**UNIVERZITA KARLOVA V PRAZE**

Přírodovědecká fakulta

Katedra demografie a geodemografie

# VYUŽITÍ DATA MININGOVÝCH METOD PŘI ZPRACOVÁNÍ DAT Z DEMOGRAFICKÝCH ŠETŘENÍ

USING DATA MINING METHODS

FOR DEMOGRAPHIC SURVEY DATA PROCESSING

Diplomová práce

Ing. David Fišer

Vedoucí diplomové práce: RNDr. Luděk Šídlo, Ph.D.

Praha, 2015

**Prohlášení:**

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 30. 7. 2015 ……

**Poděkování:**

Touto cestou bych chtěl poděkovat svému školiteli RNDr. Luďku Šídlovi, Ph.D. za trpělivost, cenné rady a odborné vedení této práce.

## Abstrakt

Cílem předkládané práce bylo popsat a následně demonstrovat na modelové úloze principy procesu dolování znalostí z databází, často označovaného jako data mining (DM). V teoretické části práce jsou popsány vybrané metodiky, na základě kterých se postupuje při DM procesu a dále jsou zjednodušeně popsány principy vybraných DM technik. V druhé části práce je pak realizována DM úloha, ve které se postupuje dle metodiky CRISP-DM. Jako modelová data pro tuto úlohu jsou vybrána data z výběrového šetření American Community Survey. Praktická část práce je rozdělena na dvě části. V první části je vyhotovena klasifikační úloha, jejíž cílem je zjistit, zda lze využít vybrané DM techniky k řešení problematiky chybějících údajů ve statistických šetřeních. Úspěšnost klasifikace a následné predikce hodnot u vybraných atributů se pohybovala v intervalu 55–80 %. Druhá část praktické části práce je pak zaměřena na hledání zajímavých znalostí ve vybraných datech pomocí asociačních pravidel a metody GUHA.



**Klíčová slova:** data mining, dolování znalostí z databází, statistická šetření, chybějící hodnoty, klasifikace, asociační pravidla, metoda GUHA, ACS



## Abstract

The goal of the thesis was to describe and demonstrate principles of the process of knowledge discovery in databases - data mining (DM). In the theoretical part of the thesis, selected methods for data mining processes are described as well as basic principles of those DM techniques. In the second part of the thesis a DM task is realized in accordance to CRISP-DM methodology. Practical part of the thesis is divided into two parts and data from the survey of American Community Survey served as the basic data for the practical part of the thesis. First part contains a classification task which goal was to determinate whether the selected DM techniques can be used to solve missing data in the surveys. The success rate of classifications and following data value prediction in selected attributes was in 55–80 % range. The second part of the practical part of the thesis was then focused of determining knowledge of interest using associating rules and the GUHA method.



**Keywords:** data mining, knowledge discovery in databases, statistic surveys, missing values, classification, association rules, GUHA method, ACS

# Obsah

# Přehled použitých zkratek

| | |
|---|---|
| ACS | American Community Survery |
| BI | Bussiness Inteligence |
| CRISP-DM | Cross-Industry Standard Process for Data Mining |
| DM | Data mining |
| DZD | Dolování znalostí z databází |
| EHIS | European Health Interview Survey |
| EU-SILC | European Union - Statistics on Income and Living Conditions |
| GUHA | General Unary Hypotheses Automaton |
| IBL | Instance-based learning |
| IPUMS | Integrated Public Use Microdata Series |
| KDD | Knowledge Discovery in Databases |
| MBL | Mission-based learning |
| PAT | Predictive Analytics Today |
| SLDB | Sčítání domů lidí a bytů |
| SRÚ | Statistika rodinných účtů |
| SVM | Support vector machines |
| TDIDT | Top Down Induction of Decision Trees |
| ÚZIS | Ústav zdravotnických informací a statistiky |
| VŠPS | Výběrové šetření pracovních sil |

# Seznam tabulek

# Seznam obrázků

# Seznam grafů

## Kapitola 1

# Úvod

Pojem Data mining (DM) se začal dostávat do povědomí veřejnosti v 90. letech minulého století. První zmínky o DM, případně tzv. Dolování znalostí z databází (DZD), se datují řádově do roku 1960. V této době byly tyto pojmy diskutovány pouze v akademické sféře (BERKA, 2003). Fayyad definuje DZD jako netriviální extrakce implicitních, dříve neznámých a potenciálně užitečných informací z dat (FAYYAD et. al., 1996). Pojmy Data mining a DZD jsou často používány jako synonyma. Vymezením těchto pojmů se zabývali přední odborníci na první DZD konferenci v roce 1995 v Montrealu. Jedním ze závěrů této konference byl fakt, že DM lze chápat dvěma způsoby a oba jsou „správné". Data mining v tzv. užším pojetí lze chápat tak, že se jedná pouze o analytickou část procesu DZD. Naopak v širším pojetí by se dal DM chápat jako synonymum k procesu DZD (BERKA, 2003). V této práci bude DM dále chápán právě v „širším pojetí", tedy jako synonymum k DZD.

Poprvé se začalo mluvit o DZD na workshopech věnovaných umělé inteligenci, resp. metodám strojového učení. DZD je kombinací tří disciplín. První z nich jsou již zmiňované metody strojového učení, pomocí nichž se snaží zakomponovat za pomoci různých algoritmů a heuristik lidské myšlenkové pochody do statistických problémů. Dále jsou to databázové technologie, protože představují osvědčený způsob uchování dat a jejich následné vyhledávání. Poslední disciplínou je statistika, protože se jedná o prostředek, pomocí nějž lze modelovat a analyzovat závislosti v datech (KANTARDZIC, 2003).

Až do roku 1990 se zmíněné vědní disciplíny stojící za rozvojem DM vyvíjely nezávisle na sobě až do fáze, kdy objem automaticky sbíraných dat začal být pro jejich uživatele neúnosný. V této době se začalo mluvit o tom, že by bylo potřeba tato data využívat pro strategická rozhodnutí ve firmě. Právě zájem, aplikace a hlavně investice velkých firem byla klíčová pro rozvoj DZD. Současně se zvedla popularita DM i u odborné veřejnosti, o čemž svědčí vznik několika pravidelných konferencí na různých kontinentech (Amerika, Asie, Evropa) a v neposlední řadě vznik odborných časopisů, např. Data Mining and Knowledge Discovery vydávaný nakladatelstvím Kluwer (BERKA, 2003).

V současnosti nachází DM využití prakticky ve všech oborech lidské činnosti, např. v lékařství, výrobě, financích, pojišťovnictví, marketingu, státní veřejné správě a v mnoha dalších. V korporátní sféře se postupem času staly DM nástroje součástí tzv. Bussiness Inteligence (BI) nástrojů. BI nástroje by měly ve firmách sloužit pro získání lepšího pochopení chování trhu a obchodních souvislostí, často se o těchto nástrojích hovoří jako o systémech pro podporu rozhodování.

O DZD (resp. Knowledge Discovery in Databases – anglický ekvivalent) a jeho spojení s demografií toho v odborné literatuře zatím moc popsáno nebylo. Většinou se jedná o spojení využití DM metod a použití klasických demografických proměnných (např. věk, pohlaví, vzdělání, velikost místa bydliště atd.) při analýze vybraného jevu. Těmito jevy mohou být např. predikce odchodu zákazníka od telefonního operátora (UMAYAPARVATHI et. al. , 2012), sledování chování uživatelů v internetových obchodech (SONG et. al. , 2001) nebo analýza faktorů ovlivňujících výkony studentů na univerzitě (AFFENDEY et. al. , 2010).

Existuje i pár relativně zajímavějších studií, ve kterých se zkoumají vybrané demografické jevy. Za příklad lze uvést studii, ve které se pomocí DM technik zkoumá demografický vývoj (RODRIGUES et. al. , 1999), analýza demografických faktorů vedoucích k předčasným porodům (GOODWIN et. al. , 2001), analýza longitiduálních dat za Ženevu za rok 1800–1880 z pohledu historické demografie (RITSCHARD, 2004), odhad míry rozvodovosti na základě DM algoritmů (CARBUREANU , 2007) nebo analýza plodnosti za pomocí DM metod (GAMS et. al. , 2008). V českých končinách lze snad zmínit pouze dizertační práci zabývající se analýzou rizikových faktorů v programu asistované reprodukce pomocí systému pro DZD (HUDEČEK, 2006).

Právě skutečnost, že neexistuje mnoho zmínek o DZD v české odborné literatuře spojených s demografií, byla hlavní motivací pro vznik této práce. Během mých demografických studií jsem pochopil jednu věc, a to že se v demografii pracuje s rozsáhlými datovými soubory, nad kterými se provádí různé typy většinou statistických analýz (testování hypotéz, regresní analýzy apod.). Pro demografii existuje celá řada datových zdrojů, většinou se jedná o agregovaná data v nějakých databázích (Eurostat, Human Mortality Database, OSN apod.), datové zdroje ČSÚ (ročenky, výsledky sčítání lidu, projekce, atd.). Zdrojem dat pro tyto databáze, projekce, ročenky apod. jsou většinou data, která se získávají z různých statistických šetření, ať už se jedná o úplná šetření (Sčítání lidu), nebo určitá výběrová šetření (EU-SILC, VŠPS, EHIS). Právě tato raw (primární) data jsou ideální typ dat pro DM úlohy, proto se tato práce zabývá primárně využitím DM metod při zpracování dat ze statistických (demografických šetření). Dle mého názoru by právě demonstrace využití DM metod nad tímto typem dat mohla být zajímavá. V ideálním případě by mohla zvýšit povědomí o data miningových metodách mezi lidmi, kteří se zabývají různorodými demografickými problémy. Ti by pak mohli využít dané metody k nalezení určitých zajímavých hypotéz v rámci svých vlastních výzkumů.

