Tato prezentace je součástí wiki-prezentace Metoda GUHA, LISp-Miner a typové úlohy
Je dostupná z této adresy
Verse 19. 8. 2019

Typ úlohy: Využití disjunkcí

Data: Traffic Maryland

Problém: Jací řidiči vykazují velkou pravděpodobnost nějaké chyby spojené s vážnou nehodou?

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

© Jan Rauch

Asociační pravidla - využití disjunkcí

- Data „Traffic violations Maryland“
- Charakteristiky řidiče
- Chyba řidiče
- Špatné počasí
- Vážná nehoda
- Analytická otázka *„Jací řidiči vykazují velkou pravděpodobnost nějaké chyby spojené s vážnou nehodou?“*

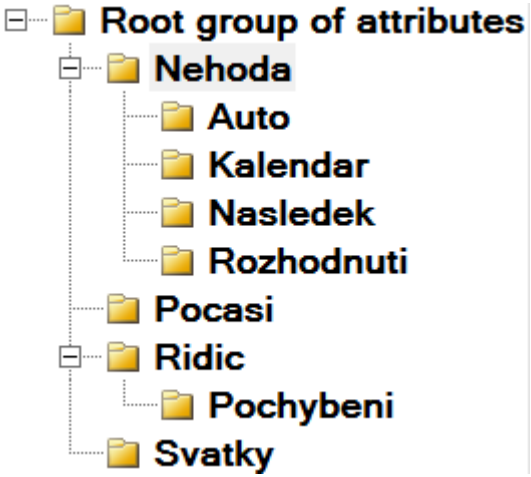
Data „Traffic violations Maryland“ (1)

Vznikla v rámci studentského projektu týmu 13 v předmětu „Pokročilé přístupy k dobývání znalostí z databází“ v zimním semestru 2018/119. Byly využity zdroje

- 2015–2016 School Calendar. *Montgomeryschoolsmd.org* [online]. [vid. 2018-10-9]. Dostupné z: https://www.montgomeryschoolsmd.org/uploadedFiles/0992.15_2015-16_SchoolCalendar%281%29.pdf
- 2016–2017 School Calendar. *Montgomeryschoolsmd.org* [online]. [vid. 2018-10-9]. Dostupné z: https://www.montgomeryschoolsmd.org/uploadedFiles/calendar/0904.16_2016-17_SchoolCalendar%20(FINAL).pdf
- 2017–2018 School Calendar. *Montgomeryschoolsmd.org* [online]. [vid. 2018-10-9]. Dostupné z: https://www.montgomeryschoolsmd.org/uploadedFiles/calendar/0700.17_2017-18_SchoolCalendar.pdf
- Business Park. *Montgomerycountymd.gov* [online]. [vid. 2018-10-9]. Dostupné z: https://data.montgomerycountymd.gov/Business/Business-Parks/rkpg-nghx
- Daily Summaries Location Details. *Noaa.gov* [online]. [vid. 2018-10-9]. Dostupné z: https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/FIPS:24031/detail
- Enterprise Zones. *Montgomerycountymd.gov* [online]. [vid. 2018-10-9]. Dostupné z: https://data.montgomerycountymd.gov/Business/Enterprise_Zones/a6xq-4hqk
- Holidays. *Excel-easy.com* [online]. [vid. 2018-10-9]. Dostupné z: https://www.excel-easy.com/examples/holidays.html
- Traffic Violations. *Data.gov* [online]. [vid. 2018-10-9]. Dostupné z: https://catalog.data.gov/dataset/traffic-violations-56dda

Data „Traffic violations Maryland“ (2)

- Celkem se jednalo o 1 386 207 záznamů o dopravních přestupcích za období 2012 až 2018, každý záznam obsahoval 35 dílčích údajů. Pro další zpracování byla využita data za roky 2016 a 2017. Krom toho byla využita data o počasí, státních svátcích a některá další.
- Data byla studenty zpracována a analyzována pomocí systému LISp-Miner. Vzniklé atributy byly rozděleny do skupin dle následujícího obrázku.



