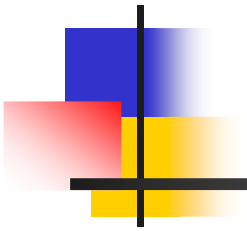
Using Rapid Miner in Computer Aided Quality



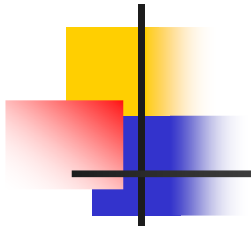
Petr Berka
(berka@vse.cz)

University of Economics
Prague

Seminar on Data mining tools and CAQ (S1)

Vysoká škola manažmentu v Trenčíne, International Workshop on Knowledge Management, IWKM'2017

Trenčín 12 - 13. 10. 2017



Computer Aided Quality as a Data Mining Problem

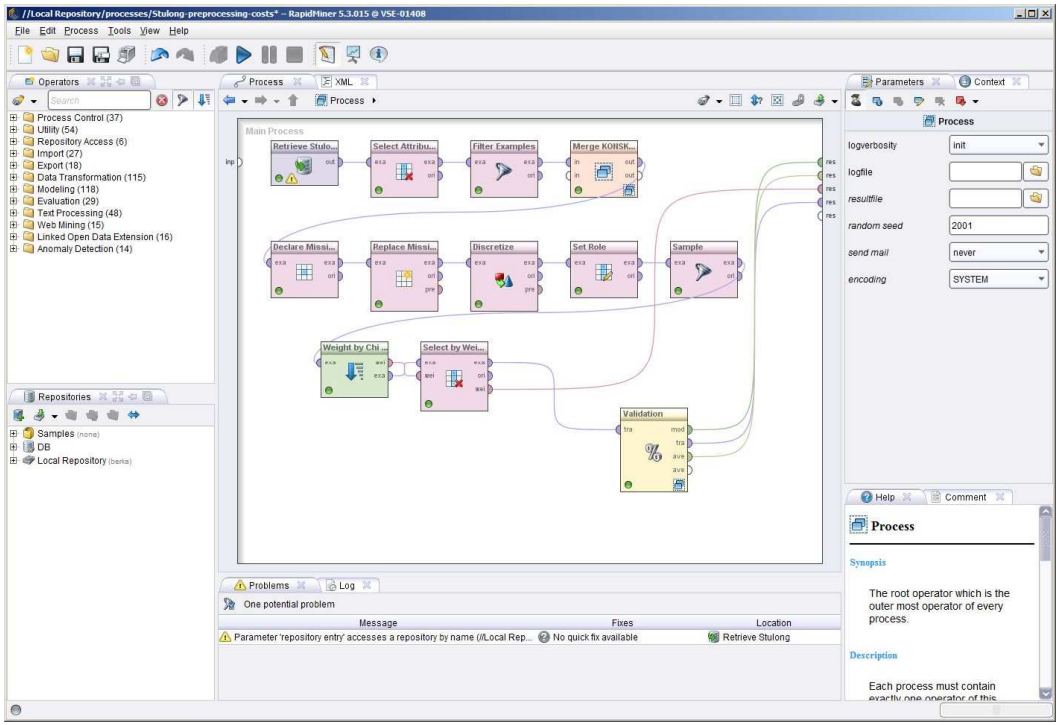
Based on measured characteristics of a product decide if the product is Risky or not

From data mining point-of-view:

- Binary classification task
- All input attributes are numerical

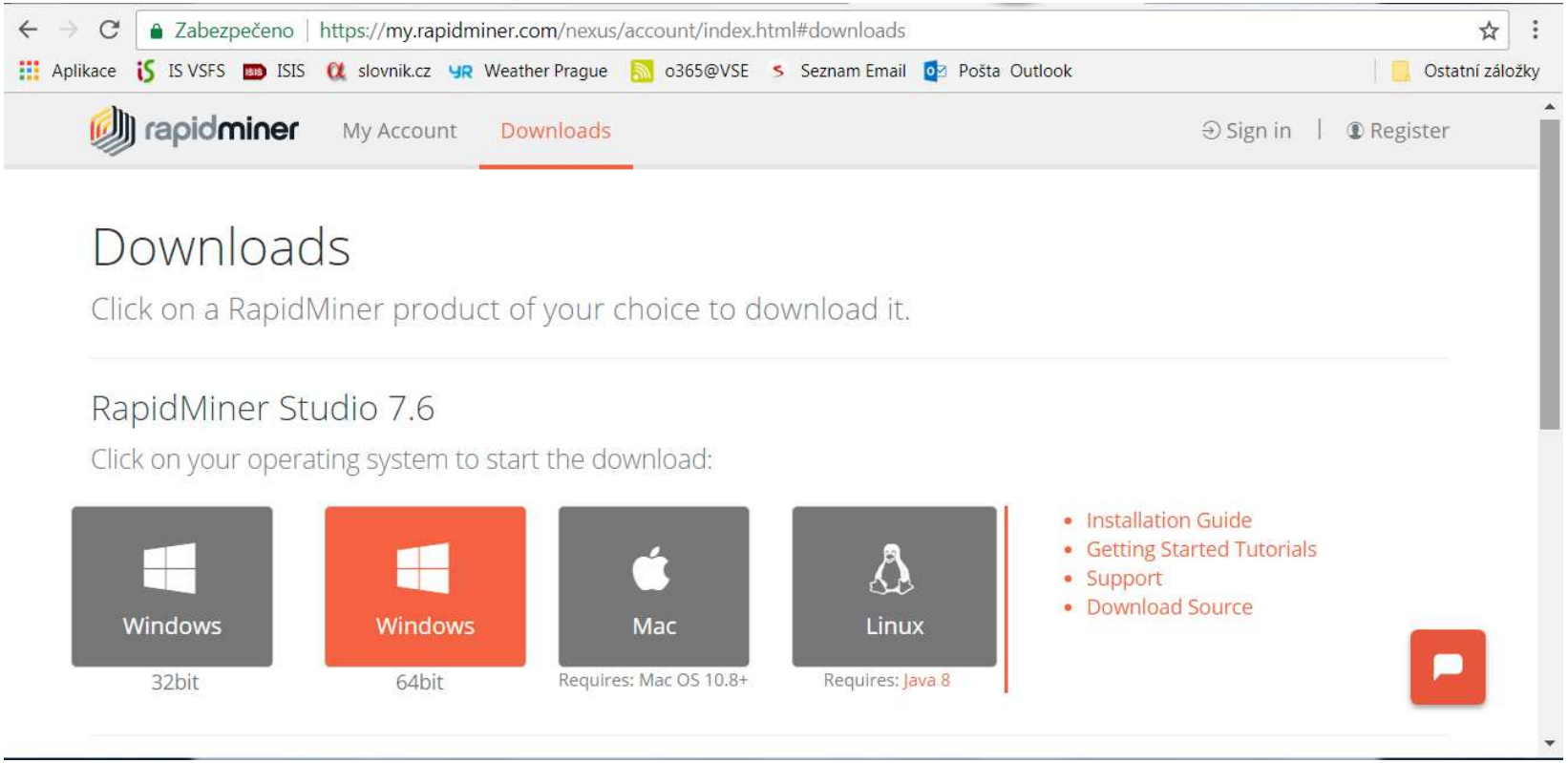
Rapid Miner (rapidminer.com)

- a leading open-source system for knowledge discovery and data mining (www.kdnuggets.com)
- a Leader in 2016 Gartner Magic Quadrant for Advanced Analytics (www.gartner.com)
- the Top 3 Rated Predictive Analytics Software for Enterprise (www.g2crowd.com)

