Katedra informatiky FEI VŠ-TUO

Analýza dat do předmětu Metody analýzy dat

Krasobruslaři

Květen 2011

Jan Górecki

1. ZADÁNÍ ANALÝZY

1.1. Popis dat

Sportovní klub krasobruslení Karviná za posledních 10 let sleduje data o mladých krasobruslařích a krasobruslařkách z Moravskoslezského kraje. Data se týkají měření výkonnostních parametrů sportovců:

- obsahují údaje o 287 sportovcích,
- sledují u nich 20 atributů několika typů.

Majitelem dat a expertem pro jejich porozumění je trenér krasobruslařů. Výsledky analýz by mu umožnily …

Popis atributů

Podle majitele dat je možno atributy rozdělit dle jejich charakteru do pěti skupin. Neovlivnitelné atributy jsou označeny znakem „n“ za jejich popisem.

Skupina Zákl

základní osobní údaje*:*

- Pohlaví {m, ž} - n
- Věk (roků) věk při měření adepta - n
- Výška (cm) - n
- Váha (kg)
- Tuk v těle (%) … standardně muž 8%, žena 16%

Skupina Dálka

charakteristika skoku do dálky:

- SDM (cm) skok daleký z místa
- TP (cm) trojskok na pravé noze
- TL (cm) trojskok na levé noze

Pokud rozdíl mezi TP a TL je větší než 50 cm, ukazuje to na výraznou nevyrovnanost obou nohou a je potřeba s tím okamžitě něco dělat, jinak je krasobruslař „nepoužitelný“ – neskáče stejně dobře z obou nohou a brzo si namáhavější nohu „odrovná“.

Skupina Běh

charakteristika rozběhu na „suché“ podlaze na 15m, měřeno speciálním přístrojem:

- t15 (…) čas na uběhnutí 15 m
- a1s (…) průměrná rychlost běhu po 1 s
- a2s (…) průměrná rychlost běhu po 2 s
- a3s (…) průměrná rychlost běhu po 3 s

Skupina Běh_led

charakteristika rozběhu na bruslích na 15m, měřeno speciálním přístrojem:

- t15 (…) čas na uběhnutí 15 m
- a1s (…) průměrná rychlost běhu po 1 s
- a2s (…) průměrná rychlost běhu po 2 s
- a3s (…) průměrná rychlost běhu po 3 s

Skupina Výška

charakteristika skoku snožmo do výšky, měřeno speciálním přístrojem:

- t opory (ms) čas odrazu … tento atribut ukazuje na styl odrazu každého jedince, zda se odráží spíše z lýtka či spíše ze stehenního svalu; ideální je 280ms - n
- P (W/kg) výkon v aktivní části odrazu
- h (cm) dosažená výška po odrazu
- h/tc = výška výskoku/doba vzletu … důležitý ukazatel na kvalitu skoku (nejlepší mají >200)

Ukázka dat

1	Pohlaví	Věk	Výška	Váha	%tuku	SDM	TP	TL	t_opory	P (W/kg)	h	h/tc	t15_běh	a1s_běh	a2s_běh	a3s_běh	t15_brus	a1s_brus	a2s_brus	a3s_brus
2	m	7	130	26		122	383	380	204	24,7	17,3	80,3	4,013	1,54	3,61	4,6				
3	m	8	129	27	5,1	130	404	390	166	30,1	20,3	109,6	3,1	2,597	5,343	6,196				
4	m	8	128	24	4,2	189	430	449	166	35,3	22,8	128,1	2,92	3,307	5,537	6,42				
5	m	8	129	28	5	165	420	445	190	33,2	24,3	116,6	3,465	2,54	4,52	5,13	3,973	1,48	3,56	4,77
6	m	9	137	27		158	458	382	194	25,8	17,6	85,8	3,572	2,37	4,21	5,09				
7	m	9	134	30	5,1	175	443	506	154	42,5	24,8	161,4	3,448	2,59	4,37	5,27	3,958	1,64	3,61	4,64
8	m	9	132	27	2,2	165	474	480	131	51,3	27,7	203,1	3,639	2,04	4,24	5,06	3,627	2,12	4,11	5,1
9	m	9	138	32	5,1	176	455	442	186	35,1	23,5	125,6	3,614	1,75	4,25	5,34	3,771	1,79	3,95	4,89
10	m	9	134	30	4,2	180	484	503	141	47,5	26,8	184,9	3,519	2,23	4,45	5,27	3,673	1,89	4,05	5,07
11	m	9	130	31	5,1	161	464	499	190	34,8	23,5	123,6	2,900	2,398	4,573	5,637	2,94	2,578	4,439	5,29
12	m	9	129	26	2,2	180	489	466	169	35,5	24,9	128,6	3,348	2,51	4,4	5,32	3,791	1,8	3,93	4,97
13	m	9	148	36	4,2	156	429	437	178	34,6	23,2	123,5	3,621	2,1	4,18	5,23	3,739	2,26	3,8	4,89
14	m	9	145	33	3,2	183	510	497	198	36,7	26,5	130,4	2,870	2,888	5,73	6,96	3,080	3,162	5,033	5,977
15	m	9	141	34	6,1	176	529	495	215	31,3	25,3	105,5	2,770	3,966	5,934	6,972	3,030	3,398	5,09	6,04
16	m	9	142	33	5,1	174	527	472	182	30,2	20,1	140,2	2,770	3,519	5,99	6,945	3,230	2,797	4,755	5,131
17	m	9	138	34	3,2	172	512	513	161	38,5	23,4	142,8	2,680	3,875	6,082	7,041	2,990	3,347	5,315	6,222
18	m	9	135	30	5,1	174	497	473	140	43,9	24,6	169	2,94	3,016	5,388	6,258	3,418	2,33	4,597	5,308
19	m	7	129	33	8	160	407	431	177	25,7	14,9	89,5	3,24	2,658	4,916	5,68	3,25	2,782	4,868	5,915
20	m	9	133	33	8	156	434	404	187	30	20,6	103,7	3,11	2,995	5,118	6,656	3,67	2,313	4,145	5,095
21	m	6	123	25	8,9	119	266	311	171	16,5	8,8	51,3	3,6	2,363	4,327	4,937	3,7	2,093	4,015	5,038
22	m	10	128	29	5,1	170	458	440	164	37,6	23,6	138,7	3,778	2,16	4,05	4,71	3,801	1,66	3,08	4,96
23	m	10	142	33	5,1	173	545	515	141	47,3	26,8	183,9	3,66	1,87	4,18	5,12	3,478	2,33	4,31	5,35
24	m	10	140	32	5,1	158	490	495	187	35,6	25,3	174,4	3,636	2,07	4,09	5,1	3,484	2,57	4,39	5,18
25	m	10	135	29	5,1	163	433	462	150	40,9	24,6	154,5	3,908	1,77	3,88	4,59	3,753	2,21	3,98	4,65

Pro analýzy pojmenujeme atributy po skupinách následovně:

pohlavi, vek, vyska, vaha, tuk,
SDM, TP, TL,
t15_beh, a1s_beh, a2s_beh, a3s_beh,
t15br_brus, a1sbr_brus, a2sbr_brus, a3sbr_brus,
t_opory, P, h, h_tc

1.2. Zadání pro analýzy

Majitele dat zajímá vše, co se z dat dá určit.

1.3. Základní charakteristiky atributů

Nejprve byly spočítány základní statistiky celé množiny dat pro každý atribut.

Název	měr.j.	Dat.typ	min	max	avg	std.odch.	NULL
pohlavi		nominal	m (75)	z (212)			0
vek	rok	integer	6	19	10.624	1.969	0
vyska	cm	integer	109	197	141.453	10.978	0
vaha	kg	real	20	79	34.538	7.698	0
tuk	%	real	2.200	20.600	10.433	3.991	10
SDM	cm	real	119.000	263.000	181.362	22.844	0
TP	cm	integer	266	830	519.491	81.793	0
TL	cm	integer	274	815	515.223	80.848	0
t_opory	ms	real	125.000	250.000	173.197	19.400	0
P	(W/kg)	real	16.500	70.100	40.921	7.909	0
h	cm	real	8.800	42.000	27.135	4.985	0
h_tc		real	51.300	288.100	153.516	35.469	0
t15_beh		real	2.470	4.013	3.204	0.350	0
a1s_beh		real	1.430	4.505	2.729	0.570	0
a2s_beh		real	3.168	6.853	4.979	0.716	0
a3s_beh		real	4.470	8.095	5.918	0.775	0
t15_brus		real	2.470	4.301	3.416	0.401	26
a1s_brus		real	0.960	3.966	2.432	0.607	26
a2s_brus		real	2.610	6.756	4.497	0.735	26
a3s_brus		real	3.770	7.600	5.466	0.784	26

Z hodnot min, max, avg, std.odch. není zřejmá žádná výrazná chyba v datech.

Z hodnot NULL je vidět několik chybějících údajů: 10 u atributu %Tuku, po 26 u skupiny Běh_led. U následných analýz bude nutné tyto chybějící údaje řešit podle konkrétní použité metody a příslušného programu, ale zatím budou ponechány všechny objekty i atributy.

1.4. Předzpracování dat

Všechny atributy mimo pohlavi jsou numerické, proto jedinou úpravou dat je binarizace atributu pohlavi na atribut pohl:

$$\text{pohlavi} \Rightarrow \text{pohl}: \quad m \Rightarrow 0, \quad z \Rightarrow 1$$

Nyní jsou všechny atributy numerické a je možno považovat všechny za reálné.

