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

Typ úlohy: vysoký rozdíl konfidencí

Data: Hotel

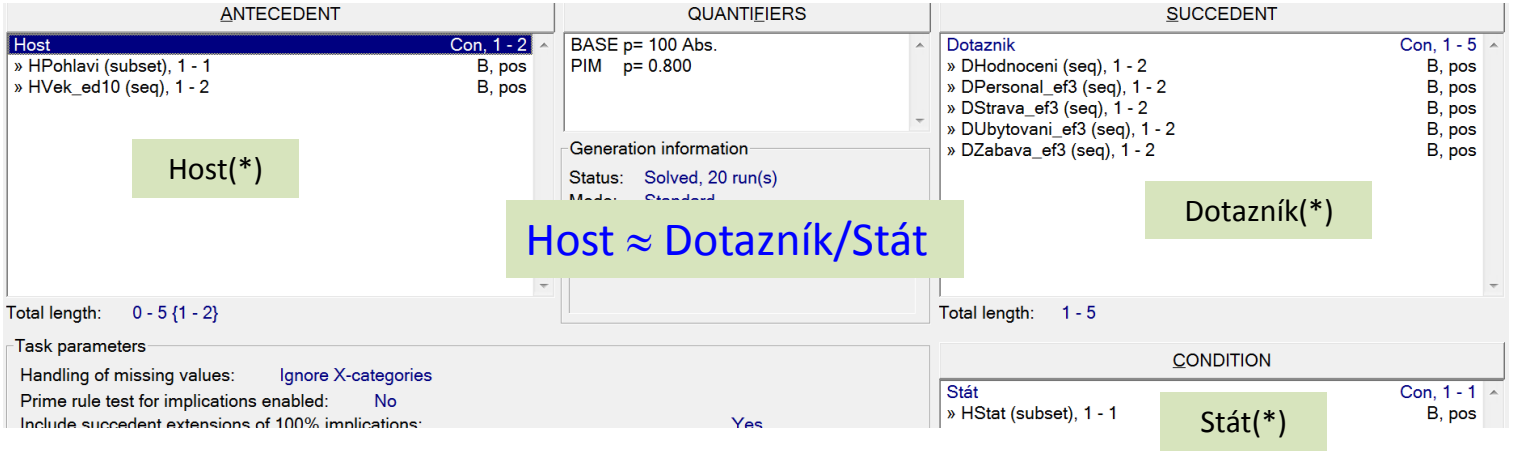
Problém: *Mezi kterými státy jsou významné rozdíly ohledně vztahů mezi charakteristikami hosta (pohlaví, věk) a typickými parametry odpovědí v dotazníku pobytu?*

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

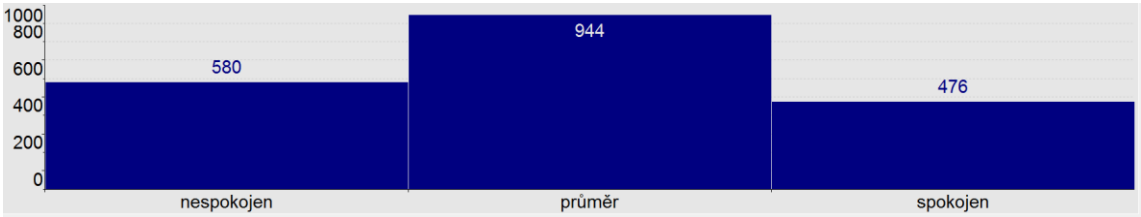
SD4ft-Miner – příklad analytické otázky

HOTEL:
Mezi kterými státy jsou významné rozdíly ohledně vztahů mezi charakteristikami hosta (pohlaví, věk) a typickými parametry odpovědí v dotazníku pobytu?
Stát(?) x Stát(?) [Host ≈ Dotazník]

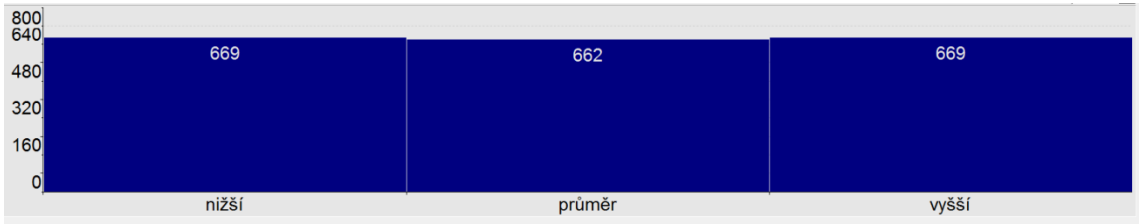
Stát(?) x Stát(?) [Host ≈ Dotazník] – pokus o řešení pomocí 4ft-Miner