- Pro demonstraci možností využití disjunkcí byla použita data za rok 2017. Byly použit dále uvedené atributy.

Data „Traffic violations Maryland“ (3)

Pro demonstraci možností využití disjunkcí byla použita data za rok 2017 s atributy:

- Stát, Pohlaví a Rasa pro charakteristiku řidiče. Na jejich základě jsou definovány dílčí cedenty Řidič jako vhodné konjunkce.
- Alkohol, Pásy a Viník/spoluviník pro charakteristiku chování řidiče. Na jejich základě jsou definovány dílčí cedenty Chyba řidiče jako vhodné disjunkce.
- Úmrtí, Zranění, a Majetková_škoda pro charakteristiku nehody. Na jejich základě jsou definovány dílčí cedenty Vážná nehoda jako vhodné disjunkce.
- Mlha, Blesky, Déšť_ sníh, Krupobití, a Vítr pro charakteristiku počasí. Na jejich základě jsou definovány dílčí cedenty Špatné počasí jako vhodné disjunkce.

Je třeba poznamenat, že data nejsou prostá chyb.

Řidič (1)

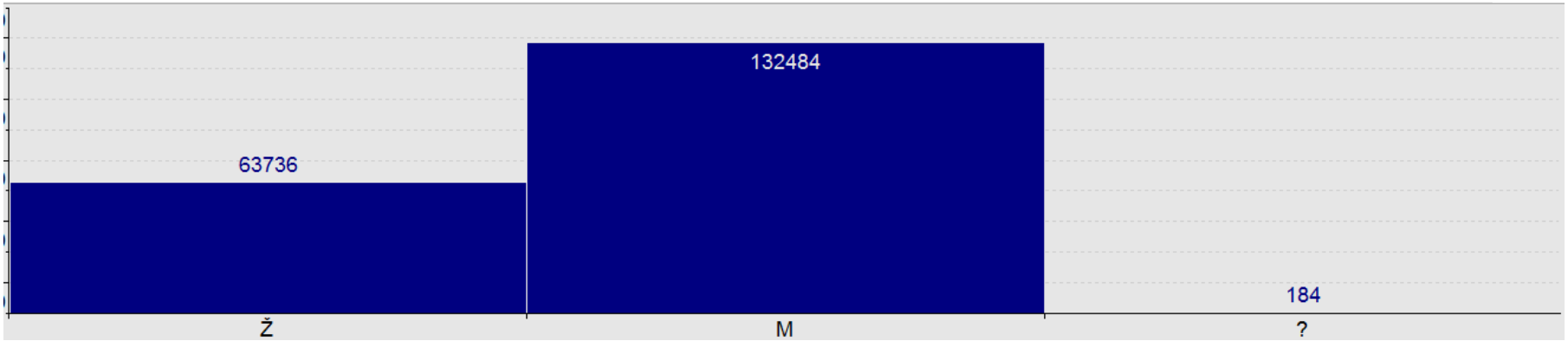
Atribut Stát

