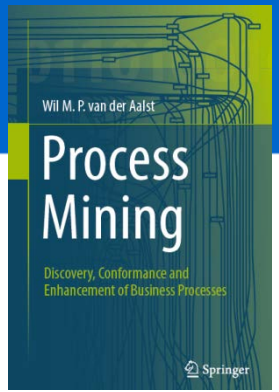
Process Mining: Data Science in Action

Applying Decision Trees

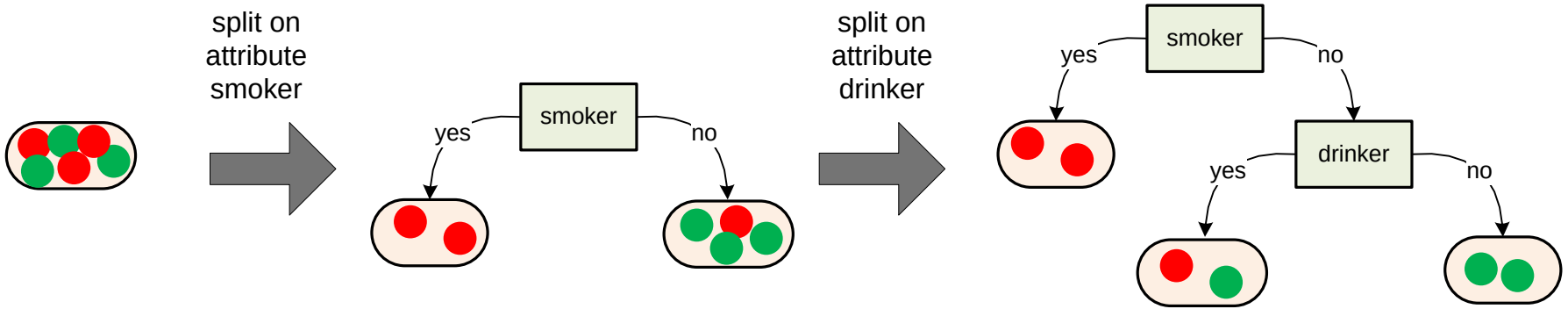
prof.dr.ir. Wil van der Aalst
www.processmining.org



Technische Universiteit Eindhoven University of Technology

Where innovation starts

Decision tree learning



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Iteratively reduce the overall level of uncertainty (entropy) using label splitting until no significant information gain is possible.

Example: 160 students (100 pass, 60 fail)

What matters?

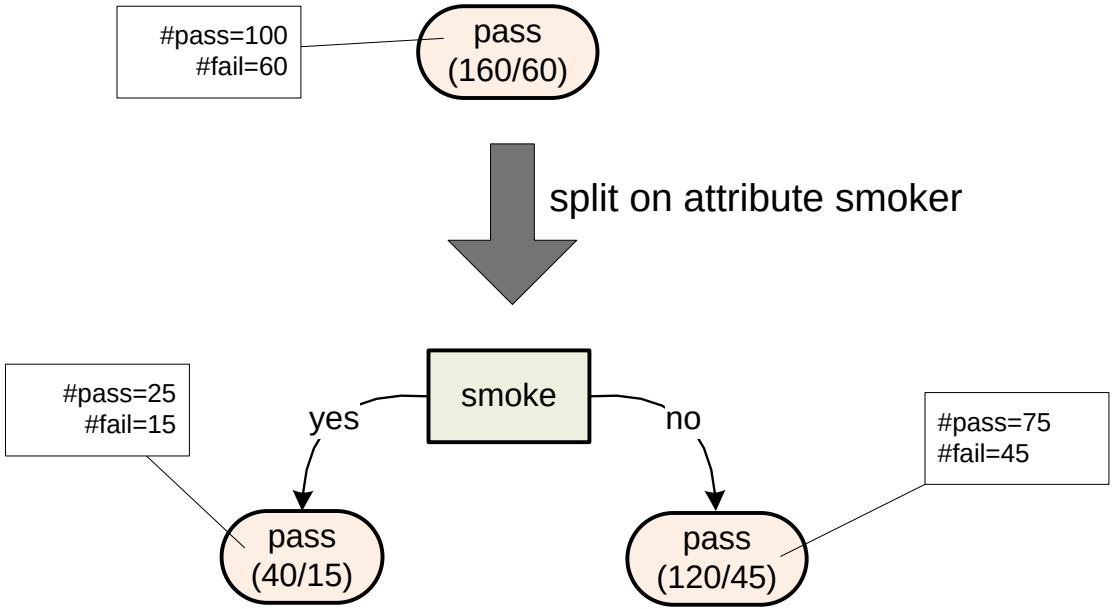
attending lectures?

gender?

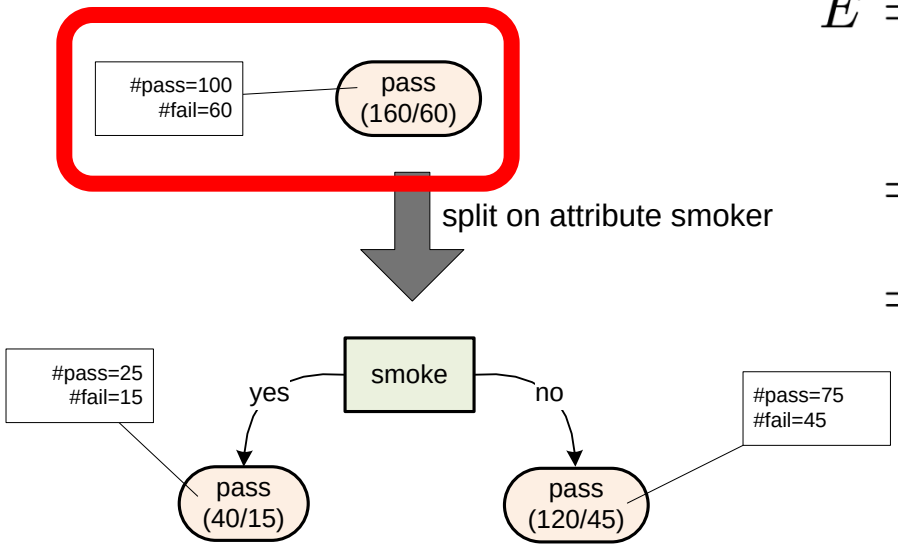
smoking?

TU/e

Question: What is the information gain?



Answer: Entropy of root node

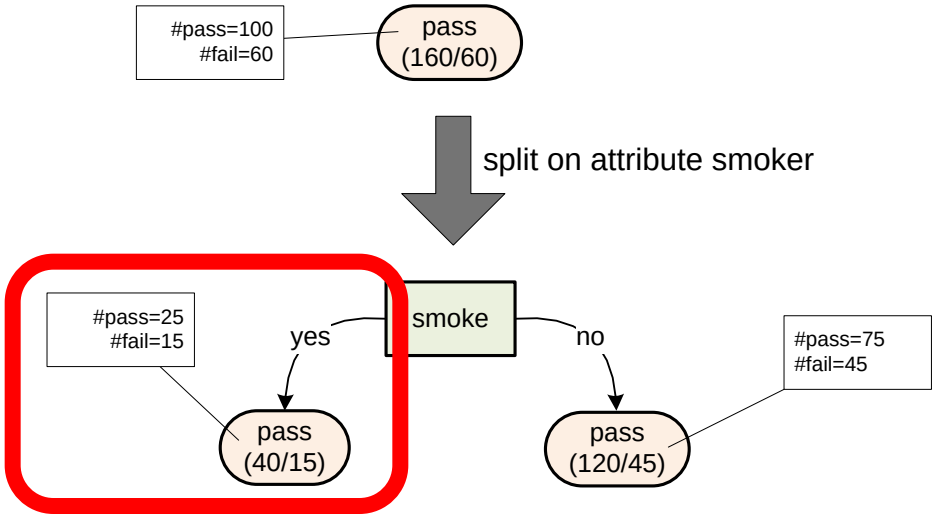


$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$

$$= 0.9544$$

Answer: Entropy of smokers



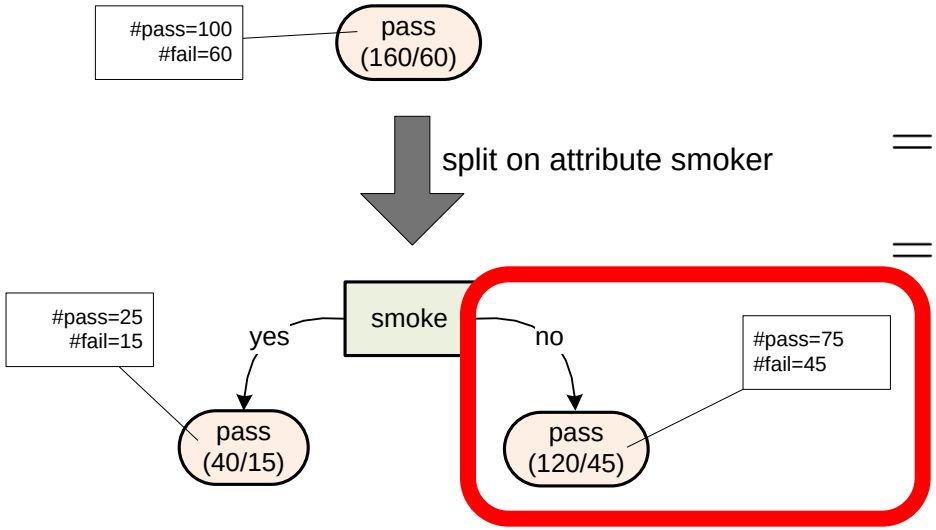
$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{25}{40}\log_2(\frac{25}{40}) + \frac{15}{40}\log_2(\frac{15}{40}))$$

$$= 0.9544$$

TU/e

Answer: Entropy of non-smokers

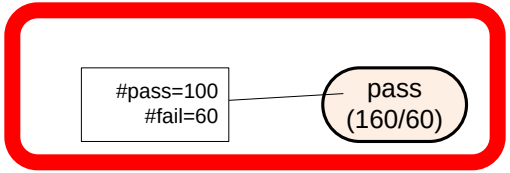


$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{75}{120}\log_2(\frac{75}{120}) + \frac{45}{120}\log_2(\frac{45}{120}))$$

