Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 20. 8. 2019

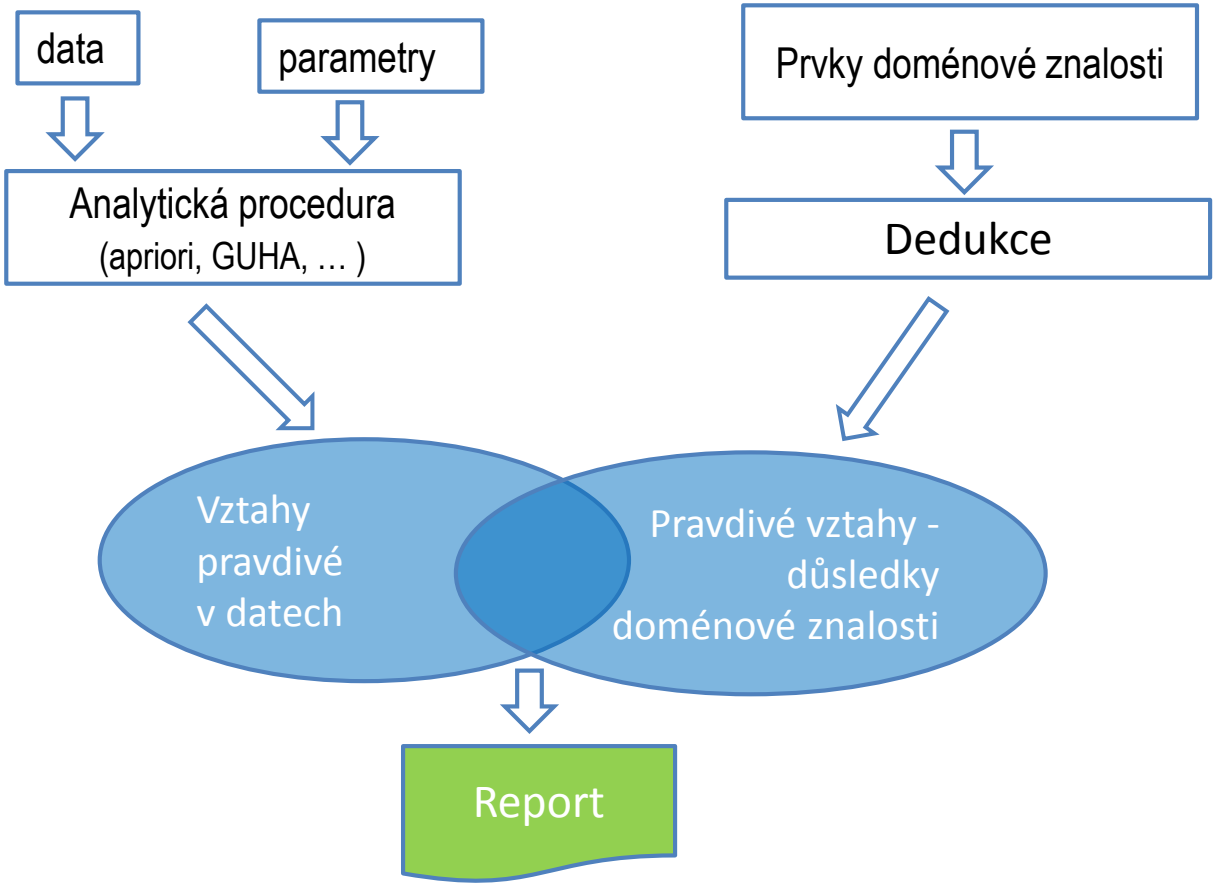
Typ úlohy: Asociační pravidla - práce s doménovou znalostí
Data: Stulong
Problém: *Automatické odfiltrování asociačních pravidel - důsledků doménové znalosti.*

Jan Rauch
Katedra informačního a znalostního inženýrství
Vysoká škola ekonomická v Praze

© Jan Rauch

Odfiltrování nezajímavých výstupů analytických procedur

Princip:



Příklad – data STULONG

Data exploration

Data matrix: Entry

Total number of rows: 1417

Filter: (empty)