DHodnoceni



DPersonal_ef3



DStrava_ef3, DUbytovani_ef3, DZabava_ef3 – analogicky k DPersonal_ef3

Stát(?) x Stát(?) [Host ≈ Dotazník] – pokus o řešení pomocí procedury 4ft-Miner

Task run

Start: 29.10.2015 21:41:07 Total time: 0h 0m 0s

Number of verifications: 39169

Number of hypotheses: 55 Mode: Standard

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 55 Shown hypotheses: 55 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	7	1.000	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) / **HStat**(*ČR*)
2	8	0.985	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) / **HStat**(*ČR*)
3	17	0.985	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DPersonal**(>=*průměr*) / **HStat**(*ČR*)
4	14	0.969	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) & **DUbytovani**(>=*průměr*) / **HStat**(*ČR*)
5	23	0.969	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DUbytovani**(>=*průměr*) / **HStat**(*ČR*)
6	10	0.954	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) & **DUbytovani**(>=*průměr*) / **HStat**(*ČR*)
7	19	0.954	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DPersonal**(>=*průměr*) & **DUbytovani**(>=*průměr*) / **HStat**(*ČR*)
8	4	0.922	**HPohlavi**(*žena*) & **HVek**(*<25;44>*) >÷< **DHodnoceni**(*nespokojen*,*průměr*) / **HStat**(*ČR*)
9	52	0.918	**HVek**(<=*34*) >÷< **DHodnoceni**(*nespokojen*,*průměr*) / **HStat**(*Německo*)
10	35	0.909	**HVek**(*<55;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) / **HStat**(*ČR*)
11	36	0.901	**HVek**(*<55;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) / **HStat**(*ČR*)
12	39	0.901	**HVek**(*<55;64>*) >÷< **DPersonal**(>=*průměr*) / **HStat**(*ČR*)
13	41	0.901	**HVek**(*<55;64>*) >÷< **DUbytovani**(>=*průměr*) / **HStat**(*ČR*)
14	16	0.892	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DHodnoceni**(*průměr*,*spokojen*) & **DZabava**(>=*průměr*) / **HStat**(*ČR*)
15	25	0.892	**HPohlavi**(*žena*) & **HVek**(*<45;64>*) >÷< **DZabava**(>=*průměr*) / **HStat**(*ČR*)
16	51	0.886	**HPohlavi**(*žena*) >÷< **DHodnoceni**(*nespokojen*,*průměr*) / **HStat**(*Německo*)

- Pokud rozdíl mezi státy vyjádříme rozdílem konfidencí, musíme zadat minimální hodnotu konfidence = 0 a pak hledat dvojice s rozdílem konfidencí nad daným prahem v rozsáhlém výstupu .
- Obdobně pro jiná vyjádření rozdílu.

Stát(?) x Stát (?) [Host ≈ Dotazník], princip řešení (1)

Jsou významné rozdíly mezi jednotlivými státy ohledně vztahů mezi charakteristikami hosta (pohlaví, věk) a typickými parametry odpovědí v dotazníku pobytu?

Princip řešení – příklad: ČR x Polsko [HVek(<=34) ≈ DHodnocení(průměr)]

ČR	DHodnocení(průměr)	¬ Dhodnocení(průměr)
HVek(<=34)	a_1	b_1
¬ HVek(<=34)	c_1	d_1

Polsko	DHodnocení(průměr)	¬ Dhodnocení(průměr)
HVek(<=34)	a_2	b_2
¬ HVek(<=34)	c_2	d_2

Rozdíl mezi ČR a Polskem ohledně vztahu mezi HVek(<=34) a DHodnocení(průměr) považujeme za významný, pokud platí (například)

$$\frac{a_1}{a_1+b_1}-\frac{a_2}{a_2+b_2}\geq 0.3 \wedge a_1\geq 50 \wedge a_2\geq 50$$

.