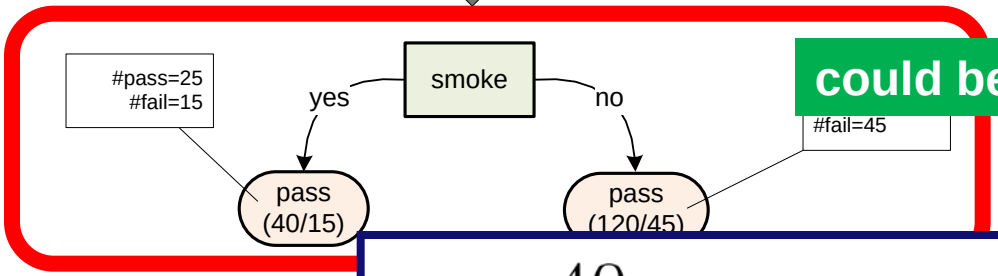
$$= 0.9544$$

Answer: No information gain



$$E = \frac{160}{160} \times 0.9544 = 0.9544$$

split on attribute smoker

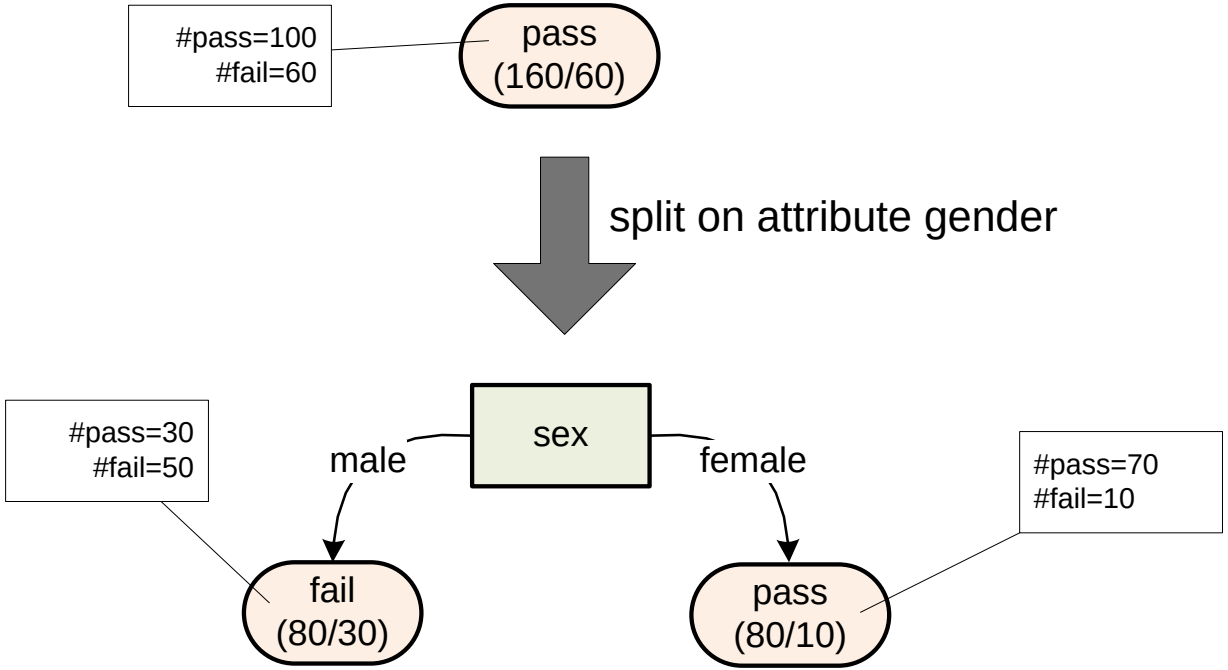


information gain = 0

could be seen without computation

$$E = \frac{40}{160} \times 0.9544 + \frac{120}{160} \times 0.9544 = 0.9544$$

Question: What is the information gain?



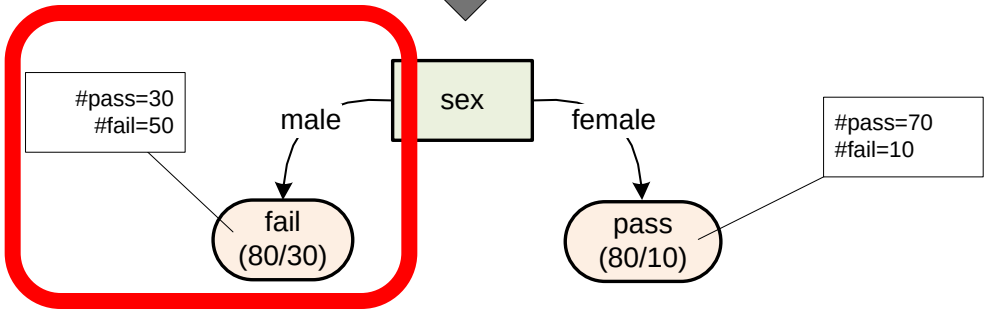
Answer: Entropy of male students

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$

$$= 0.9544$$

split on attribute gender



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{30}{80}\log_2(\frac{30}{80}) + \frac{50}{80}\log_2(\frac{50}{80}))$$

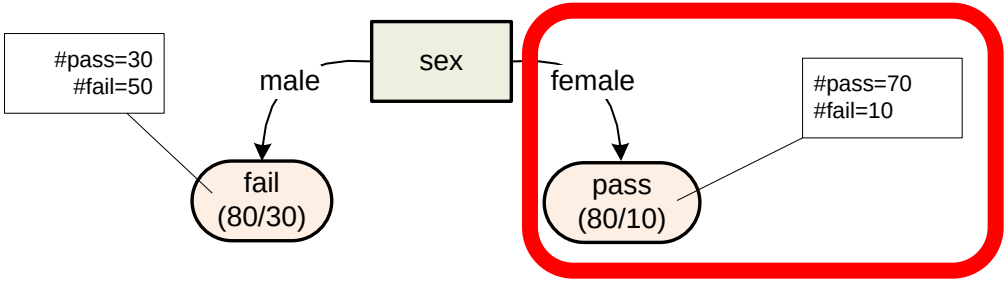
$$= 0.9544$$

TU/e

Answer: Entropy of female students

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$
$$= 0.9544$$

split on attribute gender



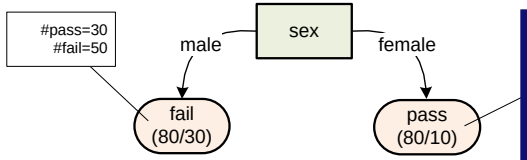
$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{70}{80}\log_2(\frac{70}{80}) + \frac{10}{80}\log_2(\frac{10}{80}))$$
$$= 0.5436$$

TU/e

Answer: Information gain

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$
$$= 0.9544$$

split on attribute gender



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{30}{80}\log_2(\frac{30}{80}) + \frac{50}{80}\log_2(\frac{50}{80}))$$
$$= 0.9544$$

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{70}{80}\log_2(\frac{70}{80}) + \frac{10}{80}\log_2(\frac{10}{80}))$$
$$= 0.5436$$

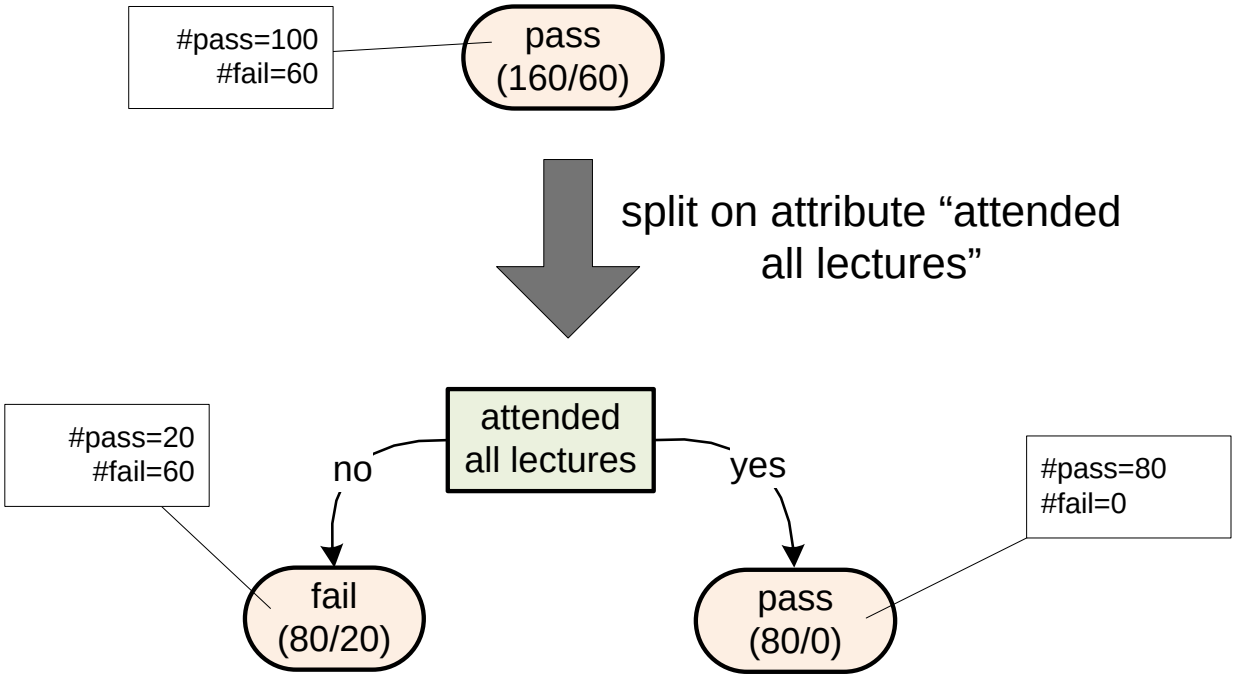
$$E = \frac{160}{160} \times 0.9544 = 0.9544$$

information gain = 0.2054

$$E = \frac{80}{160} \times 0.9544 + \frac{80}{160} \times 0.5436 = 0.7490$$

TU/e

Question: What is the information gain?



