

STULONG

Datamining

zápočtová analýza č. 1

Zpracovali: Robert Poch, Pavel Petřek

Cvičící: Mgr. Tomáš Karban

Zdrojová data: <http://euromise.vse.cz/stulong>

Použitý software: MS Access, 4ftMiner

1. Obsah

1. Obsah.....	- 1 -
2. Zkoumané charakteristiky	- 2 -
3. Zpracování atributů	- 3 -
3.1 Míra pití alkoholu.....	- 4 -
3.2 Intenzita tělesných aktivit.....	- 5 -
4. Dílčí analytické otázky	- 5 -
5. Výsledky analýz.....	- 6 -
6. Slepé uličky analyzování.....	- 10 -
7. Závěr.....	- 10 -

2. Zkoumané charakteristiky

- Konzumace alkoholu složená z následujících sub-charakteristik:
 - Pití alkoholu
 - Popisuje obecný vztah jedince k alkoholu. Je členěna do skupin: „nepije“, „pije příležitostně“ a „pije pravidelně“.
 - Konzumace 7° piva
 - Tato charakteristika říká, zda jedinec pije pivo 7°.
 - Konzumace 10° piva
 - Tato charakteristika říká, zda jedinec pije pivo 10°.
 - Konzumace 12° piva
 - Tato charakteristika říká, zda jedinec pije pivo 12°.
 - Konzumace vína
 - Tato charakteristika říká, zda jedinec pije víno.
 - Konzumace lihovin
 - Tato charakteristika říká, zda jedinec pije tvrdý alkohol.
 - Množství konzumovaného piva
 - Tato charakteristika je doplňkem informace o požívání jednotlivých druhů piv a vypovídá o množství konzumovaného piva. Rozlišované kategorie jsou „nepije“, „do jednoho litru“ a „více než jeden litr“.¹
 - Množství konzumovaného vína
 - Tato charakteristika udává množství konzumovaného vína. Jednotlivé kategorie jsou „nepije“, „do půl litru“ a „více než půl litru“.¹
 - Množství konzumovaného alkoholu
 - Tato charakteristika informuje o množství konzumovaných lihovin – tvrdého alkoholu. Je řazena do kategorií „nepije“, „do 100cl“ a „více než 100cl“.¹
- Tělesné aktivity složené z následujících sub-charakteristik:
 - Tělesná aktivita v zaměstnání
 - Charakteristika popisující množství pohybu jedince v zaměstnání je rozdělena na následující skupiny: „převážně stojící“, „převážně sedící“, „převážně chodící“, „přenášející těžká břemena“.
 - Tělesná aktivita po zaměstnání

- Charakteristika popisující množství pohybu jedince mimo jeho zaměstnání. Skupiny rozdělení jsou: „převážně sedící“, „mírná aktivita“, „velká aktivita“.
- Způsob dopravy do zaměstnání
 - Charakteristika popisující dopravní prostředek, kterým se jedinec dopravuje do svého zaměstnání. Jednotlivé kategorie jsou „pěšky“, „na kole“, „veřejným dopravním prostředkem“ a „autem“.
- Doba dopravy do zaměstnání
 - Charakteristika říká, jak dlouho trvá jedinci dopravit se z domova do zaměstnání. Doba je členěna po půl hodinách a to: „cca půl hodiny“, „cca jedna hodina“, „cca dvě hodiny“, „více než dvě hodiny“.
- Kožní řasy popisující množství tělesného tuku sestávají z těchto sub-charakteristik:
 - Kožní řasa na tricepsu
 - Určuje množství tuku na spodní straně tricepsového svalu. Hodnoty jsou v milimetrech tloušťky při nařazení kůže.
 - Kožní řasa na subscapularu
 - Určuje množství tuku na zádech v prostoru pod lopatkou. Hodnoty jsou v milimetrech tloušťky při nařazení kůže.
- Skupiny pacientů dle rizikovosti
 - Normální skupina bez přítomných rizikových faktorů
 - Riziková skupina s přítomnými některými rizikovými faktory
 - Patologická skupina s nemocným srdcem nebo jinou vážnou chorobou, která statisticky vylučuje dlouhodobé sledování.