Number of filtered rows: 1417

#	BMI	Subsc	Tric	Status	Education	Diastolic	Systolic	Beer	Liquors	Vine
1	(24;25>	<12;14)	<=4	married	secondary	<85;95)	<115;125)	he does not drink	more than 100 cc	he does not drink
2	(27;28>	<22;24)	10	married	university	<95;105)	<145;155)	he does not drink	he does not drink alcohol	up to half a litre
3	(28;29>	<14;16)	12	married	university	<75;85)	<115;125)	he does not drink	up to 100 cc	up to half a litre
4	(27;28>	<20;22)	10	married	apprentice	<65;75)	<115;125)	he does not drink	up to 100 cc	up to half a litre
5	(28;29>	<12;14)	<=4	married	university	<85;95)	<155;165)	up to 1 litre	up to 100 cc	up to half a litre
6	(31;32>	<26;28)	18-35	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
7	>32	<32;36)	15-17	single	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	up to half a litre
8	(26;27>	<36;72)	10	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
9	(25;26>	<18;20)	13-14	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
10	(22;23>	<14;16)	8	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
11	(26;27>	<20;22)	18-35	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
12	(27;28>	<22;24)	18-35	married	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	up to half a litre
13	(25;26>	<32;36)	10	married	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	up to half a litre
14	(30;31>			divorce	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	he does not drink
15	(21;22>	<32;36)	15-17	married	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	up to half a litre
16	(25;26>	<12;14)	10	married	[illegible]	[illegible]	[illegible]	[illegible]	c	he does not drink
17	(22;23>	<14;16)	8	divorce	[illegible]	[illegible]	[illegible]	[illegible]	k alcohol	he does not drink
18	(27;28>	<22;24)	18-35	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
19	(29;30>	<16;18)	15-17	married	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	up to half a litre
20	<=21	< 10	5	married	university	<85;95)	<125;135)	he does not drink	up to 100 cc	he does not drink
21	(25;26>	<16;18)	13-14	married	university	<75;85)	<115;125)	he does not drink	he does not drink alcohol	up to half a litre
22	(30;31>	<20;22)	9	married	apprentice	<75;85)	<115;125)	up to 1 litre	he does not drink alcohol	he does not drink
23	(27;28>	<16;18)	8	married	university	<75;85)	<125;135)	up to 1 litre	up to 100 cc	up to half a litre
24	(31;32>	<36;72)	13-14	married	secondary	<105;115)	<165;175)	up to 1 litre	he does not drink alcohol	he does not drink
25	(28;29>	<22;24)	7	married	university	<75;85)	<115;125)	he does not drink	he does not drink alcohol	he does not drink
26	(23;24>	< 10	6	married	secondary	<85;95)	<125;135)	he does not drink	he does not drink alcohol	up to half a litre
27	(29;30>	<20;22)	10	married	secondary	<75;85)	<135;145)	he does not drink	up to 100 cc	up to half a litre
28	(25;26>	<12;14)	<=4	divorced	apprentice	< 65	<105;115)	up to 1 litre	up to 100 cc	he does not drink
29	>32	<22;24)	12	married	apprentice	<75;85)	<135;145)	up to 1 litre	he does not drink alcohol	he does not drink
30	(23;24>	<10;12)	<=4	married	university	< 65	<105	up to 1 litre	he does not drink alcohol	up to half a litre
31	(28;29>	<24;26)	15-17	married	university	<75;85)	<125;135)	he does not drink	he does not drink alcohol	he does not drink
32	(29;30>	<14;16)	12	divorced	university	<85;95)	<135;145)	up to 1 litre	up to 100 cc	up to half a litre