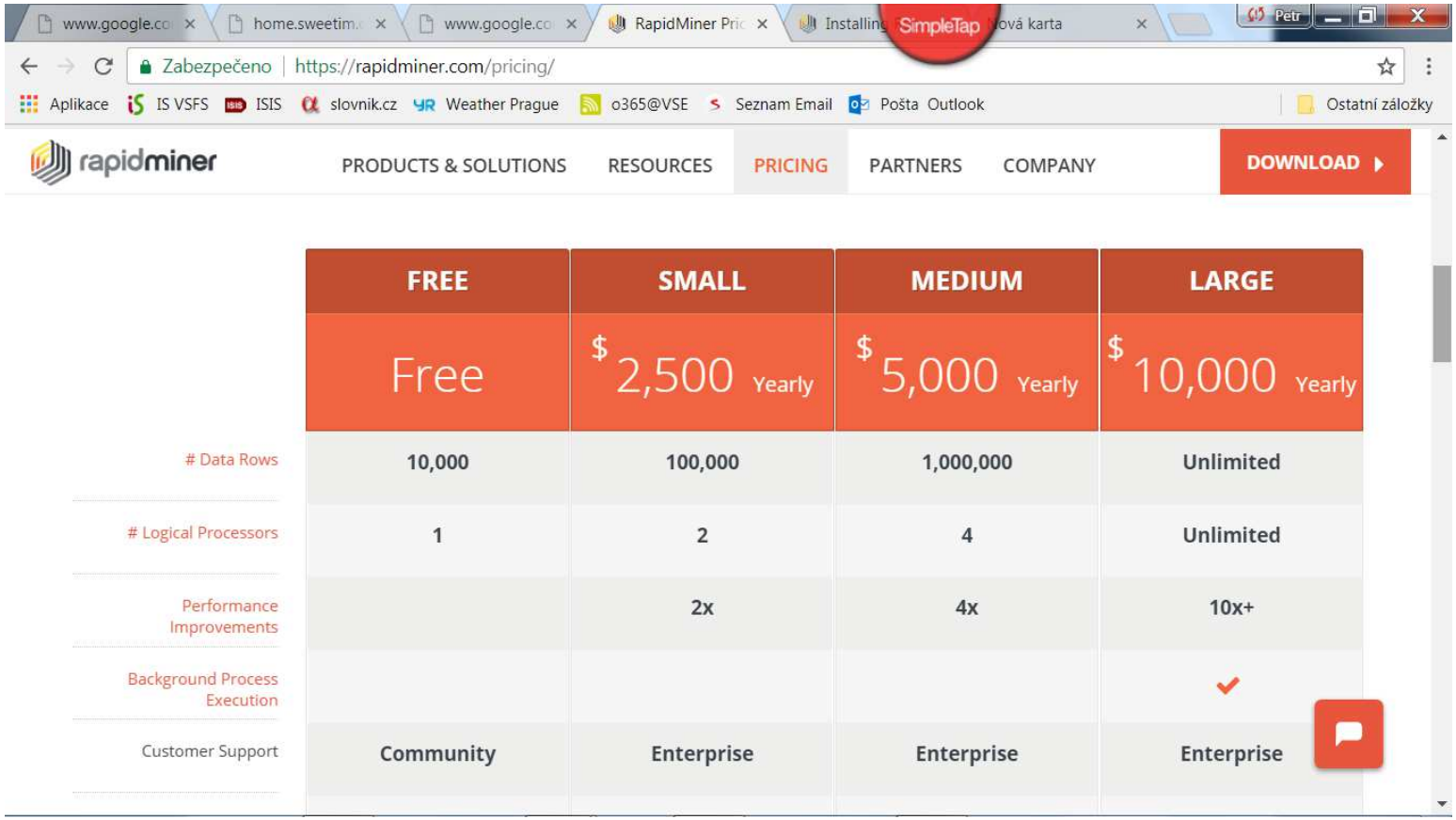


Rapid Miner Downloads