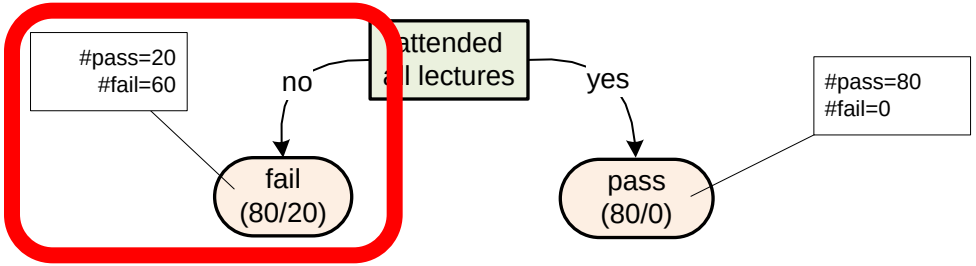
Answer: Entropy of missing students

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$

$$= 0.9544$$

split on attribute “attended all lectures”



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{20}{80}\log_2(\frac{20}{80}) + \frac{60}{80}\log_2(\frac{60}{80}))$$

$$= 0.8113$$

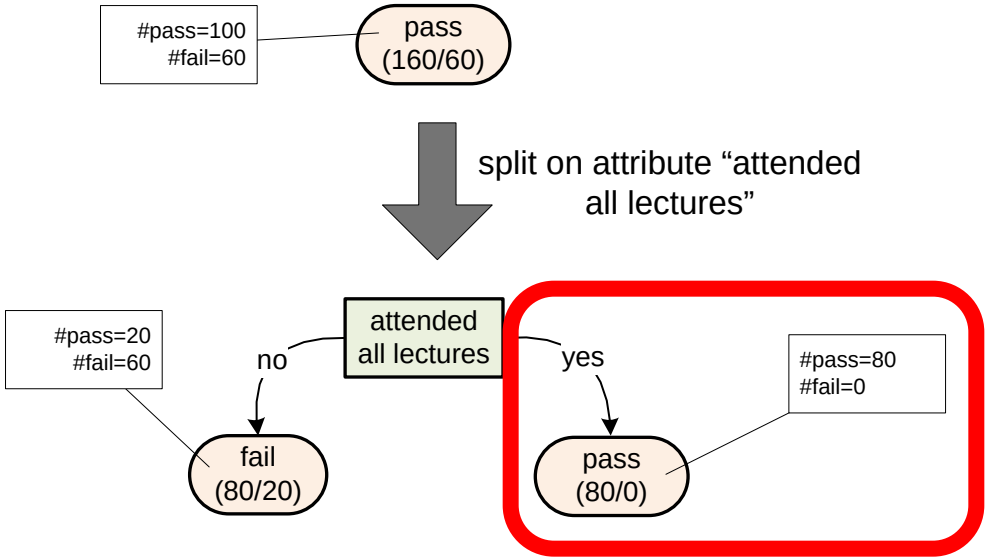
TU/e

Answer: Entropy of attending students

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$

$$= 0.9544$$



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -(\frac{80}{80}\log_2(\frac{80}{80}))$$

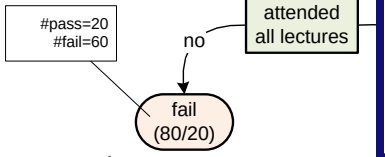
$$= 0$$

TU/e

Answer

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{100}{160}\log_2(\frac{100}{160}) + \frac{60}{160}\log_2(\frac{60}{160}))$$
$$= 0.9544$$

split on attribute “attended all lectures”



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{20}{80}\log_2(\frac{20}{80}) + \frac{60}{80}\log_2(\frac{60}{80}))$$
$$= 0.8113$$

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$
$$= -(\frac{80}{80}\log_2(\frac{80}{80}))$$
$$= 0$$

$$E = \frac{160}{160} \times 0.9544 = 0.9544$$

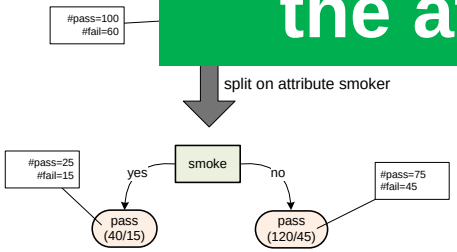
information gain = 0.5488

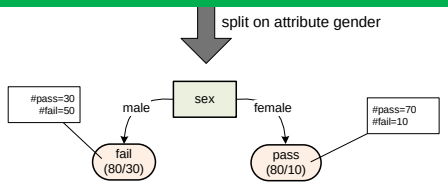
$$E = \frac{80}{160} \times 0.8113 + \frac{80}{160} \times 0 = 0.4056$$

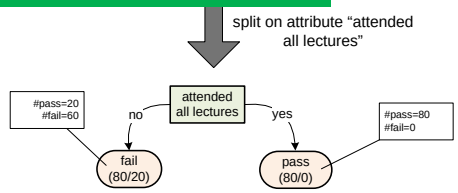
TU/e

Comparing information gains

So we should split the root node on the attribute "attend all lectures"!







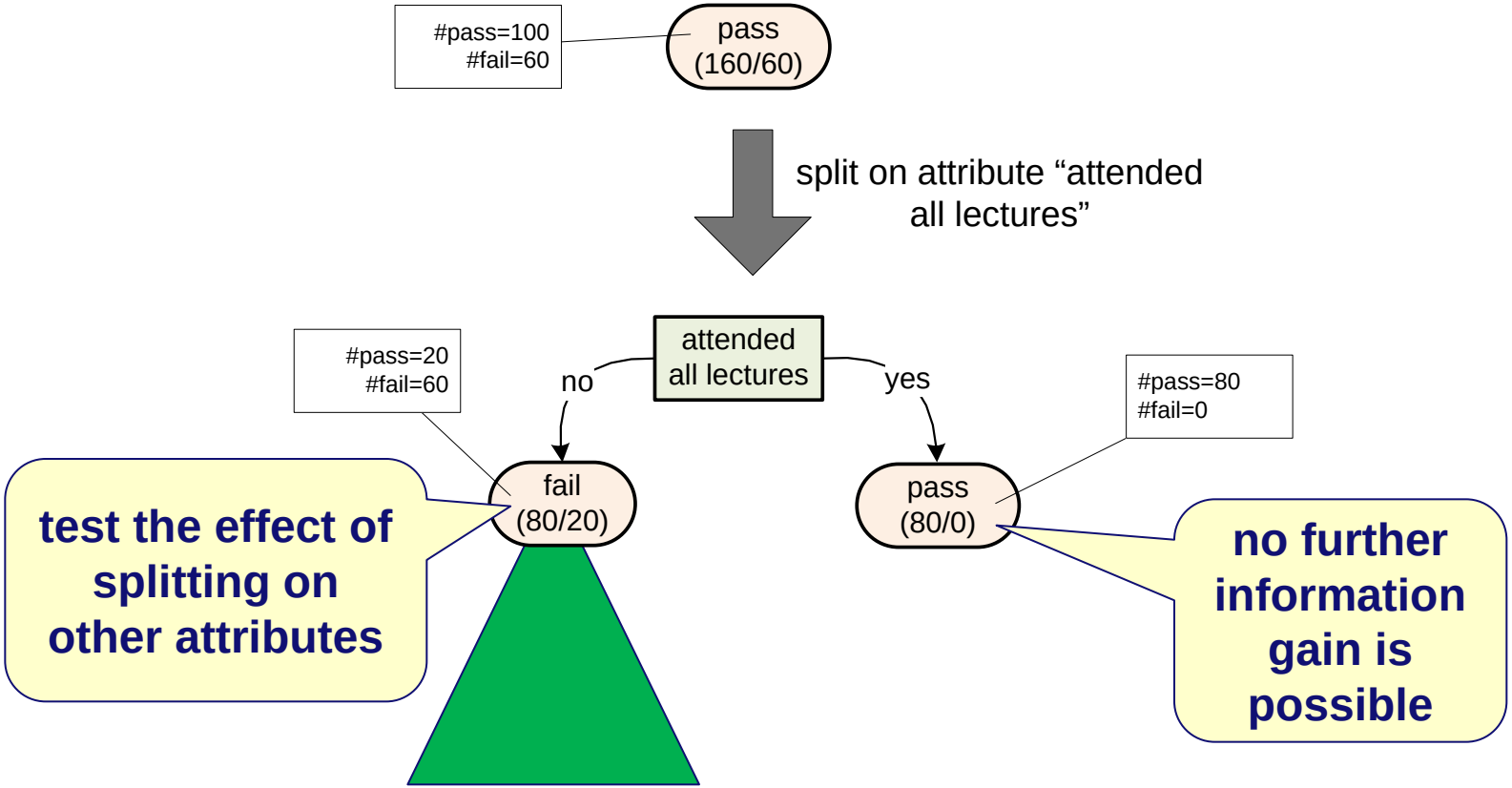
information gain = 0

information gain = 0.2054

information gain = 0.5488

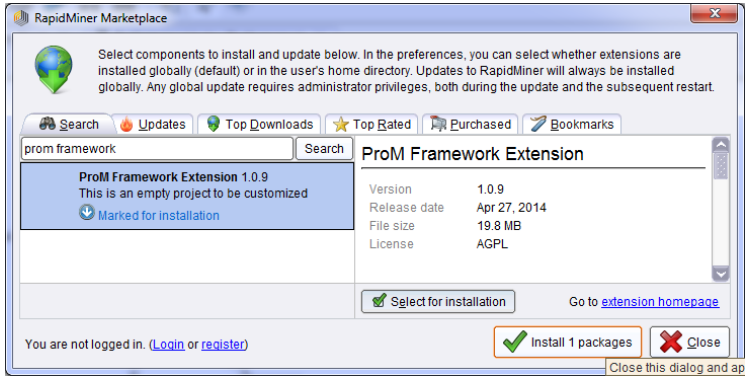
TU/e

Iterate until no significant gain is possible



RapidMiner
(installation is optional)

- An integrated extendible environment for machine learning, data mining, text mining, and predictive analytics.
- RapidMiner Marketplace also provides a ProM extension for process mining.
- Commercial and open-source versions of the software.



Decision trees in RapidMiner

gender	age	smoker	car brand	claim
female	47	yes	Volvo	no
male	31	no	Alfa Romeo	yes
ma				
ma				
male	44	no	BMW	no
fema				
ma				
...	...	..	...	...

CSV file contains information about 999 customers of an insurance company.

The company wants to know which customers claim insurance.