Stát(?) x Stát (?) [Host ≈ Dotazník], princip řešení(2)

Rozdíl mezi ČR a Polskem ohledně vztahu mezi HVek(<=34) a DHodnocení(průměr)

považujeme za významný, pokud platí (například) $\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2} \geq 0.3 \wedge a_1 \geq 50 \wedge a_2 \geq 50$

ČR	DHodnocení(průměr)	¬ Dhodnocení(průměr)
HVek(<=34)	a_1	b_1
¬ HVek(<=34)	c_1	d_1

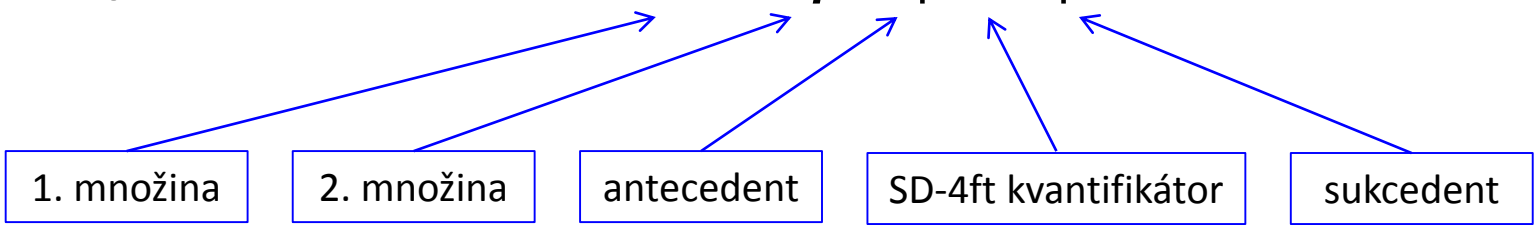
Polsko	DHodnocení(průměr)	¬ Dhodnocení(průměr)
HVek(<=34)	a_2	b_2
¬ HVek(<=34)	c_2	d_2

SD-pravidlo:

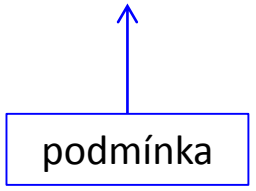
Stát(ČR) x Stát(Polsko): HVek(<=34) $\Rightarrow_{0.3,50,50}$ DHodnocení(průměr)

SD-pravidlo

SD-pravidlo: α x β: $\varphi \approx \psi$



Podmíněné SD-pravidlo: α x β: $\varphi \approx \psi$ / χ



α, β, φ, ψ, χ jsou booleovské atributy

SD4ft-tabulka

SD4ft-tabulka pro SD-pravidla α x β: $\varphi \approx \psi$ α x β: $\varphi \approx \psi / \chi$ v matici dat $\mathcal{M}$:

$$\langle T_\alpha, T_\beta, n, n_T \rangle$$

- $T_\alpha = 4ft(\varphi,\psi,\mathcal{M}/(\chi \wedge \alpha))$ $T_\beta = 4ft(\varphi,\psi,\mathcal{M}/(\chi \wedge \beta))$

$\mathcal{M}/(\alpha\wedge\chi)$	ψ	$\neg\psi$
φ	a_α	b_α
$\neg\varphi$	c_α	d_α

$\mathcal{M}/(\beta\wedge\chi)$	ψ	$\neg\psi$
φ	a_β	b_β
$\neg\varphi$	c_β	d_β

- n je počet řádků v matici dat $\mathcal{M}/\chi$
- n_T je počet řádků v matici dat $\mathcal{M}$
- pro nepodmíněné SD-pravidlo předpokládáme, že χ je identicky pravdivé

SD4ft-kvantifikátor

Matice dat $\mathcal{M}$, SD-pravidlo α x β: $\varphi \approx \psi$, α x β: $\varphi \approx \psi / \chi$

SD4ft-kvantifikátoru $\approx$ je přiřazena podmínka na $\langle T_\alpha, T_\beta, n, n_T \rangle$

$\mathcal{M}/(\alpha \wedge \chi)$	ψ	$\neg\,\psi$
φ	a_α	b_α
$\neg\,\varphi$	c_α	d_α

$\mathcal{M}/(\beta \wedge \chi)$	ψ	$\neg\,\psi$
φ	a_β	b_β
$\neg\,\varphi$	c_β	d_β

Příklad: SD4ft-kvantifikátoru $\Rightarrow_{0.3,50,50}$ je přiřazena podmínka

$$\frac{a_\alpha}{a_\alpha + b_\alpha} - \frac{a_\beta}{a_\beta + b_\beta} \geq 0.3 \wedge a_\alpha \geq 50 \wedge a_\beta \geq 50$$

SD-pravidlo je pravdivé v matici dat $\mathcal{M}$

SD-pravidlo α x β: $\varphi \approx \psi$, α x β: $\varphi \approx \psi / \chi$ je pravdivé v matici dat $\mathcal{M}$: v matici dat $\mathcal{M}$ je pro SD4ft-tabulku $\langle T_\alpha, T_\beta, n, n_T \rangle$ splněna podmínka přiřazená SD4ft-kvantifikátoru $\approx$.