Matice dat Entry

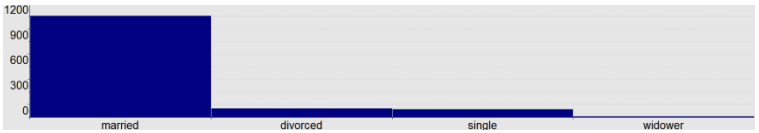
1417 řádků – pacientů, 64 sloupců

Část dat STULONG

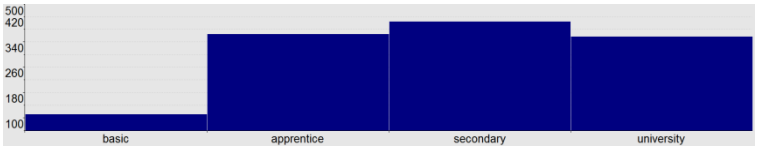
http://euromise.vse.cz/challenge2004/,

Data STULONG - použité atributy a jejich skupiny

Osobní

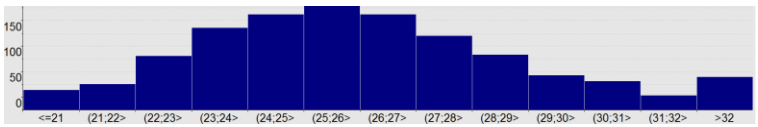


Stav – ženatý, rozvedený, svobodný, vdovec

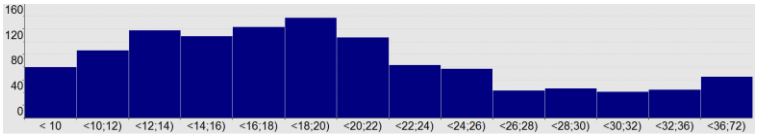


Vzdělání – základní, vyučen, maturita, vysoká škola

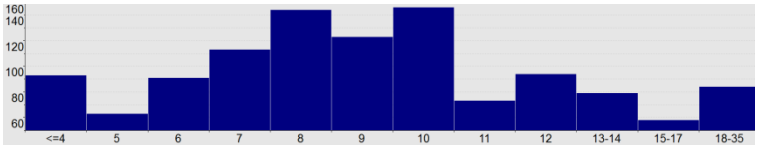
Měření



BMI – Body Mass Index

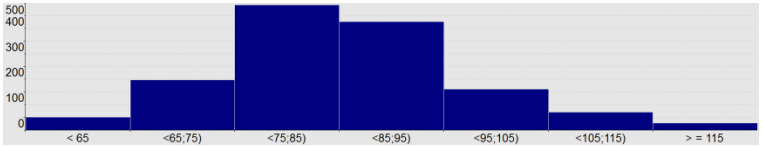


Subsc – kožní řasa subscapular v mm

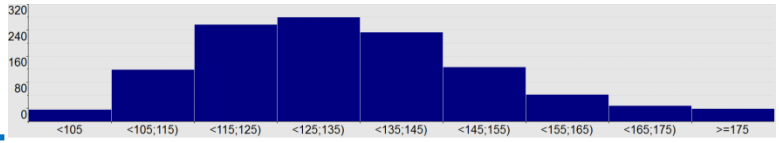


Tric – kožní řasa triceps v mm

Krevní tlak



Diast – diastolický krevní tlak

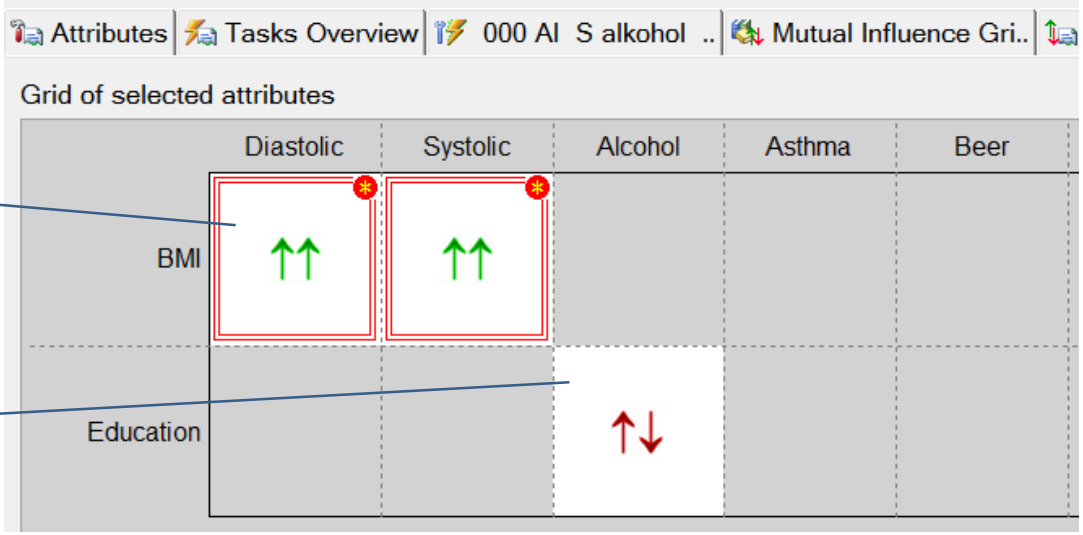


Syst – systolický krevní tlak

Prvky doménové znalosti – příklady, data STULONG

Jestliže roste BMI, tak roste i diastolický krevní tlak

Jestliže roste vzdělání, tak klesá spotřeba alkoholu



Příklad analytické otázky