## 1.1 Struktura práce

Práce je členěna na teoretickou a na praktickou část. Práce je rozdělena celkem na čtyři hlavní kapitoly. První kapitola se věnuje úvodu do tématu, cílům práce a relevantní literatuře. Následující kapitola bude věnována obecně problematice DZD. Bude provedeno základní vymezení základních pojmů, data miningu a DZD, popsány typické úlohy, které se pomocí DM metod řeší, vypsáno několik metodik postupu při DM procesu, budou popsány principy nejpoužívanějších DM technik a na závěr této kapitoly bude zmíněno několik SW nástrojů, které lze využít při DZD. Další kapitola se bude věnovat statistickým šetřením. V této kapitole budou popsány základní pojmy a rozdělení statistických šetření, samostatná podkapitola bude věnována výběrovým šetřením, dále budou popsány způsoby sběru dat, bude upozorněno na problémy při práci s chybějícími údaji a nakonec budou zmíněna vybraná statistická šetření probíhající pravidelně na českém území. Poslední kapitola této práce a jediná kapitola praktické části práce bude věnována vyhotovení modelové DM úlohy. Tato modelová úloha bude vyhotovena dle metodiky CRISP-DM, jež bude popsána v teoretické části práce.

## 1.2 Relevantní literatura vztahující se k tématu

Za 25 let, kdy je data mining ve všeobecném povědomí, byla již napsána celá řada knih o této problematice[1]. Mezi průkopníky v této oblasti lze zařadit (MICHIE et. al., 1994), který se ve své knize zabývá v té době „moderními" metodami strojového učení. Zabývá se celou řadou DM technik, např. rozhodovacími stromy, rozhodovacími pravidly, bayesovskými klasifikátory, statistickými metodami. V druhé části této knihy se snaží o klasifikaci za pomoci vybraných DM algoritmů a snaží se vybrat nejhodnější klasifikátory pro různé typy dat. O tři roky později vyšla kniha Machine Learning (MITCHELL, 1997), kde se autor zabývá pokročilejšími DZD technikami, mezi které se řadí neuronové sítě, evoluční algoritmy a techniky založené na analogii.

Na českém území se dobýváním znalostí z databází zabývá primárně (BERKA, 2003). Kniha vyšla jako skripta k předmětu 4IZ450 – Dobývání znalostí z databází na VŠE. Ve své knize se zabývá kompletně celým procesem dobývání znalostí, tedy od popisu dat, přes přípravu dat, modelování až k evaluaci výsledků. Popisuje zde, jaké úkony mají být provedeny v jednotlivých fázích. Ve své knize se také věnuje vybraným metodikám. Součástí této knihy je také reálná úloha a jsou zde popsány určité SW nástroje využívané k DZD. Zajímavostí je, že data mining se v ČR vyučuje na celé řadě univerzit, ale ve většině prezentacích, se kterými jsem se setkal, byly citovány příklady právě z této knihy.

Další kniha zabývající se DZD vyšla poměrně nedávno (RAUCH et. al., 2015). Tato kniha shrnuje 20 let práce a výzkumu v oblasti DZD. Teoreticky navazuje na (HÁJEK, 1978) a jeho metodu GUHA. Rauch v rámci svého výzkumu tuto metodu neustále rozvíjí. Jedná se o metodu, která využívá především asociační pravidla. Praktickým výsledkem jejich výzkumu je pak SW nástroj Lisp-Miner, což je akademický systém pro DZD, zaměřený jak na podporu výuky, tak výzkumu. Kniha slouží mj. jako podklad pro předmět na VŠE 4IZ460 – pokročilé přístupy k DZD. V praktické části práce bude postupováno podle metodiky CRISP-DM. O této metodice se zmiňují např. (WIRTH et. al., 2001).

Problematikou statistických šetření se v cizojazyčné literatuře zabývá celá řada autorů. Jako příklady lze uvést (GROVES et. al., 2009) a (SINGH et. al., 1996). Groves se ve své knize zabývá obecně problematikou statistických šetření, zatímco Singh se specializuje na výběrové šetření, zejména pak na volbu ideálního výběru. Z české literatury zabývající se touto problematikou lze zmínit např. (VALENTOVÁ, 2013), která se ve své knize zabývá obecně problémem statistických šetření, problémy výběrových šetření, způsoby získávání dat a také problematikou chybějících údajů. O možnostech využití metod strojového učení při práci s chybějícími daty se zmiňuje (PEJČOCH , 2011).

[1] Pro ilustraci, na Google books se jedná o 3440 knih při vyhledání spojení „knowledge discovery in databases" a 36 700 při vyhledání spojení „data mining"

## Kapitola 2

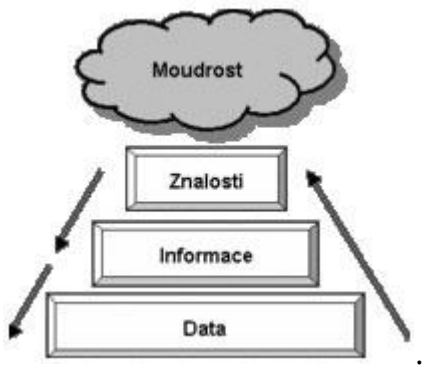
# Data mining

V této kapitole bude nejprve vymezen pojem Data mining (DM). Dále bude provedena klasifikace úloh, ke kterým lze data miningové úlohy využít. V další podkapitole budou zmíněny používané DM algoritmy a na závěr této kapitoly budou uvedeny příklady SW nástrojů, které lze využít k DM.

## 2.1 Vymezení základních pojmů

V práci se často bude pracovat s pojmy data, informace a znalosti, proto zde bude jasně vymezen rozdíl mezi danými výrazy. Bude k tomu využito pyramidy znalostí, která je na obrázku 1.

***Obr. 1: Pyramida znalostí***




***Zdroj: (BROŽOVÁ, 2007)***

Data jsou souborem faktů, měření a statistik a sama o sobě nemají vypovídající hodnotu. Data neobsahují porovnání a není možné na základě jednotlivých dat provádět rozhodnutí. Informace představují uspořádaná nebo zpracovaná data, u nichž je kontrolována jejich aktuálnost a přesnost. Znalosti obsahují informace sdělované v určitém kontextu, významné a použitelné v dané situaci. Znalosti je možné přímo použít k řešení problémů. Znalosti společně se zkušenostmi tvoří moudrost (BROŽOVÁ, 2007).

Již v úvodní kapitole byla zmíněna problematika ohledně nejasných rozdílů mezi pojmy data mining a dolování znalostí z databází. V této části kapitoly bude tento problém rozebrán podrobněji. Na úvod zde budou zmíněny vybrané definice DM a DZD.

**DZD je analýza, často obrovských souborů dat, za účelem nalezení netušených vztahů a shrnutí dat novým způsobem tak, aby byly tyto nalezené vztahy a nová shrnutí pro vlastníka užitečné a aby jim byl vlastník dat schopen porozumět.**

***Autor: (BERKA, 2003)***

**Data mining zahrnuje aplikaci vybraných analytických metod pro vyhledávání zajímavých vztahů v datech.**

***Autor: (BERKA, 2003)***

**DZD je netriviální extrakce implicitních, dříve neznámých a potenciálně užitečných informací z dat.**

***Autor: (FAYYAD et. al., 1996)***

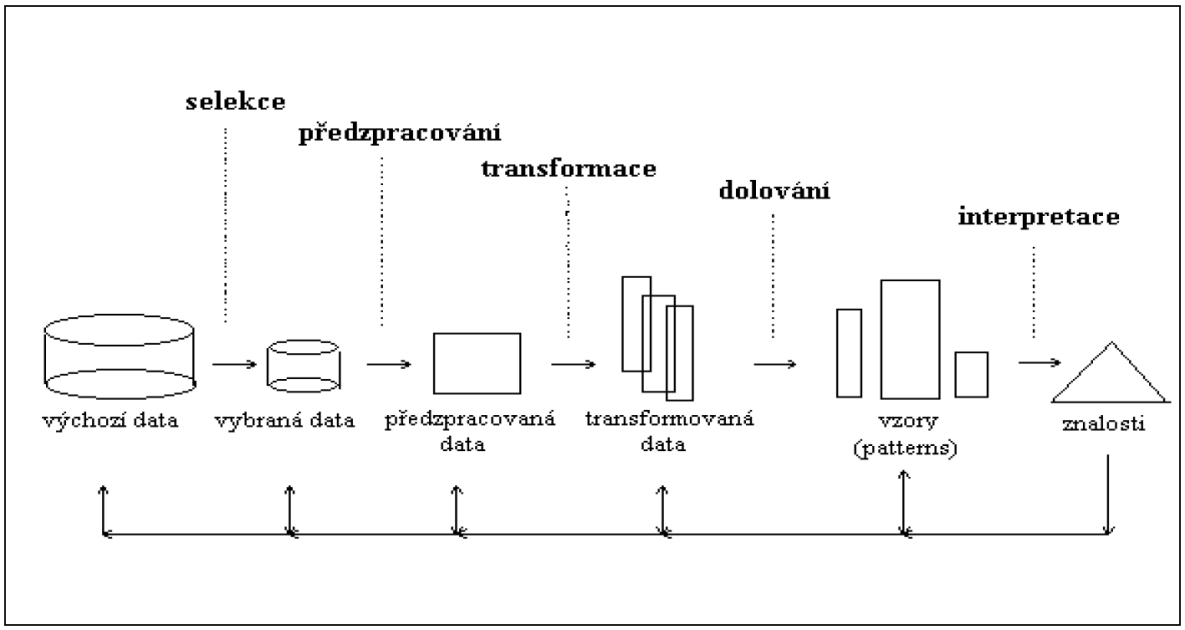
**Data Mining je proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty za účelem získání obchodní výhody.**

***Autor: Fayyad, Zdroj: (MITCHELL, 1997)***

Již z vybraných definic předního světového odborníka Fayyada a předního českého odborníka Berky si lze všimnout určitého rozporu. Zatímco Fayyadovi definice DZD a DM jsou si velmi podobné, Berka považuje za DM pouze analytickou část procesu DZD. Obecné schéma procesu DZD je znázorněno na obrázku 2 a Berka z celého procesu považuje za data mining pouze fázi „dolování“.

Jak již bylo zmíněno v úvodní kapitole, tímto problémem vymezení definice KDD a DM se zabývali také účastníci mezinárodní konference v Montrealu v roce 1995. Tato konference byla věnována výhradně KDD. Jedním z výstupů této konference byl fakt, že lze data mining chápat dvěma způsoby. A to buď v užším pojetí, tedy že DM je pouze jeden z procesů KDD, nebo v pojetí širším, které chápe data mining jako synonymum k KDD. V této práci se bude dále o pojmu data mining uvažovat právě v širším pojetí, a to zejména z důvodu, že pokud se v českém prostředí mluví o data miningu, je v tom ve většině případů spatřován právě celý proces DZD a ne pouze jeho analytická část. O tomto faktu jsem byl přesvědčen při prohledávání odborné literatury během tvorby úvodní kapitoly.