¹ Poznámky k charakteristikám:

- Uváděná množství alkoholů jsou počítána na jeden den.

3. Zpracování atributů

Před samotných analyzováním dat bylo třeba data patřičně předpřipravit, aby byla jednak použitelná pro zpracování v aplikaci 4ftMiner a abychom zároveň eliminovali anomálie vzniklé při pořizování informací od sledovaných jedinců.

Základní informaci o vztahu jedince k alkoholu (*Pití alkoholu*) jsme z dalšího výzkumu eliminovali, neboť fakta, která vykazovala, se neshodovala s dalšími charakteristikami. Místo

toho jsme zavedli vlastní charakteristiku, která hovoří o míře pití alkoholu a která věrněji vystihuje ostatní data uváděná u jedince. Je možné se domnívat, že zmiňované neshody vznikly jako snaha jedinců zlehčovat svou konzumaci alkoholu určením sebe sama jako nepijáka, zatímco u otázek na konzumaci jednotlivých druhů alkoholů svorně uvedli jejich užívání.

Další odstranění informace jsme provedli u pití piva 7°. Informaci jsme kvůli malému vzorku jedinců sloučili spolu s pitím piva 10°.

U množství alkoholů konzumovaných jedinci bylo také uváděno, že např. nepijí pivo, zatímco v množství spotřebovaného piva uváděli skupinu „do jednoho litru“. Takovýmto jedincům jsme indikátor o pití piva nastavili na ano. Analogicky jsme postupovali u ostatních druhů alkoholu.

Další charakteristikou, kterou bylo třeba explicitně zavést (*stejně jako Míra pití alkoholu*) z pořízených dat byla informace o tělesných aktivitách. Aktivity jsme rozdělili do tří skupin podle jejich intenzity. Detaily jsou uvedeny níže.

Množství kožních řas jsou číselné hodnoty, proto jsme se na ně pokusili nahlížet více způsoby. Jako první krok to bylo dělení do stejně dlouhých intervalů po 2mm a později to bylo rozdělení do tří větších skupin dle počtu jedinců tak, aby byly skupiny přibližně stejně velké. Při dělení do skupiny po 2mm jsme byli nuceni několik nejvyšších skupin sloučit dohromady, protože nízké počty jedinců ve skupinách by nevykázaly žádnou relevantní informaci.

Jako poslední úpravu jsme spojili nejvyšší dva časy uváděné u dopravy do zaměstnání, neboť nejvyšší skupina byla málo početná a navíc je-li doprava alespoň dvě hodiny, je možné ji již považovat za časově náročnou.

Je třeba uvést, že u všech zpracovávaných charakteristik se více či méně často objevovaly neuvedené či nevyplněné informace. V takových případech nebyl jedinec s neuvedenou informací při analýze brán v potaz.

3.1 Míra pití alkoholu

- Úroveň 1
 - Je-li u každého druhu alkoholu shodně uvedena informace, že nepije, je jedinec zařazen do první skupiny, tedy skupiny „nepije“

- Úroveň 2
 - Je-li uvedeno alespoň u jednoho druhu alkoholu, že jej jedinec pije a nevyskytuje-li se u žádného alkoholu informace o nadměrném požívání, je jedinec zařazen do druhé skupiny „mírné konzumace“.
- Úroveň 3
 - Je-li alespoň u jednoho z alkoholů uvedena informace o nadměrné konzumaci, je jedinec zařazen do třetí skupiny „nadměrná konzumace“.

