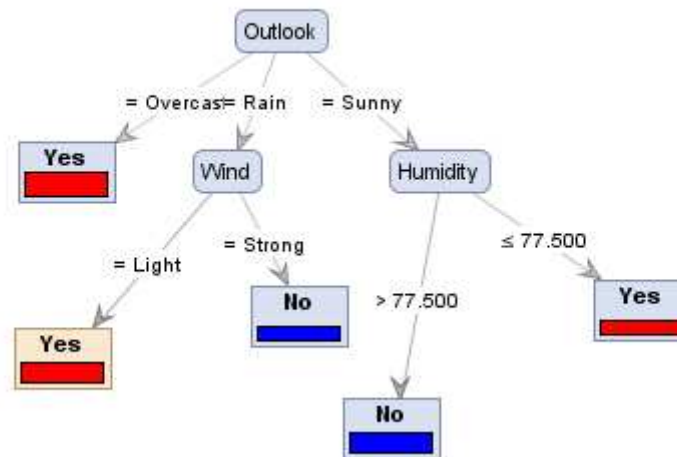
Decision Tree

Synopsis

Generates decision trees to classify nominal data.

Description

ID3 does not support numerical attributes



Decision Tree

criterion: gain_ratio
 minimal size for split: 4
 minimal leaf size: 3
 minimal gain: 0.1
 maximal depth: 20
 confidence: 0.1
 number of prepruning altern...: 3

no pre pruning
 no pruning

Mode

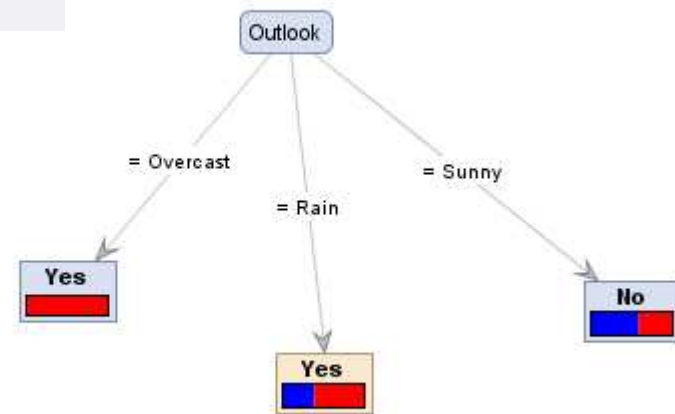
Tree

Node Labels
 Edge Labels

Save Image...

ExampleSet (Set Role (2)) x I Tree (Decision Tree) x

Annotations



risk example

ID	AGE	INCOME	GENDER	MARITAL	NUMKIDS	NUMCARDS	HOWPAID	MORTGAGE	STORECAR	LOANS	RISK
100756	44	59944	m	married	1	2	monthly	y	2	0	good risk
100668	35	59692	m	married	1	1	monthly	y	1	0	bad loss
100418	34	59508	m	married	1	1	monthly	y	2	1	good risk
100416	34	59463	m	married	0	2	monthly	y	1	1	bad loss
100590	39	59393	f	married	0	2	monthly	y	1	0	good risk
100657	41	59276	m	married	1	2	monthly	y	1	1	good risk
100702	42	59201	m	married	0	1	monthly	y	2	0	good risk
100319	31	59193	f	married	1	2	monthly	y	1	1	good risk
100666	28	59179	m	married	1	1	monthly	y	2	1	bad loss
100389	30	59036	m	married	1	1	monthly	y	2	1	good risk
100758	38	58914	m	married	0	1	monthly	y	1	1	bad profit
100695	36	58878	f	married	1	1	monthly	y	1	0	bad profit
100698	42	58785	f	married	0	2	monthly	y	1	0	good risk
100769	44	58529	m	married	0	1	monthly	y	1	0	bad loss
100376	33	58505	f	married	0	2	monthly	y	1	0	good risk
100796	45	58381	m	married	1	1	monthly	y	1	0	good risk
100414	34	58026	m	married	0	1	monthly	y	2	0	good risk
100354	32	57718	m	married	1	2	monthly	y	1	1	bad profit
100452	35	57689	m	married	1	1	monthly	y	2	1	good risk
100567	38	57683	f	married	1	1	monthly	y	2	1	bad loss