***Obr. 2: Obecné schéma procesu dolování dat z databází***




***Zdroj: (BERKA, 2003)***

## 2.2 Typy úloh

V této podkapitole bude na základě analýzy (BERKA, 2003) a (STATSOFT, 2013) sestavena tabulka 1 s klasifikací typů úloh, jež lze teoreticky řešit pomocí DM metod. U každého typu úlohy bude uveden zdroj, ve kterém o něm byla zmínka.

***Tab. 1: Typy úloh řešených pomocí data miningových metod***

| Úloha | Popis | Zdroj |
|---|---|---|
| Klasifikace | Klasifikační metody lze využít ve velkém množství oborů. Jedná se o zařazení objektů do tříd. Těmito objekty mohou být např. zákazníci, dlužníci, pacienti nebo účastníci nějakého průzkumu. Třidami pak mohou být „splatí / nesplatí“, „nemocný / zdravý“, „odpoví / neodpoví“, „registruje se / neregistruje se“. V praxi se jedná o historická data, nad nimiž je vystavěný model, který klasifikuje nové objekty. | (STATSOFT, 2013) |
| Predikce | Predikce proces určení dodatečných, případně chybějících hodnot analyzovaných záznamů (RYCHLÝ, 2008). Predikce je velmi podobná klasifikaci, nicméně zde existuje zásadní rozdíl, a to že u predikce hraje důležitou roli čas, protože se ze starších hodnot určité veličiny snaží odhadnout hodnoty budoucí (BERKA, 2003). | (BERKA, 2003), (STATSOFT, 2013) |

| | | |
|---|---|---|
| Shlukování / Segmentace | Cílem shlukování je nalézt objekty, které jsou si vzájemně podobné, případně skupiny podobných objektů (zákazníků). Je to opět velmi podobné klasifikaci, akorát v této úloze není žádná cílová proměnná. Využití lze nalézt v marketingu, např. cílování určitých reklamních kampaní pouze na určitou skupinu zákazníků, a snížit tak náklady na kampaň a optimalizovat případný zisk. | (STATSOFT, 2013) |
| Regrese | Regrese se v DM úlohách prakticky neodlišuje od statistických úloh. Regrese je „specifický typ“ predikční úlohy. Jedná se o úlohy, jež slouží pro vysvětlení, případně pro předpověď hodnot spojitých proměnných na základě dostupných informací z historických dat. Od klasifikace se regrese odlišuje zásadně v tom, že se jedná o jiný typ výsledku. V klasifikaci se odhaduje předem vybraná kategorie (třída) a v regresi je výsledek spojitá číselná hodnota. | (STATSOFT, 2013) |
| Detekce odchylek | Tato úloha slouží k objevení určitých neobvyklých jevů. V podstatě k nalezení jedince, který se nějakým způsobem odlišuje např. od nějakého standardu, případně od chování ostatních jedinců. Typickým příkladem může být odhalování podvodů. | (BERKA, 2003) (STATSOFT, 2013) |
| Analýza vztahů / Asociační pravidla | Jedná se o velmi specifické úlohy, pomocí nichž lze z velkého počtu záznamů stanovit asociační pravidla typu IF fakt THEN fakt, jako nejznámější příklad těchto úloh může posloužit analýza nákupních košíků, ve které se analyzuje zboží, které se nejčastěji v supermarketech prodává společně. Příkladem takového pravidla může být například to, že pokud je zákazník muž a koupí si pivo, tak na 70 % si koupí také párek. Na základě výsledků těchto analýz poté supermarkety kupříkladu upraví polohu regálů, případně vytváří speciální slevové nabídky, aby mohly navýšit prodeje. | (BERKA, 2003) |

***Zdroj: uveden v tabulce, vlastní zpracování***

V následující tabulce uvedu příklady reálných úloh řešených pomocí DM metod, v této tabulce bude použito kategorizace z tabulky 1. Tabulka 2 bude sestavena na základě analýzy následujících zdrojů (BERKA, 2003) a (STATSOFT, 2013).

***Tab. 2: Praktické příklady úloh řešených pomocí DM metod***

| Typ úloha | Příklady | Zdroj |
|---|---|---|
| Klasifikace / predikce | Určení příčin poruch automobilů,<br>Detekce spam emailu | (BERKA, 2003)<br>(STATSOFT, 2013) |
| Shlukování / Segmentace | Segmentace a klasifikace klientů pojišťovny (např. rozpoznání problémových nebo naopak vysoce bonitních klientů nebo zacílení určitých reklamních kampaní pouze na určitou skupinu zákazníků) | (BERKA, 2003) |
| Regrese | Predikce vývoje kursů akcií,<br>Predikce spotřeby elektrické energie | (STATSOFT, 2013) |
| Detekce odchylek | Detekce podvodů v pojišťovnictví | (STATSOFT, 2013) |
| Analýza vztahů / Asociační pravidla | Analýza nákupního košíku (Market Basket Analysis),<br>Analýza příčin poruch v telekomunikačních sítích,<br>Analýza důvodů změny poskytovatele služeb (internet, mobilní telefony),<br>Rozbor databáze pacientů v nemocnici,<br>Analýza webových stránek (typické průchody, překážky, stránky, ze kterých se nejčastěji kliká na reklamy). | (BERKA, 2003)<br>(STATSOFT, 2013) |

***Zdroj: uveden v tabulce, vlastní zpracování***

## 2.3 Metodiky

Postupem času, jak se vyvíjely DM metody a neustále se zvyšoval počet SW řešení a firem zapojujících se do vývoje, začaly se vytvářet „doporučené“ postupy, jak realizovat data miningové úlohy. Obecný postup byl znázorněn na obrázku 2 v kapitole 2.1. Stále neexistuje žádný všemi stranami uznávaný postup. Většinou si každá firma navrhne nějaké vlastní řešení, a snaží se tak sdílet a přenášet zkušenosti z jimi realizovaných projektů.

V této kapitole budou představeny tři metodiky. Byly vybrány dvě metodiky, za kterými stojí právě producenti velkých programových systémů. První metodikou je 5A, jejíž autoři jsou lidé, kteří stojí za vývojem SW nástroje SPSS. Další metodikou je SEMMA, jež je z dílny SASu. Poslední metodikou, která bude představena, je metodika, které vznikla v akademických kruzích, a je tak SW nezávislá. Jedná se o metodiku CRISP-DM. Tato metodika bude představena velmi podrobně, protože pomocí ní bude posléze vyhotovena praktická část této práce (BERKA, 2003).

### 2.3.1 5A

Jak již bylo zmíněno, s touto metodikou přišla na trh firma SPSS. byla vyvinuta řádově v roce 2000. Název 5A je akronym pro jednotlivé kroky, jichž je celkem pět.

- *Assess* – posouzení potřeb projektu,
- *Access* – shromáždění potřebných dat,
- *Analyze* – provedení analýz,
- *Akt* – přeměna znalostí na akční znalosti,
- *Automate* – převedení výsledků analýzy do praxe.

Prvním krokem procesu na základě této metodiky by tedy mělo být stanovení strategie resp. cílů, jichž je potřeba dosáhnout, stanovení analytických otázek, na které je nutné sehnat odpovědi. K tomu je nutné nejprve určit data, která jsou potřeba k provedení daných analýz. Dále je potřeba porozumět oboru, jehož se daná analýza týká, je potřeba na základě toho vybrat analytické nástroje a případně zaškolit analytiky (BERKA, 2003).

Druhým krokem by měl být samotný sběr a příprava dat. Je potřeba zajistit vhodné soubory z datových skladů či jiných podnikových informačních systémů. Autoři této metodiky také doporučují získat co nejvíce dat ohledně dané problematiky z veřejných zdrojů (oficiální statistiky, rezortní data, apod.). Třetím krokem je využití různých analytických nástrojů a postupů k dosažení odpovědí na otázky stanovených v kroku 1. Firma SPSS doporučuje využít co nejvíce metod a porovnat jejich výsledky a vhodnost tak, aby bylo získáno co nejlepší řešení (BERKA, 2003).

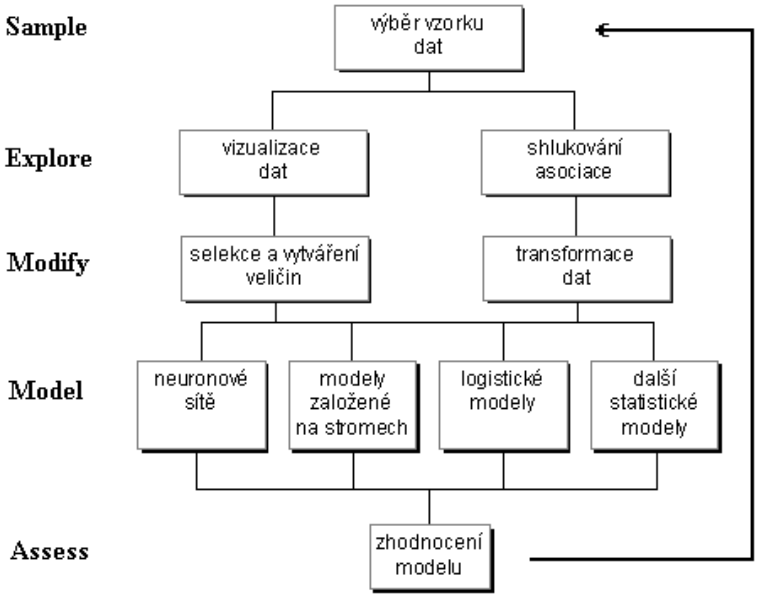
Poslední dva kroky se týkají zejména komunikace se zadavatelem dané analýzy. Ve čtvrtém kroku by mělo být snahou předložit výsledky v nějaké jednoduché a srozumitelné podobě zadavateli, který by měl na jejich základě vytvořit konkrétní závěry (převést výsledky do praxe). Součástí posledního kroku by teoreticky mělo být určité zautomatizování jednotlivých analýz tak, aby bylo možné zpětně monitorovat výsledky a také to, zda nastal v dané problematice určitý posun (BERKA, 2003).