3.2 Intenzita tělesných aktivit

- Mírné pohybové aktivity
 - Je-li jedinec v zaměstnání spíše sedící nebo stojící nebo mimo zaměstnání nejvíce sedí, je zařazen do skupiny „mírných pohybových aktivit“.
- Střední pohybové aktivity
 - Je-li jedinec z těch, kteří v práci chodí, má-li po zaměstnání mírně pohybu nebo dopravuje-li se do zaměstnání pěšky alespoň hodinu nebo na kole alespoň půl hodiny, je zařazen do skupiny středních pohybových aktivit.
- Velké pohybové aktivity
 - Zvedá-li jedinec v zaměstnání těžká břemena, má-li po zaměstnání spoustu pohybu nebo dopravuje-li se do práce pěšky nebo na kole alespoň dvě hodiny, je zařazen do „velkých pohybových aktivit“.

4. Dílčí analytické otázky

- Jako základní směr analýzy jsme se snažili orientovat na určení závislosti množství kožních řas na ostatních charakteristikách, tzn. množství pohybu jedince a množství a druhu konzumovaného alkoholu, neboť tato závislost působí nejvíce logicky vzhledem ke zpracovávaným informacím.
- Dalším analyzováním jsme se pokusili ukázat opačný fakt a to závislost jakékoliv charakteristiky na množství tělesného tuku.
- Následný výzkum se ubíral k určení možných závislostí množství podkožního tuku na jednotlivých místech na těle.

- Na závěr hledání jsme se zaměřili na vzájemnou závislost mezi jednotlivými charakteristikami a rizikovými faktory určujícími zařazení pacienta do normální / rizikové / patologické skupiny, přestože zkoumané charakteristiky samy nejsou rizikovými faktory rozhodujícími o zařazení.

5. Výsledky analýz

- Skupina silnějších jedinců jako skupina s nedostatkem pohybu
 - Jako jeden z prvních výsledků jsme zjistili závislost, kterou lze zařadit do skupiny obvyklých známých faktů. U téměř 13% lidí s mírnou aktivitou bylo zjištěno, že spadají do skupiny s kožní řasou na tricepsu od 17mm do 20mm. Přestože se 13% nezdá být mnoho, je to hodnota o více než 160% vyšší než je výskyt celé skupiny s tricepsovými tuky 17mm až 22mm ve všech sledovaných jedincích.

	Tricepsová řasa 17mm až 22mm	Tricepsová řasa jiná
Aktivity mírné	22	169
Aktivity jiné	43	1183

* Rozdělení jedinců pro uvedené zjištění

- při rozčlenění jedinců do tří velkých skupin dle jejich tricepsových řas (*malé / střední / velké řasy, jak bylo uvedeno výše*), jsme dostali mnohem přesnější výsledky.

	Tric. řasa vysoká	Ostatní Subsc. a Tric.
Aktivity mírné, žádné lihoviny	44	21
Ostatní	549	803

Výskyt jedinců s danými řasami ve skupině je 68%.

To je o 62% více než daných řas ve všech jedincích.

* Rozdělení jedinců pro uvedené zjištění

	Tric. řasa vysoká	Ostatní Subsc. a Tric.
Aktivity mírné, pivo do jednoho litru, cesta do práce asi půl h.	44	27
Ostatní	549	797

Výskyt jedinců s danými řasami ve skupině je 62%.

To je o 48% více než daných řas ve všech jedincích.

* Rozdělení jedinců pro uvedené zjištění

- Skupina se spoustou tělesné aktivity a pozitivním vztahem k alkoholu a její kolísavé tělesné tuky

- U skupiny s výraznou pohybovou aktivitou jsme v závislosti na drobných změnách jejich vztahu k alkoholu zjišťovali zajímavé odchylky v tělesných tucích.
- Jako první výsledek to bylo zjištění, že z jedinců s velkým množstvím pohybu, kteří pijí průměrně do 100cl lihovin denně a do půllitru vína denně je 30% těch, kteří mají subscapulární kožní řasy od 12mm do 17mm a tricepsově kožní řasy od 9mm do 14mm. Šetření se vztahuje na rizikovou spolu s patologickou skupinou. Jedince si lze představit jako lehce obéznější nikoliv ale s příliš tučnými pažemi, kteří se hodně pohybují. Opět je třeba říci, že jakkoliv se nezdá 30% vysoká hodnota, je poměr jedinců s takovými kožními řasami ke všem specifikovaným jedincům o 150% vyšší než poměr všech jedinců s danými řasami ke všem sledovaným jedincům.