Uvedeno 12 nejfrekventovanějších států USA, které zahrnují 98,9 % řádků

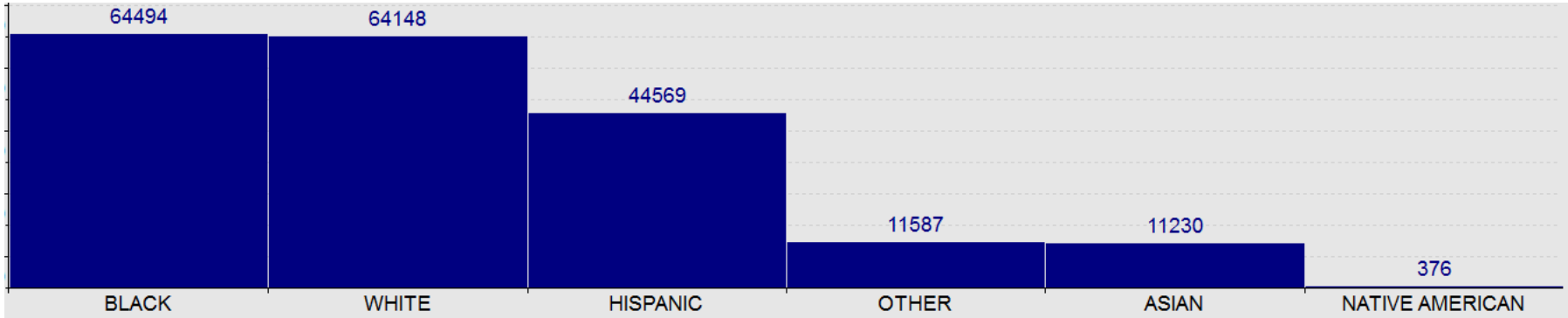
Driver_State	%	Frequency	Cummul. %	Cummul. Frequency
MD	89.9 %	176561	89.9 %	176561
DC	3.6 %	7072	93.5 %	183633
VA	3.2 %	6237	96.7 %	189870
PA	0.5 %	951	97.2 %	190821
FL	0.4 %	715	97.5 %	191536
NY	0.3 %	578	97.8 %	192114
NC	0.3 %	514	98.1 %	192628
WV	0.2 %	413	98.3 %	193041
CA	0.2 %	385	98.5 %	193426
NJ	0.2 %	318	98.6 %	193744
TX	0.2 %	306	98.8 %	194050
GA	0.1 %	284	98.9 %	194334

Řidič (2)

Atribut Pohlaví



Atribut Rasa



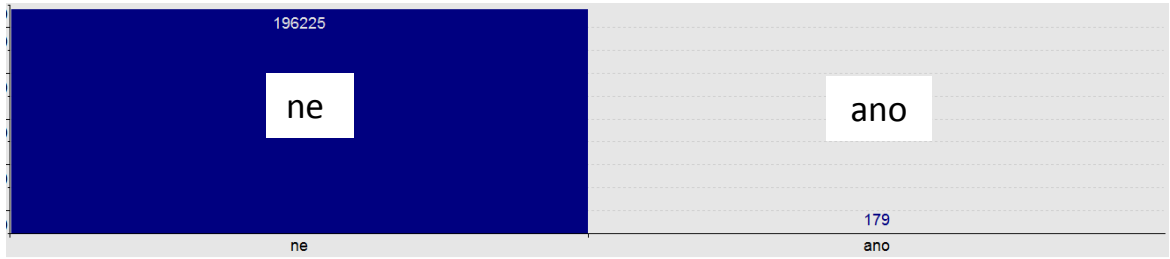
Zadání relevantních dílčích cedentů Řidič

Relevantní dílčí cedenty Řidič jsou charakteristiky řidiče dané konjunkcemi jednoho až tří literálů Stát(*), Pohlaví(*), Rasa(*) kde * značí jednu z kategorií atributu.

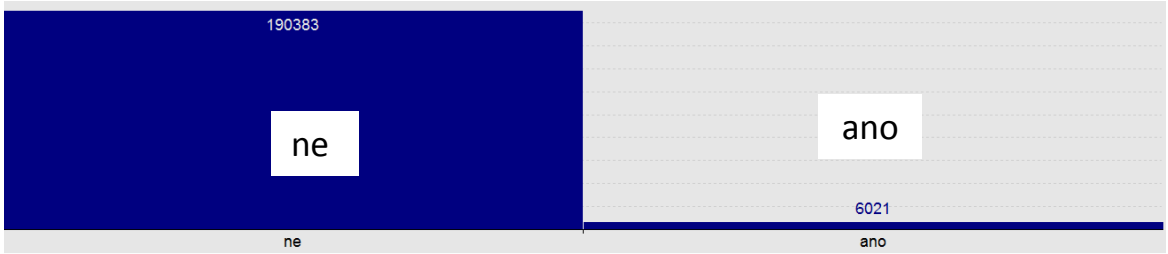
Řidič	Con, 1 - 3
» Stát (subset), 1 - 1	B, pos
» Pohlaví (subset), 1 - 1	B, pos
» Rasa (subset), 1 - 1	B, pos