Jsou v matici dat **Entry** nějaké zajímavé vztahy mezi kombinacemi atributů ze skupin **Osobní** a **Měření** na straně jedné a kombinacemi atributů ze skupiny **Krevní tlak** na straně druhé ?

Nezajímají nás však důsledky známého vztahu že když roste BMI, tak roste i diastolický krevní tlak.

Obecně:

Entry? (BMI $\uparrow\uparrow$ Diastolic) $\not\rightarrow$ Osobní $\wedge$ Měření $\approx$ Krevní tlak

4ft-Miner:

Entry? (BMI $\uparrow\uparrow$ Diastolic) $\not\rightarrow$ $\mathcal{B}$(Osobní) $\wedge$ $\mathcal{B}$(Měření) $\Rightarrow_{0.9,\ 30}$ $\mathcal{B}$(Krevní tlak)

Zadání procedury 4ft-Miner

Entry? $\mathcal{B}$(Osobní) $\wedge$ $\mathcal{B}$(Měření) $\Rightarrow_{0.9,\ 30}$ $\mathcal{B}$(Krevní tlak)

ANTECEDENT	
Osobní	Con, 0 - 2
» Stav (subset), 1 - 1	B, pos
» Vzdělání (subset), 1 - 1	B, pos
Měření	Con, 1 - 3
» BMI (seq), 1 - 3	B, pos
» Subscapular (seq), 1 - 3	B, pos
» Triceps (seq), 1 - 3	B, pos

QUANTIFIERS
PIM p= 0.900
BASE p= 30 Abs.

Generation information

Status: Solved, 7 run(s)

SUCCEDENT	
Krevní tlak	Con, 1 - 2
» Diastolic (seq), 1 - 3	B, pos
» Systolic (seq), 1 - 3	B, pos