RapidMiner@feup-i023

File Edit Process Tools View Help

Overview Process XML

Parameters

Decision Tree

Read Excel Split Data Decision Tree Apply Model Performance

Performance Measurement (19)

- Classification and Regression (6)
 - Performance (Regression)
 - Performance (Classification)
 - Performance (Binominal Classification)
 - Performance (Costs)
 - Performance (Ranking)
 - Performance (Support Vector Count)
- Attributes (3)
- Clustering (5)
 - Performance
 - Extract Performance
 - Combine Performances
 - Performance (User-Based)
 - Performance (Min-Max)

Decision Tree

critterion information_...

minimal size for s... 4

minimal leaf size 2

minimal gain 0.1

maximal depth 20

confidence 0.25

number of prepru... 3

no pre pruning

no pruning

Decision Tree

Synopsis

Generates decision trees to classify nominal data.

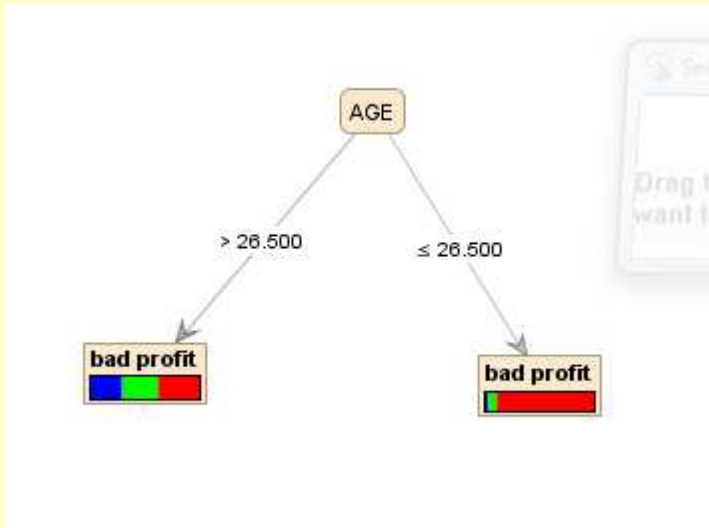
Description

This operator learns decision trees from both nominal and numerical

Problems Log

No problems found

Message	Fixes	Location



accuracy: 58.22%

	true good risk	true bad loss	true bad profit	class precision
pred. good risk	0	0	0	0.00%
pred. bad loss	0	0	0	0.00%
pred. bad profit	253	263	719	58.22%
class recall	0.00%	0.00%	100.00%	

RapidMiner@feup-i023

File Edit Process Tools View Help

Overview Process XML

Process

Repositories Operators

performan

- Import (1)
- Export (1)
- Evaluation (19)
 - Performance Measurement
 - Classification and Regression
 - Performance (Regression)
 - Performance (Classification)
 - Performance (Binomial)
 - Performance (Costs)
 - Performance (Ranking)
 - Performance (Support)
 - Attributes (3)
 - Clustering (5)
 - Performance
 - Extract Performance
 - Combine Performances
 - Performance (User-Based)
 - Performance (Min-Max)

Read Excel Split Data Decision Tree Apply Model Performance

Parameters

Decision Tree

criteria informatio...