Pro asociace a rozhodovací stromy byly reálné atributy kategorizovány a binarizovány takto:

$\text{vek} \Rightarrow \text{vek\_k} \Rightarrow \text{vek\_b}$,
$\text{vyska} \Rightarrow \text{vyska\_k} \Rightarrow \text{vyska\_b}$,
…
$\text{h\_tc} \Rightarrow \text{h\_tc\_k} \Rightarrow \text{h\_tc\_b}$

Intervaly kategorizace:

Kat	vek_k	vyska_k	vah_k	tuk_k	SDM_k	TP_k	TL_k
1							
2							
3							
4							
5							

Kat	t15_beh_k	a1s_beh_k	a1s_beh_k	a1s_beh_k
1				
2				
3				
4				
5				

Kat	t15_brus_k	a1s_brus_k	a1s_brus_k	a1s_brus_k
1				
2				
3				
4				
5				

Kat	t_opory_k	P_k	h_k	h_tc_k
1				
2				
3				
4				
5				

Provedl jsem diskretizaci všech reálných atributů do 5 ekvifrekvenčních intervalů. Dále byl nutný převod všech kategoriálních atributů na binární.

Pro shlukování byly dále původní atributy mimo pohl standardizovány:

vek $\Rightarrow$ vek_s, …

2. ANALÝZY DAT

Poznámka: Červený text obsahuje komentáře a poznámky majitele dat k výsledkům analýzy.

Vzhledem k charakteru dat bylo rozhodnuto o provedení následujících analýz:

1. *Korelační matice* – většina atributů je reálných, má smysl zjišťovat lineární závislosti.
2. *Hlavní komponenty* – u reálných atributů je možno očekávat skryté faktory.
3. *Asociační pravidla* – po kategorizaci reálných atributů možno použít hledání asociací.
4. *Shlukování* – pro hledání skupin podobných si objektů
5. *Rozhodovací stromy* – pro předvídání zajímavých hodnot některých atributů (h, h/tc).

Pro analýzy byl použit Data-miningový systém Rapid Miner.

2.1. Korelační matice

První analýzou byla vypočtena korelační matici nad všemi reálnými daty (viz. Obrázek 1: Korelační matice). V korelační matici jsou spočteny míry korelace mezi jednotlivými atributy. Korelační koeficienty jsou počítány mezi všemi atributy (tedy i mezi atributy s chybějícími údaji – program v případě chybějícího údaje příspěvek ke koeficientu korelace ignoruje).

Ve výsledné matici lze vidět vysokou míru korelace - lineární závislosti mezi atributy uvnitř jednotlivých podskupin atributů (žluté obdélníky podél diagonály). Intuitivní podobnost (závislost) mezi atributy uvnitř podskupin se tedy potvrdila.

U podskupin atributů Běh a Běh_Led lze dokonce vidět lineární závislost nejen uvnitř podskupin samotných, ale i mezi atributy z jedné a druhé skupiny (atributy obou skupin navzájem korelují – žlutý obdélník vpravo dole).

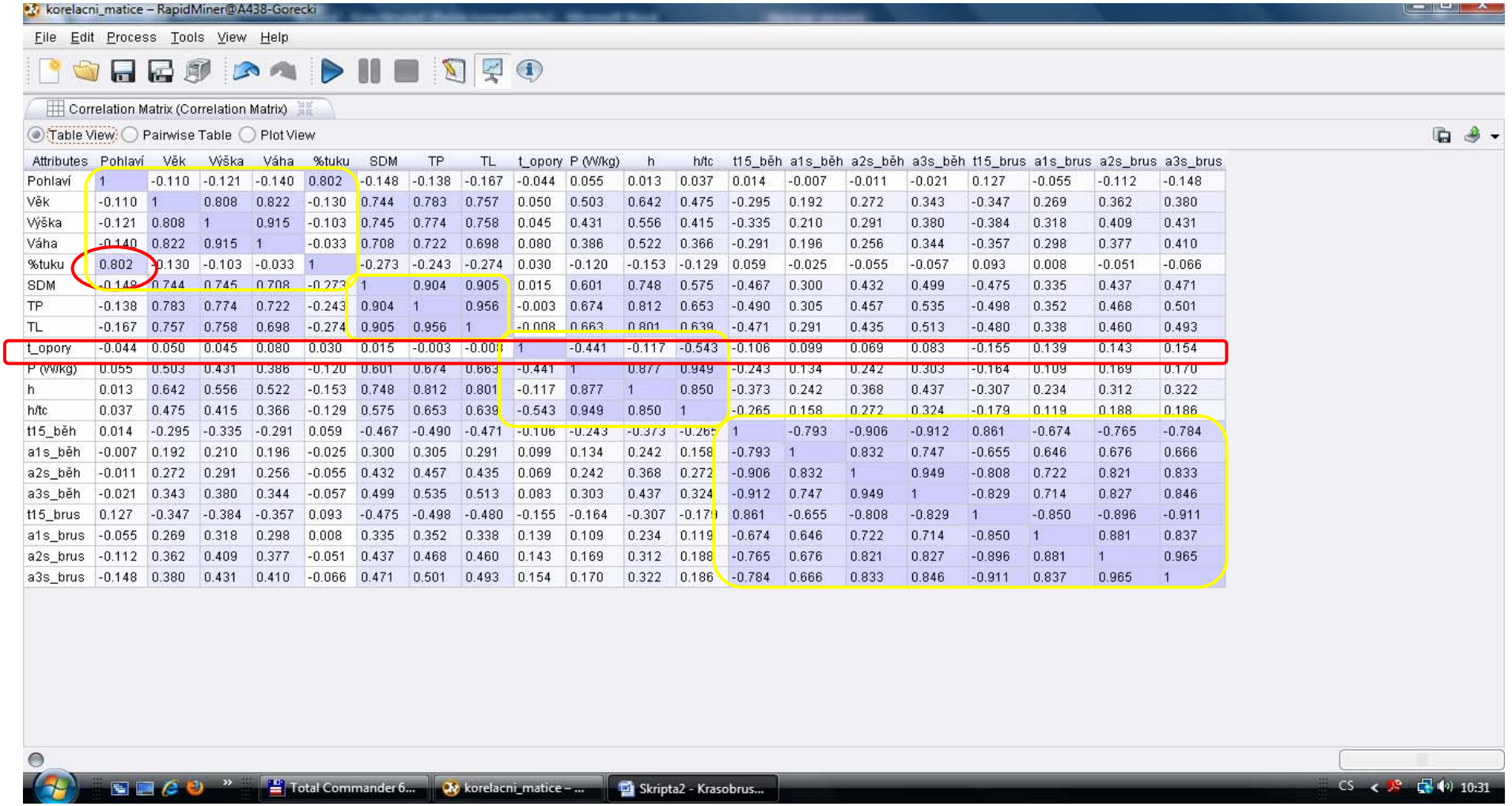
Výjimečné jsou (v první podskupině atributů Zákl – levý horní žlutý obdélník) atributy Pohlaví a %Tuku, které korelují jen mezi sebou navzájem (červená elipsa). Detailnější vztah mezi těmito atributy se ukáže pomocí asociačních pravidel.

Další výjimečný atribut je t-opory (červený obdélník), který převážně s ničím nekoreluje. Je to tím, že atribut t-opory (čas opory) popisuje vlastnost, která je pro každého krasobruslaře charakteristická a nedá se pozitivně příliš ovlivnit tréninkem; lépe řečeno špatným tréninkem se dá tato vlastnost zhoršit, kdežto dobrým tréninkem lze vlastnost spíše jen udržet a popř. velmi mírně korigovat.

Výpočet korelační matice tedy odhalil množství lineárních závislostí mezi jednotlivými atributy, které se projevují i ve výsledcích dalších analýz.

Výsledek získaný výpočtem korelační matice ukazuje, že záměry konstrukce baterie (= sady měření vlastností) testů byly splněny. Šlo o to vytvořit sadu testů, kde celá sada bude zachycovat dynamicko-silové vlastnosti měřených jedinců a v této sadě bude vždy nějaká skupina atributů popisovat jednotlivé typy těchto dynamicko-silových vlastností (lokomoce neboli pohyb vpřed, vertikální zdvih těžiště, kombinace obou pohybů).

Obrázek 1: Korelační matice



korelacni_matice – RapidMiner@A438-Gorecki

File Edit Process Tools View Help

Correlation Matrix (Correlation Matrix)