TU/e

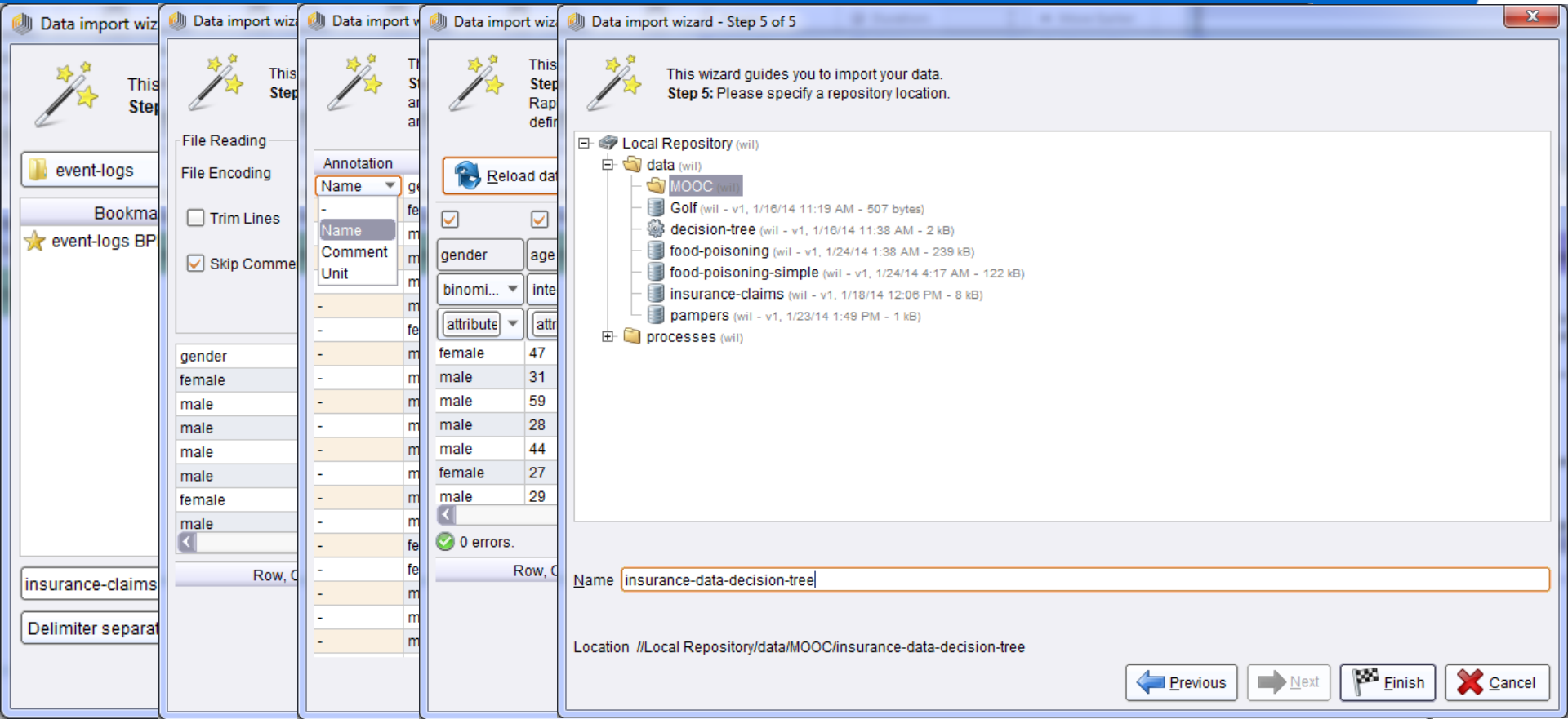
Decision trees in RapidMiner

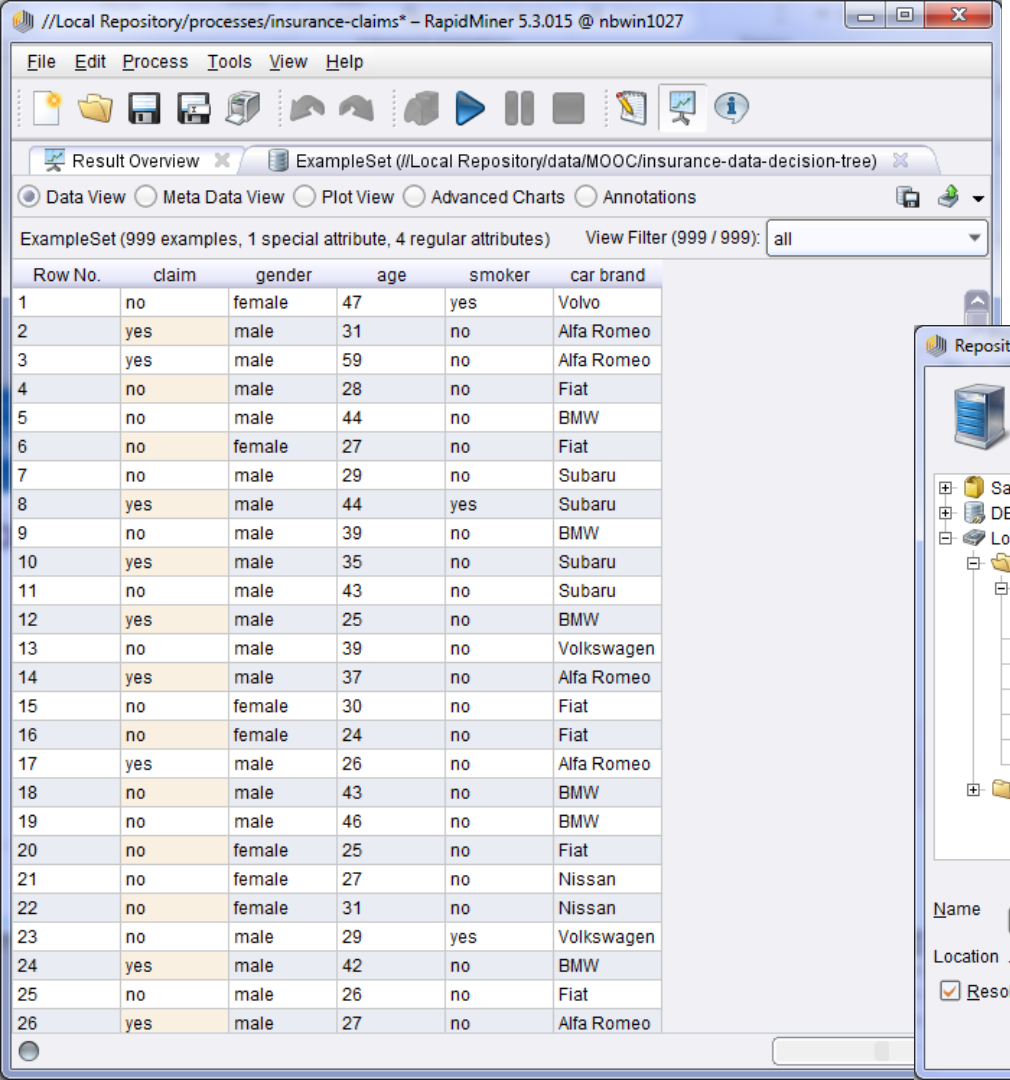
gender	age	smoker	car brand	claim
female	47	yes	Volvo	no
male	31	no	Alfa Romeo	yes
male	59	no	Alfa Romeo	yes
male	28	no	Fiat	no

Response variable (dependent variable): claim.

Predictor variables (independent variables): gender, age, smoker, car brand.