Výstup procedury 4ft-Miner

Task run

Start: 20.11.2017 23:34:33 Total time: 0h 3m 8s

Number of verifications: 13977840

Number of hypotheses: 205 Mode: Standard

13.9 * 10^6 verifikací

205 pravidel

Edit group

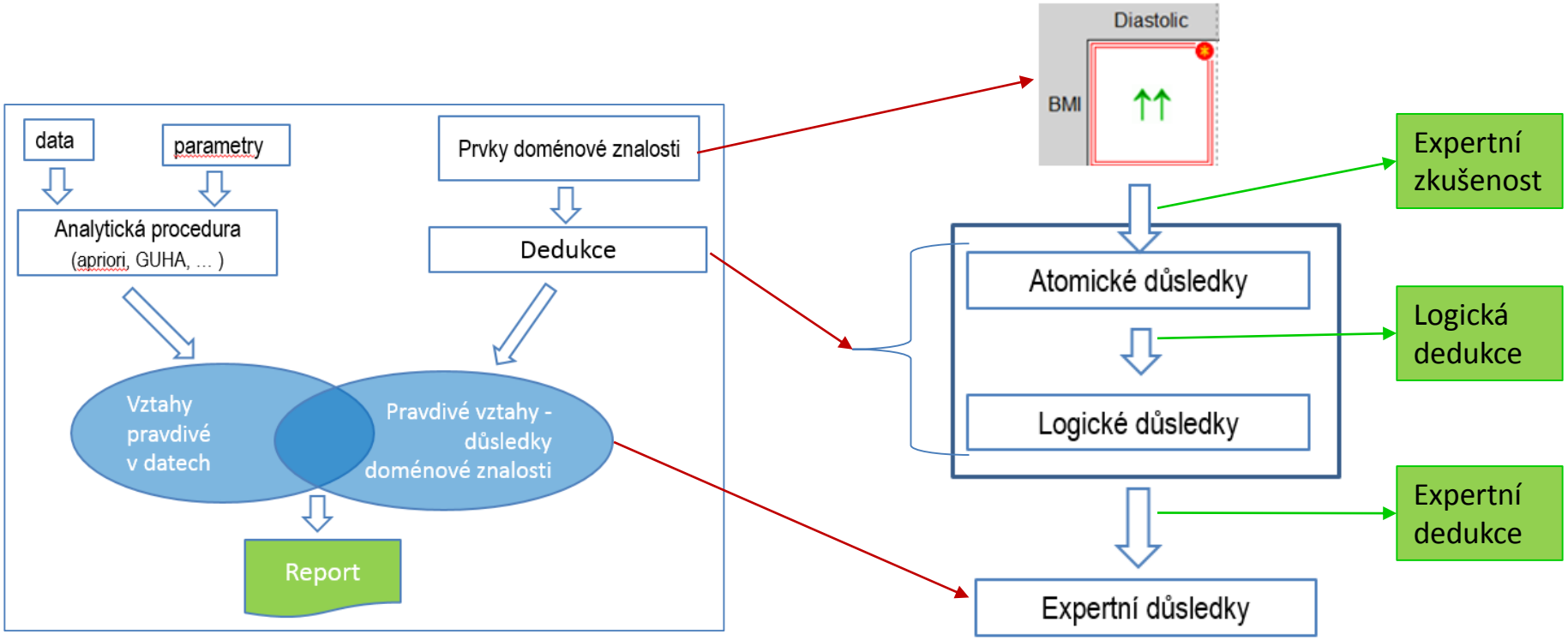
Actual group of hypotheses: All hypotheses

Hypotheses in group: 205 Shown hypotheses: 205

Důsledky *BMI* ↑↑ *Diastolic* ?

Nr.	Id	Conf	Hypothesis
1	21	0.971	**BMI**(*(21;22>*) & **Subscapular**(<*14*) >÷< **Diastolic**(*<65;95)*)
2	166	0.970	**Status**(*married*) & **Education**(*university*) & **Triceps**(*5,6*) >÷< **Diastolic**(*<75;105)*)
3	192	0.969	**Education**(*university*) & **BMI**(*(22;25>*) & **Subscapular**(*<18;24)*) >÷< **Diastolic**(*<65;95)*)
4	61	0.968	**BMI**(*(27;28>*) & **Subscapular**(*<16;22)*) & **Triceps**(*7..9*) >÷< **Diastolic**(*<75;105)*)
5	33	0.968	**BMI**(*(23;26>*) & **Subscapular**(*<18;24)*) & **Triceps**(*12..15-17*) >÷< **Diastolic**(*<65;95)*)
6	12	0.968	**BMI**(<=*22*) & **Triceps**(*5..7*) >÷< **Diastolic**(*<65;95)*)
7	20	0.960	**BMI**(*(21;22>*) >÷< **Diastolic**(*<65;95)*)
8	83	0.957	**Status**(*married*) & **BMI**(<=*22*) & **Subscapular**(<*14*) >÷< **Diastolic**(*<65;95)*)
9	81	0.956	**Status**(*married*) & **BMI**(<=*22*) >÷< **Diastolic**(*<65;95)*)
10	2	0.955	**BMI**(<=*22*) >÷< **Diastolic**(*<65;95)*)
11	139	0.953	**Status**(*married*) & **Subscapular**(*<26;32)*) & **Triceps**(*8..10*) >÷< **Diastolic**(*<75;105)*)

