Úloha 2 – Rapid Miner

Zadání:

1. Na stránkách předmětu si vyberete data, která budete zpracovávat v programu Rapidminer a z výstupů Rapidmineru vytvoříte krátký report.
2. Váš proud by měl dělat zhruba následující:
   - Načte vaše vybraná a stažená data.
   - Rozdělí data na trénovací a testovací množinu v poměru 2:1 (buď pomocí uzlu Rapidmineru nebo vytvořte v matlabu skript, který to za vás udělá. Pak jen načtete do Rapidmineru 2 množiny).
   - Vytvoříte rozhodovací strom z trénovacích dat.
   - Zjistěte chybu vytvořeného stromu na trénovacích a testovacích datech.

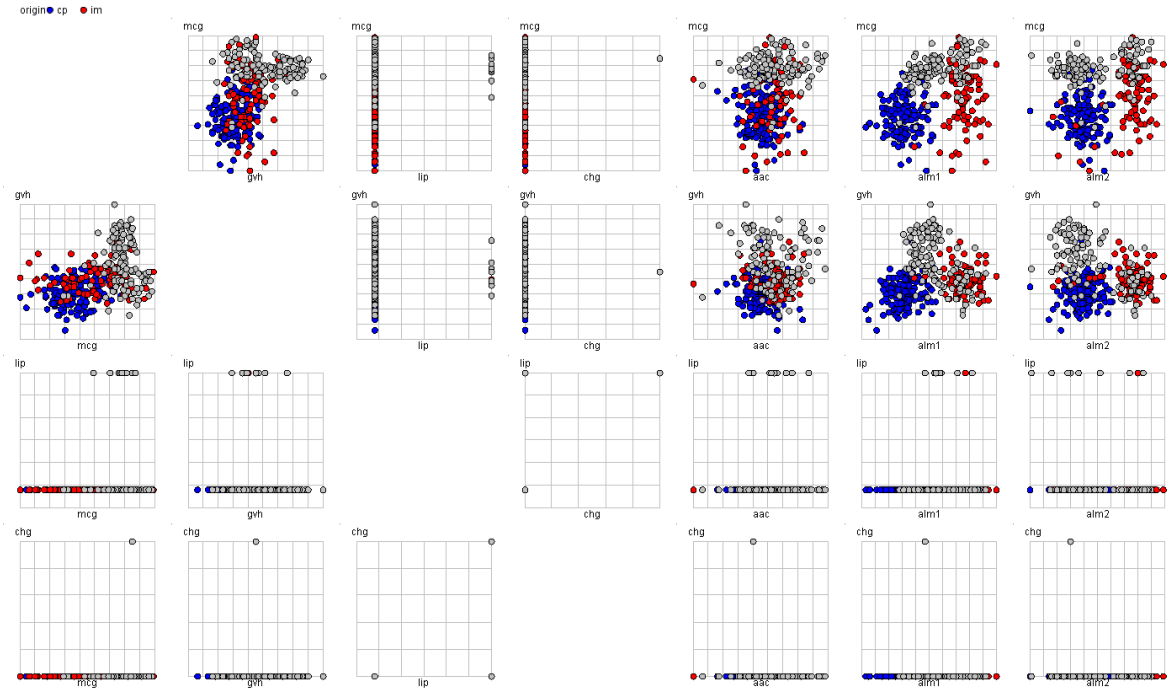
Váš report by měl obsahovat následující výstupy z Rapidmineru:

- Základní statistiku vstupních dat (pro každý sloupec průměr, rozptyl pro číselné atributy, počty hodnot pro nominální). Můžete přidat i bodové grafy nebo matici bodových grafů (scatter plot nebo scatter plot matrix), případně jiné grafy, pokud se vám budou zdát užitečné.
- Vizualizaci rozhodovacího stromu (obrázek) a tento strom přepsaný do formy if-then podmínek (použijte Javovskou nebo Matlabovskou syntaxi).
- Matici záměn (confusion matrix) pro trénovací a testovací data, přesnost klasifikace a krátký komentář, jestli se vám zdá přesnost (accuracy) dostatečná, případně která třída přesnost kazí.