Table View | Pairwise Table | Plot View

Attributes	Pohlaví	Věk	Výška	Váha	%tuku	SDM	TP	TL	t_opory	P (W/kg)	h	h/tc	t15_běh	a1s_běh	a2s_běh	a3s_běh	t15_brus	a1s_brus	a2s_brus	a3s_brus
Pohlaví	1	-0.110	-0.121	-0.140	0.802	-0.148	-0.138	-0.167	-0.044	0.055	0.013	0.037	0.014	-0.007	-0.011	-0.021	0.127	-0.055	-0.112	-0.148
Věk	-0.110	1	0.808	0.822	-0.130	0.744	0.783	0.757	0.050	0.503	0.642	0.475	-0.295	0.192	0.272	0.343	-0.347	0.269	0.362	0.380
Výška	-0.121	0.808	1	0.915	-0.103	0.745	0.774	0.758	0.045	0.431	0.556	0.415	-0.335	0.210	0.291	0.380	-0.384	0.318	0.409	0.431
Váha	-0.140	0.822	0.915	1	-0.033	0.708	0.722	0.698	0.080	0.386	0.522	0.366	-0.291	0.196	0.256	0.344	-0.357	0.298	0.377	0.410
%tuku	0.802	0.130	-0.103	-0.033	1	-0.273	-0.243	-0.274	0.030	-0.120	-0.153	-0.129	0.059	-0.025	-0.055	-0.057	0.093	0.008	-0.051	-0.066
SDM	-0.148	0.744	0.745	0.708	-0.273	1	0.904	0.905	0.015	0.601	0.748	0.575	-0.467	0.300	0.432	0.499	-0.475	0.335	0.437	0.471
TP	-0.138	0.783	0.774	0.722	-0.243	0.904	1	0.956	-0.003	0.674	0.812	0.653	-0.490	0.305	0.457	0.535	-0.498	0.352	0.468	0.501
TL	-0.167	0.757	0.758	0.698	-0.274	0.905	0.956	1	-0.008	0.663	0.801	0.639	-0.471	0.291	0.435	0.513	-0.480	0.338	0.460	0.493
t_opory	-0.044	0.050	0.045	0.080	0.030	0.015	-0.003	-0.008	1	-0.441	-0.117	-0.543	-0.106	0.099	0.069	0.083	-0.155	0.139	0.143	0.154
P (W/kg)	0.055	0.503	0.431	0.386	-0.120	0.601	0.674	0.663	-0.441	1	0.877	0.949	-0.243	0.134	0.242	0.303	-0.164	0.109	0.169	0.170
h	0.013	0.642	0.556	0.522	-0.153	0.748	0.812	0.801	-0.117	0.877	1	0.850	-0.373	0.242	0.368	0.437	-0.307	0.234	0.312	0.322
h/tc	0.037	0.475	0.415	0.366	-0.129	0.575	0.653	0.639	-0.543	0.949	0.850	1	-0.265	0.158	0.272	0.324	-0.179	0.119	0.188	0.186
t15_běh	0.014	-0.295	-0.335	-0.291	0.059	-0.467	-0.490	-0.471	-0.106	-0.243	-0.373	-0.265	1	-0.793	-0.906	-0.912	0.861	-0.674	-0.765	-0.784
a1s_běh	-0.007	0.192	0.210	0.196	-0.025	0.300	0.305	0.291	0.099	0.134	0.242	0.158	-0.793	1	0.832	0.747	-0.655	0.646	0.676	0.666
a2s_běh	-0.011	0.272	0.291	0.256	-0.055	0.432	0.457	0.435	0.069	0.242	0.368	0.272	-0.906	0.832	1	0.949	-0.808	0.722	0.821	0.833
a3s_běh	-0.021	0.343	0.380	0.344	-0.057	0.499	0.535	0.513	0.083	0.303	0.437	0.324	-0.912	0.747	0.949	1	-0.829	0.714	0.827	0.846
t15_brus	0.127	-0.347	-0.384	-0.357	0.093	-0.475	-0.498	-0.480	-0.155	-0.164	-0.307	-0.179	0.861	-0.655	-0.808	-0.829	1	-0.850	-0.896	-0.911
a1s_brus	-0.055	0.269	0.318	0.298	0.008	0.335	0.352	0.338	0.139	0.109	0.234	0.119	-0.674	0.646	0.722	0.714	-0.850	1	0.881	0.837
a2s_brus	-0.112	0.362	0.409	0.377	-0.051	0.437	0.468	0.460	0.143	0.169	0.312	0.188	-0.765	0.676	0.821	0.827	-0.896	0.881	1	0.965
a3s_brus	-0.148	0.380	0.431	0.410	-0.066	0.471	0.501	0.493	0.154	0.170	0.322	0.186	-0.784	0.666	0.833	0.846	-0.911	0.837	0.965	1

Total Commander 6... | korelacni_matice – ... | Skripta2 - Krasobrus... | CS | 10:31

2.2. Hlavní komponenty

Program pro hlavní komponenty neumí pracovat s chybějícími údaji, proto bylo odstraněno 26 zmíněných záznamů.

Výpočet hlavních komponent byl proveden pro všechny reálné atributy, tedy pro

pohl, vek, vyska, vaha, tuk,
SDM, TP, TL,
t15_beh, a1s_beh, a2s_beh, a3s_beh,
t15br_brus, a1sbr_brus, a2sbr_brus, a3sbr_brus,
t_opory, P, h, h_tc

Obrázek 2: Vlastní čísla

Eigenvalues Eigenvectors Cumulative Variance Plot

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	3.101	0.481	0.481
PC 2	1.962	0.192	0.673
PC 3	1.410	0.099	0.773
PC 4	1.277	0.082	0.854
PC 5	0.846	0.036	0.890
PC 6	0.751	0.028	0.918
PC 7	0.580	0.017	0.935
PC 8	0.492	0.012	0.947
PC 9	0.458	0.010	0.958
PC 10	0.411	0.008	0.966
PC 11	0.395	0.008	0.974
PC 12	0.343	0.006	0.980
PC 13	0.329	0.005	0.985
PC 14	0.250	0.003	0.988
PC 15	0.236	0.003	0.991
PC 16	0.214	0.002	0.993
PC 17	0.202	0.002	0.995
PC 18	0.192	0.002	0.997
PC 19	0.184	0.002	0.999
PC 20	0.152	0.001	1.000

Na obrázku je první část výsledku, vlastní čísla korelační matice. První čtyři komponenty mají 85%ní podíl na velikosti stopy matice vlastních čísel (červená elipsa). Budeme se jimi tedy dále zabývat. Další komponenty mají již nevýrazné podíly na velikosti stopy matice, dále se jimi tedy zabývat nebudeme.

Obrázek 3: Vlastní vektory

Eigenvalues ◉ Eigenvectors ○ Cumulative Variance Plot

Attribute	PC 1	PC 2	PC 3	PC 4
Pohlaví	-0.050	0.048	-0.520	0.452
Věk	0.228	-0.214	0.161	0.235
Výška	0.232	-0.178	0.192	0.267
Váha	0.222	-0.172	0.220	0.310
%tuku	-0.051	0.089	-0.428	0.549
SDM	0.262	-0.199	0.115	0.062
TP	0.277	-0.210	0.063	0.046
TL	0.271	-0.211	0.089	0.025
t_opory	0.012	0.194	0.364	0.404
P (W/kg)	0.174	-0.320	-0.311	-0.167
h	0.228	-0.276	-0.169	-0.012
h/tc	0.174	-0.306	-0.336	-0.225
t15_běh	-0.257	-0.223	0.082	0.047
a1s_běh	0.213	0.244	-0.093	-0.063
a2s_běh	0.259	0.237	-0.116	-0.084
a3s_běh	0.272	0.202	-0.096	-0.047
t15_brus	-0.260	-0.240	-0.005	0.063
a1s_brus	0.225	0.262	-0.037	-0.024
a2s_brus	0.260	0.243	-0.010	-0.041
a3s_brus	0.265	0.233	0.015	-0.039

První komponenta PC1 (48%) se dá charakterizovat jako faktor popisující *celkový výkon a fyzickou stavbu jedince*. Tento faktor ovlivňují s poměrně vyrovnaným podílem všechny atributy **až na atributy** odpovídající červeným elipsám (tedy pohlaví, %tuku a t_opory).

Druhá komponenta PC2 (19%) má téměř identický charakter jako PC1 s jediným rozdílem, že tento faktor je ovlivněn i atributem t_opory.

Dá se tedy říci, že PC1 a PC2 tvoří hlavní dvojici skrytých faktorů, které popisují celkový výkon a fyzickou stavbu jedince.

Další dvě komponenty PC3 (10%) a PC4 (8%) si jsou taktéž mezi sebou podobné. Tyto komponenty jsou ovlivněny atributy, které první dvě komponenty neovlivňovaly, a to atributy Pohlaví, %tuku a t_opory (viz. výrazné koeficienty v oranžových elipsách) plus některými dalšími atributy (opět označeno oranžovou elipsou). Lze vidět, že atributy, které ovlivňují tyto dvě komponenty, jsou všechny z podskupin atributů Zákl a Výška. Tyto dvě komponenty tedy popisují skrytý faktor, který se dá interpretovat jako *kvalitativní popis skoku do výšky a fyzické stavby jedince* (s tím že u skoku do výšky jde spíše o neovlivnitelné atributy a u fyzické stavby jde spíše o vysoce korelované atributy pohlaví a %tuku).

2.4. Asociační pravidla

Pro získání asociačních pravidel byly použity kategoriální atributy, tedy:

pohl, vek_k, vyska_k, vaha_k, tuk_k,
SDM_k, TP_k, TL_k,
t15_beh_k, a1s_beh_k, a2s_beh_k, a3s_beh_k,
t15br_brus_k, a1sbr_brus_k, a2sbr_brus_k, a3sbr_brus_k,
t_opory_k, P_k, h_k, h_tc_k

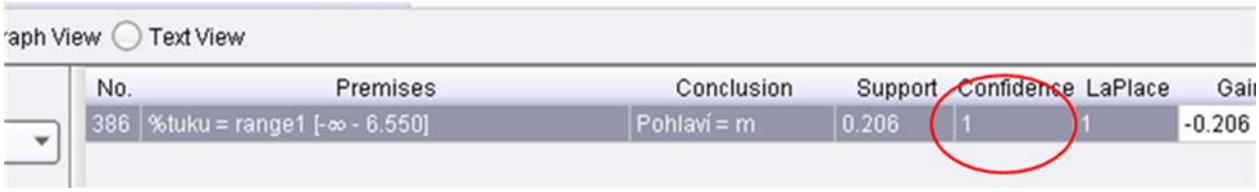
Pro analýzu byly zvoleny tyto parametry:

- V programu nebylo nutné nastavit, které atributy zařadím mezi antecedenty a které mezi sukcedenty. Stačilo pouze zadat, mezi kterými atributy chci hledat asociace, a metoda sama vyzkoušela všechny možné kombinace atributů jak v antecedentu, tak v sukcedentu. Zadané atributy jsou uvedeny u jednotlivých výsledků.
- Program umožnil pracovat i s atributy s chybějícími údaji.
- Kvantifikátor – fundovaná implikace FI
- Minimální podpora Pmin = 8% … nastavení na základě několika experimentů (tak, abych dostal přiměřené množství výsledků)
- Minimální spolehlivost Smin = 90% … mohl jsem si dovolit nastavit takto vysoko, jelikož data jsou získána přesným měřením objektivní skutečnosti (neobsahují subjektivní veličiny, jako např. u dotazníků)

Výsledky asociací

Pro 1. analýzu jsem zvolil do množiny {ancetedent, sukcedent} atributy z podskupin Zákl, Výška, Dálka.