### 2.3.2 SEMMA

Metodika SEMMA byla představena spolu se SW nástrojem Enterprise Miner, který byl vyvíjen firmou SAS. SEMMA je zkratka vytvořená z počátečních písmen jednotlivých kroků. Obecné schéma této metodiky je znázorněné na obrázku 3. Jednotlivé kroky této metodiky jsou:

- *Sample* (vybrání vhodných objektů),
- *Explore* (vizuální explorace a redukce dat),
- *Modify* (seskupování objektů a hodnot atributů, datové transformace),
- *Model* (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- *Assess* (porovnání modelů a interpretace) (BERKA, 2003).

***Obr. 3: Metodika SEMMA***




***Zdroj: (BERKA, 2003)***

Tato metodika je velmi specifická, v podstatě vytvořena na míru SW nástroje od firmy SAS. V podstatě by se dalo říct, že je do velké míry „omezená“ možnostmi tohoto SW nástroje. Dle mého názoru je velmi technicky orientovaná, chybí tam komunikace se zadavatelem a je orientovaná jenom na pouhou analýzu vzorku dat. Autoři tohoto nástroje kladou údajně velký důraz na snadnou interpretaci výsledků ve formě, která by měla být srozumitelná i laikům (příležitostným uživatelům tohoto SW nástroje).

### 2.3.3 CRISP-DM

Zkratka CRISP-DM pochází z anglického Cross-Industry Standard Process for Data Mining. Tato metodika vznikla jako výsledek Evropského výzkumného projektu, jehož cílem bylo vytvořit univerzální postup pro dobývání znalostí z databází. Na projektu spolupracovaly mj. firma NCR (přední dodavatel datových skladů), ISL (tvůrce systému Clementine, později SPSS), OHRA (holandská pojišťovna). Všechny tyto firmy měly zkušenosti s řešením rozsáhlých DM úloh (BERKA, 2003).

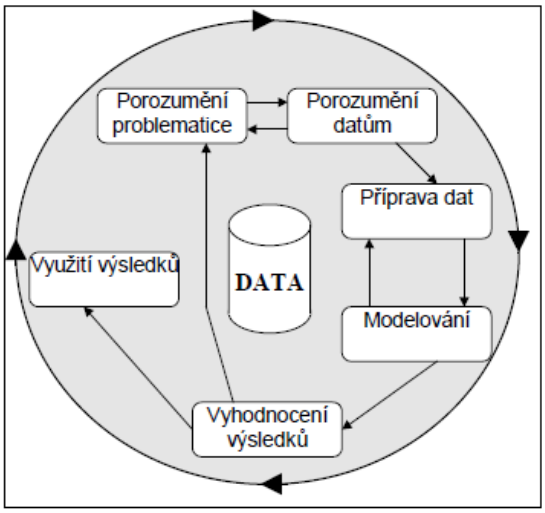
Tato metodika je i v roce 2014 dle ankety na specializovaném webu Kdnuggets nejpoužívanější. Využívá ji 42 % respondentů dané ankety. V anketě odpovědělo 200 návštěvníků tohoto specializovaného webu. Celkem 27,5 % respondentů uvedlo, že využívá svoji vlastní metodiku, což je 7% nárůst oproti roku 2007, kdy anketa proběhla poprvé. Třetí nejpoužívanější metodikou je SEMMA s 8,5 %, což je naopak pokles o 4,5% v porovnání s rokem 2007 (PIATETSKY, 2014).

Metodika CRISP-DM je složena z celkem šesti fází. Dle autorů této metodiky není přesně určeno pořadí jednotlivých fází a některé fáze lze provádět i víckrát během daného projektu. Podle dílčích výsledků je doporučeno vracet se k jednotlivým fázím a optimalizovat je. V následujícím seznamu jsou vypsány jednotlivé kroky v nejčastěji prováděném pořadí.

1. Porozumění problematice
2. Porozumění datům
3. Příprava dat
4. Modelování
5. Hodnocení výsledků
6. Implementace vytvořeného modelu

Na schématu, které je na obrázku 4, jsou pomocí šipek znázorněny zmiňované možnosti návratu v rámci jednotlivých fází. Zároveň je na tomto diagramu znázorněna nejpoužívanější a zároveň i doporučená následnost jednotlivých fází. Je zde také nastíněna cyklická povaha celého procesu, resp. na základě výsledků jednoho projektu se doporučuje založit projekt další a využít znalostí a informací z již dokončeného projektu (BERKA, 2003).

***Obr. 4: Metodika CRISP-DM***




***Zdroj: (Berka, 2003)***

V následující části kapitoly budou podrobněji rozepsány jednotlivé fáze CRISP-DM. Budou zde rozepsány jednotlivé akce, jež autoři této procedury doporučují provést.

## Porozumění problematice

Tato fáze je zaměřena primárně na pochopení základních cílů dané úlohy. Součástí by měla být komunikace se zadavatelem dané úlohy, a tak zjištění základních požadavků z „manažerského“ hlediska. Tato formulace by měla být pak následně převedena do zadání úlohy pro DZD. V této fázi by se měl stanovit i předběžný plán prací, měly by být zhodnoceny přínosy a rizika využití DM metod, případně je možné provést inventuru lidských i výpočetních zdrojů (sestavit tým analytiků atd.) (BERKA, 2003) (WIRTH et. al., 2001).

## Porozumění datům

Tato fáze začíná sběrem dostupných dat, která jsou klíčová k provedení následných analýz. Dále následují činnosti, které by měly dát základní představu o datech. Jedná se zejména o posouzení datové kvality (chybějící záznamy, špatně vyplněná pole, pole s nepřesnými údaji, atd.) nebo

vytipování zajímavých podmnožin v databázi. Často se během této fáze využívá metod deskriptivní statistiky (četnosti jednotlivých atributů, minima, maxima, průměrné hodnoty, apod.). Autoři doporučují využít nástroje, které využívají pokročilé vizualizační techniky (BERKA, 2003).

### Příprava dat

Hlavním cílem této fáze je transformovat data do formátu, který je ideální pro zpracování vybranými analytickými metodami. Cílový datový soubor by měl obsahovat pouze relevantní data k dané úloze a mít podobu vyžadovanou jednotlivými algoritmy (případně by měla být optimalizována pro vybraný SW nástroj). Tato fáze zahrnuje selekci dat, čištění dat (jedna z disciplín datové kvality), transformaci dat, integraci dat a formátování dat. Dle autorů se jedná o tu nejpracnější část z celého procesu, údajně se jedná o 80 % práce a pouze 20 % významu (v porovnání modelování zabere pouze 5 % času a má 5 % význam) (BERKA, 2003).

Často testovaný datový soubor obsahuje velké množství chybějících hodnot, tudíž je potřeba provést rozhodnutí, jak se s těmito hodnotami bude během analýzy nakládat. Zda se tyto hodnoty „domodelují“, případně zda budou během analýz úplně ignorovány. Je potřeba toto pečlivě zdokumentovat a zejména při interpretování výsledků analýz to neopomenout. Většina datových souborů nepočítá s tím, že nad nimi bude prováděn data mining, tudíž neobsahují veškeré potřebné atributy, proto je potřeba tyto atributy odvodit, případně vytvořit nové (BERKA, 2003), (WIRTH et. al., 2001).

### Modelování

V této fázi dochází k využití analytických metod (DM algoritmů viz kapitola 2.4). Většinou neexistuje pouze jeden algoritmus pro řešení dané úlohy, proto je důležité využít více různých metod a následně vybrat ty nejvhodnější a jejich výsledky kombinovat, např. na základě kvality modelu. Navíc použití některých algoritmů může vést k potřebě modifikovat data, a tedy k návratu k datovým transformací z fáze „Příprava dat“ (BERKA, 2003).

### Hodnocení výsledků

Během této fáze by opět mělo dojít ke konzultaci se zadavatelem úlohy. Ve fázi modelování byly získány výsledky, které se zdají být v pořádku z pohledu metod dobývání znalostí. Nyní je potřeba vyhodnotit, zda jsou tyto výsledky v pořádku i z pohledu zadavatele úlohy, tedy zda byly splněny cíle, které byly zformulovány během formulování úlohy ve fázi „Porozumění problematice“. Mělo by dojít k rozhodnutí, zda tento projekt ukončit jako úspěšný a přejít k poslední fázi celého procesu (využití výsledků), nebo je potřeba některé fáze projektu zopakovat, či rovnou založit nový projekt DZD. V tomto případě je potřeba zvážit rozpočtové možnosti a také možnosti lidských zdrojů (BERKA, 2003).

### Využití výsledků

Vytvořením modelů většinou projekt nekončí, a to i přestože je cílem pouhý „popis dat“. Získané znalosti je ještě potřeba transformovat do podoby vhodné pro zadavatele úlohy. Buď se jedná pouze o sepsání závěrečné zprávy, nebo o zavedení HW nebo SW systému pro klasifikaci nových případů, např. v případě klasifikace potenciálních bonitních klientů v bance. Ve většině případů totiž není analytik ten, kdo implementuje daná řešení, nýbrž zadavatel úlohy. Zejména proto je důležité, aby pochopil, co je potřeba udělat proto, aby mohly být výsledky analýz efektivně využívány v praxi (WIRTH et. al., 2001).

### 2.3.4 Porovnání vybraných metodik

Jednotlivé vybrané metodiky jsou si velmi podobné a obsahují jen nepatrné odlišnosti, případně se liší v doporučení autorů, jaké akce provádět v jednotlivých fázích. V následující tabulce budou metodiky porovnány, na stejnou úroveň (řádek) budou posazeny podobné fáze jednotlivých procesů. Tabulka je horizontálně rozdělena na tři úrovně, a to: zjištění stavu, modelování a implementace.