Příklad dedukce – asociační pravidla (1)



Atomické důsledky BMI ↑↑ Diastolic – příklad (1)

Atomické důsledky:

BMI(nízké) $\Rightarrow_{p,\ Base}$ Diastolic(nízké)

BMI(střední) $\Rightarrow_{p,\ Base}$ Diastolic(střední)

BMI(vysoké) $\Rightarrow_{p,\ Base}$ Diastolic(vysoké)

Expertní zkušenost

Atomic Association Rules

Antecedent	Succedent
<= (24;25>	<= <75;85)
(23;24>...(28;29>	<75;85), <85;95)
>= (27;28>	>= <85;95)

3

Add | Del | Join | Split

Detail | Autocreate

Contingency table — Blind — Values

	< 65	<65;75)		<85;95)			> = 115
<=21	1	9	21	7	1	0	0
(21;22>	0	8	27	13	1	0	1
(22;23>	9	19	39	26	7	4	2
(23;24>	8	27	65	38	16	7	0
(24;25>	9	27	66	60	18	3	4
(25;26>	8	38	67	72	18	7	0
(26;27>	10	18	65	56	24	12	2
(27;28>	2	18	52	47	17	6	3
(28;29>	0	13	34	35	11	12	3
(29;30>	0	8	19	23	9	6	3
(30;31>	1	3	12	18	13	7	2
(31;32>	0	2	9	9	4	1	3
>32	1	5	12	19	20	4	3

0 36.00 72

Atomické důsledky BMI ↑↑ Diastolic – příklad (2)

Expertní zkušenost

Atomické důsledky:

BMI(nízké) $\Rightarrow_{p,\ Base}$ Diastolic(nízké) $p \geq 0.9$, Base ≥ 30

Antecedent	Succedent	3
<= (24;25>	<= <75;85)	
(23;24>...(28;29>	<75;85), <85;95)	
>= (27;28>	>= <85;95)	

Contingency table ○ Blind ● Values

	< 65	<65;75)		<85;95)			> = 115
<=21	1	9	21	7	1	0	0
(21;22>	0	8	27	13	1	0	1
(22;23>	9	19	39	26	7	4	2
(23;24>	8	27	65	38	16	7	0
(24;25>	9	27	66	60	18	3	4

BMI(≤21) $\Rightarrow_{p,\ Base}$ Diastolic(<65), … BMI(21;22⟩ $\Rightarrow_{p,\ Base}$ Diastolic(<75),

BMI(≤22) $\Rightarrow_{p,\ Base}$ Diastolic(<65), … BMI(21;23⟩ $\Rightarrow_{p,\ Base}$ Diastolic(<75)

BMI(≤23) $\Rightarrow_{p,\ Base}$ Diastolic(<65), … BMI(21;24⟩ $\Rightarrow_{p,\ Base}$ Diastolic(<75)

BMI(≤24) $\Rightarrow_{p,\ Base}$ Diastolic(<65), … BMI(21;25⟩ $\Rightarrow_{p,\ Base}$ Diastolic(<75)

BMI(≤25) $\Rightarrow_{p,\ Base}$ Diastolic(<65), …

….

BMI(21;22⟩ $\Rightarrow_{p,\ Base}$ Diastolic⟨65;75) … BMI(23;25⟩ $\Rightarrow_{p,\ Base}$ Diastolic⟨65;85), …

… BMI(24;25⟩ $\Rightarrow_{p,\ Base}$ Diastolic⟨75;85)

Atomické důsledky BMI ↑↑ Diastolic – příklad (3)

Ve výstupu nejsou atomické důsledky BMI ↑↑ Diastolic

Task run

Number of verifications: 13977840

Number of hypotheses: 205 Mode: Standard

Contingency table ○ Blind ● Values

	< 65	<65;75)		<85;95)			> = 115
<=21	1	9	21	7	1	0	0
(21;22>	0	8	27	13	1	0	1
(22;23>	9	19	39	26	7	4	2
(23;24>	8	27	65	38	16	7	0
(24;25>	9	27	56	60	18	3	4
(25;26>	8	38	67	72	18	7	0
(26;27>	10	11	65	56	24	12	2
(27;28>	2	18	52	47	17	6	3
(28;29>	0	13	34	35	11	12	3
(29;30>	0	8	19	23	9	6	3
(30;31>	1	3	12	18	13	7	2
(31;32	0	2	9	9	4	1	3
32	1	5	12	19	20	4	3