Obrázek 4: Asociační pravidla 1

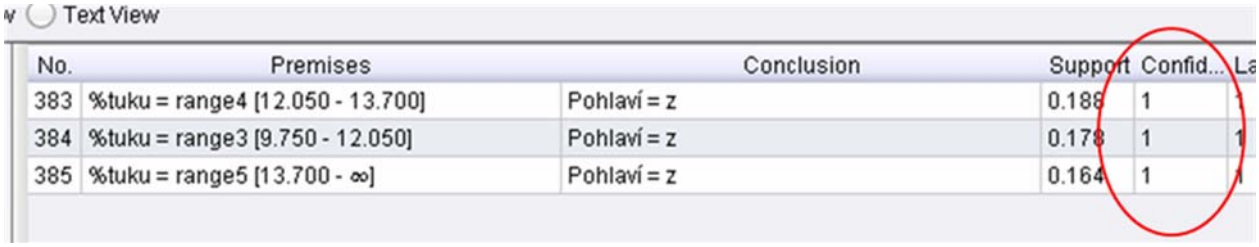


Prvním výsledkem bylo pravidlo (viz. Obrázek 4: Asociační pravidla 1)

Málo tučný => muž

Toto pravidlo bylo očekávané vzhledem k vysoké korelaci atributů %tuku a Pohlaví, nicméně bylo překvapivé, že hodnota spolehlivosti tohoto pravidla je 100% (viz. červená elipsa výše).

Obrázek 5: Asociační pravidla 2



Obdobným výsledkem bylo pravidlo (viz. Obrázek 5: Asociační pravidla 2)

Tučný => žena

Opět byla překvapivá 100%ní spolehlivost (viz. červená elipsa výše). Je vidět, že trenéři velmi dbají na množství tuku v těle u každého krasobruslaře (viz. informace v popisu tohoto atributu v kapitole Data).

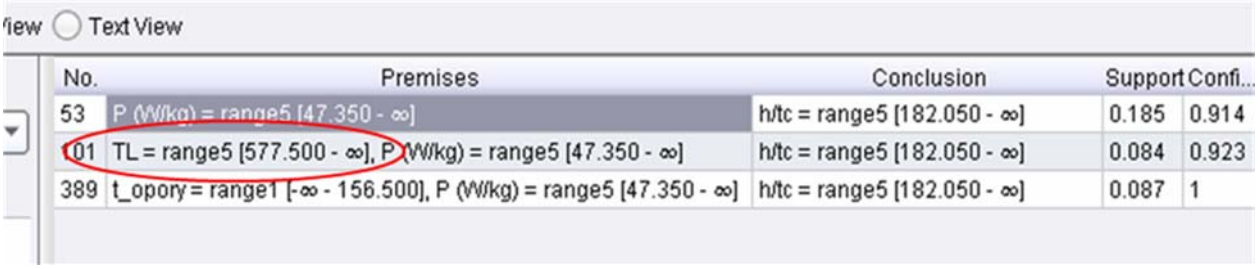
Množství tuku v těle je první číslo, které chtějí trenéři vědět o novém adeptovi. Tato hodnota je extrémně důležitá u žen, jelikož (pokud je hodnota tuku příliš velká):

1) chce-li adeptka jezdit sólo, tak není schopná dosti skákat
2) chce-li jezdit v páru, partner má potom problémy ji unést.

Každý reálný atribut byl rozdělen do 5ti ekvifrekvenčních intervalů a na obrázcích výše vidíme, že pro atribut %tuku asociační pravidla hovoří o čtyřech z těchto pěti intervalů (range1, 3, 4, 5). Chybí pravidlo pro interval range2 [6.55 – 9.75]- důvod je takový, že hodnoty %tuku v tomto intervalu jsou na rozmezí mezi malými a velkými hodnotami, a těmto hodnotám odpovídají jak muži (ti více tuční), tak ženy (ty méně tučné), tudíž nelze s dostatečnou spolehlivostí na tomto intervalu tvrdit něco o závislosti %tuku a jednoho druhu pohlaví (jinak řečeno, v počtu objektů odpovídajících tomuto intervalu výrazně nepřevažují ani muži, ani ženy).

Dále jsem se zaměřil na atribut h/tc (odpovídající „kvalitě“ skoku do výšky) – konkrétně na interval s nejvyššími hodnotami tohoto atributu.

Obrázek 6: Asociační pravidla 3



No.	Premises	Conclusion	Support	Confi...
53	P (W/kg) = range5 [47.350 - ∞]	h/tc = range5 [182.050 - ∞]	0.185	0.914
101	TL = range5 [577.500 - ∞], P (W/kg) = range5 [47.350 - ∞]	h/tc = range5 [182.050 - ∞]	0.084	0.923
389	t_opory = range1 [-∞ - 156.500], P (W/kg) = range5 [47.350 - ∞]	h/tc = range5 [182.050 - ∞]	0.087	1

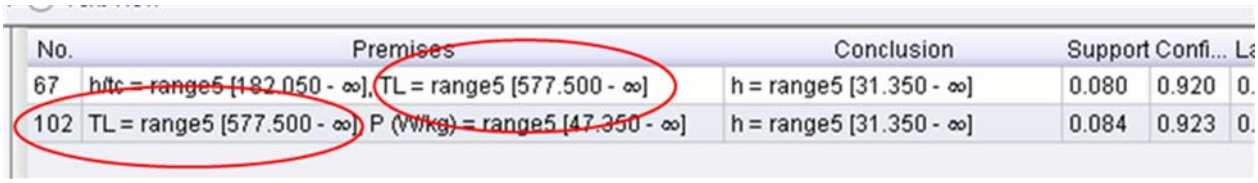
Z obrázku výše (v části Premises jsou **konjunkce** podmínek) lze vidět, že vysokých hodnot h/tc dosahují krasobruslaři s vysokými hodnotami atributu P (výkon v aktivní části odrazu) – tento předpoklad se objevuje ve všech podmínkách na obrázku (a není to nic překvapivého).

Zajímavé je, že se v podmínce objevuje atribut TL (viz. červená elipsa) a to opět pro jeho nejvyšší hodnoty. Zdá se, jakoby kvalita skoku byla dána délkou trojskoku na levé noze – dalo by se usuzovat na, že většina lidí jsou praváci a ti se odráží přirozeně z levé nohy a skok z této nohy mají lepší, což určuje i kvalitu skoku h/tc.

Trenér potřebuje sadu jednoduše měřitelných ukazatelů, které mu budou dobře popisovat výkon daného adepta. Délku TL lze tedy díky výše uvedeným výsledkům zařadit mezi takovéto ukazatele.

Dále jsem se zaměřil na atribut h (odpovídající výšce skoku do výšky) – konkrétně na interval s nejvyššími hodnotami tohoto atributu (range5).

Obrázek 7: Asociační pravidla 4



No.	Premises	Conclusion	Support	Confi...	La
67	h/tc = range5 [182.050 - ∞), TL = range5 [577.500 - ∞]	h = range5 [31.350 - ∞]	0.080	0.920	0.
102	TL = range5 [577.500 - ∞), P (W/kg) = range5 [47.350 - ∞]	h = range5 [31.350 - ∞]	0.084	0.923	0.

Vysoké výšky výskoku dosahují krasobruslaři s vysokou kvalitou skoku h/tc a s vysokým P – to se dalo očekávat. Překvapující je opět výskyt atributu TL u obou pravidel – vysvětlení je obdobné vysvětlení u předchozího obrázku (Obrázek 6: Asociační pravidla 3).

Metoda generovala velké množství dalších pravidel (viz. příloha Analýza_dat_příloha.docx), nicméně z různých důvodů mi už nepřišla zajímavá (buď redundantní, nebo zřejmá, nebo prakticky nezajímavá).

2.5. Shlukování

Další metodou, kterou jsem při analýze dat použil, bylo shlukování.

Jelikož hledáme přirozené shluky (α-shluky) v datech a neznáme vzdálenost α, použijeme aglomerativní shlukování se strategií nejbližšího souseda.