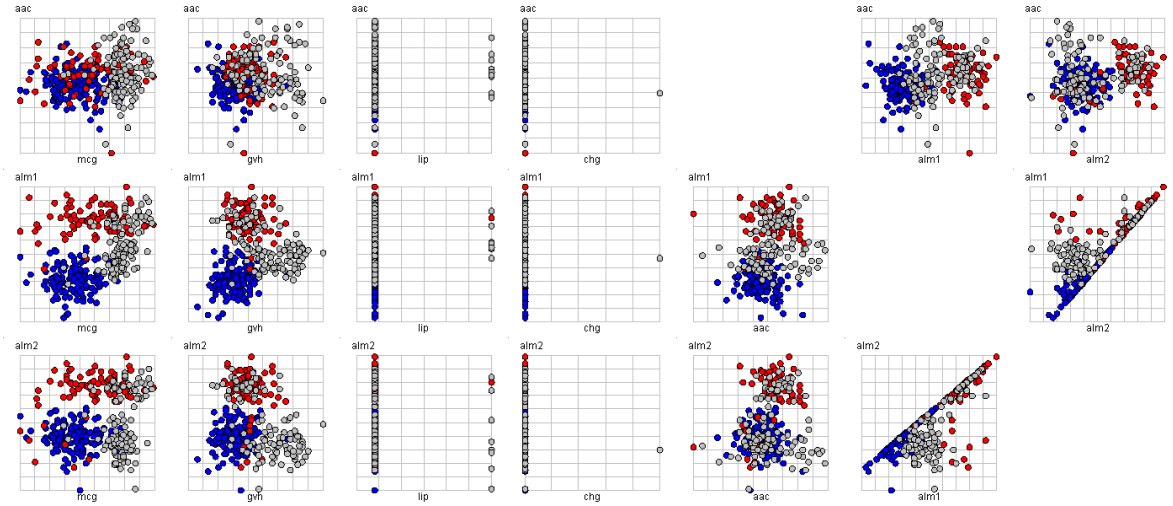
Pro úlohu 2. jsem si vybrala vzorek dat **`ecoli.csv`**. Takto byla reprezentována vstupní data:

Role	Name	Type	Statistics	Range	Missings
label	origin	binominal	mode = cp (143), least = im (7	cp (143), im (77)	116
regular	mcg	real	avg = 0.500 +/- 0.195	[0.000 ; 0.890]	0
regular	gvh	real	avg = 0.500 +/- 0.148	[0.160 ; 1.000]	0
regular	lip	real	avg = 0.495 +/- 0.088	[0.480 ; 1.000]	0
regular	chg	real	avg = 0.501 +/- 0.027	[0.500 ; 1.000]	0
regular	aac	real	avg = 0.500 +/- 0.122	[0.000 ; 0.880]	0
regular	alm1	real	avg = 0.500 +/- 0.216	[0.030 ; 1.000]	0
regular	alm2	real	avg = 0.500 +/- 0.209	[0.000 ; 0.990]	0

Matice bodových grafů (scatter plot matrix)