Actual group of hypotheses: All hypotheses

Hypotheses in group: 205 Shown hypotheses: 205

Nr.	Id	Conf	Hypothesis
1	21	0.971	**BMI**(*(21;22>*) & **Subscapular**(<*14*) >÷< **Diastolic**(*<65;95)*)
2	166	0.970	**Status**(*married*) & **Education**(*university*) & **Triceps**(*5,6*) >÷< **Diastolic**(*<75;105)*)
3	192	0.969	**Education**(*university*) & **BMI**(*(22;25>*) & **Subscapular**(*<18;24)*) >÷< **Diastolic**(*<65;95)*)
4	61	0.968	**BMI**(*(27;28>*) & **Subscapular**(*<16;22)*) & **Triceps**(*7..9*) >÷< **Diastolic**(*<75;105)*)
5	33	0.968	**BMI**(*(23;26>*) & **Subscapular**(*<18;24)*) & **Triceps**(*12..15-17*) >÷< **Diastolic**(*<65;95)*)
6	12	0.968	**BMI**(<=*22*) & **Triceps**(*5..7*) >÷< **Diastolic**(*<65;95)*)
7	20	0.960	**BMI**(*(21;22>*) >÷< **Diastolic**(*<65;95)*)
8	83	0.957	**Status**(*married*) & **BMI**(<=*22*) & **Subscapular**(<*14*) >÷< **Diastolic**(*<65;95)*)
9	81	0.956	**Status**(*married*) & **BMI**(<=*22*) >÷< **Diastolic**(*<65;95)*)
10	2	0.955	**BMI**(<=*22*) >÷< **Diastolic**(*<65;95)*)
11	139	0.953	**Status**(*married*) & **Subscapular**(*<26;32)*) & **Triceps**(*8..10*) >÷< **Diastolic**(*<75;105)*)

?

Logický důsledek atomického důsledku – příklad (1)

Logická dedukce

LM STULONG_RuleML_obnovena MB - LISp-Miner Workspace module - 25.28.00

File Data Introduction Preprocessing Interactive Analysis Data-mining Tasks Domain Knowledge Window Help

460 příklad Doménové.. | Mutual influence (7) | 460 příklad Doménové.. | 460 příklad Doménové.. | Hypothesis (16982)

Antecedent: BMI((21;22>)
Succedent: Diastolic(<65;75)...<85;95))
Condition: (empty)

TEXT | DATA | FREQUENCIES | PIE | BAYES

	Succedent	¬Succedent
Antecedent	48	2
¬Antecedent	1058	301

0 529.00 1058

Value: Abs | Rel sum | Rel max | Rel row | Rel col

Reset | Y-Max Lock

Export | Add as column | Attributes

BMI(21;22⟩ $\Rightarrow_{0.96,48}$ Diastolic ⟨65;95)

Není to atomický důsledek, ale

BMI(21;22⟩ $\Rightarrow_{0.96,48}$ Diastolic ⟨65;85)

je atomický důsledek .

	< 65	<65;75)		<85;95)
<=21	1	9	21	7
(21;22>	0	8	27	13
(22;23>	9	19	39	26
(23;24>	8	27	65	38
(24;25>	9	27	66	60

Logický důsledek atomického důsledku – příklad (2)

Nalezené pravidlo:

Antecedent			
Měření	Conjunction		
	BMI	(21;22>	
Succedent			
Krevní tlak	Conjunction		
	Diast	<65;75)...<85;95)	
Contingency table			
	Succedent	NOT Succedent	
Antecedent	48	2	50
NOT Antecedent	1065	302	1367
	1113	304	1417

$\mathrm{BMI}\langle 21;22\rangle \Rightarrow_{0.96,48} \mathrm{Diastolic}\ \langle 65;95)$