***Tab. 3: Porovnání vybraných procesních metodik***

| | 5A | SEMMA | CRISP-DM |
|---|---|---|---|
| Zadání projektu zjištění stavu | *Assess* – posouzení potřeb projektu | | Porozumění problematice |
| | *Access* – shromáždění potřebných dat | *Sample* (vybrání vhodných objektů) | Porozumění datům |
| | | *Explore* (vizuální explorace a redukce dat) | |
| Úprava dat a modelování | | *Modify* (seskupování objektů a hodnot atributů, datové transformace), | Příprava dat |
| | *Analyze* – provedení analýz | *Model* (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování) | Modelování |
| Evaluace výsledků implementace | *Akt* – přeměna znalostí na akční znalosti | *Assess* (porovnání modelů a interpretace) | Hodnocení výsledků |
| | *Automate* – převedení výsledků analýzy do praxe | | Implementace vytvořeného modelu |

***Zdroj: vlastní zpracování***

Obecný pohled na to, jak by měl probíhat DM proces, je dle mého názoru ustálený. Vybrané metodiky se zásadně liší pouze v pojmenování jednotlivých částí procesu, případně v tom, že se v

nějaké části procesu řeší více problémů. Z tabulky je patrné, že v rámci metodiky SEMMA není řešeno porozumění problému v širším kontextu a převedení výsledků analýz do praxe. Je to zejména proto, že se jedná o podpůrnou metodiku při využití SW nástroje od SASu. Metodika 5A je jediná metodika, která nemá jako samostatnou fázi přípravy dat a je sloučena do fáze Access, kde je provedena společně s určitou úvodní datovou analýzou.

## 2.4 DM techniky

V této kapitole budou obecně představeny nejpopulárnější DM metody. Jednotlivé techniky budou rozděleny do skupin tak, jak je ve svých publikacích uvádí (BERKA, 2003), (MITCHELL, 1997) a (MICHIE et. al., 1994). Budou vybrány ty metody, které v praxi mají úspěch a často se využívají. Ke každé metodě budou uvedeny příklady algoritmů a příklad praktického využití. Většina příkladů bude přebrána z vybrané odborné literatury, tudíž se bude jednat o příklady z bankovního sektoru. Jedná se pouze o ilustrační příklady, které mají za úkol přiblížit princip jednotlivých metod.

### 2.4.1 Rozhodovací stromy

Rozhodovací stromy jsou jednou z nejoblíbenějších DM technik. Hlavními důvody jejich oblíbenosti je zejména jejich přehlednost a snadná interpretace. Mezi nevýhody patří zejména obtížné zpracování při chybějících údajích a to, že se nebere v úvahu vzájemná korelace údajů (Mrázová, 2011).

Při tvorbě rozhodovacích stromů se nejčastěji postupuje metodou „rozděl a panuj“, tedy indukcí stromu metodou shora dolů (v praxi se používá zkratka TDIDT z anglického Top Down Induction of Decision Trees). Cílem je nalézt strom konzistentní s trénovacími daty [2]. Data se postupně rozdělují na menší podmnožiny (uzly stromu) s cílem, aby v jednotlivých podmnožinách převládaly příklady jedné třídy. Pro ilustraci rozhodovacích stromů je uveden obrázek 5 (BERKA, 2003). Na tomto obrázku je znázorněn příklad rozhodovacího stromu. Jedná se o příklad, kde se na základě dat výše příjmu, výše konta a zaměstnání rozhoduje o tom, zda bude poskytnut klientovi banky úvěr. Na tomto stromu si lze všimnout jednoho omezení, a to že rozhodující stromy pracují ve většině případů pouze s kategoriálními daty, tudíž je potřeba pro tento typ stromu data přetransformovat do této podoby.

Pro jednodušší interpretaci výsledků lze daný strom převést na jednoduchá pravidla (více o rozhodovacích pravidlech v kapitole 2.4.3) typu:

IF příjem(vysoký) THEN úvěr(ano)

IF příjem(nízký) ∧ konto(vysoké) THEN úvěr(ano)

IF příjem(nízký) ∧ konto(střední) ∧ nezaměstnaný(ano) THEN úvěr(ne)

IF příjem(nízký) ∧ konto(střední) ∧ nezaměstnaný(ne) THEN úvěr(ano)

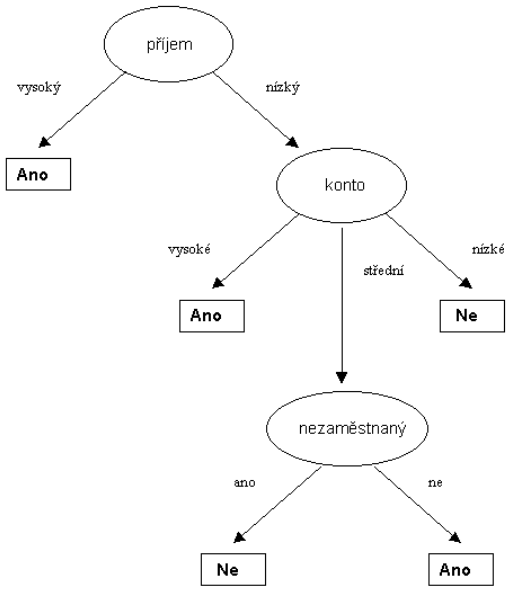
IF příjem(nízký) ∧ konto(nízké) THEN úvěr(ne)

Slovní interpretace jednoho z těchto pravidel je následující. Pokud uchazeč o úvěr má příjem z kategorie „vysoký“, tak je mu úvěr poskytnut ve všech případech. Pokud má příjem nízký, tak je mu

[2] Při klasifikačních úlohách se často rozdělí výchozí dataset na trenovací a testovací podmnožiny. Např. 40 % objektů tvoří trenovací množinu, na které se algoritmus „zaučí“, a poté se pravidla vzniklá na této množině testuje na zbylých datech. Kvalita modelu se pak hodnotí na základě úspěšnosti predikce na testovacích datech.

úvěr poskytnut pouze v případech, že má vysoký zůstatek na kontu, případně středně vysoký zůstatek na kontu a má zaměstnání.

***Obr. 5: Příklad obecného rozhodovacího stromu***




***Zdroj: (BERKA, 2003)***

**Obecný algoritmus pro konstrukci rozhodovacího stromu (algoritmus TDIDT)**

1. zvol jeden atribut jako kořen dílčího stromu,

2. rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného

atributu a přidej uzel pro každou podmnožinu,

3. existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči.

***Zdroj: (BERKA, 2003)***

Klíčovou otázkou celého algoritmu je výběr vhodného atributu pro větvení stromu. Měl by být vybrán atribut, který od sebe odliší příklady jednotlivých tříd. Vodítkem pro volbu takovéhoto atributu jsou charakteristiky převzaté z teorie informace nebo pravděpodobnosti, např. entropie, informační zisk, chí-kvadrát nebo Giniho Index. Za předpokladu, že se postupuje důsledně dle algoritmu uvedeného v předchozí části textu, pak větvení stromu skončí ve fázi, kdy všechny příklady odpovídající jednotlivým listovým uzlům patří do téže třídy. V tomto případě může dojít k tzv. přeučení neboli bezchybné klasifikaci dat na trénovací množině, což je většinou jev nežádoucí (BERKA, 2003), (MICHIE et. al., 1994).

Dalším problémem je „košatost“ stromu, tedy jeho nepřehlednost a nesrozumitelnost. Proto je potřeba strom „prořezat“. Existují dvě řešení, jak toto provést:

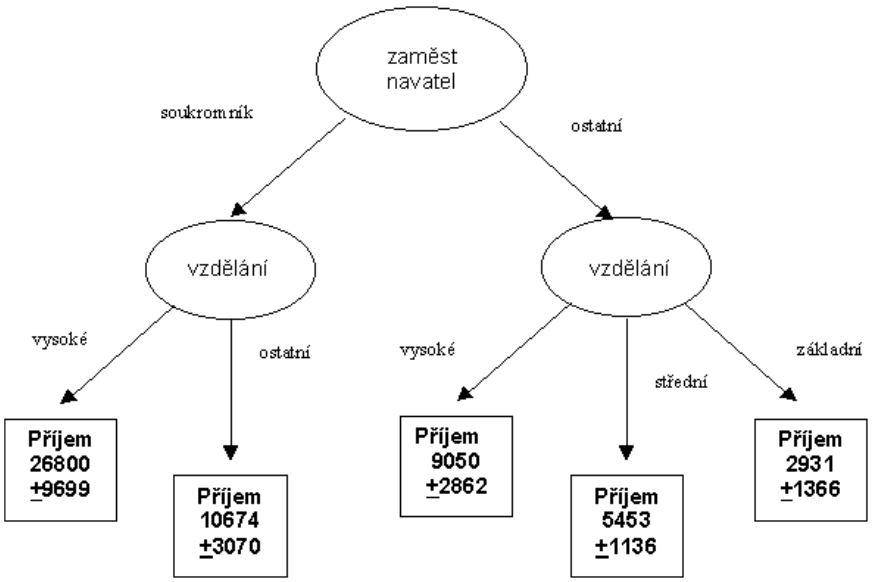
- Modifikací původního algoritmu (redukovaný strom se vytvoří přímo),
- následným prořezáním (post-pruning) úplného stromu.

V praxi se používá spíš druhá možnost, protože je velmi obtížné určit, kdy je potřeba ukončit růst stromu. Při tomto způsobu je nejprve nutné sestavit úplný strom a posléze se tento strom redukuje. Ve fázi prořezávání se pro jednotlivé nelistové uzly posuzuje, do jaké míry zhorší náhrada tohoto uzlu listem úspěšnost klasifikace. Vhodnost tohoto kroku se posuzuje buď na nových datech, tzv. validační data, nebo se posuzuje pouze na základě statistického testu z trénovacích dat (BERKA, 2003).

Rozhodovací stromy je možné využít i za předpokladu, že máme numerické hodnoty vybraných atributů. V tomto případě je nutné řešit problém s velkým počtem hodnot, protože není možné vytvářet pro každou hodnotu ve stromu zvláštní větev. Většinou se tento problém řeší ve fázi předzpracování dat vytvořením několika intervalů, nicméně některé algoritmy mívají metody diskretizace zabudované přímo v sobě. Při hledání dělících bodů se může vycházet např. z entropie (míra neuspořádanosti dat), vybírá se právě takový bod, jenž má nejnižší entropii. Toto probíhá přímo při vytváření stromu (MITCHELL, 1997).