	Subsc. 12mm až 17mm Tric. 9mm až 14mm	Ostatní Subsc. a Tric.
Vysoké aktivity, do 100cl lihovin, do ½ litru vína	27	63
Ostatní	81	733

* Rozdělení rizikových a patologických jedinců pro uvedené zjištění

- Jeden z dalších výsledků pouze zaměnil vztah k lihovinám za vztah k pivu a posunul tím subscapulární kožní řasy směrem nahoru na 14mm až 17mm. Jedinců s takovými řasami je ve vybraných jedincích 23% a to je o 174% více než je všech jedinců s danými řasami ve všech sledovaných jedincích. Opět se jedná o rizikovou a patologickou skupinu.

	Subsc. 14mm až 17mm Tric. 9mm až 14mm	Ostatní Subsc. a Tric.
Vysoké aktivity, do litru piva, do ½ litru vína	22	72
Ostatní	55	755

* Rozdělení rizikových a patologických jedinců pro uvedené zjištění

- V dalších výsledcích, které se již nevztahovali pouze na rizikovou a patologickou skupinu, se objevily variace předchozích zjištění pouze s drobnými odchylkami.

	Subsc. 12mm až 17mm Tric. 9mm až 10mm	Ostatní Subsc. a Tric.
Vysoké aktivity, do 100cl lihovin, do ½ litru vína	22	114
Ostatní	69	1212

Výskyt jedinců s danými řasami ve skupině je 16%.
To je o 152% více než daných řas ve všech jedincích.

* Rozdělení jedinců pro další zjištění

	Subsc. 14mm až 17mm Tric. 9mm až 10mm	Ostatní Subsc. a Tric.
Vysoké aktivity, míra alkoholu mírná, do ½ litru vína	22	169
Ostatní	43	1183

Výskyt jedinců s danými řasami ve skupině je 12%.
To je o 151% více než daných řas ve všech jedincích.

* Rozdělení jedinců pro další zjištění

- Další zástupci velmi se pohybující skupiny tentokrát ale z průměrně „tučných“ jedinců
 - Jedněmi z výsledků s výraznou pohybovou aktivitou byly také ukázky jedinců s průměrnou velikostí tricepsových řas a se „zdravým“ vztahem k lihovinám. Vzhledem k nízkému rozdílu procentních výskytů daných tricepsů v rámci vybrané skupiny a rámci všech sledovaných jedinců je třeba říci, že se nejedná o příliš významný výsledek.

	Tric. 5mm až 10mm	Ostatní Tric.
Vysoké aktivity, lihovin do 100cl	122	62
Ostatní	616	617

Výskyt jedinců s danými řasami ve skupině je 66%.
To je o pouhých 27% více než daných řas ve všech jedincích.

* Rozdělení jedinců v prvním uvedeném zjištění

	Tric. 5mm až 10mm	Ostatní Tric.
Vysoké aktivity, lihovin do 100cl, míra alkoholu mírná	114	56
Ostatní	624	623

Výskyt jedinců s danými řasami ve skupině je 67%.
To je o pouhých 29% více než daných řas ve všech jedincích.