minimal size fo... 4

minimal leaf si... 2

minimal gain 0.01

maximal depth 20

confidence 0.25

number of pre... 3

no pre pruning

Problems Log

No problems found

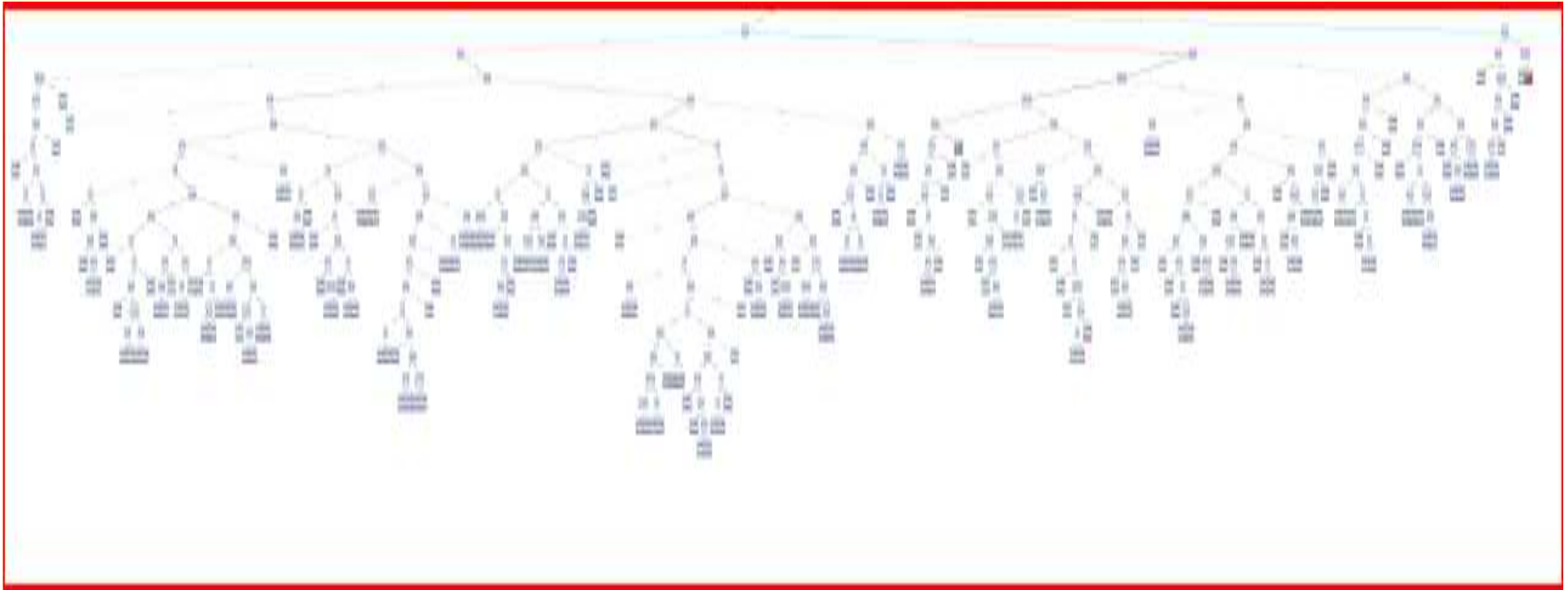
Message	Fixes	Location
---------	-------	----------

Decision Tree

Synopsis

accuracy: 56.68%

	true good risk	true bad loss	true bad profit	class precision
pred. good risk	162	195	190	29.62%
pred. bad loss	33	17	8	29.31%
pred. bad profit	58	51	521	82.70%
class recall	64.03%	6.46%	72.46%	



The image displays a software interface for a data mining process. On the left, a 'Main Process' window contains a workflow diagram with the following steps: 'Read Excel' (input 'inp', output 'out'), 'Set Role' (input 'exa', output 'exa', role 'on'), 'Set Role (2)' (input 'exa', output 'exa', role 'on'), 'Split Data' (input 'exa', outputs 'par', 'par', 'par'), 'Decision Tree' (input 'tra', outputs 'mod', 'exa'), 'Apply Model' (inputs 'mod', 'unl', outputs 'lab', 'mod'), and 'Performance' (inputs 'lab', 'per', outputs 'per', 'exa'). On the right, a 'Parameters' window for the 'Decision Tree' model is shown. The 'criterion' parameter is set to 'information_gain'. A red rectangle highlights a section of the parameter list, which includes 'minimal size for split', 'minimal leaf size', 'minimal gain', 'maximal depth', 'confidence', and 'number of prepruning alternatives'. Below these are two unchecked checkboxes: 'no pre pruning' and 'no pruning'.