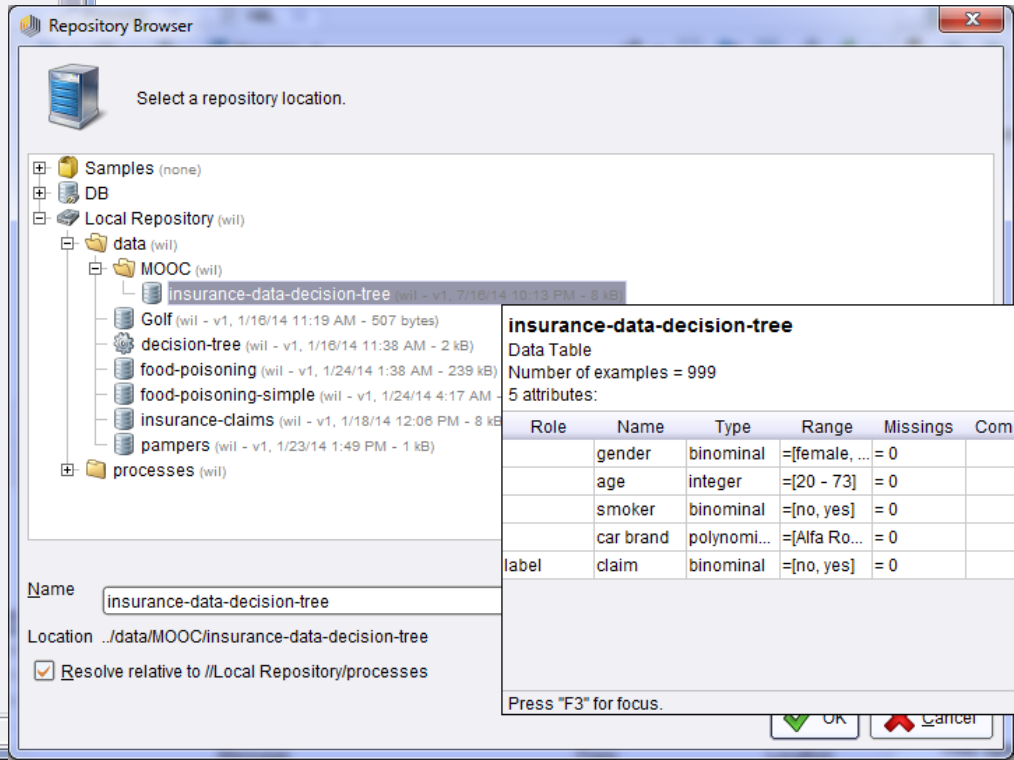
Data in RapidMiner

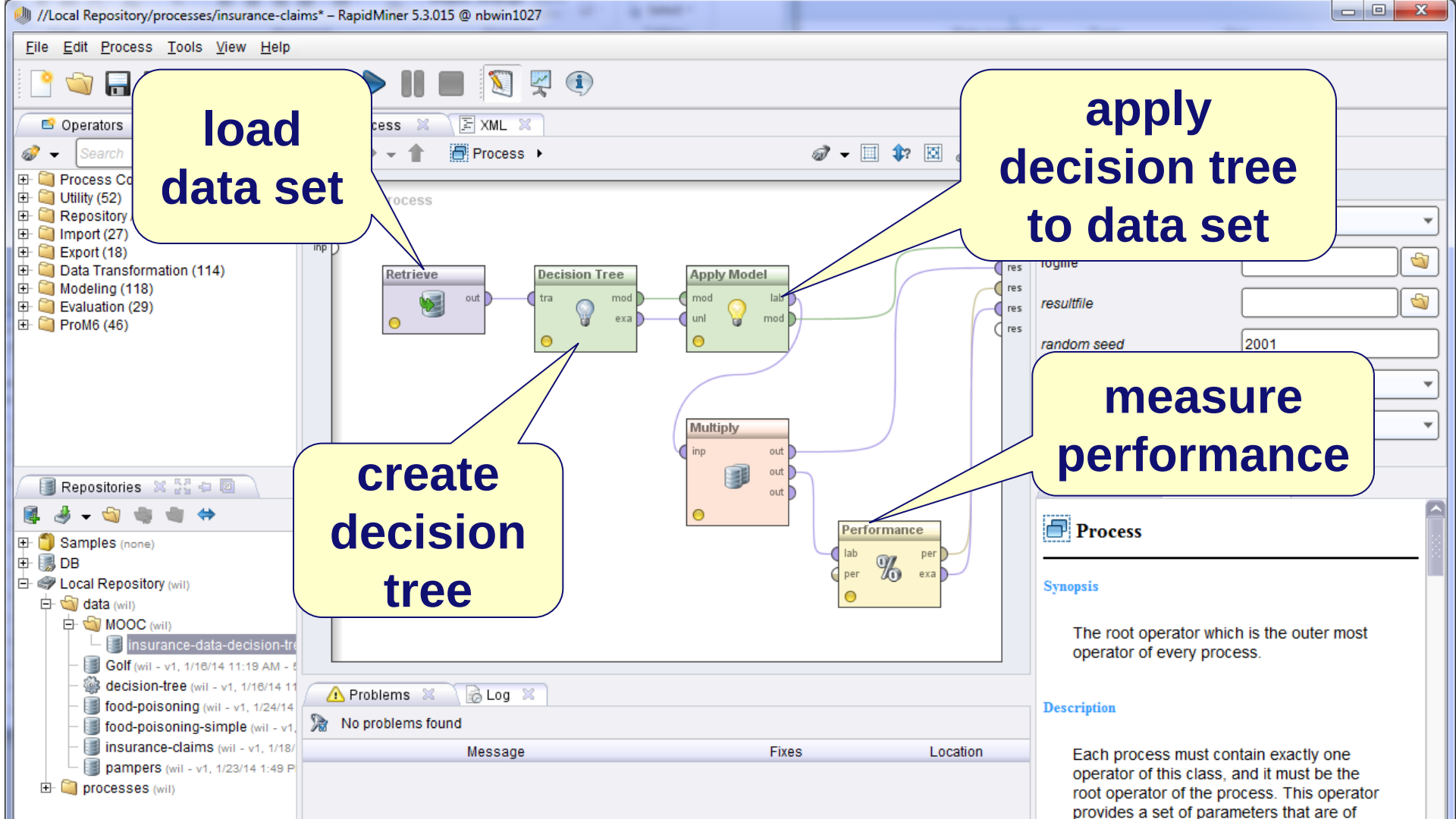




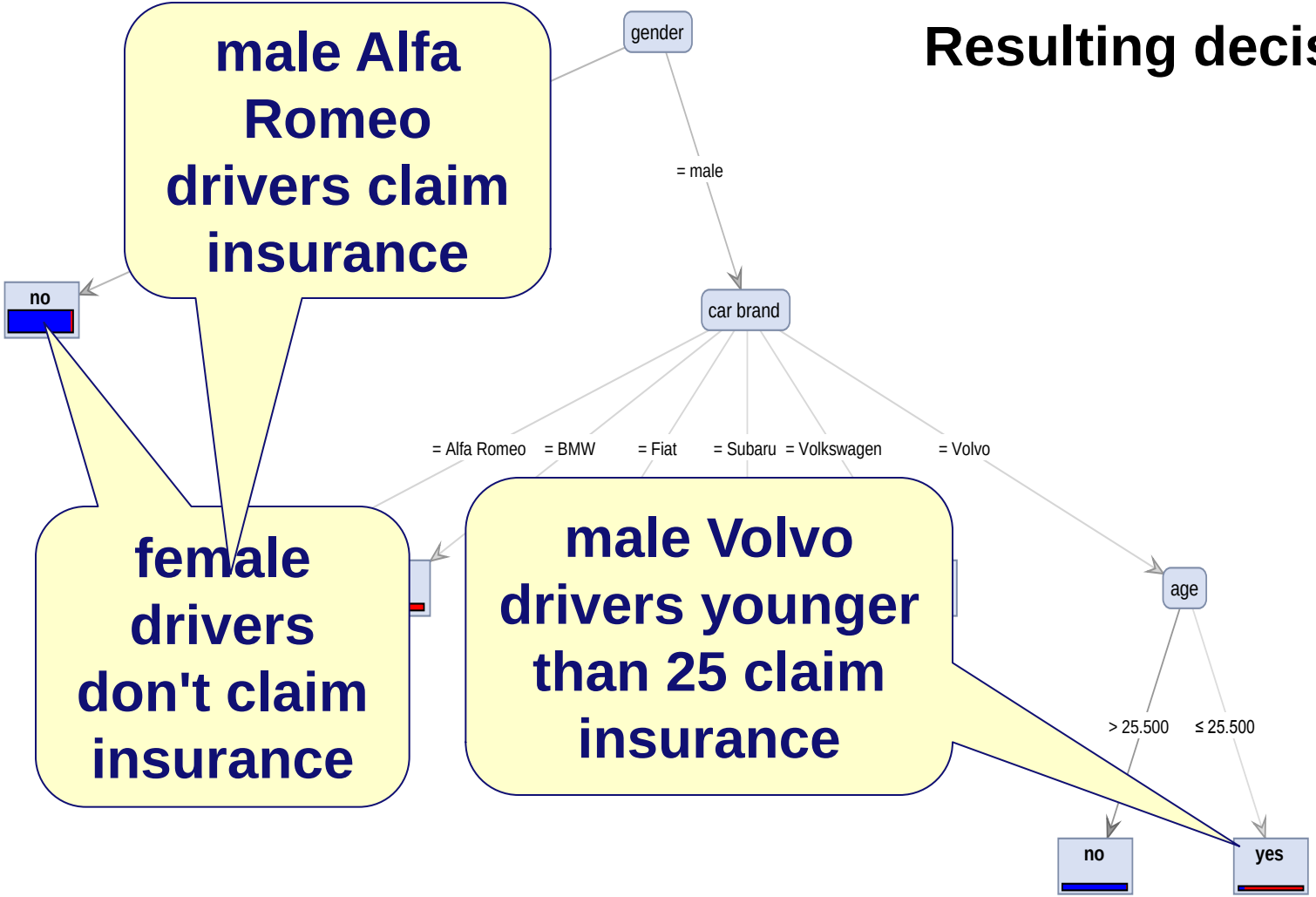
Row No.	claim	gender	age	smoker	car brand
1	no	female	47	yes	Volvo
2	yes	male	31	no	Alfa Romeo
3	yes	male	59	no	Alfa Romeo
4	no	male	28	no	Fiat
5	no	male	44	no	BMW
6	no	female	27	no	Fiat
7	no	male	29	no	Subaru
8	yes	male	44	yes	Subaru
9	no	male	39	no	BMW
10	yes	male	35	no	Subaru
11	no	male	43	no	Subaru
12	yes	male	25	no	BMW
13	no	male	39	no	Volkswagen
14	yes	male	37	no	Alfa Romeo
15	no	female	30	no	Fiat
16	no	female	24	no	Fiat
17	yes	male	26	no	Alfa Romeo
18	no	male	43	no	BMW
19	no	male	46	no	BMW
20	no	female	25	no	Fiat
21	no	female	27	no	Nissan
22	no	female	31	no	Nissan
23	no	male	29	yes	Volkswagen
24	yes	male	42	no	BMW
25	no	male	26	no	Fiat
26	yes	male	27	no	Alfa Romeo

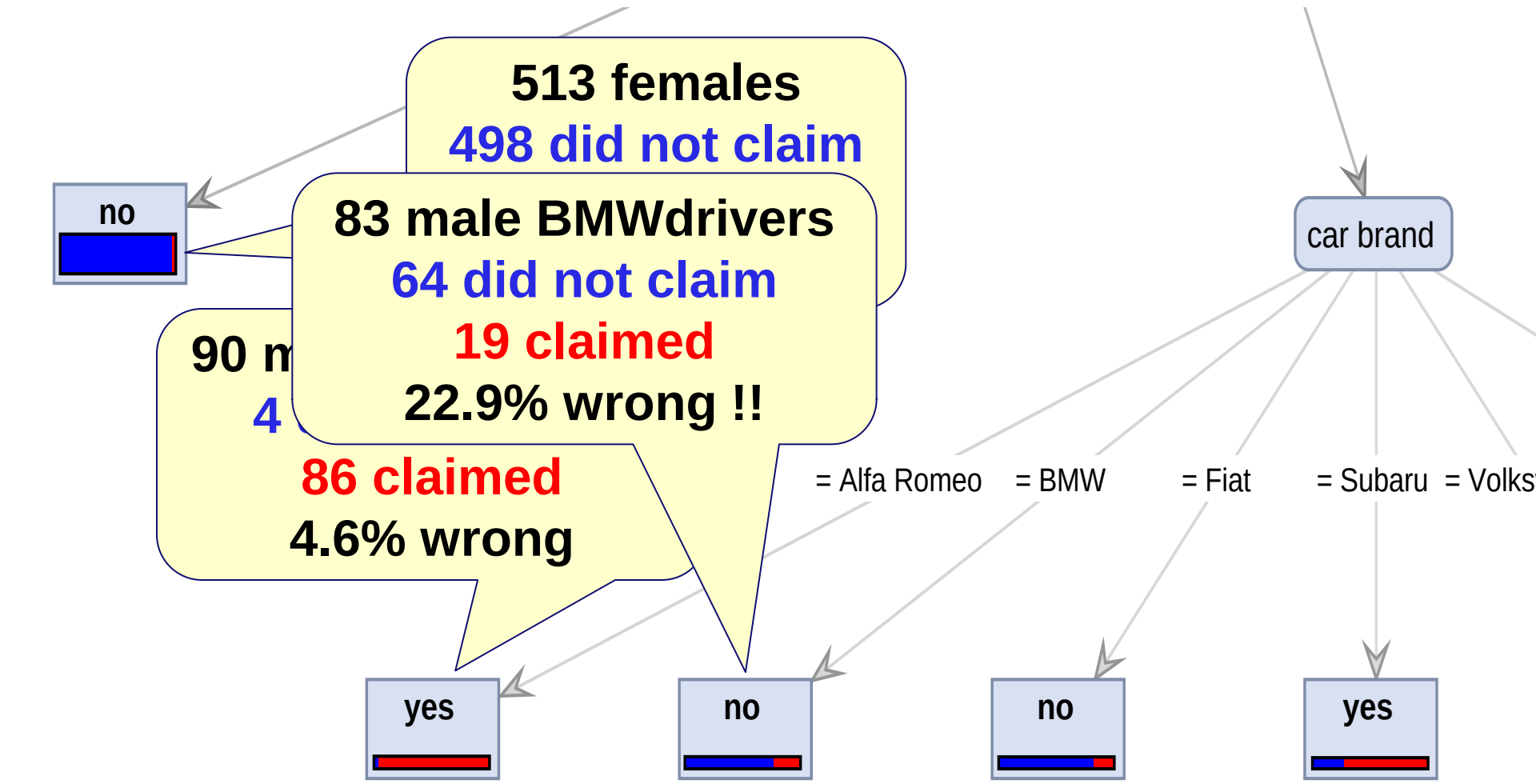
Data is stored in repository.
Now we can apply an analysis workflow to it.