* Rozdělení jedinců ve druhém uvedeném zjištění

- Skupina lihových existencí
 - Další z výsledků lze považovat za jeden z výrazných, ale zároveň očekávaných. Jedná se totiž o skupinu výrazných konzumentů tvrdého alkoholu. V této skupině je 64% jedinců s podprůměrnými tricepsovými řasami od 3mm do 8mm. Zároveň je na tomto výsledku vidět způsob volby atributu Míry konzumace alkoholu, neboť je-li skupina navíc omezena ještě výraznou

konzumací alkoholu, vychází naprosto totožný číselný výsledek. Tyto jedince si lze typicky představit jako muže s dlouhými vlasy a koženou bundou, kteří obchází hospodská zařízení a konzumují obvykle jeden rum za druhým.

	Tric. 3mm až 8mm	Ostatní Tric.
Lihoviny nad 100cl	47	27
Ostatní	503	840

* Rozdělení jedinců v prvním uvedeném zjištění

Výskyt jedinců s danými řasami ve skupině je 64%.

To je o 64% více než daných řas ve všech jedincích.

	Tric. 3mm až 8mm	Ostatní Tric.
Lihoviny nad 100cl, vysoká konzumace alkoholu	47	27
Ostatní	503	840

* Rozdělení jedinců ve druhém uvedeném zjištění

Výskyt jedinců s danými řasami ve skupině je 64%.

To je o 64% více než daných řas ve všech jedincích.

- Závislost zařazení do skupiny na základě jiných než rizikových charakteristik
 - Jako zajímavý výsledek, ne však neočekávaný, lze zmínit závislost zařazení do rizikové skupiny na velikosti kožních řas, ačkoliv kožní řasy nejsou přímým rizikovým faktorem. Zjistili jsme, že více než 90% jedinců s tricepsovou řasou od 15mm do 16mm a více než 93% se subscapulární řasou od 26mm do 31mm a s tricepsovou řasou od 5mm do 10mm patří do rizikové skupiny. Nalezenou závislost lze považovat jako tranzitivní přenos BMI, který je rizikovým faktorem a velikost kožních řas s ním jednoznačně souvisí.

	Riziková skupina	Ostatní
Tric. 15mm až 17mm	40	4
Ostatní Tric.	819	554

* Rozdělení jedinců v prvním uvedeném zjištění

Výskyt rizikových jedinců mezi jedinci s danými řasami je 91%.

To je o 50% více než rizikových jedinců ve všech jedincích.

	Riziková skupina	Ostatní
Subsc. 26mm až 31mm Tric. 5mm až 10mm	43	3
Ostatní	816	555

* Rozdělení jedinců v druhém uvedeném zjištění

Výskyt rizikových jedinců mezi jedinci s danými řasami je 94%.

To je o 54% více než rizikových jedinců ve všech jedincích.

6. Slepé uličky analyzování

- Jak je možné nahlédnou z dílčích analytických otázek, nebyla mezi sledovanými charakteristikami (*vyjma zařazení do rizikové skupiny*) zjištěna žádná závislost na tělesných řasách. Je pochopitelné, že spíše tělesné řasy jsou důsledkem životního stylu jedince, než aby životní styl byl závislý na přítomnosti podkožního tuku. Tedy takové analýzy nepřinesla žádné výsledky.
- Dalším neúspěšným krokem bylo sledování závislosti rozložení podkožního tuku po těle, což je překvapivý výsledek, neboť by bylo logické se domnívat, že silnější (*co do podkožního tuku*) jedinci budou mít vysoké kožní řasy jak na tricepsu, tak v podlopatkové části.

7. Závěr

Analýzy ukázaly jak výsledky, které by bylo možné zařadit do kategorie obvyklých známých jevů, tak některé zajímavější skutečnosti. Nejobvyklejší jevy prezentující závislost množství podkožního tuku na ostatních charakteristikách, které by člověk očekával nejčastěji, bylo přesto obtížné získat, protože kvalita dat, co do nevyplnění informací o kožních řasách, je velmi bídná. Přesto se však po řádné úpravě dat podařilo tyto výsledky získat a snad vypovídají i cosi zajímavého.