Byly použity atribut pohl a všechny standardizované vek_s, …, h_tc_s

Program neumí pracovat s chybějícími údaji, bylo nutné se s tím vypořádat. Udělal jsem to takto (buď a) nebo b)):

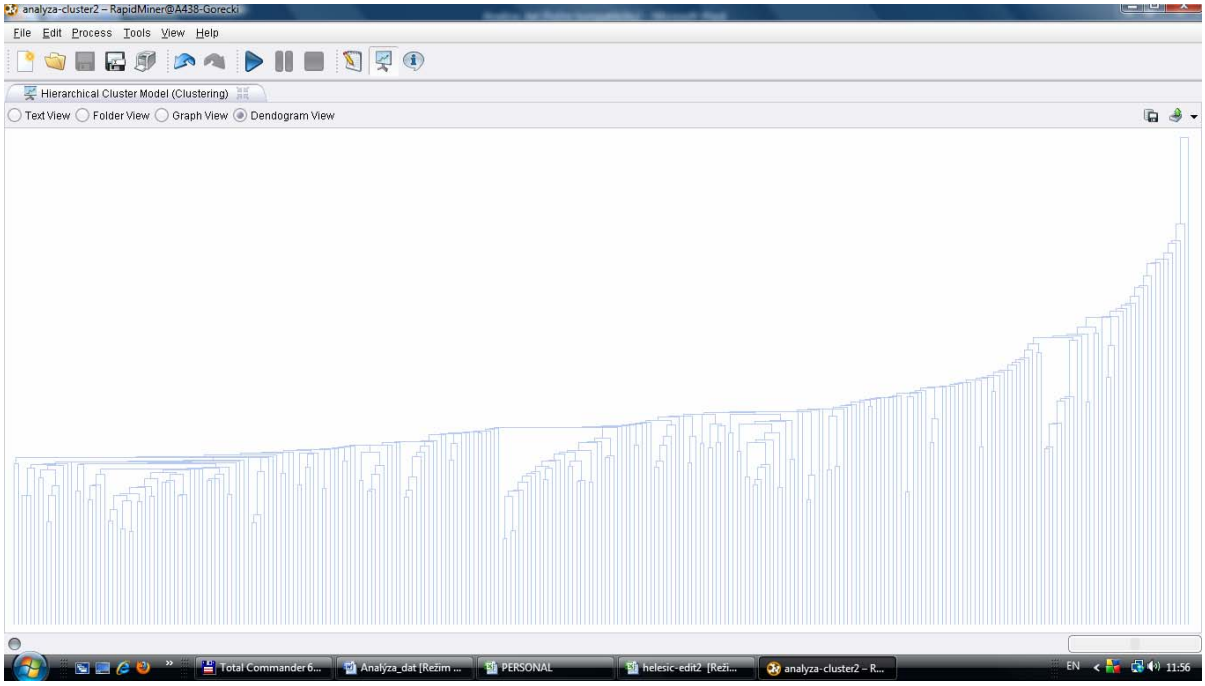
a) Odstranil jsem atributy, ve kterých chyběly údaje

b) Odstranil jsem záznamy, ve kterých chyběly údaje

Výsledky shlukování

XXX … pro která data?

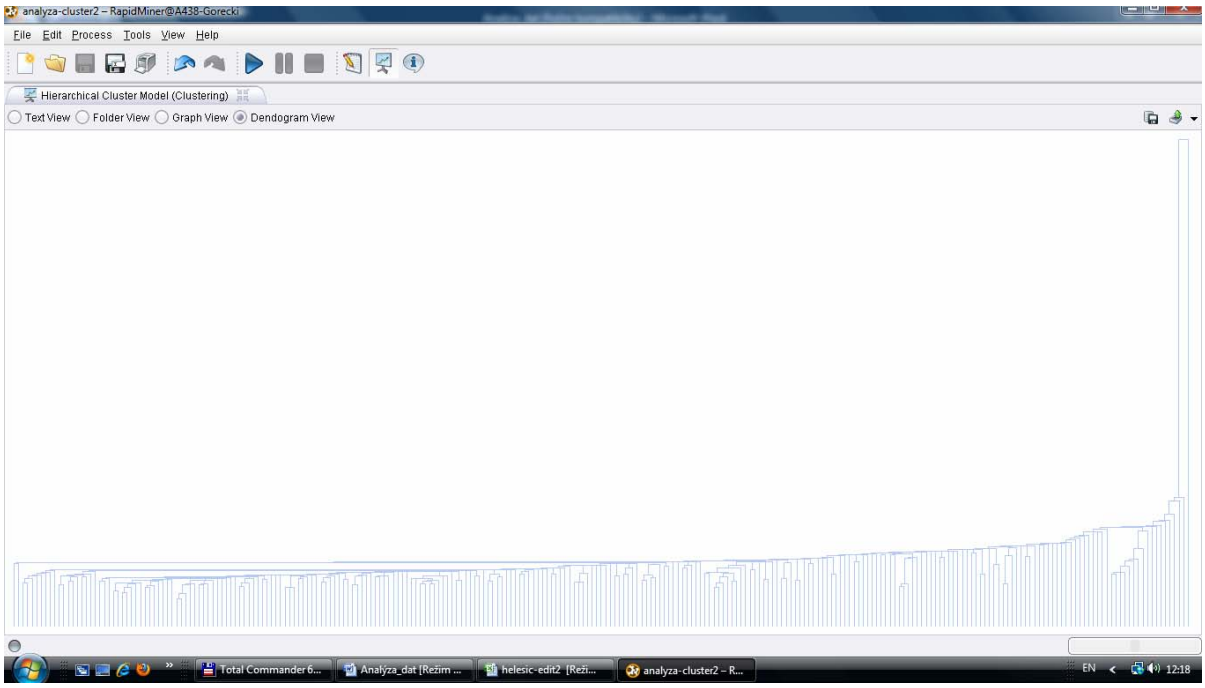
Obrázek 8: Shlukování a)



Vidíme, že v datech nejsou žádné výrazné shluky. Lze to interpretovat tak, že mezi krasobruslaři neexistují nějaké skupinky výrazně lepších (uvažme to, že se zabýváme převážně jejich výkonnostními parametry) nebo horších jedinců, ale je to tak, jak už to v životě bývá, že většina lidí jsou průměrní, občas má někdo lepší jednu vlastnost a horší jinou vlastnost, a tudíž ani nemůžeme čekat nějaké výrazně oddělené skupinky superdobrých, superšpatných nebo superprůměrných krasobruslařů.

XXX … pro která data?

Obrázek 9: Shlukování b)

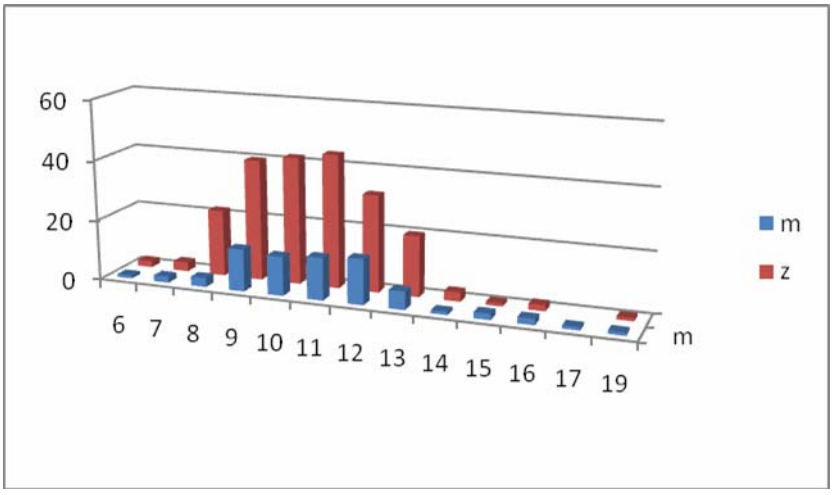


Výsledek, který vidíme zde je obdobný jak v bodě a) – opět žádné výrazné shluky.

Věková kategorie 9 až 12 let

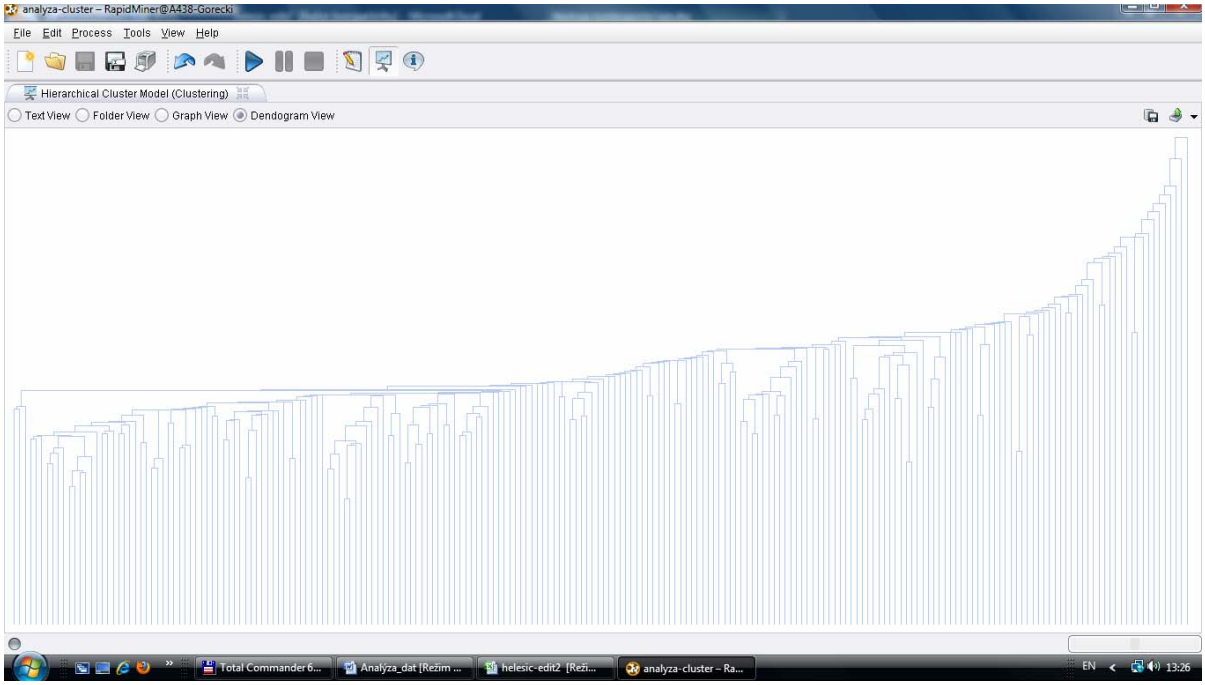
Dále mě tedy zajímalo, jak by dopadlo shlukování, kdybych vyjmul z dat ty nejstarší a nejmladší jedince, abych tím omezil vliv věku na jednotlivé vlastnosti (evidentně na věku závislých). Vybral jsem si co nejpočetnější, avšak věkově omezenou skupinku podle histogramu (věk – počty mužů a žen - obr. níže) a to jedince ve věku 9-12 let včetně.