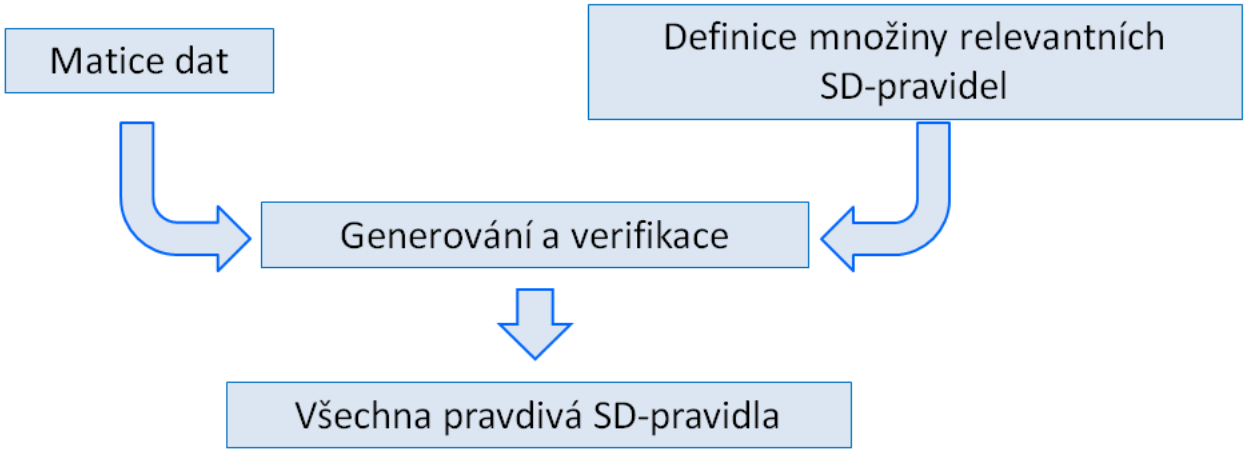
- $T_\alpha = 4ft(\varphi,\psi,\mathcal{M}/(\chi \wedge \alpha))$ $\quad T_\beta = 4ft(\varphi,\psi,\mathcal{M}/(\chi \wedge \beta))$

$\mathcal{M}/(\alpha\wedge\chi)$	ψ	$\neg\,\psi$
φ	a_α	b_α
$\neg\,\varphi$	c_α	d_α

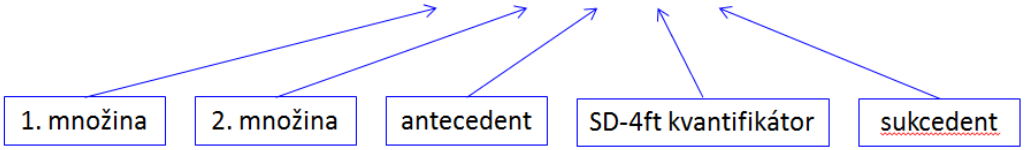
$\mathcal{M}/(\beta\wedge\chi)$	ψ	$\neg\,\psi$
φ	a_β	b_β
$\neg\,\varphi$	c_β	d_β

- n je počet řádků v matici dat $\mathcal{M}/\chi$
- n_T je počet řádků v matici dat $\mathcal{M}$