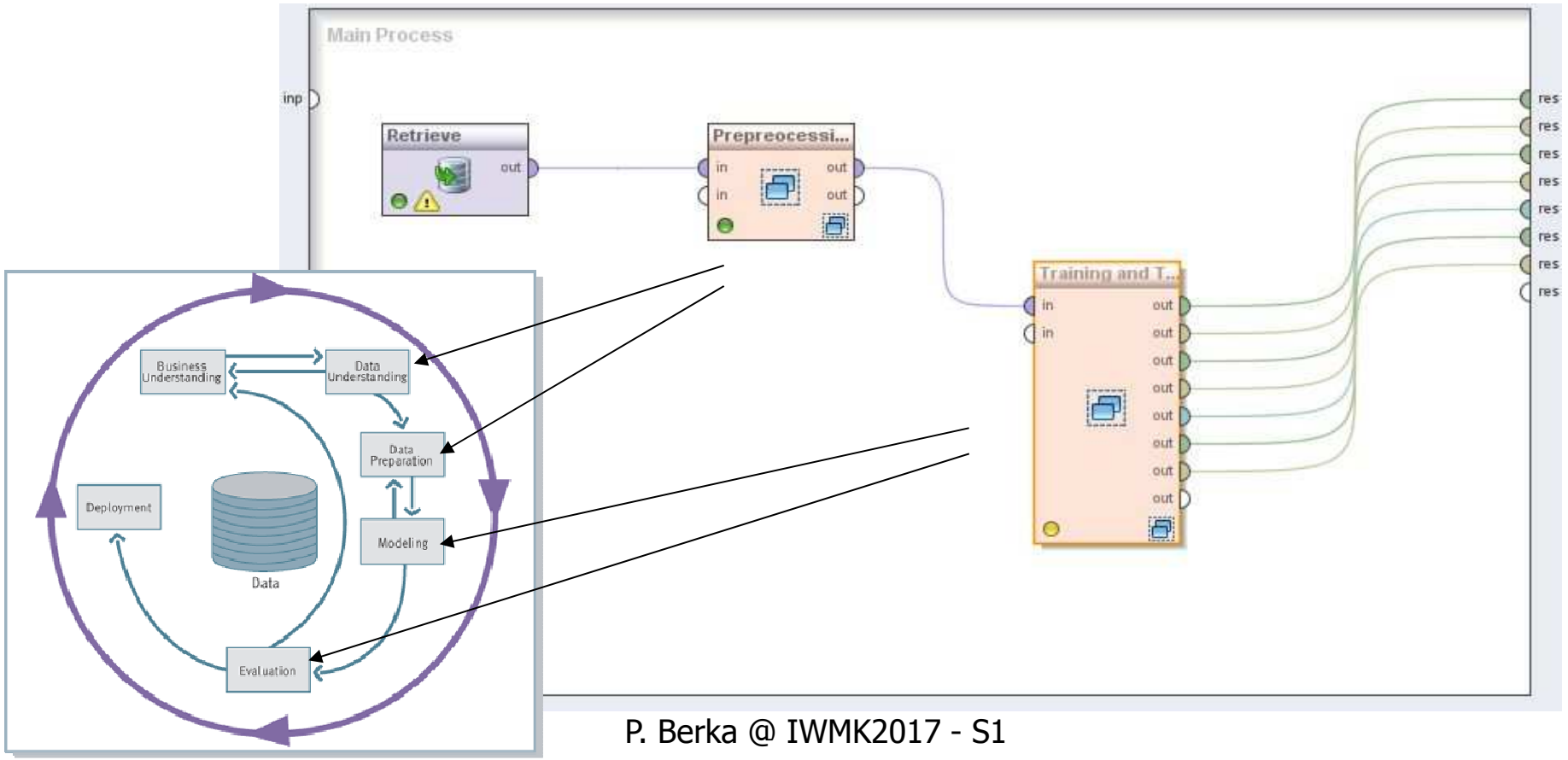
https://my.rapidminer.com/nexus/account/index.html#downloads