Chování řidiče

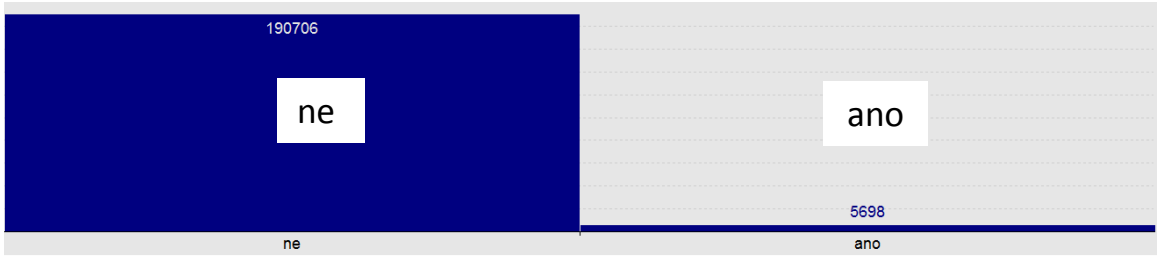
Alkohol



Pásy (ano = nezapnuté)



Viník/spoluviník



Chyba řidiče

Alkohol	Viník/spoluviník	Pásy	Počet řádků
ano	ano	ano	13
ano	ano	ne	36
ano	ne	ano	17
ano	ne	ne	113
ne	ano	ano	628
ne	ano	ne	5021
ne	ne	ne	5363
ne	ne	ne	185213

Za chybu řidiče považujeme situaci, kdy platí disjunkce jednoho až tří literálů Alkohol(ano), Viník/spoluviník(ano) a Pásy(ne).
To se týká celkem 13 + 36 + 17 + 113 + 628 + 5021 + 5363 = 11 191 řádků, tedy 5,7 % všech dat.

Zadání relevantních dílčích cedentů
Chyba řidiče

Chyba řidiče	Dis, 1 - 3
» Alkohol(ano)	R, pos
» Pásy(ano)	R, pos
» Viník/spoluviník(ano)	R, pos

4ft Succedent Partial cedent Settings

Basic parameters

Name: Chyba řidiče

Min. length: 1 Max. length: 3 Literals boolean operation type: Disjunction Edit

Comment: -

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
Alkohol	2	No	One category	1	pos	Remaining	-
Pásy	2	No	One category	1	pos	Remaining	-
Viník/spoluviník	2	No	One category	1	pos	Remaining	-

Špatné počasí

Mlha	Blesky	Déšť_ sníh	KrupoBití	Vítr	Počet
ne	ne	ne	ne	ne	176869
ne	ne	ne	ne	ano	789
ne	ne	ano	ne	ne	1254
ne	ano	ne	ne	ne	11802
ne	ano	ne	ano	ne	948
ano	ne	ne	ne	ne	4742

Za špatné počasí považujeme situaci, kdy platí disjunkce jednoho až pěti literálů Mlha(ano), Blesky(ano), Déšť_sníh(ano), Krupobití(ano) a Vítr(ano). To se týká celkem 789 + 1 254 + 11 802 + 948 + 4 742 = 19 535 řádků, tedy 9,9 % všech dat.