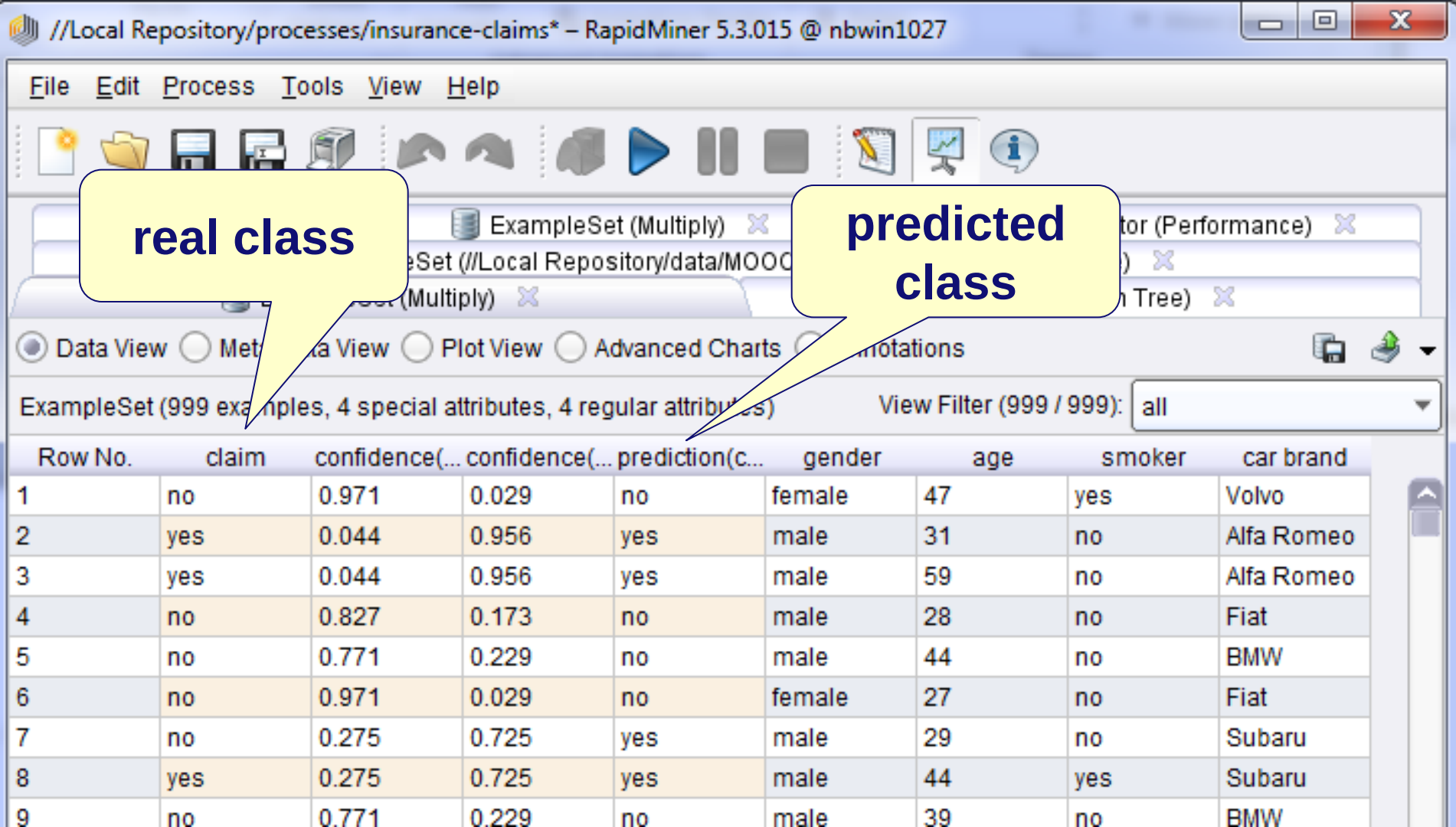



//Local Repository/processes/insurance-claims* – RapidMiner 5.3.015 @ nbwin1027
File Edit Process Tools View Help
load data set
apply decision tree to data set
create decision tree
measure performance
Retrieve
Decision Tree
Apply Model
Multiply
Performance
Operators
Repositories
Process
XML
Problems
Log
No problems found
Message
Fixes
Location
Synopsis
The root operator which is the outer most operator of every process.
Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of
resultfile
random seed
2001

Resulting decision tree




513 females
498 did not claim
83 male BMWdrivers
64 did not claim
19 claimed
22.9% wrong !!
86 claimed
4.6% wrong
no
car brand
= Alfa Romeo
= BMW
= Fiat
= Subaru
yes
no
no
yes


//Local Repository/processes/insurance-claims* – RapidMiner 5.3.015 @ nbwin1027
File Edit Process Tools View Help
real class
predicted class
ExampleSet (Multiply)
Data View
Plot View
Advanced Charts
ExampleSet (999 examples, 4 special attributes, 4 regular attributes)
View Filter (999 / 999): all
Row No. claim confidence(... confidence(... prediction(c... gender age smoker car brand
1 no 0.971 0.029 no female 47 yes Volvo
2 yes 0.044 0.956 yes male 31 no Alfa Romeo
3 yes 0.044 0.956 yes male 59 no Alfa Romeo
4 no 0.827 0.173 no male 28 no Fiat
5 no 0.771 0.229 no male 44 no BMW
6 no 0.971 0.029 no female 27 no Fiat
7 no 0.275 0.725 yes male 29 no Subaru
8 yes 0.275 0.725 yes male 44 yes Subaru
9 no 0.771 0.229 no male 39 no BMW

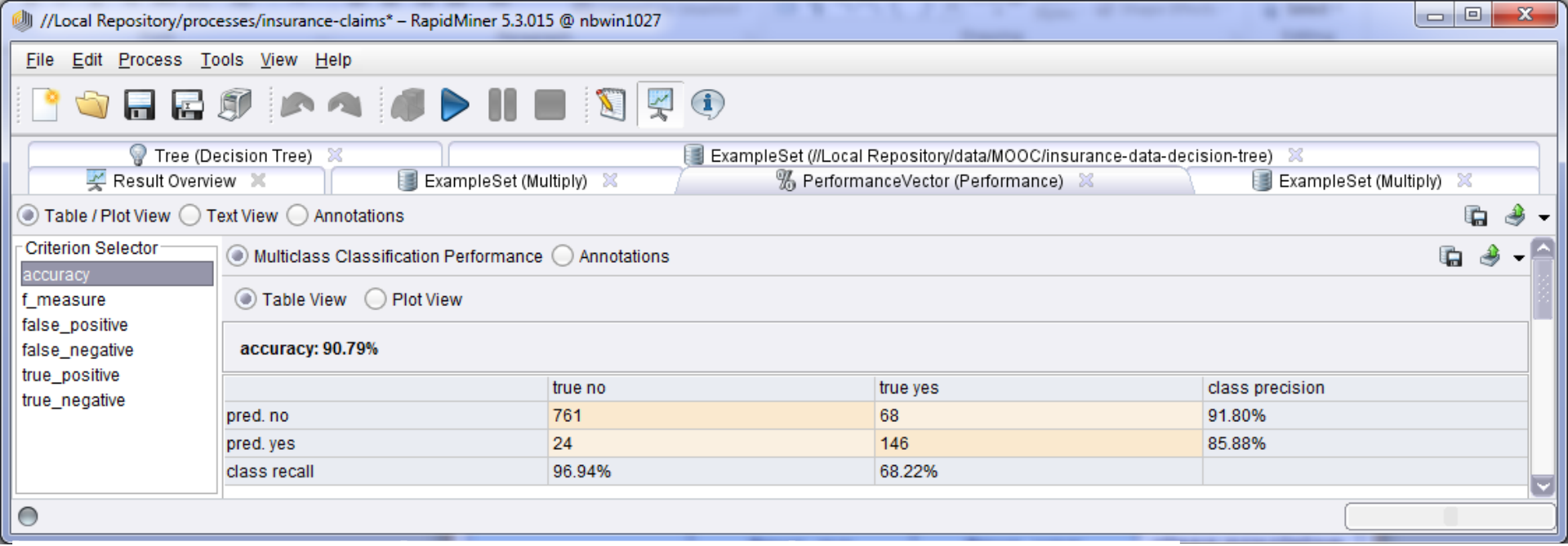
ExampleSet (999 examples, 4 special attributes, 4 regular attributes) View Filter (999 / 999

Row No.	claim	confidence(...	confidence(...	prediction(c...	gender	age	smoker	car brand
9	no	0.771	0.229	no	male	39	no	BMW
10	yes	0.275	0.725	yes	male	35	no	Subaru
11	no	0.275	0.725	yes	male	43	no	Subaru
12	yes	0.771	0.229	no	male	25	no	BMW
13	no	0.740	0.260	no	male	39	no	Volkswagen
14	yes	0.044	0.956	yes	male	37	no	Alfa Romeo

Which instances are classified incorrectly?

11: A male 43-year old non-smoking Subaru driver was predicted to claim but did not.

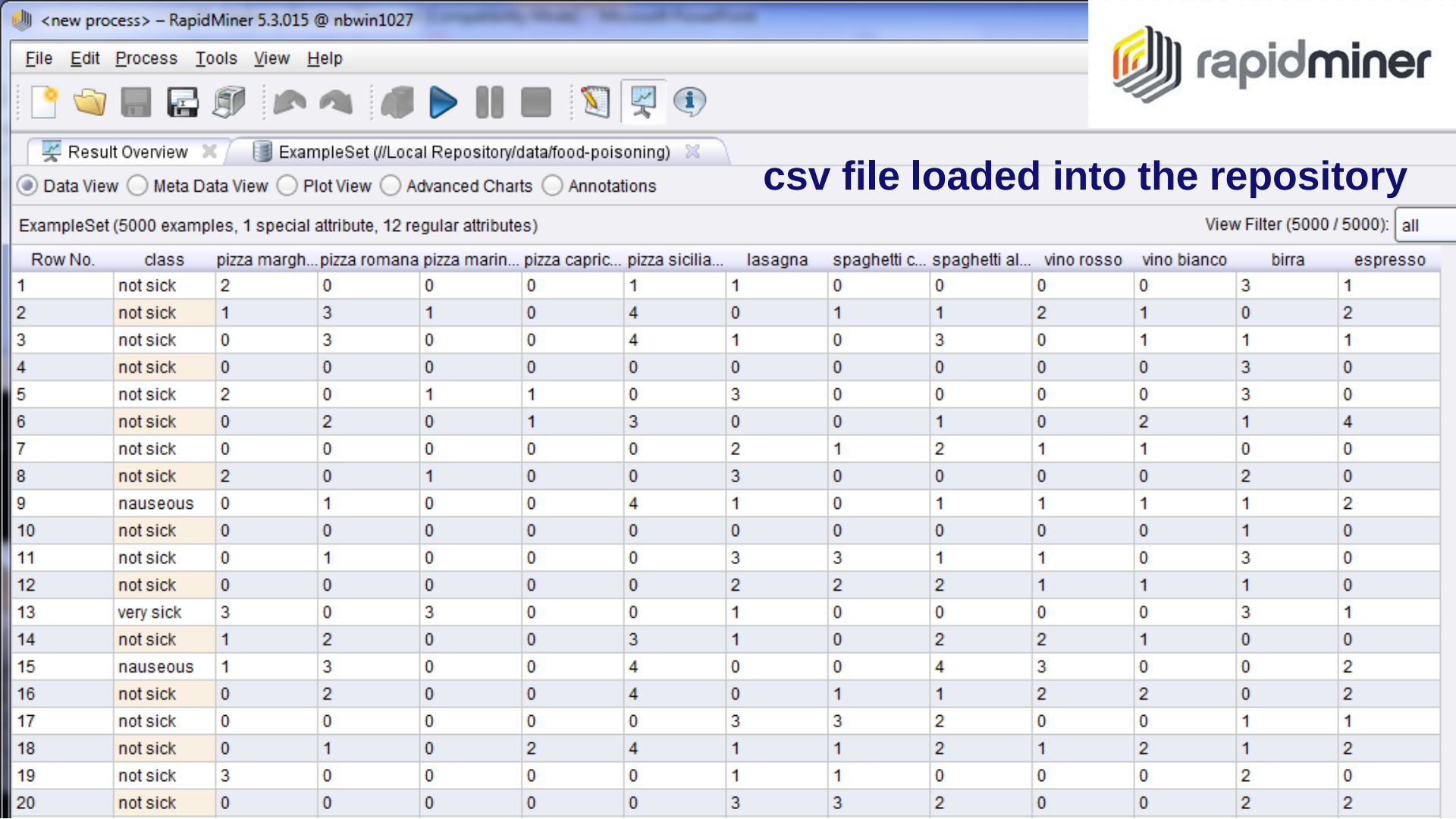
12: A male 25-year old non-smoking BMW driver was predicted to not claim, but actually did claim insurance.