Logická dedukce

atomický důsledek: $\mathrm{BMI}\langle 21;22\rangle \Rightarrow_{0.96,48} \mathrm{Diastolic}\ \langle 65;85)$

$\mathrm{BMI}\langle 21,22\rangle \Rightarrow_{0.96,48} \mathrm{Diastolic}\ \langle 65;95)$ logicky plyne z $\mathrm{BMI}\langle 21;22\rangle \Rightarrow_{0.96,48} \mathrm{Diastolic}\ \langle 65;85)$

matice dat	Diastolic ⟨65;85)	¬ Diastolic ⟨65;85)
BMI(21;22⟩	*a*	*b*
¬BMI(21;22⟩	*c*	*d*

matice dat	Diastolic ⟨65;95)	¬ Diastolic ⟨65;95)
BMI(21;22⟩	*a'*	*b'*
¬BMI(21;22⟩	*c'*	*d'*

protože a' ≥ a ∧ b' ≤ b a tudíž $\frac{a}{a+b} \geq 0.96 \wedge a \geq 48 \quad \Rightarrow \quad \frac{a'}{a'+b'} \geq 0.96 \wedge a' \geq 48$

Logický důsledek atomického důsledku – příklad (3)

Ve výstupu jsou 3 logické důsledky atomických důsledků BMI ↑↑ Diastolic

Logická dedukce

Actual group of hypotheses: 02 Logické důsledky atomických důsledků

Hypotheses in group: 3 Shown hypotheses: 3

Nr.	Id	Conf	Hypothesis
1	20	0.960	**BMI**(*(21;22>*) >÷< **Diastolic**(*<65;95)*)
2	2	0.955	**BMI**(<=*22*) >÷< **Diastolic**(*<65;95)*)
3	1	0.949	**BMI**(*<=21*) >÷< **Diastolic**(*<65;95)*)

Contingency table ○ Blind ● Values

	< 65	<65;75)		<85;95)			> = 115
<=21	1	9	21	7	1	0	0
(21;22>	0	8	27	13	1	0	1
(22;23>	9	19	39	26	7	4	2
(23;24>	8	27	65	38	16	7	0
(24;25>	9	27	66	60	18	3	4
(25;26>	8	38	67	72	18	7	0
(26;27>	10	18	65	56	24	12	2
(27;28>	2	18	52	47	17	6	3
(28;29>	0	13	34	35	11	12	3
(29;30>	0	8	19	23	9	6	3
(30;31>	1	3	12	18	13	7	2
(31;32>	0	2	9	9	4	1	3
>32	1	5	12	19	20	4	3

Logický důsledek – poznámka

Asociační pravidlo $\varphi' \approx \psi'$ je logickým důsledkem asociačního pravidla $\varphi \approx \psi$ (logicky plyne z asociačního pravidla $\varphi \approx \psi$) pokud pro každou matici dat M platí:

Je-li $\varphi \approx \psi$ pravdivé v M, pak je také $\varphi' \approx \psi'$ pravdivé v M.

Více viz v přednášce o observačních kalkulech.

Příklad – viz předchozí slide:

$\boxed{\text{BMI}(21;22\rangle \Rightarrow_{0.96,48} \text{Diastolic}\ \langle 65;95)}$ $\boxed{\text{logicky plyne z}}$ $\boxed{\text{BMI}(21;22\rangle \Rightarrow_{0.96,48} \text{Diastolic}\ \langle 65,85)}$

matice dat	Diastolic ⟨65;85)	¬ Diastolic ⟨65;85)
BMI(21;22⟩	*a*	*b*
¬BMI(21;22⟩	*c*	*d*

matice dat	Diastolic ⟨65;95)	¬ Diastolic ⟨65;95)
BMI(21;22⟩	*a'*	*b'*
¬BMI(21;22⟩	*c'*	*d'*

protože $a' \geq a \wedge b' \leq b$ a tudíž $\dfrac{a}{a+b} \geq 0.96 \wedge a \geq 48 \quad \Rightarrow \quad \dfrac{a'}{a'+b'} \geq 0.96 \wedge a' \geq 48$