Zadání relevantních dílčích cedentů
Špatné počasí

Špatné počasí	Dis, 1 - 5
» Blesky(ano)	B, pos
» Dest_Snih(ano)	B, pos
» Krupobiti(ano)	B, pos
» Mlha(ano)	B, pos
» Vitr(ano)	B, pos

4ft Succedent Partial cedent Settings

Basic parameters

Name: Špatné počasí

Min. length: 1 Max. length: 5 Literals boolean operation type: Disjunction

Edit

Comment: -

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
Blesky	2	No	One category	1	pos	Basic	-
Dest_Snih	2	No	One category	1	pos	Basic	-
Krupobiti	2	No	One category	1	pos	Basic	-
Mlha	2	No	One category	1	pos	Basic	-
Vitr	2	No	One category	1	pos	Basic	-

Vážná nehoda

Zranění	Majetková_škoda	Úmrtí	Počet řádků
ano	ano	ano	0
ano	ano	ne	0
ano	ne	ano	0
ano	ne	ne	2561
ne	ano	ano	0
ne	ano	ne	4563
ne	ne	ano	33
ne	ne	ne	189247

Za vážnou nehodu považujeme situaci, kdy platí disjunkce jednoho až tří literálů Zranění(ano), Majetková_škoda(ano) a Úmrtí(ano).
To se týká celkem 2561 + 4 563 + 33 = 7 157 řádků, tedy 3,6 % všech dat.

Zadání relevantních dílčích cedentů
Vážná nehoda

Vážná nehoda	Dis, 1 - 3
» Úmrtí(ano)	R, pos
» Úraz(ano)	R, pos
» Majetková_škoda(ano)	R, pos

4ft Succedent Partial cedent Settings

Basic parameters

Name: Vážná nehoda

Min. length: 1 Max. length: 3 Literals boolean operation type: Disjunction

Edit

Comment: -

Options

Allow only a consecutive sequence of literals in cedent (only neighbouring literals): No

Linked coefficients (all literals must have the same coefficient as in the first one): No

Literals Settings

Underlying attribute	Categories	X-cat	Coefficient type	Length	+/-	B/R	Class of equiv.
Úmrtí	2	No	One category	1	pos	Remaining	-
Úraz	2	No	One category	1	pos	Remaining	-
Majetková_škoda	2	No	One category	1	pos	Remaining	-

Analytické otázka

Jací řidiči vykazují velkou pravděpodobnost nějaké chyby spojené s vážnou nehodou?

$$\text{Řidič}(?) \rightarrow \text{Chyba řidiče}(?) \wedge \text{Vážná nehoda}(?)$$

Řidič(?) → Chyba řidiče(?) ∧ Vážná nehoda(?)

Zadání procedury 4ft-Miner

ANTECEDENT

Řidič	Con, 1 - 3
» Stát (subset), 1 - 1	B, pos
» Pohlaví (subset), 1 - 1	B, pos
» Rasa (subset), 1 - 1	B, pos

Total length: 1 - 3

QUANTIFIERS

BASE	p= 100 Abs.
AAD	p= 0.250

Generation information

Status: Solved, 71 run(s)

Mode: Standard

SUCCEDENT

Chyba řidiče	Dis, 1 - 3
» Alkohol(ano)	R, pos
» Pásy(ano)	R, pos
» Viník/spoluviník(ano)	R, pos
Vážná nehoda	Dis, 1 - 3
» Úmrtí(ano)	R, pos
» Úraz(ano)	R, pos
» Majetková_škoda(ano)	R, pos

Total length: 1 - 6 {2 - 6}

Task parameters

Handling of missing values: Pesimistic fill up

CONDITION

Přehledný výstup

Uvedeno 12 pravidel s nejvyšší hodnotou AA-míry (= lift - 1)

Task run

Start: 17.1.2019 20:40:32 Total time: 0h 0m 2s

Number of verifications: 6664

Number of hypotheses: 113 Mode: Standard

Add group | Del group | Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 113 Shown hypotheses: 113 Highlighted: 0