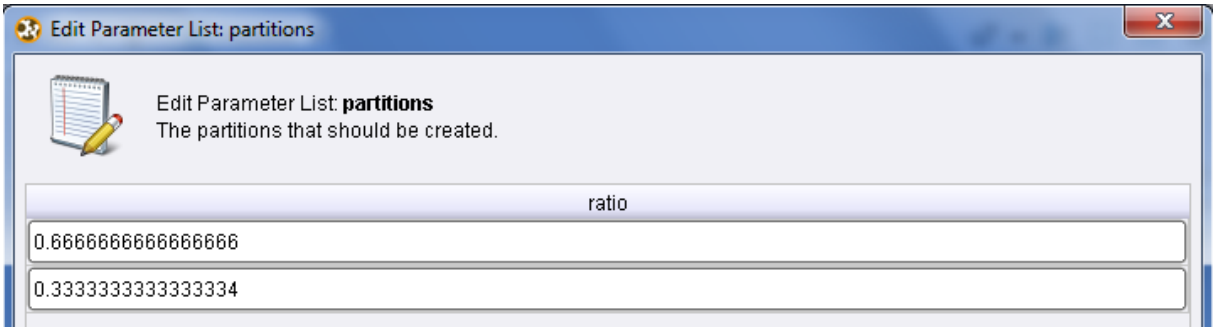


Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

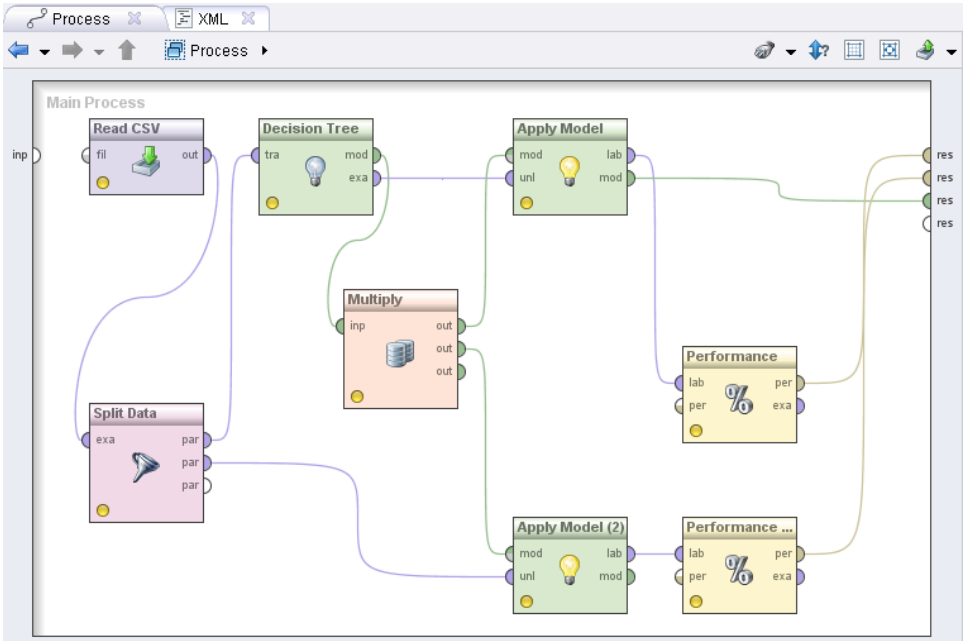


Sestavení procesu:

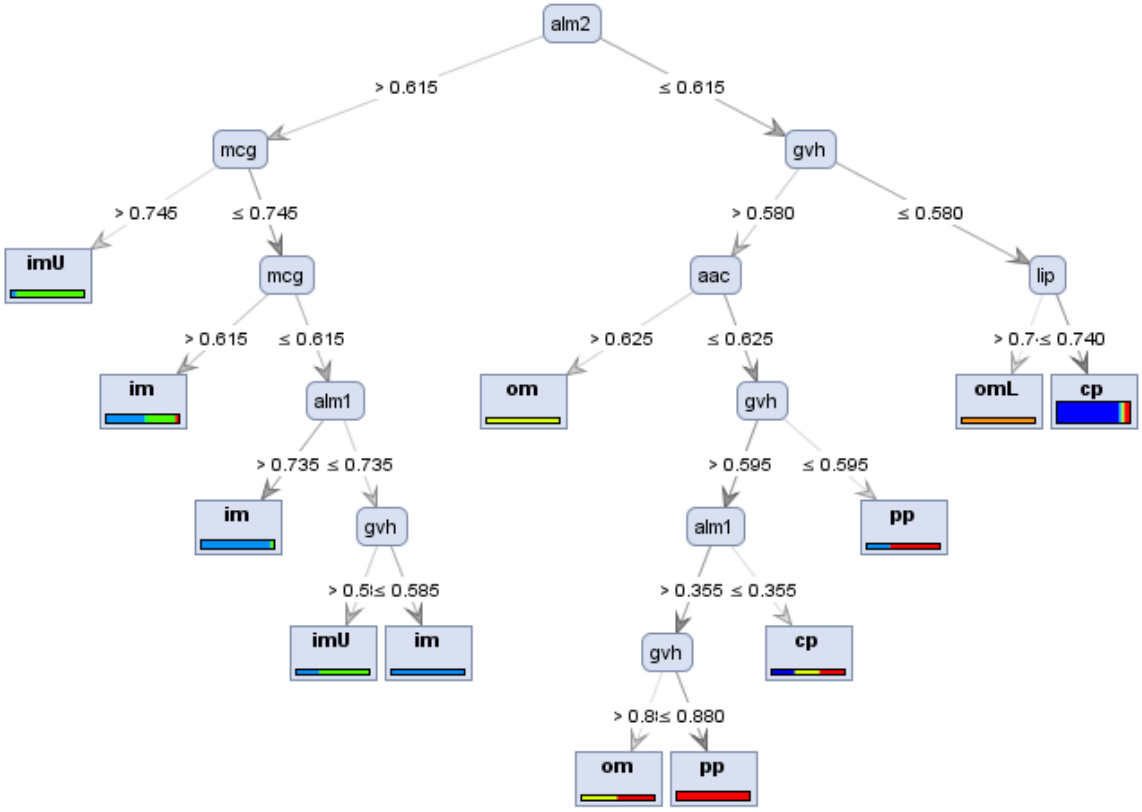
- Read CSV = načtení dat
- Split Data = rozdělení dat na trénovací a testovací množinu v poměru 2:1