Obrázek 10: Histogram atributů Věk x Pohlaví



Dendrogram dopadl takto (z dat jsou vyjmuty atributy s chybějícími údaji):

Obrázek 11: Dendrogram 9-12let



Lze vidět, že vzdálenosti mezi objekty jsou menší, dendrogram je více „sešlápnutý“. Nicméně, žádné výrazné shluky se opět neobjevily.

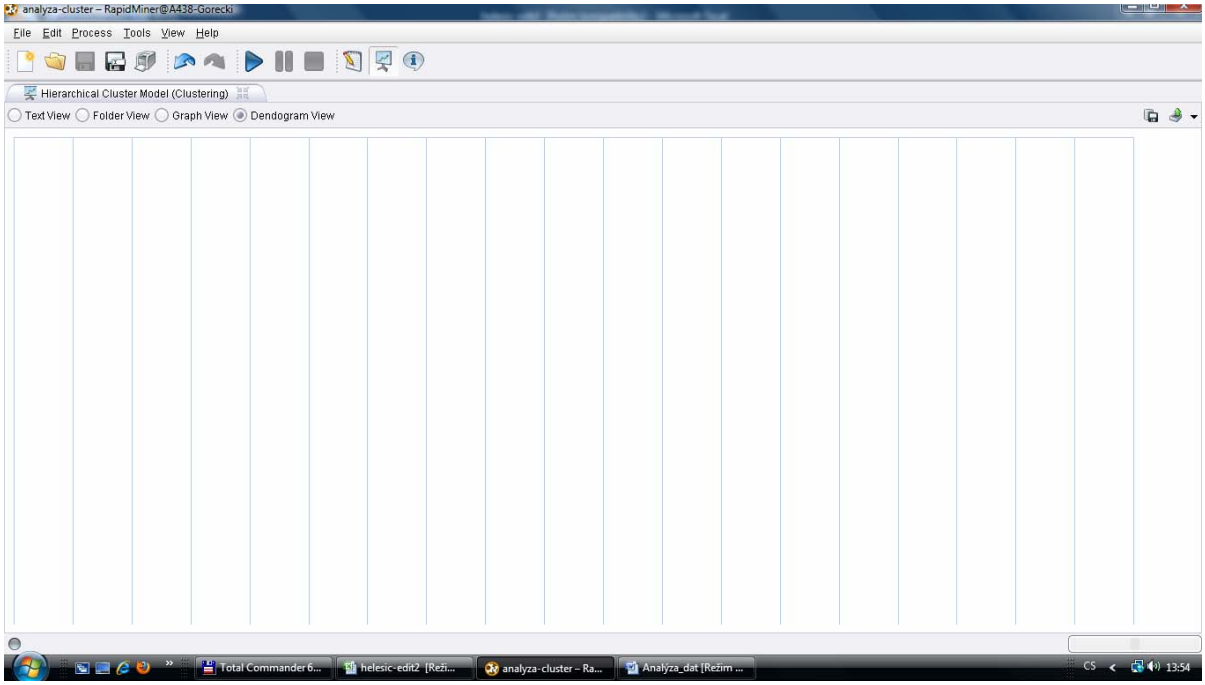
Dále jsem shlukl data po vyjmutí pouze záznamů s chybějícími údaji, nicméně výsledek dopadl opět velmi podobně.

Shlukování atributů

Jelikož jsem zatím žádné shluky nedostal, zkusil jsem, jak dopadne situace při shlukování transponovaných dat. Opět aglomerativní shlukování se strategií nejbližšího souseda, věk 9-12 let, bez chybějících záznamů.

Shlukování všech atributů:

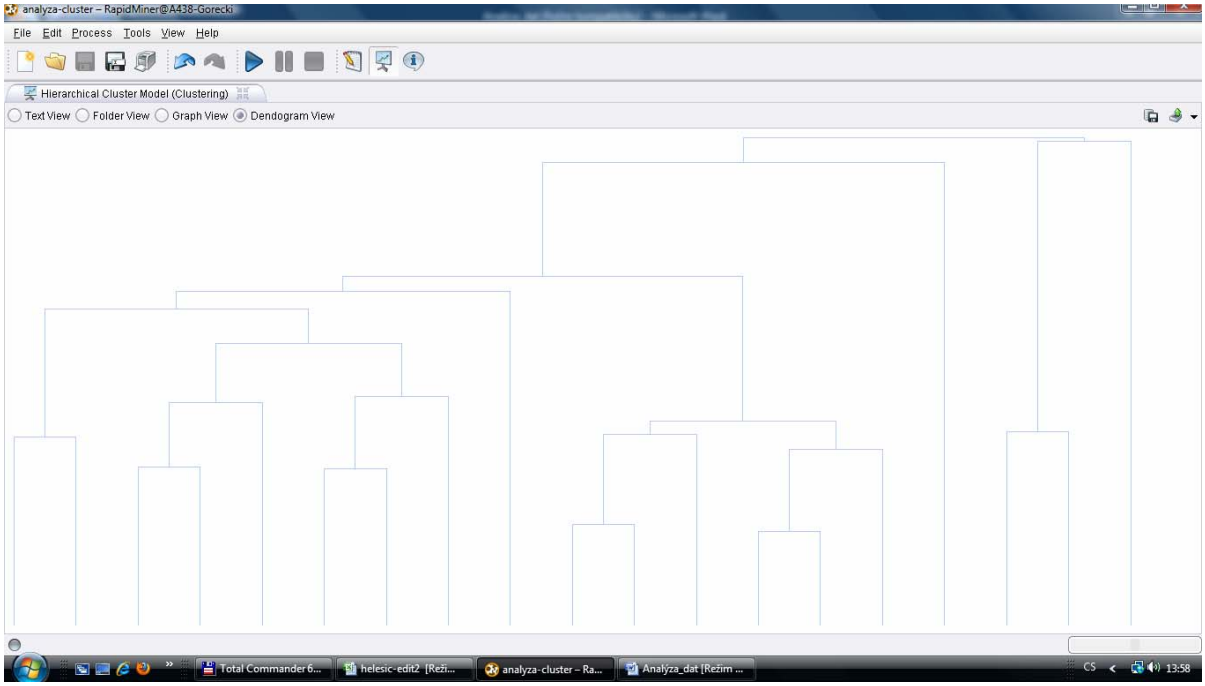
Obrázek 12: Dendrogram shlukování všech atributů



Tento výsledek byl pro mě zklamáním, čekal jsem, že se aspoň nějaké shluky objeví.

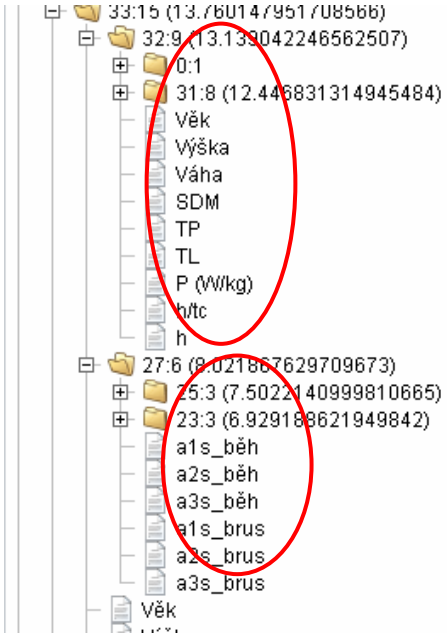
Zkusil jsem vyjmout z analýzy atribut pohlaví (jediný binární atribut, ostatní jsou reálné). Následující výsledek mě překvapil. Dopadl takto:

Obrázek 13: Dendrogram shlukování všech atributů bez atributů Pohlaví



Konečně se objevily nějaké shluky. Podívejme se, které atributy se shlukly pod oranžovou hladinou.

Obrázek 14: Shluky atributů – první část



Shluky atributů:

Zde jsou atributy, které jsou v dendrogramu vlevo (9 a 6 atributů).

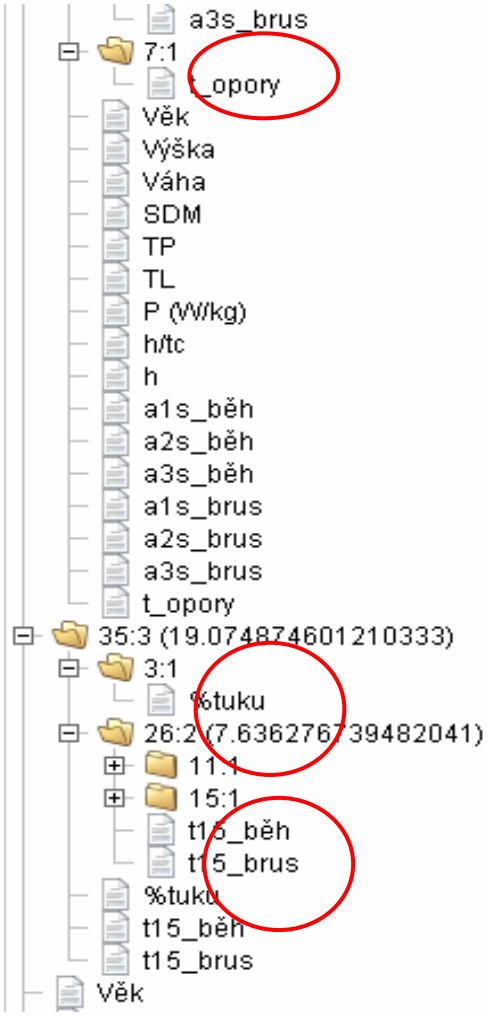
První skupinka (elipsa nahoře) jsou atributy ze skupiny Zákl, Skok_dálka a 3 atributy ze Skok_Výška.