Nr.	Id	AAD	Hypothesis
1	4	0.700	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & **Majetková_škoda**(*ano*)
2	2	0.690	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
3	71	0.661	**Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & **Majetková_škoda**(*ano*)
4	69	0.651	**Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
5	46	0.543	**Stát**(*MD*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & **Majetková_škoda**(*ano*)
6	44	0.534	**Stát**(*MD*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
7	101	0.503	**Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & **Majetková_škoda**(*ano*)
8	99	0.494	**Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
9	16	0.486	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)
10	14	0.477	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
11	3	0.464	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< [**Alkohol**(*ano*) | **Pásy**(*ano*)] & [**Úraz**(*ano*) | **Majetková_škoda**(*ano*)]
12	79	0.462	**Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)

Viz příklad detailního výstupu

Příklad detailního výstupu

Stát(MD) $\wedge$ Pohlaví(M) $\wedge$ Rasa(HISPANIC) $\sim_{0,69;193}$
(Alkohol(ano) $\vee$ Pásy(ano)) $\wedge$ (Úmrtí(ano) $\vee$ Majetková_škoda(ano))

- Četnost nehod pod vlivem alkoholu nebo s nezapnutými pásy při kterých došlo k úmrtí nebo k majetkové škodě je u řidičů - mužů hispánské rasy z Marylandu o 69 procent vyšší, než četnost takových nehod v celých datech.
- Takových nehod je u mužů hispánské rasy z Marylandu celkem 193.

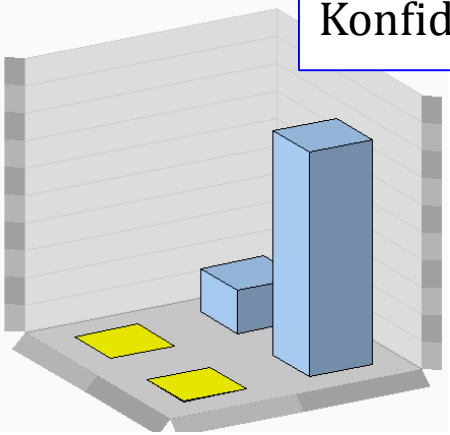
Antecedent: Stát(MD) & Pohlaví(M) & Rasa(HISPANIC)
Succedent: [Alkohol(ano) | Pásy(ano)] & [Úmrtí(ano) | Majetková_škoda(ano)]
Condition: (empty)

TEXT | DATA | FREQUENCIES | 9-FOLD TABLE | FREQDIFF | PIE | BAYE

	Succedent	¬Succedent
Ant.	193	31755
¬Ant.	509	163947

$$\frac{193}{193 + 31\,755} = (1 + 0.69)\frac{193 + 509}{193 + 31\,755 + 509 + 163\,947}$$

$$\text{Konfidence} = \frac{193}{193+31\,755} = 0.006$$



Modifikace zadání - konjunkce místo disjunkcí

ANTECEDENT	
Řidič	Con, 1 - 3
» Stát (subset), 1 - 1	B, pos
» Pohlaví (subset), 1 - 1	B, pos
» Rasa (subset), 1 - 1	B, pos

QUANTIFIERS	
BASE	p= 100 Abs.
AAD	p= 0.250

Generation information

Status: Solved, 69 run(s)

Mode: Standard

SUCCEDENT	
Chyba řidiče	Con, 1 - 3
» Alkohol(ano)	R, pos
» Pásy(ano)	R, pos
» Viník/spoluviník(ano)	R, pos
Vážné nehody	Con, 1 - 3
» Úmrtí(ano)	R, pos
» Úraz(ano)	R, pos
» Majetková_škoda(ano)	R, pos

Výstup: 13 pravidel místo 113:

Hypotheses in group: 13 Shown hypotheses: 13 Highlighted: 0

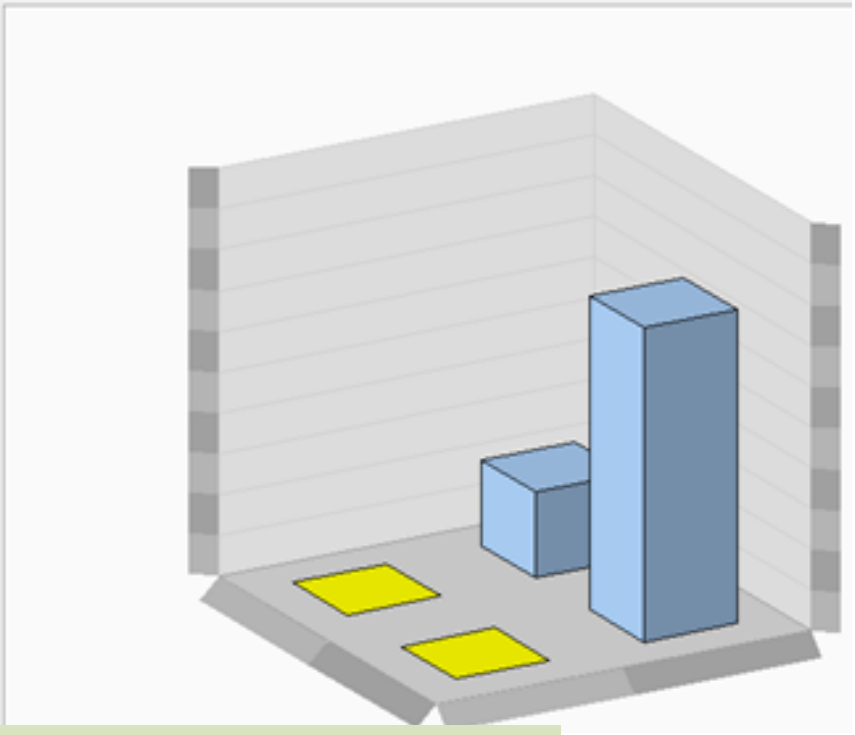
Nr.	Id	AAD	Hypothesis
1	1	0.486	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)
2	6	0.483	**Stát**(*MD*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
3	8	0.462	**Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)
4	12	0.449	**Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
5	5	0.396	**Stát**(*MD*) & **Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)
6	2	0.389	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
7	11	0.366	**Rasa**(*HISPANIC*) >÷< **Pásy**(*ano*) & **Majetková_škoda**(*ano*)
8	9	0.345	**Pohlaví**(*M*) & **Rasa**(*HISPANIC*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
9	4	0.337	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
10	7	0.313	**Stát**(*MD*) & **Rasa**(*HISPANIC*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
11	10	0.297	**Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
12	13	0.281	**Rasa**(*HISPANIC*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
13	3	0.275	**Stát**(*MD*) & **Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< **Viník/spoluviník**(*ano*) & **Úraz**(*ano*)

Stát(MD) $\wedge$ Rasa(HISPANIC) $\sim_{0,69;193}$
Pásy(ano) $\wedge$ Viník/spoluviník(ano) $\wedge$ Majetková_škoda(ano))

Antecedent: Stát(MD) & Rasa(HISPANIC)
Succedent: Pásy(ano) & Viník/spoluviník(ano) & Majetková_škoda(ano)
Condition: (empty)