GUHA procedura SD4ft-Miner

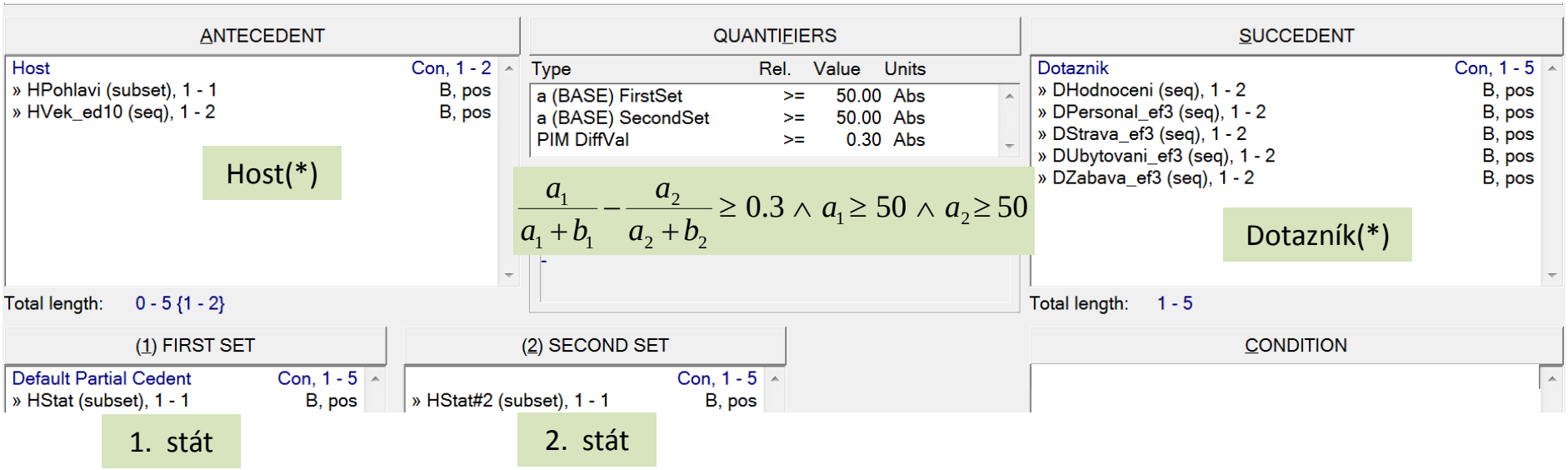


SD-pravidlo: α x β: $\varphi \approx \psi$



α, β, φ, ψ, χ jsou relevantní cedenty, viz
http://lispminer.vse.cz/wiki/doku.php?id=lmtask:settings:ftcedenthierarchy

Stát(?) x Stát(?) [Host ≈ Dotazník] – SD4ft-Miner, vstup



Stát(?) x Stát(?) [Host ≈ Dotazník] – SD4ft-Miner, výstup

25 vteřin
2 997 100 verifikací
92 vztahů

Task run
Start: 30.10.2015 08:40:01 Total time: 0h 0m 25s
Number of verifications: 2997100
Number of hypotheses: 92 Mode: Standard

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses
Hypotheses in group: 92 Shown hypotheses: 92 Highlighted: 0

Delete hypotheses

Nr.	Id	Df-Conf	1:Conf	2:Conf	Hypothesis
1	83	0.429	0.885	0.457	**HVek**(>=*65*) >+< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) & **DUbytovani**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
2	88	0.429	0.885	0.457	**HVek**(>=*65*) >+< **DPersonal**(>=*průměr*) & **DUbytovani**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
3	6	0.426	0.981	0.556	**HPohlavi**(*žena*) & **HVek**(*<55;74>*) >+< **DHodnoceni**(*nespokojen*,*průměr*) : **HStat**(*Německo*) × **HStat**(*ČR*)
4	84	0.410	0.852	0.442	**HVek**(>=*65*) >+< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) & **DZabava**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
5	89	0.410	0.852	0.442	**HVek**(>=*65*) >+< **DPersonal**(>=*průměr*) & **DZabava**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
6	82	0.409	0.967	0.558	**HVek**(>=*65*) >+< **DHodnoceni**(*průměr*,*spokojen*) & **DPersonal**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
7	87	0.409	0.967	0.558	**HVek**(>=*65*) >+< **DPersonal**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)
8	25	0.403	0.984	0.581	**HPohlavi**(*žena*) & **HVek**(<=*34*) >+< **DPersonal**(<=*průměr*) : **HStat**(*Německo*) × **HStat**(*Rakousko*)
9	12	0.403	0.762	0.359	**HVek**(*<25;34>*) >+< **DHodnoceni**(*průměr*) : **HStat**(*Německo*) × **HStat**(*ČR*)
10	3	0.398	1.000	0.602	**HPohlavi**(*muž*) & **HVek**(*<25;34>*) >+< **DHodnoceni**(*průměr*,*spokojen*) : **HStat**(*Německo*) × **HStat**(*ČR*)
11	24	0.397	0.967	0.570	**HPohlavi**(*žena*) & **HVek**(<=*34*) >+< **DHodnoceni**(*nespokojen*,*průměr*) & **DPersonal**(<=*průměr*) : **HStat**(*Německo*) × **HStat**(*Rakousko*)
12	31	0.393	0.515	0.123	**HPohlavi**(*žena*) >+< **DHodnoceni**(*nespokojen*,*průměr*) & **DStrava**(*nižší*) & **DUbytovani**(<=*průměr*) & **DZabava**(<=*průměr*) : **HStat**(*Slovensko*) × **HSt**
13	38	0.393	0.515	0.123	**HPohlavi**(*žena*) >+< **DStrava**(*nižší*) & **DUbytovani**(<=*průměr*) & **DZabava**(<=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*ČR*)
14	28	0.388	0.515	0.127	**HPohlavi**(*žena*) >+< **DHodnoceni**(*nespokojen*,*průměr*) & **DPersonal**(<=*průměr*) & **DStrava**(*nižší*) & **DZabava**(<=*průměr*) : **HStat**(*Slovensko*) × **HSta**
15	35	0.388	0.515	0.127	**HPohlavi**(*žena*) >+< **DPersonal**(<=*průměr*) & **DStrava**(*nižší*) & **DZabava**(<=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*ČR*)
16	85	0.387	0.902	0.514	**HVek**(>=*65*) >+< **DHodnoceni**(*průměr*,*spokojen*) & **DUbytovani**(>=*průměr*) : **HStat**(*Slovensko*) × **HStat**(*Rakousko*)