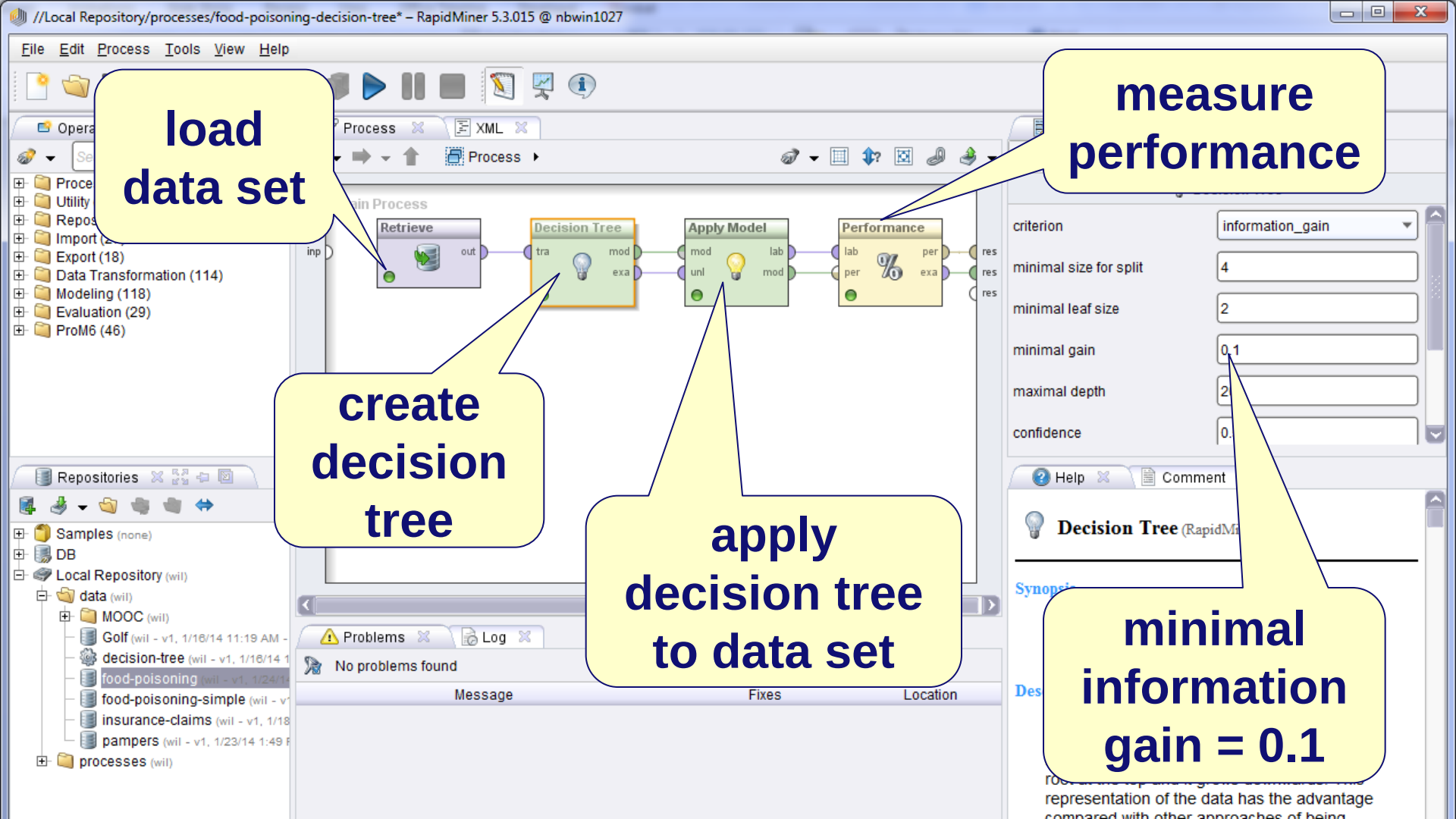
	true no	true yes	class precision
pred. no	761	68	91.80%
pred. yes	24	146	85.88%
class recall	96.94%	68.22%	

5000 parties ate at an Italian restaurant.

Menu includes: pizza margherita, pizza romana, pizza marinara, pizza capricciosa, pizza siciliana, lasagna, spaghetti carbonara, spaghetti alla diavola, vino rosso,vino bianco, birra, and espresso.

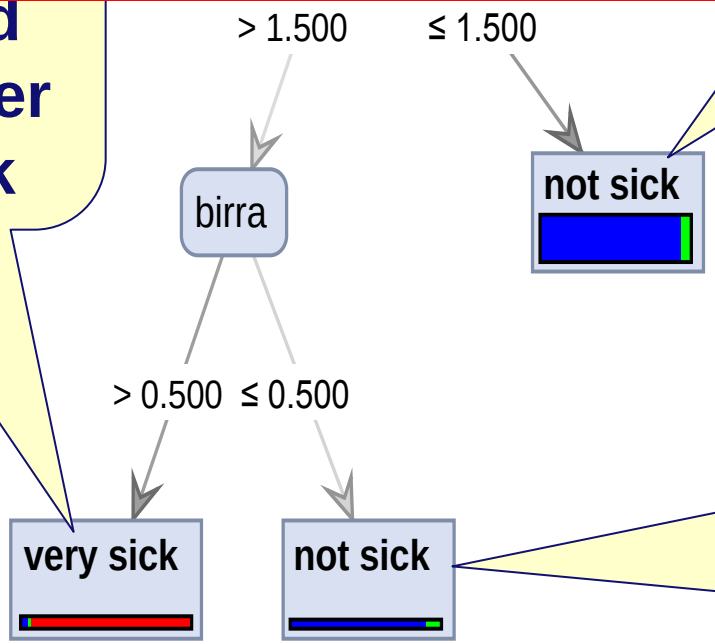


Row No.	class	pizza margh...	pizza romana	pizza marin...	pizza capric...	pizza sicilia...	lasagna	spaghetti c...	spaghetti al...	vino rosso	vino bianco	birra	espresso
1	not sick	2	0	0	0	1	1	0	0	0	0	3	1
2	not sick	1	3	1	0	4	0	1	1	2	1	0	2
3	not sick	0	3	0	0	4	1	0	3	0	1	1	1
4	not sick	0	0	0	0	0	0	0	0	0	0	3	0
5	not sick	2	0	1	1	0	3	0	0	0	0	3	0
6	not sick	0	2	0	1	3	0	0	1	0	2	1	4
7	not sick	0	0	0	0	0	2	1	2	1	1	0	0
8	not sick	2	0	1	0	0	3	0	0	0	0	2	0
9	nauseous	0	1	0	0	4	1	0	1	1	1	1	2
10	not sick	0	0	0	0	0	0	0	0	0	0	1	0
11	not sick	0	1	0	0	0	3	3	1	1	0	3	0
12	not sick	0	0	0	0	0	2	2	2	1	1	1	0
13	very sick	3	0	3	0	0	1	0	0	0	0	3	1
14	not sick	1	2	0	0	3	1	0	2	2	1	0	0
15	nauseous	1	3	0	0	4	0	0	4	3	0	0	2
16	not sick	0	2	0	0	4	0	1	1	2	2	0	2
17	not sick	0	0	0	0	0	3	3	2	0	0	1	1
18	not sick	0	1	0	2	4	1	1	2	1	2	1	2
19	not sick	3	0	0	0	0	1	1	0	0	0	2	0
20	not sick	0	0	0	0	0	3	3	2	0	0	2	2


//Local Repository/processes/food-poisoning-decision-tree* – RapidMiner 5.3.015 @ nbwin1027
File Edit Process Tools View Help
load data set
create decision tree
apply decision tree to data set
measure performance
minimal information gain = 0.1
Retrieve
Decision Tree
Apply Model
Performance
criterion
information_gain
minimal size for split
4
minimal leaf size
2
minimal gain
0.1
maximal depth
confidence
Repositories
Samples (none)
DB
Local Repository (wil)
data (wil)
MOOC (wil)
Golf (wil - v1, 1/16/14 11:19 AM -
decision-tree (wil - v1, 1/16/14 1
food-poisoning
food-poisoning-simple
insurance-claims (wil - v1, 1/18
pampers (wil - v1, 1/23/14 1:49 F
processes (wil)
Problems
Log
No problems found
Message
Fixes
Location
Help
Comment
Decision Tree
representation of the data has the advantage

Decision tree indicates that a combination of pizzas marinara and beer caused sickness.

partie
multip
marinara and
that drank beer
got very sick



pizzas marinara
did not get sick

parties that did
not drink beer
did not get sick

blue = not sick
red = very sick
green = nauseous

...al information gain = 0.1

307 of the 313 parties that were nauseous were classified as "not sick"

accuracy: 93.26%

	true not sick	true ...eous	true very sick	class precision
pred. not sick	4193	307		
pred. nauseous	0	0		
pred. very sick	24	6		
class recall	99.43%	0.00%		

6 of the 313 parties that were nauseous were classified as "very sick"

> 0.500 ≤ 0.500

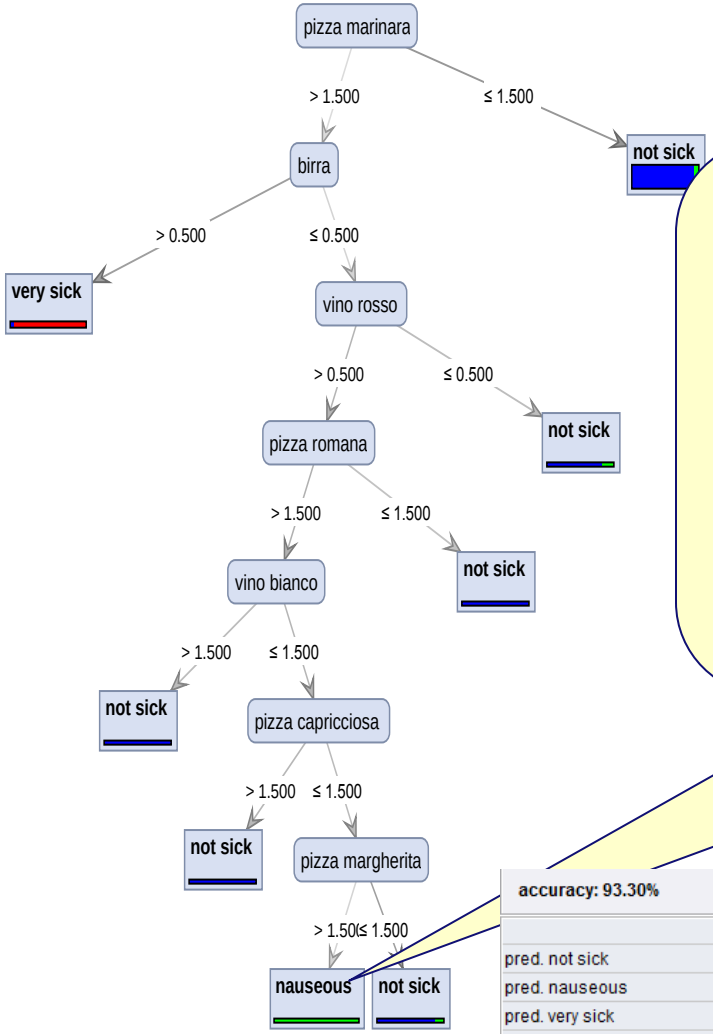
The decision tree does not explain why some parties were nauseous.

blue = ...

red = ...

green = nauseous

minimal information gain = 0.05

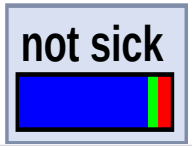


people that ate multiple pizzas marinara, pizzas romana, pizzas margherita, but at most one pizza capricciosa, and drank red wine but not multiple glasses of white wine and did not drink any beer got nauseous.