TEXT | DATA | FREQUENCIES | 9-FOLD TABLE | FREQDIFF | PIE | BAYES

	Succedent	¬Succedent
Ant.	118	42007
¬Ant.	253	154026



Není relevantní pro zadání s disjunkcemi

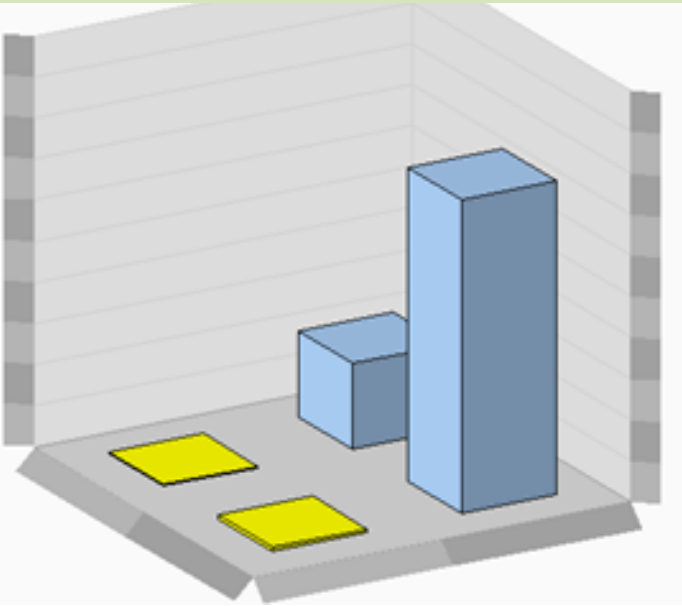
$$\text{Stát(MD)} \wedge \text{Rasa(HISPANIC)} \sim_{0,69;675} [\text{Pásy(ano)} \vee \text{Viník/spoluviník(ano)}] \wedge \text{Majetková_škoda(ano)})$$

Antecedent: Stát(MD) & Rasa(HISPANIC)
Succedent: [Pásy(ano) | Viník/spoluviník(ano)] & Majetková_škoda(ano)
Condition: (empty)

TEXT | DATA | FREQUENCIES | 9-FOLD TABLE | FREQDIFF | PIE | BAYES

	Succedent	¬Succedent
Ant.	675	41450
¬Ant.	1728	152551

$$\frac{675}{675+41\,450} = (1+0.31)\,\frac{675+1728}{675+41450+1\,728+152551}$$



Poznámky ke struktuře výsledků

- Při použití disjunkcí nalezeno 113 pravidel, nejvyšší hodnota AAD = 0.70
- Při použití konjunkcí nalezeno 13 pravidel, nejvyšší hodnota AAD = 0.486 (deváté pravidlo mezi pravidly pro disjunkce)
- Každé pravidlo nalezené pro konjunkce nemusí být ve výstupu pro disjunkce, viz šipky na předchozím slide (tato pravidla nejsou relevantní)
- Třeba pečlivá interpretace, výskyt „hroznů příbuzných pravidel“ ve výstupu pro disjunkce, ale disjunkce poskytují více informací
- Nejvyšší hodnota konfidence u pravidel pro disjunkce = 0.026
- Ne všechna pravidla se týkají jenom Marylandu nebo Hispánců - viz příklady

AAD	Hypothesis
0.301	**Stát**(*VA*) & **Pohlaví**(*M*) >÷< [**Alkohol**(*ano*) | **Viník/spoluviník**(*ano*)] & [**Úraz**(*ano*) | **Majetková_škoda**(*ano*)]
0.301	**Stát**(*VA*) & **Pohlaví**(*M*) >÷< **Viník/spoluviník**(*ano*) & [**Úmrtí**(*ano*) | **Úraz**(*ano*) | **Majetková_škoda**(*ano*)]
0.299	**Stát**(*VA*) >÷< **Viník/spoluviník**(*ano*) & [**Úraz**(*ano*) | **Majetková_škoda**(*ano*)]
0.297	**Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< **Viník/spoluviník**(*ano*) & **Majetková_škoda**(*ano*)
0.297	**Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< **Viník/spoluviník**(*ano*) & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]
0.295	**Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< [**Alkohol**(*ano*) | **Viník/spoluviník**(*ano*)] & **Majetková_škoda**(*ano*)
0.295	**Pohlaví**(*M*) & **Rasa**(*WHITE*) >÷< [**Alkohol**(*ano*) | **Viník/spoluviník**(*ano*)] & [**Úmrtí**(*ano*) | **Majetková_škoda**(*ano*)]