SD4ft-Miner – detail výstupu prvního SD-pravidla

LM Hotel MB - LISp-Miner Workspace module - 25.26.00

HVek⟨65;85⟩ ≈ DHodnoceni(průměr, spokojen) ∧ DPersonal(průměr, vyšší) ∧ DUbytovani(průměr, vyšší)

Antecedent: HVek(<65;74>, <75;85>)
Succedent: DHodnoceni(průměr, spokojen) & DPersonal(průměr, vyšší) & DUbytovani(průměr, vyšší)
First set: HStat(Slovensko)
Second set: HStat(Rakousko)
Condition: (empty)

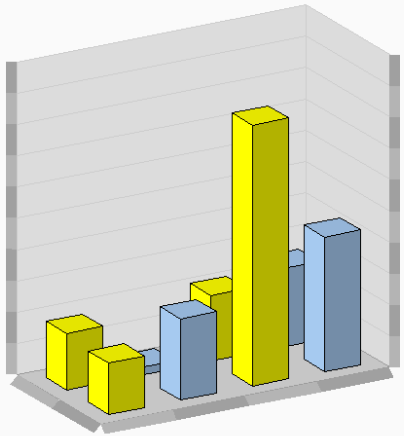
First set: HStat(Slovensko)

Second set: HStat(Rakousko)

TEXT | DATA | FIRST SET | SECOND SET | F+S SET | DIFF ABS | DIFF REL

	Succedent	¬Succedent	Succedent	¬Succedent
Antecedent	54	7	63	75
¬Antecedent	49	78	250	129

0 125.00 250



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

Reset

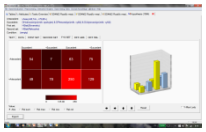
Y-Max Lock

Export

Ready

NUM

CS 100% 8:48



SD4ft-Miner – detail výstupu, komentář

OK = DHodnoceni(průměr, spokojen) ∧ DPersonal(průměr, vyšší) ∧ DUbytovani(průměr, vyšší)

Slovensko	OK	¬OK
HVek⟨65;85⟩	54	7
¬HVek⟨65;85⟩	49	78

Rakousko	OK	¬OK
HVek⟨65;85⟩	63	75
¬HVek⟨65;85⟩	250	129

$$\frac{54}{54+7} = 0.885$$

$$\frac{54}{54+7} - \frac{63}{63+75} = 0.429$$

$$\frac{63}{63+75} = 0.457$$

Podmínka $\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2} \geq 0.3 \wedge a_1 \geq 50 \wedge a_2 \geq 50$ je splněna

Rozdíl konfidencí pravidla HVek⟨65;85⟩ ≈ OK mezi Slovenskem a Rakouskem je 0.429

Slovensko x Rakousko: HVek⟨65;85⟩ $\Rightarrow_{0.429,54,63}$ OK