Rapid Miner Pricing



Overview of a DM Project



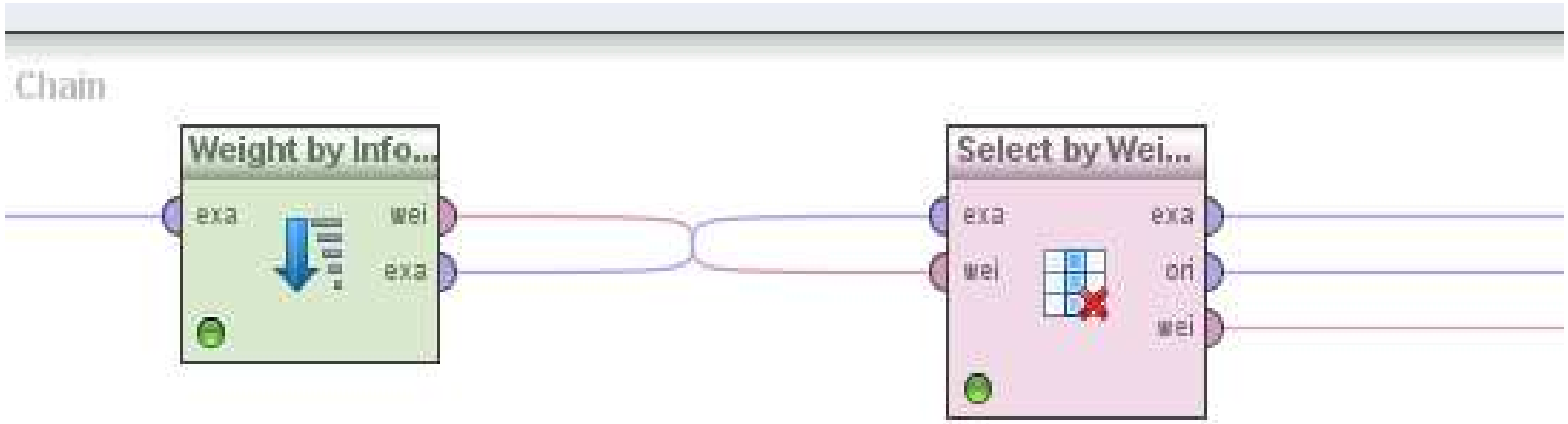
Retrieve original data

ExampleSet (376 examples, 1 special attribute, 45 regular attributes)

View Filter (376 / 376): all

Row No.	Test	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Pa
1	OK	437.200	0.070	4491.100	0.276	0.291	0.628	0.308	0.303	0.636	0.139
2	OK	351.100	0.080	4049	0.256	0.257	0.622	0.275	0.267	0.547	0.135
3	OK	382.400	0.070	4774.100	0.262	0.266	0.630	0.284	0.267	0.571	0.133
4	RISK	396.400	0.090	4088.600	0.264	0.263	0.611	0.266	0.290	0.512	0.154
5	OK	426.300	0.090	4321.500	0.263	0.307	0.636	0.297	0.272	0.567	0.148
6	RISK	373.300	0.070	4963.600	0.283	0.254	0.624	0.266	0.281	0.526	0.129
7	RISK	335.400	0.080	5363.500	0.267	0.261	0.606	0.265	0.292	0.561	0.138
8	RISK	343.100	0.080	4612.400	0.253	0.287	0.637	0.307	0.285	0.616	0.144
9	OK	383	0.080	4555.600	0.246	0.275	0.643	0.297	0.301	0.641	0.140
10	OK	375.900	0.080	4203	0.282	0.272	0.614	0.283	0.280	0.563	0.142
11	RISK	372.700	0.070	3471.900	0.286	0.271	0.615	0.296	0.265	0.586	0.130
12	RISK	368.800	0.080	4650.100	0.248	0.314	0.612	0.291	0.300	0.534	0.138
13	OK	368.500	0.070	4843.900	0.243	0.277	0.629	0.278	0.288	0.532	0.130
14	RISK	371.500	0.070	3959.700	0.308	0.297	0.627	0.285	0.276	0.582	0.129
15	OK	376.700	0.090	4201.700	0.253	0.292	0.643	0.318	0.281	0.590	0.145
16	OK	396.800	0.080	5718.500	0.265	0.263	0.629	0.284	0.272	0.580	0.137
17	RISK	388.200	0.080	4609.400	0.253	0.281	0.618	0.330	0.269	0.571	0.140
18	OK	315.200	0.080	3725.600	0.267	0.291	0.639	0.331	0.272	0.610	0.141
19	OK	371.700	0.060	4430.600	0.259	0.294	0.626	0.306	0.284	0.581	0.123
20	RISK	340.300	0.090	4098.400	0.283	0.292	0.649	0.317	0.269	0.597	0.146