Rozhodovací stromy lze využít i pro jiný typ úloh než pro klasifikaci. Pomoci tzv. regresních stromů lze řešit regresní typ úloh. Algoritmus pro sestavení regresního stromu odpovídá TDIDT algoritmu s jedním rozdílem a to že místo entropie se pro výběr uzlů používá směrodatná odchylka hodnot cílového atributu. Cílem těchto úloh je odhad nějakého atributu. Na obrázku 6 je příklad regresního stromu, kde bylo cílem vypočíst průměrný příjem osoby na základě atributů typu vzdělání a typu zaměstnavatele (BERKA, 2003).

***Obr. 6: Příklad regresního rozhodovacího stromu***




***Zdroj: (BERKA, 2003)***

Existují desítky různých algoritmů pro tvorbu stromů. Proto je potřeba při jeho výběru vybrat ten, který se hodí právě pro analyzovaná data. Nejznámější algoritmy pro tvorbu rozhodovacích stromů jsou ID3, C4.5, CART, CHAID a SPRINT (MRÁZOVÁ, 2011).

### 2.4.2 Asociační pravidla

Spolu s rozhodovacími stromy se řadí k nejčastěji používaným DM technikám. Termín asociační pravidla zpopularizoval v 90. letech Agrawal analýzou nákupních košíků. Jde o hledání

vzájemných vazeb mezi různými atributy, kde není dopředu upřednostňován žádný druh atributu jako závěr pravidla. Jedná se o použití IF – THEN pravidel. Ve většině případů se používají tato pravidla pro klasifikační typ úloh. (Berka, 2003)

U pravidel, která jsou vytvořena z dat, jsou zajímavá zejména ta pravidla, která splňují buď předpoklad nebo závěr, předpoklad i závěr současně a nebo splňují předpoklad, ale nesplňují závěr. Obecné pravidlo vypadá takto:

***Ant ⇒ Suc***

Ant (Antecedent, často se také používá znak „φ“) je předpoklad a Suc (Succedent, často se používá znak „ψ“) je závěr. Předpoklad i závěr mohou být kombinace konjunkcí několika kategoriálních atributů. Příklad pravidla z úlohy analýzy nákupního košíku je:

***{párky, hořčice} ⇒ {rohlíky}***

Tedy zda za předpokladu, že někdo nakoupí párek s hořčicí, pak k tomu koupí i rohlíky. Pro interpretaci těchto pravidel lze využít čtyřpolní tabulku, její podoba pro n příkladů je znázorněna na obrázku 7.

***Obr. 7: Čtyřpolní tabulka***

| $\mathcal{M}$ | $\psi$ | $\neg\psi$ |
|---|---|---|
| $\varphi$ | $a$ | $b$ |
| $\neg\varphi$ | $c$ | $d$ |

***Vysvětlivky:***

$n(\varphi \wedge \psi)$ = **a - počet objektů pokrytých současně předpokladem i závěrem,**

$n(\varphi \wedge \neg\psi)$ = **b - počet objektů pokrytých předpokladem a nepokrytých závěrem,**

$n(\neg\varphi \wedge \psi)$ = **c - počet příkladů nepokrytých předpokladem, ale pokrytých závěrem,**

$n(\neg\varphi \wedge \neg\psi)$ = **d - počet příkladů nepokrytých ani předpokladem, ani závěrem.**

***Zdroj: (RAUCH et. al., 2015)***

Na základě této tabulky lze vypočítat základní charakteristiky a kvalitu jednotlivých nalezených pravidel. Dvěma základními ukazateli jsou podpora a spolehlivost, které lze vypočítat pomocí následujících vzorců. V případě podpory se jedná o podíl objektů splňujících předpoklad i závěr a spolehlivost je podmíněná závislost závěru, pokud platí předpoklad. (Rauch, a další, 2015)

***Vzorec podpory:*** $\boldsymbol{P}(\varphi \wedge \psi) = \frac{a}{a+b+c+d}$

***Vzorec spolehlivosti:*** $\boldsymbol{P(\varphi|\psi) = \frac{a}{a+b}}$

## Generování kombinací

Základem všech algoritmů pro vyhledávání asociačních pravidel je generování kombinací (konjunkcí) hodnot atributů. Při generování se prakticky prochází prostor všech přípustných hodnot. Existují tři možnosti, jak daný prostor prohledávat, do hloubky, do šířky a heuristicky. Při prvních dvou způsobech se prohledá „slepě" celý prostor, tudíž se generují i kombinace, které se v datech nevyskytují. Bere se v potaz pouze seznam atributů a vůbec se nezohledňují testovaná data. V případě heuristického prohledávání lze nastavit určitá omezení během vyhledávání, např. berou se v potaz kombinace, které mají v datech četnost vyšší než daná hodnota (BERKA, 2003).

Samotné generování kombinací je výpočetně náročný proces. Lze si představit, že při velkém počtu kombinací jednotlivých atributů může tento proces zabrat i několik hodin nebo dní. Počet generovaných kombinací lze omezit nastavením vstupním parametrů, např. omezit délku kombinace, případně nastavit minimální četnost pro kombinaci. Nicméně lze prohlásit, že počet generovaných kombinací je exponenciálně závislý na počtu atributů (MRÁZOVÁ, 2010).

## Algoritmus Apriori

Mezi vůbec nejznámější algoritmus lze zařadit algoritmus Apriori. Tento algoritmus navrhnul v roce 1996 R. Arghawal v souvislosti s analýzou nákupních košíků. Algoritmus je založen na hledání často opakujících množin položek. Jedná se o kombinace kategorií, jež dosahují předem zadané četnosti v datech. Apriori využívá generování kombinací do šířky s jedním rozšířením. Pro danou kombinaci délky je požadováno, aby všechny podkombinace délky k-1 měli danou minimální četnost. Tedy např. ze tříčlenných kombinací $\{A_1A_2A_3, A_1A_2A_4, A_1A_3A_4, A_1A_3A_5, A_2A_3A_4\}$ dosahujících požadované četnosti vytvoříme pouze jedinou čtyřčlennou kombinaci $\{A_1A_2A_3A_4\}$. Existuje i rozšíření tohoto algoritmu, ve kterém lze využít taxonomie, a je tak možné hledat určitá zobecněná pravidla, protože nás nezajímají pouze pravidla na nejnižší možné úrovni hierarchie. Díky tomuto vylepšení tak lze např. pravidlo {lahůdkový párek ⇒ kremžská hořčice} upravit na zobecnění pravidlo s využitím taxonomie{párek ⇒ hořčice}. Každopádně je potřeba upravit vstupní data a tyto vztahy tam zaznamenat (BERKA, 2003), (MRÁZOVÁ, 2010).

## Metoda GUHA

S dalším poměrně zajímavým konceptem asociačních pravidel přišla česká skupina vědců kolem Petra Hájka zhruba v 60. letech minulého století. Základní myšlenkou metody, již nazvali GUHA (General Unary Hypotheses Automaton), bylo nalézt v datech veškeré zajímavé souvislosti a nabídnout je uživateli. V době vzniku této metody se ještě v podstatě nic netušilo o DZD, a tak se tato metoda původně řadila k explorativním statistickým metodám. Na rozdíl od konfirmační statistické analýzy, kdy je cílem ověřit statistické hypotézy. Cílem metody GUHA nebylo pouze ověřovat tyto hypotézy, ale také je generovat. Postupem času bylo vytvořeno několik typů hypotéz a s tím souvisejících typů algoritmů pro jejich generování. Jako příklad zde uvedu proceduru 4FT – Miner, jež vznikla v rámci výzkumu na VŠE a navazuje na původní proceduru GUHA, jež se jmenovala ASSOC (BERKA, 2003), (RAUCH et. al., 2015). Hypotézy generované a testované procedurou 4FT - Miner mají podobu:

***Ant ~ Suc / Cond,***

Kde Ant (antecedent), Suc (sukcedent) a Cond (podmínka) jsou konjunkce literálů[3] a symbol „~“ značí zobecněný kvantifikátor charakterizující typ vztahu mezi Ant a Suc na podmatici objektů, které splňují podmínku Cond. Díky této koncepci lze hledat i složitější pravidla než prostá implikace, jako tomu bylo v případě algoritmu Apriori. Procedura 4FT – Miner obsahuje 14 různých kvantifikátorů. Příklady základních kvantifikátorů včetně jejich interpretace je pro ilustraci uvedena v tabulce 4 (RAUCH et. al., 2015).

***Tab. 4: Základní 4FT - Miner kvantifikátory***

| Název | Obecný zápis | Interpretace |
|---|---|---|
| Fundovaná implikace (FUI) | $\varphi \Rightarrow_{p,Base} \psi$ $\frac{a}{a+b} \geq p \wedge a \geq Base$ | Zajímá nás, zda platnost nějaké kombinace φ znamená s vysokou pravděpodobností i platnost nějaké jiné kombinace ψ (RAUCH et. al., 2015). |
| Dvojitá fundovaná implikace (DFUI) | $\varphi \Leftrightarrow$p,Base $\psi$ $\frac{a}{a+b+c} \geq p \wedge$ $\wedge a \geq Base$ | Zajímá nás, zda pro nějaké dvě kombinace platí, že když je splněna alespoň jedna z nich, tak jsou velmi často splněny obě dvě (RAUCH et. al., 2015). |
| Fundovaná ekvivalence (FUE) | $\varphi \equiv_{p,Base} \psi$ $\frac{a+d}{a+b+c+d} \geq p \wedge$ $\wedge a \geq Base$ | Zajímá nás, zda platnost nějaké kombinace je téměř ekvivalentní platnosti nějaké jiné kombinace (RAUCH et. al., 2015). |
| AA kvantifikátor | $\varphi \Rightarrow^{+}_{p,Base} \psi$ $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge$ $\wedge a \geq Base$ | Zajímá nás, zda platnost nějaké kombinace znamená výrazné zvýšení relativní četnosti nějaké jiné kombinace (RAUCH et. al., 2015). |

***Vysvětlení vzorců: parametry „a“, „b“, „c“, „d“ jsou hodnoty ze čtyřpolní tabulky (obrázek 7).***

***P a Base jsou dva parametry pomocí níž lze omezit počet nalezených hypotéz***

***Zdroj: (Rauch, a další, 2015), vlastní úpravy***

Příklady výsledných pravidel, které lze nalézt pomocí asociačních pravidel za pomocí metody GUHA z (BERKA, 2003):

[3] Základním stavebním kamenem pro konstrukci hypotéz je takzvaný *literál* (pozitivní nebo negativní), definovaný jako *atribut (koeficient) v* případě *pozitivního literálu* resp. jako ¬*atribut(koeficient) v* případě negativního literálu (Rauch, a –další, 2015).