- Decision Tree = vytvoří rozhodovací strom
- Apply Model pro trénovací data
- Apply Model pro testovací data
- Multiply
- Performance = pro zpracování matice záměn trénovacích dat
- Performance = pro zpracování matice záměn testovacích dat



Vizualizace rozhodovacího stromu:



Přepis rozhodovacího stromu do java syntaxe:

```
string eColi
if (alm2 > 0.615)
   if (mcg > 0.745)
          eColi = "imU"
   else
      if (mcg > 0.615)
         eColi = "im"
      else
         if (alm1 > 0.735)
            eColi = "im"
         else
            if (gvh > 0.585)
               eColi = "imU"
            else
               eColi = "im"

else
   if (gvh > 0.580)
      if (aac > 0.625)
         eColi = "om"
      else
         if (gvh > 0.595)
            if (alm1 > 0.355)
               if (gvh > 0.880)
                  eColi = "om"
               else
                  eColi = "pp"
            else
               eColi = "cp"
```

Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

```
      else
        eColi = "pp"
  else
    if (lip > 0.740)
      eColi = "omI."
    else
      eColi = "cp"
```

Matice záměn pro trénovací data:

accuracy: 85.27%

	true cp	true im	true imS	true imL	true imU	true om	true omL	true pp	class precision
pred. cp	92	3	1	0	1	4	0	8	84.40%
pred. im	0	45	0	1	10	0	0	1	78.95%
pred. imS	0	0	0	0	0	0	0	0	0.00%
pred. imL	0	0	0	0	0	0	0	0	0.00%
pred. imU	0	2	0	0	12	0	0	0	85.71%
pred. om	0	0	0	0	0	11	0	1	91.67%
pred. omL	0	0	0	0	0	0	2	0	100.00%
pred. pp	0	1	0	0	0	0	0	29	96.67%
class recall	100.00%	88.24%	0.00%	0.00%	52.17%	73.33%	100.00%	74.36%	

Matice záměn pro testovací data:

accuracy: 83.93%

	true cp	true im	true imS	true imL	true imU	true om	true omL	true pp	class precision
pred. cp	51	5	0	0	0	1	0	2	86.44%
pred. im	0	17	1	0	3	0	0	0	80.95%
pred. imS	0	0	0	0	0	0	0	0	0.00%
pred. imL	0	0	0	0	0	0	0	0	0.00%
pred. imU	0	3	0	0	9	0	0	0	75.00%
pred. om	0	0	0	0	0	4	0	1	80.00%
pred. omL	0	0	0	1	0	0	3	0	75.00%
pred. pp	0	1	0	0	0	0	0	10	90.91%
class recall	100.00%	65.38%	0.00%	0.00%	75.00%	80.00%	100.00%	76.92%	

Na trénovacích datech je přesnost určení bakterie eColi na 85.27%, u testovacích dat na 83.93%. Celková přesnost určení bakterie eColi je cca 84,5%. S ohledem na množství vstupních dat považuji výsledek za uspokojivý.

Zpracovala**: Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**