Process XML

Parameters

Decision Tree

criterion information_gain

minimal size for split

minimal leaf size

minimal gain

maximal depth

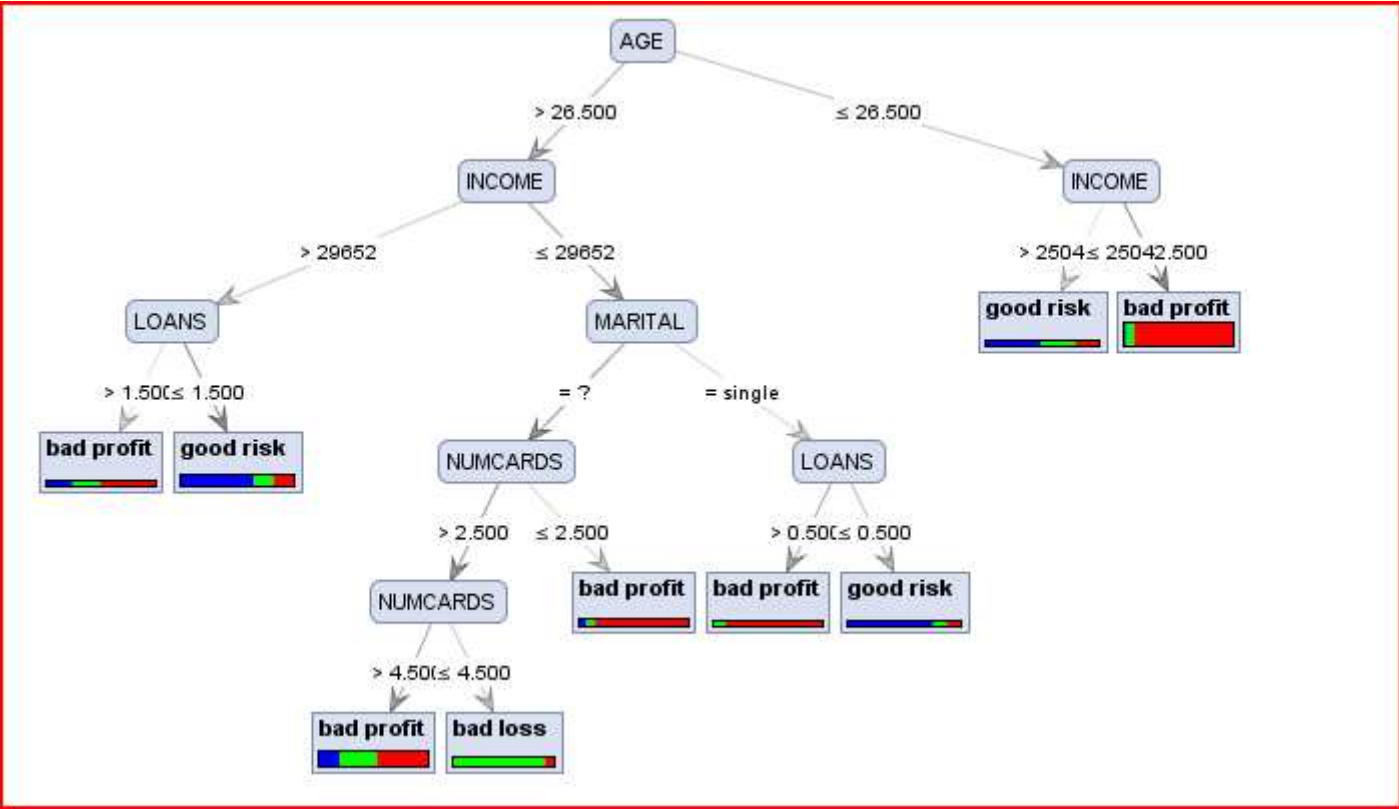
confidence

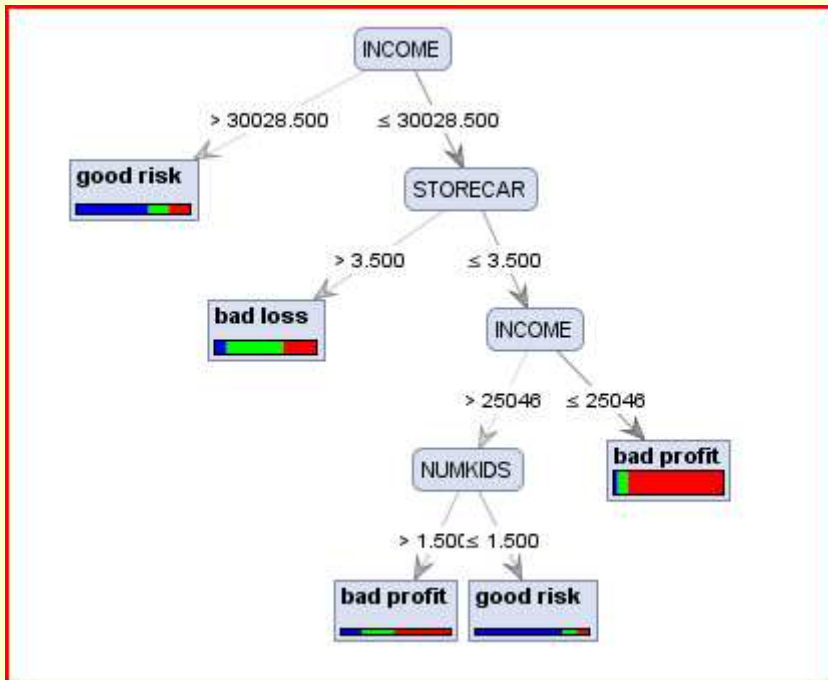
number of prepruning alternatives

no pre pruning

no pruning

accuracy: 60.00%				
	true good risk	true bad loss	true bad profit	class precision
pred. good risk	227	205	205	35.64%
pred. bad loss	0	0	0	0.00%
pred. bad profit	26	58	514	85.95%
class recall	89.72%	0.00%	71.49%	





Decision Tree

criterion:

minimal size for split:

minimal leaf size:

minimal gain:

maximal depth:

confidence:

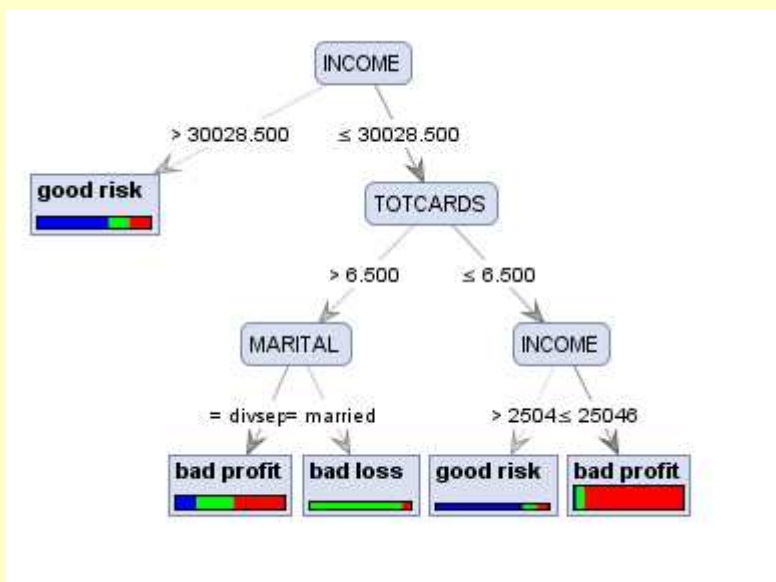
number of prepruning alternatives:

no pre pruning

no pruning

accuracy: 74.98%

	true good risk	true bad loss	true bad profit	class precision
pred. good risk	171	27	43	70.95%
pred. bad loss	43	156	77	56.52%
pred. bad profit	39	80	599	83.43%
class recall	67.59%	59.32%	83.31%	



accuracy: 75.79%

	true good risk	true bad loss	true bad profit	class precision
pred. good risk	171	27	43	70.95%
pred. bad loss	5	101	12	85.59%
pred. bad profit	77	135	664	75.80%
class recall	67.59%	38.40%	92.35%	