Výška, váha, věk a dálka a výška skok spolu jistě v realitě souvisí (jde to částečně vidět i z korelační matice). Tudíž tento závěr není příliš překvapivý.

Druhá skupinka (elipsa dole) jsou atributy průměrných rychlostí ze skupin Běh a Běh_led. Opět ne příliš překvapující výsledek – atributy mají téměř identický charakter.

Shlukování dalších atributů směrem zleva doprava na dendrogramu lze vidět níže.:

Obrázek 15: Shluky atributů – druhá část



K druhé skupince z předchozího obrázku se doshlukl (až nad oranžovou hladinou) poslední atribut t_opory ze skupiny Skok_Výška. Je to dáno tím, že tento hodnota tohoto atributu není ideální, pokud je extrémní (co největší nebo co nejmenší), ale pokud je někde „uprostřed“ – ani moc velká, ani moc malá (viz. popis atributů v kapitole Data).

Nejvzdálenější atribut je %tuku (shlukl se až na nejvyšší hladině – tedy výše než oranžová hladina). Opět – toto není výkonnostní atribut, na rozdíl od ostatních.

A poslední dva atributy se shlukly na nízké hladině (pod oranžovou), jak by se dalo očekávat (časy na běh na 15m).

2.6. Rozhodovací stromy

V analýze pomocí rozhodovacích stromů jsem:

- použil atributy Pohlaví, Váha, %tuku, SDM, TP, TL, t_opory, h, t15_běh, t15_brus, a1s_běh, a1s_brus – jsou to zástupci jednotlivých podskupin atributů a vybral jsem je na základě výsledků korelační matice a konzultaci s autorem dat
- použil vybrané atributy bez předzpracování, pouze jsem diskretizoval klasifikační atributy do 5 ekvifrekvenčních intervalů
- pracoval i s atributy s chybějícími údaji
- jako klasifikační atribut vybral výšku skoku h, jelikož hodnota tohoto atributu je pro krasobruslaře zásadní

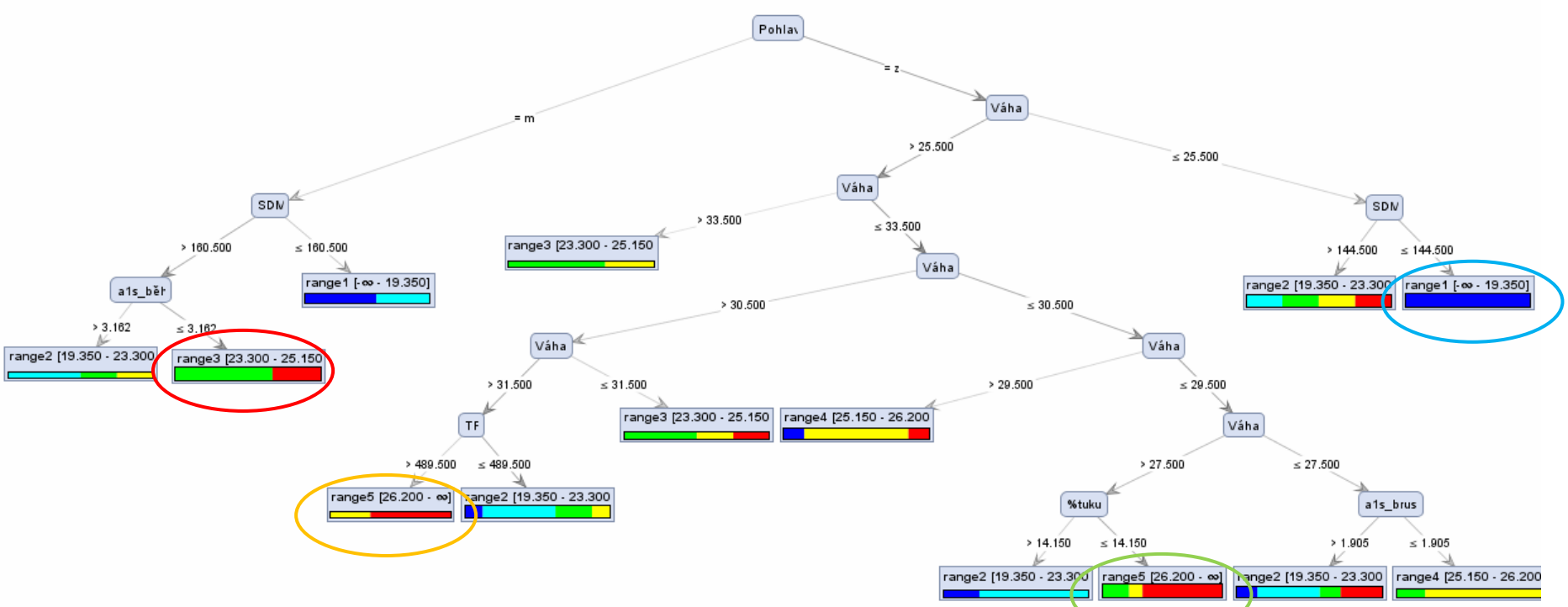
Abych se dozvěděl něco více o vztahu výšky skoku k ostatním atributům, použil jsem právě analýzu pomocí rozhodovacích stromů, která je pro tyto účely velmi vhodná. Pokud jsem však pro tvorbu stromu použil všechny objekty z dat (analyzoval jsem celý vzorek dat najednou), dozvěděl jsem se pouze to, že těžší (a tedy i silnější a zřejmě i starší) jedinci skáčou převážně vysoko a lehčí jedinci nízko, ale už nic moc dalšího. Abych tedy omezil vliv váhy na výšku skoku, rozdělil jsem krasobruslaře do několika skupin. Vzhledem k tomu, že váha u takto mladých jedinců je silně závislá na věku, rozdělil jsem tedy jedince do několika věkových kategorií. Navíc mi bylo od autora dat doporučeno rozdělení, které se běžně při tréninku používá. Použil jsem tedy kategorie:

1) do 9 let včetně (20 chlapců, 67 dívek),
2) 10 až 12 let (43 chlapců, 118 dívek),
3) 13 až 15 let (9 chlapců, 24 dívek),
4) 16 let a více (4 chlapců, 3 dívky).

Při takovémto rozdělení se potom objevilo množství vztahů mezi atributy, o kterých mi předchozí analýzy nic neřekly (např. netriviální vztah výšky výskoku h a %tuku, který jaksi tušíme, nicméně zatím se nikde neobjevil (např. korelace těchto dvou atributů je velmi nízká, téměř nulová).

Pozn.: Kategorii číslo 4) neuvádím, je příliš málo zastoupená a výsledky neobsahují nic zajímavého.

Obrázek 16: Věková kategorie do 9 let



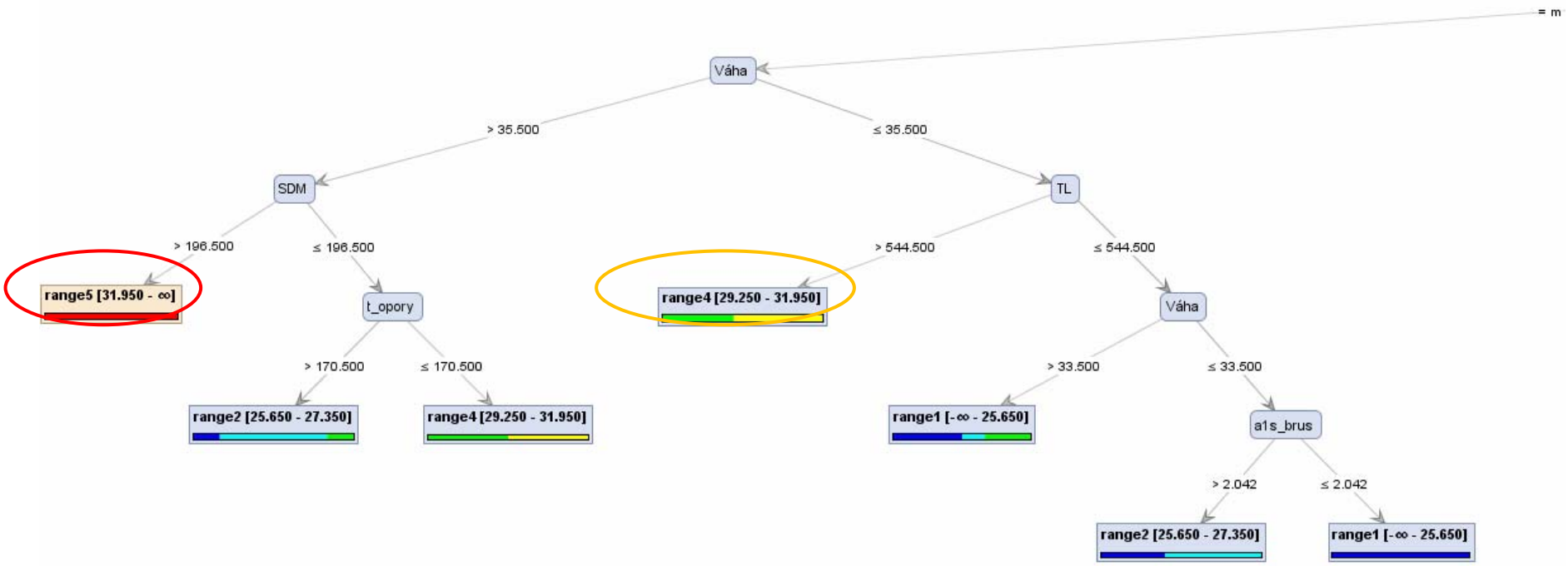
Zajímavé je, že kořenový atribut je pohlaví, který dělí strom na dvě větve odpovídající mužům a ženám (v této věkové kategorii spíše chlapcům a dívkám) – poprvé se tedy ukazuje vztah Pohlaví a výšky skoku.