Data Preprocessing

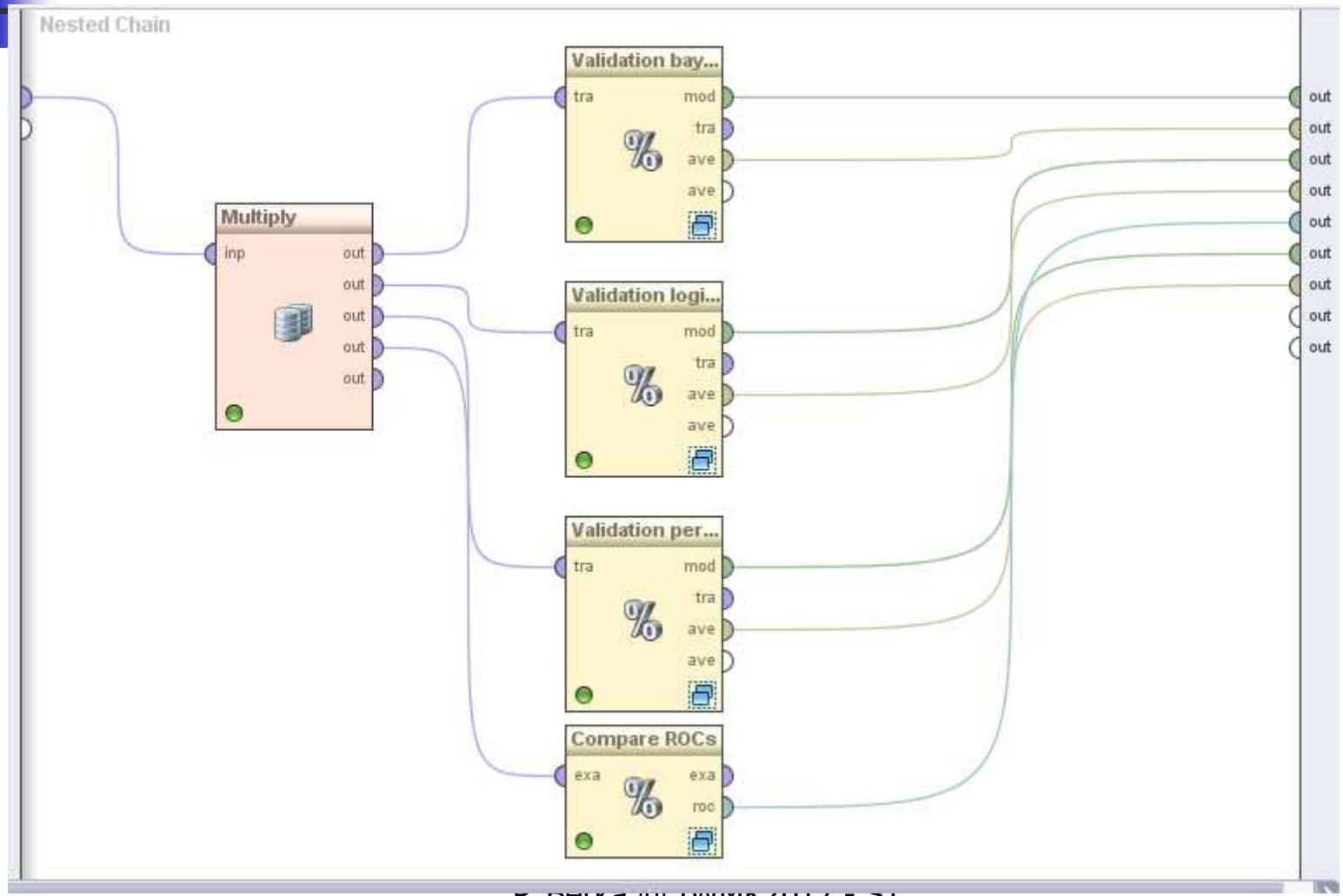


Data used for Training and Testing

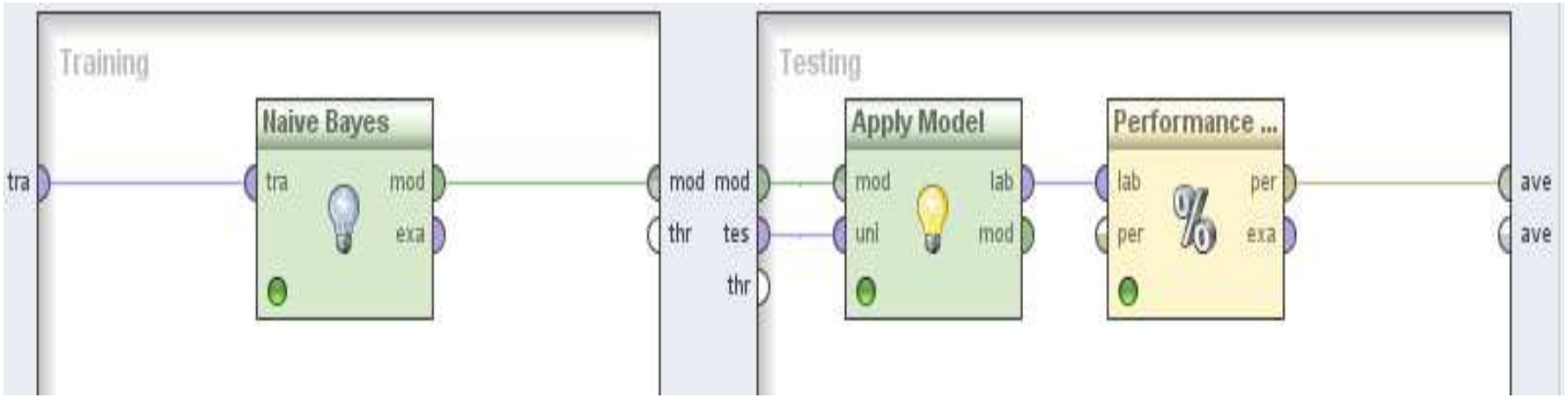
ExampleSet (376 examples, 1 special attribute, 25 regular attributes) View Filter (376 / 376): all

Row No.	Test	Par1	Par7	Par8	Par9	Par10	Par14	Par15	Par16	Par17	Pa
1	OK	437.200	0.308	0.303	0.636	0.139	1.100	1.920	5.410	7.680	12.64
2	OK	351.100	0.275	0.267	0.547	0.135	1.460	2.390	4.120	6.630	11.65
3	OK	382.400	0.284	0.267	0.571	0.133	1.350	2.170	4.150	6.700	10.76
4	RISK	396.400	0.266	0.290	0.512	0.154	1.410	2.320	4.870	6.980	10.97
5	OK	426.300	0.297	0.272	0.567	0.148	1.560	2.450	5.140	5.720	10.98
6	RISK	373.300	0.266	0.281	0.526	0.129	1.100	2.620	5.050	7.030	10.65
7	RISK	335.400	0.265	0.292	0.561	0.138	1.130	2.420	4.620	7.160	10.98
8	RISK	343.100	0.307	0.285	0.616	0.144	1.200	3	4.720	7.480	10.84
9	OK	383	0.297	0.301	0.641	0.140	0.540	2.250	4.310	7.160	10.92