Extremely small improvement at the cost of overfitting.

accuracy: 93.30%

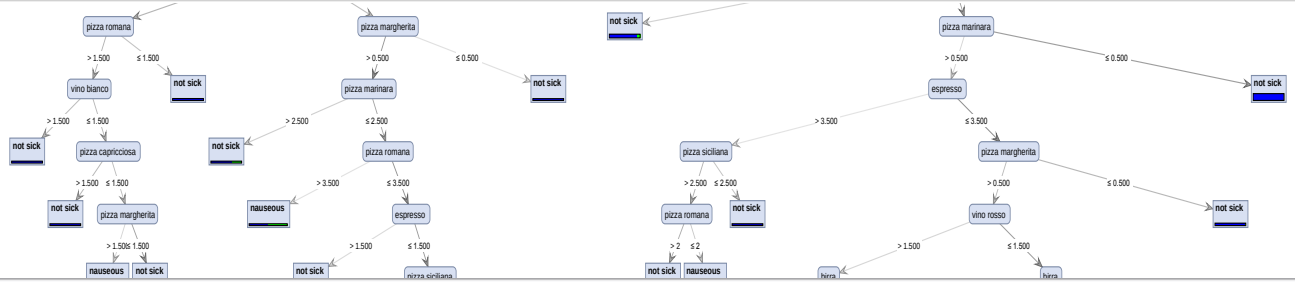
	true not sick	true nauseous	true very sick	class precision
pred. not sick	4193	305	0	93.22%
pred. nauseous	0	2	0	100.00%
pred. very sick	24	6	470	94.00%
class recall	99.43%	0.64%	100.00%	



underfitting

accuracy: 84.34%

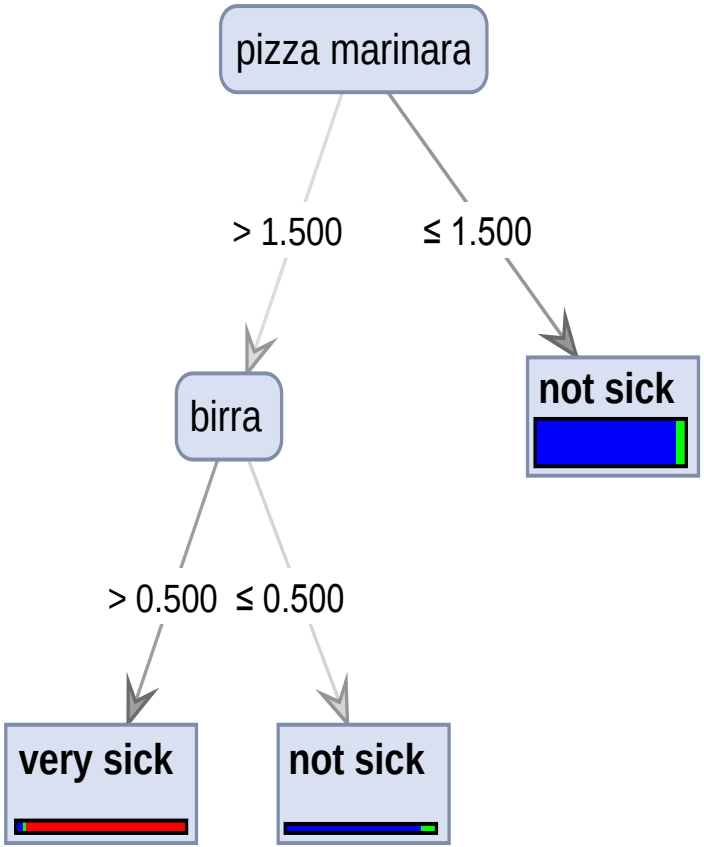
	true not sick	true nauseous	true very sick	class precision
pred. not sick	4217	313	470	84.34%
pred. nauseous	0	0	0	0.00%
pred. very sick	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	



overfitting

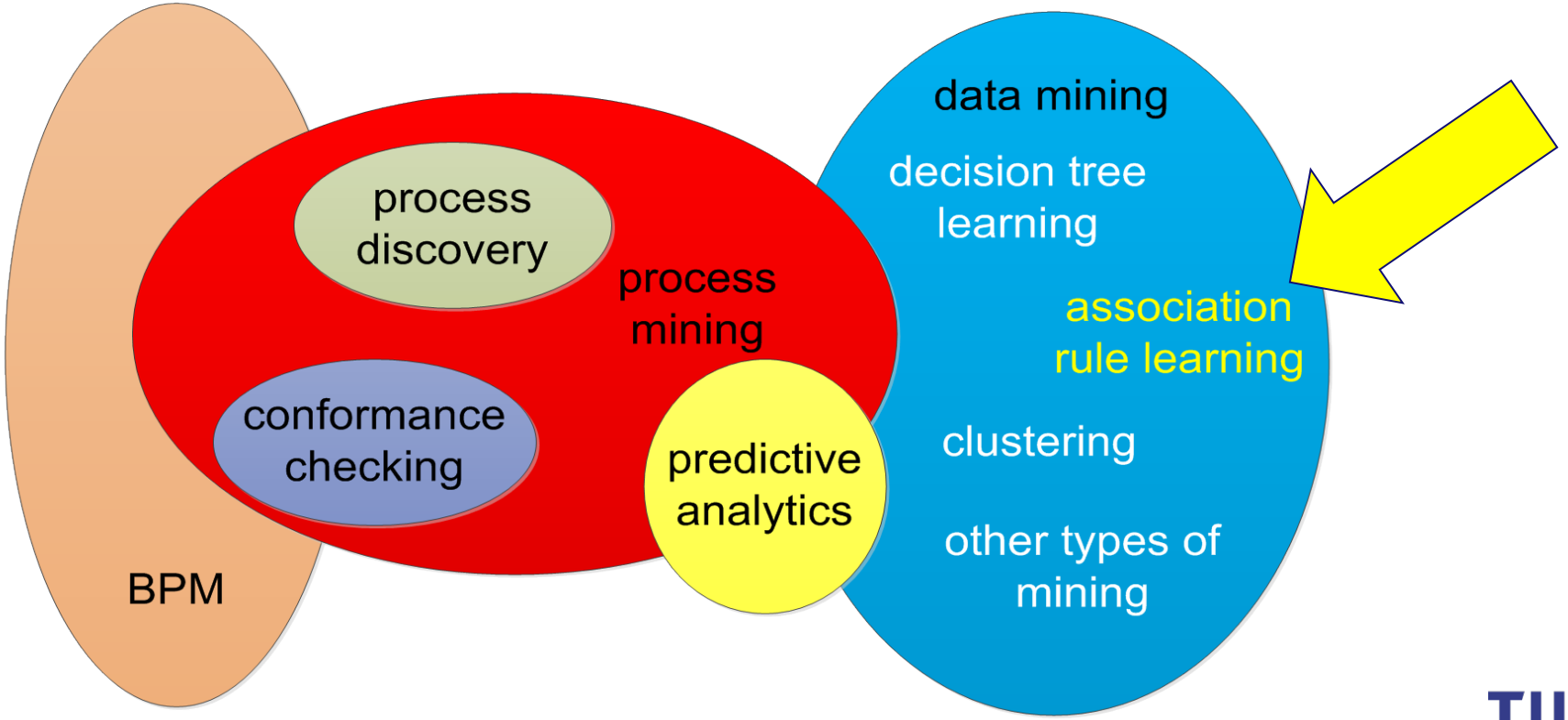
accuracy: 93.48%

	true not sick	true nauseous	true very sick	class precision
pred. not sick	4190	293	0	93.46%
pred. nauseous	3	14	0	82.35%
pred. very sick	24	6	470	94.00%
class recall	99.36%	4.47%	100.00%	



- **Reasonable balance between underfitting and overfitting.**
- **Can be used to understand what is happening.**
- **Can be used for predictions and recommendations.**

Next



TU/e

Part I: Preliminaries

- Chapter 1: Introduction
- Chapter 2: Process Modeling and Analysis
- Chapter 3: Data Mining

Part II: From Event Logs to Process Models

- Chapter 4: Getting the Data
- Chapter 5: Process Discovery: An Introduction
- Chapter 6: Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

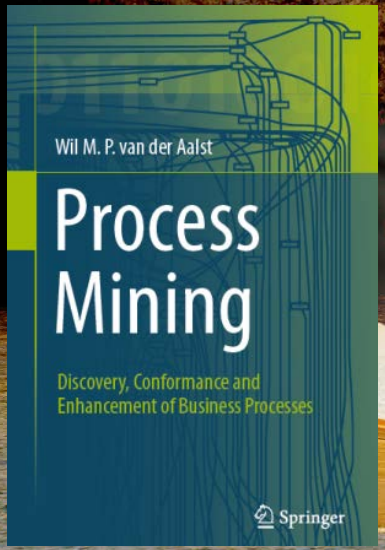
- Chapter 7: Conformance Checking
- Chapter 8: Mining Additional Perspectives
- Chapter 9: Operational Support

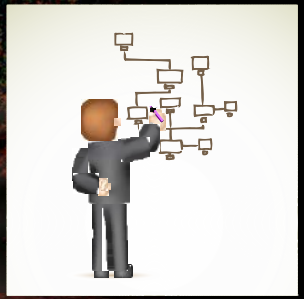
Part IV: Putting Process Mining to Work

- Chapter 10: Tool Support
- Chapter 11: Analyzing "Lasagna Processes"
- Chapter 12: Analyzing "Spaghetti Processes"

Part V: Reflection

- Chapter 13: Cartography and Navigation
- Chapter 14: Epilogue





TU/e