Pro chlapce je pak rozhodující SDM, který je dělí na lepší a na horší. Zarážející je však to, že ti lepší z těch lepších mají a1s_běh < 3,162 (červená elipsa vlevo). *Vypadá to tedy tak, že rychlost rozběhu po 1 vteřině má na výšku výskoku u chlapců přesně opačný vliv, než bychom čekali.*

U žen rozhoduje nejdříve váha a potom:

1) TP pro ty nejtěžší dívky (oranžová elipsa)
2) %tuku pro méně těžké dívky (zelená elipsa)
3) SDM pro nejlehčí dívky (modrá elipsa) – pro malé SDM skáčou dívky jen nízko

Obrázek 17: Věková kategorie 10 až 12 let - větev muži

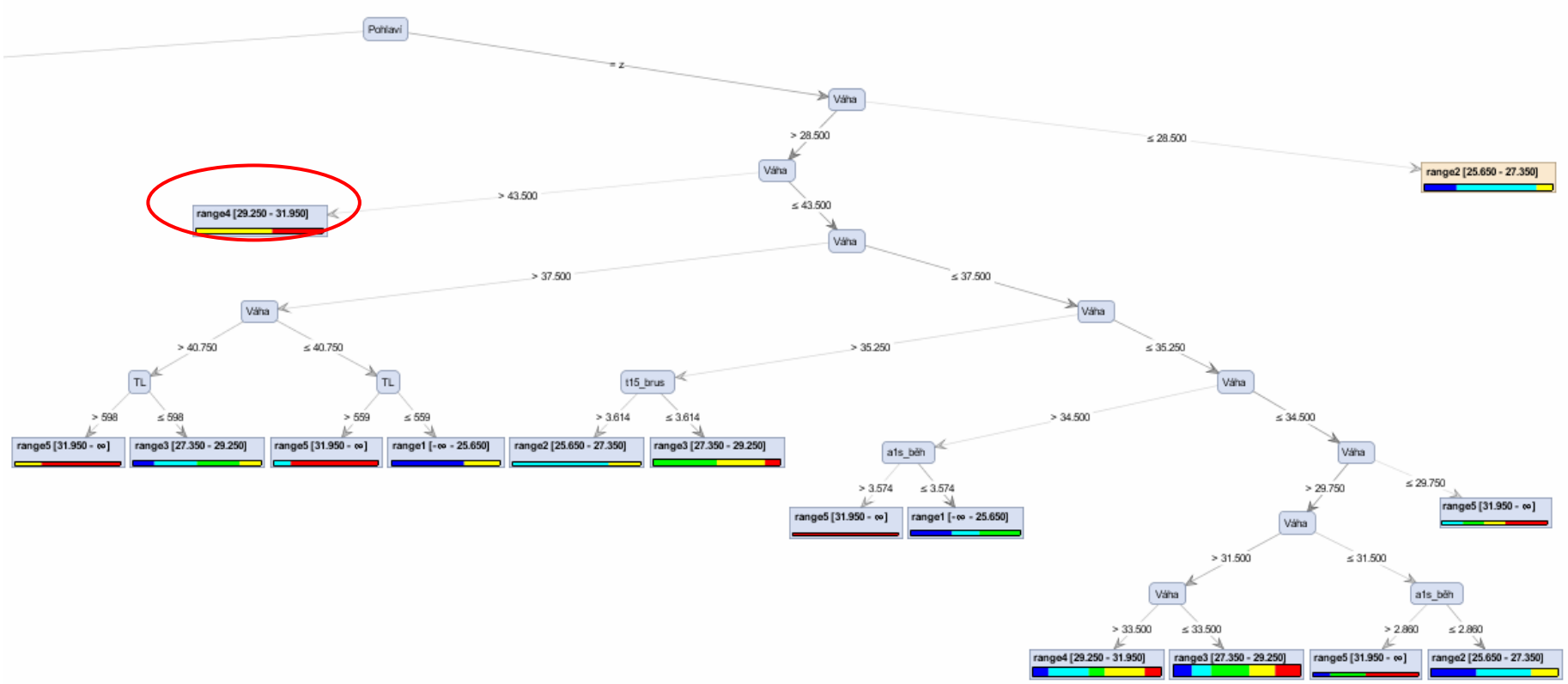


Pro tuto početně nejvíc zastoupenou kategorii byl vygenerován poměrně košatý strom, který jsem musel rozdělit na dvě části. Kořenový atribut byl Pohlaví, horní obrázek tedy odpovídá chlapcům a obrázek na další stránce odpovídá dívkám.

U chlapců se objevila skupina o počtu 7, která je těžká a SDM má vysoký (červená elipsa).

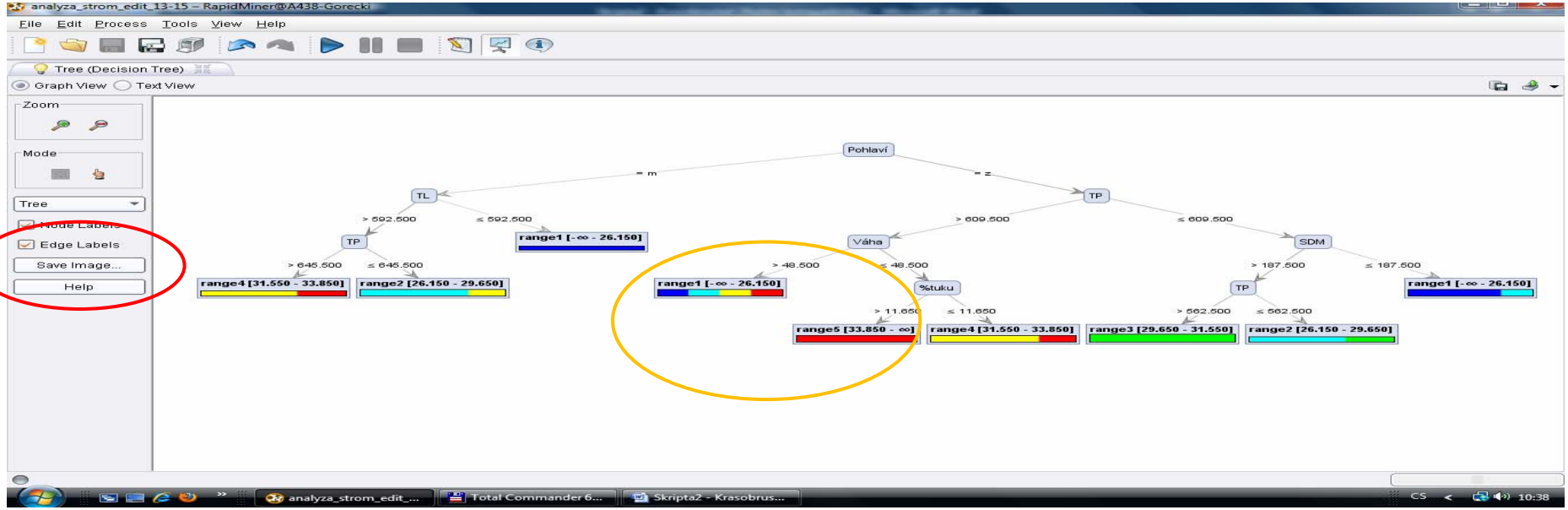
Pro lehčí chlapce (pravá větev) se ukazuje, že ti lepší z nich mají vysoké TL (oranžová elipsa).

Obrázek 18: Věková kategorie 10 až 12 let - větev ženy



Pro ženy v této věkové kategorii lze tvrdit pouze to, že ty těžší (a zřejmě tedy silnější) skáčou nejvýše (červená elipsa). Zbylá pravidla (odpovídající jednotlivým listům) mají buď příliš nízkou podporu, nebo příliš nízkou spolehlivost, aby se dalo na jejich základě něco tvrdit.

Obrázek 19: Věková kategorie 13 až 15 let



Kořenovým atributem je opět pohlaví.

Chlapci, kteří skáčou dobře na obě nohy (vysoké TP a TL), jsou nejlepší (červená elipsa). Lze tedy soudit, že pro vysoký výskok do výšky je třeba mít obě nohy „dobré“ (pokud je jedna horší, hned se skáče méně)

U dívek dělí ty lepší od těch horších délka trojskoku TP. Pokud navíc ve skupině těch lepších (TP > 609,5) mají děvčata menší váhu než 48,5 kg, pak skáčou nejvýše (oranžová) - ještě jemněji pak lze tyto nejlepší děvčata rozdělit podle %tuku, kde ty s menším množstvím tuku jsou opravdu ty „nejlepší z nejlepších“.

Seznam obrázků