10	OK	375.900	0.283	0.280	0.563	0.142	0.640	3.080	4.310	8.720	10.90
11	RISK	372.700	0.296	0.265	0.586	0.130	1.400	2.360	4.670	8.060	11.49
12	RISK	368.800	0.291	0.300	0.534	0.138	0.920	2.350	4.440	6.650	13.09
13	OK	368.500	0.278	0.288	0.532	0.130	0.830	2.680	4.920	7.690	10.85
14	RISK	371.500	0.285	0.276	0.582	0.129	0.950	2.560	3.840	8.960	10.09
15	OK	376.700	0.318	0.281	0.590	0.145	1.460	2.870	4.380	7.970	12.44
16	OK	396.800	0.284	0.272	0.580	0.137	1.150	2.460	4.930	8.120	12.12
17	RISK	388.200	0.330	0.269	0.571	0.140	0.980	2.460	3.960	6.440	10.85
18	OK	315.200	0.331	0.272	0.610	0.141	0.940	1.940	4.880	5.940	11.45
19	OK	371.700	0.306	0.284	0.581	0.123	1.100	2.540	4.840	6.980	10.40
20	RISK	340.300	0.317	0.269	0.597	0.146	0.960	2.290	4.590	7.740	10.27

Training and Testing



Validation


Training
Naive Bayes
tra
tra
mod
exa
mod mod
thr tes
thr
Testing
Apply Model
mod
unl
lab
mod
Performance ...
lab
per
per
exa
ave
ave

Comparing ROC