***Konto (nízké) ∧ Nezaměstnaný (ano) ⇒ Pohlaví (žena)***
***Konto (nízké) ∧ Nezaměstnaný (ano) ∧ Pohlaví (žena) ∧ Příjem (nízký) ⇒ Úvěr (ne)***
***Konto (nízké) ∧ Nezaměstnaný (ano) ∧ Pohlaví (žena) ∧ Příjem (vysoký) ⇒ Úvěr (ano)***

### 2.4.3 Rozhodovací pravidla

Rozhodovací pravidla slouží primárně pro klasifikaci případů do tříd, na rozdíl od asociačních pravidel. Syntaxe rozhodovacího pravidla je následující:

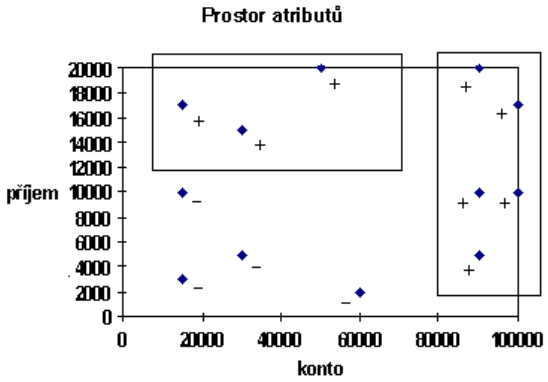
***IF Ant THEN Class***

„Ant" je kombinace vytvořená z kategorií vstupních atributů a „Class" je zařazení příkladu do třídy. Rozhodovací pravidla se díky jednoduchosti a srozumitelnosti využívají pro interpretaci výsledků jiných DM technik, viz příklad v kapitole 2.4.1 o rozhodovacích stromech (BERKA, 2003).

#### Pokrývání množin

Pokrývání množin je jeden z algoritmů, jenž se používá pro tvorbu rozhodovacích pravidel. Na rozdíl od TDIDT algoritmu, který se využívá pro tvorbu rozhodovacích stromů (princip rozděl a panuj), funguje algoritmus pokrývání množin na principu „seperate and conquer" neboli odděl a panuj. Při pokrývání množin je totiž hlavní cíl nalézt taková pravidla, která pokrývají příklady určitého konceptu, a tyto příklady pak oddělit od jiných příkladů téhož konceptu a od příkladů jiných tříd. V prostoru atributů se tak vybírají pouze oblasti, které obsahují příklady pouze jedné třídy. Nalezené oblasti mají v prostoru většinou podobu mnohorozměrných hranolů rovnoběžných s osami (BERKA, 2003).

***Obr. 8: Příklad rozdělení prostoru pomocí algoritmu pokrývání množin***




***Zdroj: (BERKA, 2003)***

Na obrázku 8 je uveden příklad rozdělení prostoru pomocí tohoto algoritmu. Jedná se opět o příklad, zda banka poskytne žadateli úvěr. Na tomto obrázku se berou pro jednodušší znázornění pouze dvě proměnné, a to výše příjmu a aktuální stav konta. Pokud byl žadateli poskytnut úvěr, je uveden

symbol „+“. Lze si všimnout, že byla nalezena dvě pravidla, která na tomto jednoduchém příkladu pokrývají všechny pozitivní příklady (úvěr byl poskytnut). Výsledná pravidla tak mohou mít takovou podobu:

***IF konto (vysoké[80000+]) THEN úvěr (ano)***
***IF příjem (vysoký[12000+]) THEN úvěr (ano)***

Využití těchto pravidel při následné klasifikaci nových příkladů je velmi jednoduché. Postupně se prochází soubor pravidel nalezených ve fázi učení, až se nalezne takové pravidlo, které lze využít. Závěr pravidla pak určí třídu, do které lze uvažovaný příklad zařadit. Obecný algoritmus pokrývání množin obsahuje tři následující kroky (BERKA, 2003).

***Algoritmus pokrývání množin:***
*1. najdi pravidlo, které pokrývá nějaké pozitivní příklady a žádný negativní*
*2. odstraň pokryté příklady z trénovací množiny DTR*
*3. pokud v DTR zbývají nějaké nepokryté pozitivní příklady, vrať se k bodu 1, jinak skonči*

***Zdroj: (BERKA, 2003)***

Klíčový bod tohoto algoritmu je bod č. 1, tedy nalezení pravidla. Existují právě dva postupy, jak tato pravidla nalézt. Buď se lze pohybovat zdola nahoru (algoritmy AQ a FindS), nebo shora dolů (algoritmy CN2 a CN4). Při pohybu zdola nahoru se jedná o metodu generalizace neboli o odstraňování kategorií z kombinace. Počáteční hypotéza v tomto postupu obsahuje jeden příklad a postupně se zobecňuje tak, aby pokrývala co nejvíce případů a žádný negativní. Naopak při pohybu shora dolů se jedná o metodu specializace, tedy přidávání kategorií do kombinace, na tomto principu funguje mj. i algoritmus pro tvorbu rozhodovacích stromů TDIDT (BERKA, 2003).

## Rozhodovací seznamy

Pomocí jednoduché úpravy lze tvořit i tzv. rozhodovací seznamy. Rozhodovací seznamy se často nazývají jako „uspořádaný“ soubor pravidel. Jedná se o strukturu typu:

***IF Ant1 THEN Classi,***
***ELSE IF Ant2 THEN Classj***
***ELSE IF Ant3 THEN Classk***

Kde se v závěrech „THEN“ mohou objevovat různé třídy. V každé podmínce „ELSE IF“ se implicitně skrývá negace všech podmínek předcházejících pravidel. Díky tomu nelze pravidla chápat jako vzájemně nezávislá. Tyto systémy jsou použity v algoritmech CN2 a CN4. Principem je nalézt takové první pravidlo, jež pokrývá velký počet objektů třídy „Class“ a malý počet objektů dalších tříd. Tvorba pravidel končí, když už se nepodaří nalézt žádné další vyhovující pravidlo (BERKA, 2003).

Rozhodovací pravidla jsou primárně určena pro práci s kategoriálními daty. Je tedy pro práci s těmito algoritmy data potřeba diskretizovat. Nabízejí se dvě možnosti. Buď data manuálně upravit ve fázi předzpracování, tedy po konzultaci s odborníkem sestavit nějaké intervaly. Druhou možností je diskretizace „automatická“ během běhu algoritmu. Tato diskretizace probíhá většinou dynamicky během tvorby pravidel.

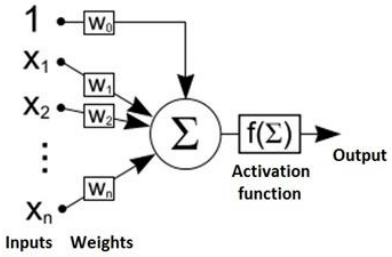
### 2.4.4 Neuronové sítě

Počítačové neuronové sítě fungují na podobném principu jako neuronové sítě v přírodě. Podobně jako mají neurony několik dendritů, jimiž přijímají signál a pouze jeden axon, jímž signál posílají dál. Na obrázku 9 je znázorněno schéma obecného počítačového neuronu. Činnost neuronu lze popsat matematicky a to takto:

***Zachycení signálu a vedení dovnitř neuronu, kde vznikne potenciál P:***
***P = w1*x1 + w2*x2 + ... + wn*xn***
***Jestliže je potenciál dostatečně velký, neuron vyšle signál y:***
***y = 1, jestliže P > w0, jinak y = 0. (w0 je nějaká prahová hodnota)***

***Zdroj: (CHALUPNÍK, 2012)***

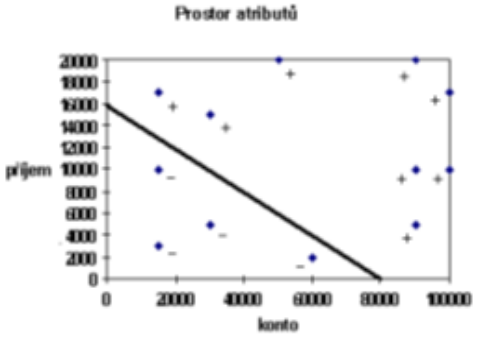
***Obr. 9: Obecný počítačový neuron***




***Zdroj: (CHALUPNIK, 2012)***

Vstupem do neuronu jsou tedy nějaké numerické hodnoty $x_1$ – $x_N$, které jsou násobeny příslušnými váhami $w_1$ – $w_N$. V případě, že tento součin přesáhne určitou prahovou hodnotu (w0), je výstup z neuronu roven 1 (v praxi, při klasifikační úloze to tak může znamenat „poskytne se žadateli" půjčka v bance), pokud tuto hodnotu nepřesáhne, tak pak je výstup roven 0 (půjčka se neposkytne žadateli). Toto byl příklad tzv. adaptivního lineárního modelu, kde jsou výstupní hodnoty pouze 0 nebo 1. Lze tak mluvit o lineárním klasifikátoru do dvou tříd (poskytne / neposkytne úvěr). Takovýto příklad je znázorněn na obrázku 10. Často se provádí normalizace hodnot jednotlivých proměnných vstupujících do neuronu na interval [-1,1]. Díky této normalizaci je poté jednodušší interpretace jednotlivých vah. Důležitou vlastností neuronů je jejich schopnost učit se. V praxi to znamená, že existují algoritmy pro nastavení vah, na základě předložených příkladů, aby systém následně co nejpřesněji klasifikoval případy nové (BERKA, 2003).

***Obr. 10: Praktický příklad rozdělení prostoru pomocí počítačového neuronu***

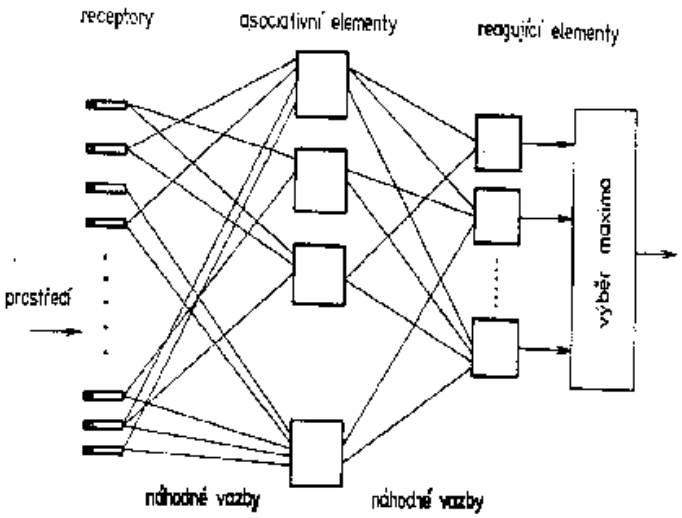


***Zdroj: (BERKA, 2003)***

## Perceptron

V předchozí části byl popsán princip modelu, jenž využívá pouze jednoho neuronu. Jako první přišel s modelem neuronové sítě Rosenblatt již v roce 1957, tento model pojmenoval perceptron a vychází z modelu zrakové soustavy. Model perceptronu je znázorněn na obrázku 11.

***Obr. 11: Model perceptronu***




***Zdroj: (Berka, 2003)***

Perceptron je složen ze tři vrstev. První vrstva je tvořena receptory, jež přijímají informace z prostředí. Výstupy z receptorů jsou poté následně převedeny do asociačních elementů (tyto elementy fungují na podobné bázi jako lineární adaptivní neuron popsaný výše), těchto elementů je řádově desítky tisíc, záleží na počtu kombinací jednotlivých atributů. Výstupy z těchto elementů jsou potom náhodně propojeny na reagující elementy, jejichž počet odpovídá počtu tříd, do kterých se provádí klasifikace (MICHIE et. al., 1994).

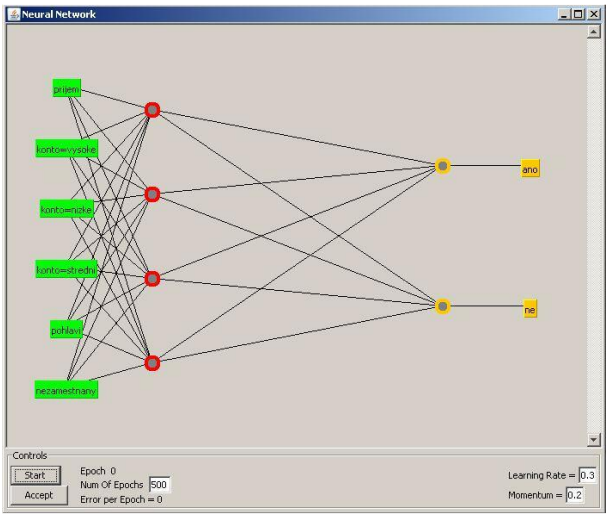
Nicméně tento model se v 70. letech stal terčem kritiky mnoha expertů, protože se nedokázal vypořádat s tak jednoduchým pojmem, jako je nonekvivalence. Většina expertů to v tehdejší době pochopila jako kritiku celého principu neuronových sítí, a tak se jejich výzkum a zdokonalování výrazně zpomalil. V 80. letech se podařilo vyvrátit nemožnost nonekvivalence v perceptronech. Dále se podařilo v této době sestavit algoritmy na možnost učení v neuronových sítích, což stálo za jejich návratem na výsluní (BERKA, 2003).

V současné době je používána celá řada neuronových sítí. Dvě nejpoužívanější jsou vrstevná síť (někdy nazývaná jako vícevrství perceptron) a Kohonova mapa. Vrstevná síť je zjednodušení obecného perceptronu. Obvykle obsahuje několik vrstev neuronů, které nejsou v rámci vrstvy propojeny, ale každý neuron z jedné vrstvy je propojen se všemi neurony z vrstvy sousední (BERKA, 2003).

### Využití v praxi

V praxi se používají neuronové sítě pro klasifikační a predikční typy úloh. Jsou vhodnou alternativou k rozhodovacím stromům a pravidlům zejména v případech, kdy se netrvá na srozumitelnosti nalezených výsledků. Výsledky nelze totiž nelze snadno interpretovat pomocí sady nějakých pravidel, jako tomu bylo u předchozích metod. Velkou výhodou neuronových sítí je zejména to, že jsou koncipovány na práci s numerickými atributy, na rozdíl od rozhodovacích stromů a pravidel, které jsou určeny primárně na práci s kategoriálními daty. Práce s kategoriálními daty je naopak problém v neuronových sítích, nicméně ho lze obejít použitím tzv. binarizace[4] (BERKA, 2003).

***Obr. 12: Příklad implementace neuronové sítě v SW nástroji Weka***




***Zdroj: (BERKA, 2003)***

Na obrázku 12 je příklad implementace neuronové sítě v SW nástroji Weka. Lze si všimnout tří vrstev, kde v první jsou vstupní parametry, pak je zde skrytá vrstva a vrstva výstupní, která má počet neuronů nestejný jako počet klasifikačních tříd.

## 2.4.5 Evoluční algoritmy

Evoluční algoritmy vycházející z biologických principů. Hlavní inspirací se staly mechanismy evoluce, zejména pak Darwinův přirozený výběr. Živočišný druh se během vývoje zdokonaluje tak, že se z generace na generaci přenáší pouze genetická informace těch nejsilnějších jedinců. Příklady evolučních metod jsou: genetické algoritmy, evoluční programování, evoluční strategie a genetické programování. Všechny tyto metody mají společné některé základní pojmy: výběr (selekce), mutace a reprodukce jedinců z určité populace (BERKA, 2003).

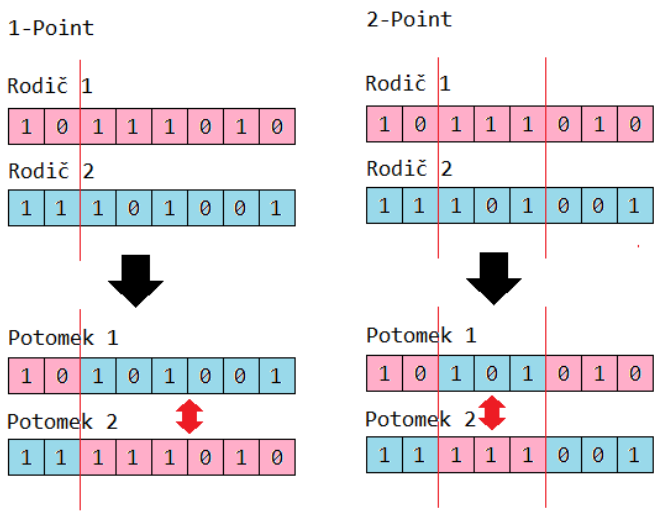
V případě algoritmické realizace je podstatný zejména způsob reprezentování jedinců. V nejjednodušším případě se může jednat o řetězec symbolů složených pouze z 0 a 1 („genů“). Pro

[4] Pro jeden kategoriální atribut vytvoříme tolik nových binárních atributů, kolik měl původní atribut různých hodnot. Hodnota 1 u nového atributu $A'_k$ vyjadřuje, že původní atribut $A$ měl pro daný objekt hodnotu $v_k$.; ostatní atributy $A'q$ , $q \neq k$ pak mají hodnotu 0. (Berka, 2003)

zjednodušení lze předpokládat, že všechny řetězce budou mít stejnou délku. Tito jedinci pak představují nějaké potenciální řešení dané úlohy, např. jedinci pokrývající příklady konceptů při klasifikační úloze. Hodnocení kvality určitého jedince v dané populaci lze pak vyjádřit pomocí nějaké funkce (např. fitness function), jíž může být třeba přesnost jedince při klasifikaci (CHALUPNIK, 2012).

Základní podoba evolučních algoritmů je taková, že algoritmus začíná pracovat s určitou výchozí podobou jedince, která je naprosto náhodná. A tuto podobu postupně zdokonaluje na základě hodnoty fit funkce. Lze využít tří operátorů selekce, křížení a mutace. Selekce se provádí právě na základě hodnoty fit funkce (typy selekce, ruletové kolo, pořadová selekce, turnajová selekce). Křížení prakticky vytváří ze dvou rodičů dva potomky, kde každý potomek zdědí část chromozomu rodiče (znázorněno na obrázku 13). Mutace je pouze drobná úprava chromozomu, většinou změna 1 bitu. Algoritmus poté skončí za předpokladu, že dosáhne nějaké předem stanovené maximální hodnoty funkce fit nebo pokud se vyčerpá maximální počet možných kombinací. Problémem těchto algoritmů je tendence nalézt lokální msto globálního optima. V podstatě to znamená, že nejlepí jedinec se reprodukuje nejvíce, což vede k málo diverzifikované populaci. Toto lze vyřešit pomocí modifikace fit funkce (BERKA, 2003), (CHALUPNIK, 2012).

***Obr. 13: Příklad křížení v genetickém algoritmu***




***Zdroj: (CHALUPNIK, 2012)***

Genetické algoritmy nalezly uplatnění v řadě oblastí, např. numerická optimalizace a rozvrhování, strojové učení, tvorba ekonomických, populačních a sociálních modelů. Z hlediska DM technik se jedná o velmi specifické využití, primárně se jedná o unikátní učení se novým konceptům, protože se provádí paralelní náhodné prohledávání prostoru hypotéz. Tyto algoritmy se dají využít i pro optimalizaci neuronových sítí. Pro snazší pochopení této DM techniky bude uveden příklad. Pro interpretaci lze opět využít hypotézy v podobě nějakých pravidel, např.:

***IF konto (nízké) ∧ příjem (nízký) THEN úvěr(ne) - interpretace genetického zápisu 100 10 01***

V praxi to znamená, že počet genů pro každý atribut je roven počtu hodnot, jakých může nabývat. Díky tomu lze interpretovat i disjunkce, tudíž pokud je zapotřebí vyjádřit, že atribut nabývá

druhé nebo třetí hodnoty, tak pak je zápis „011“, případně pokud na hodnotě atributu nezaleží, tak pak je to „111“ (BERKA, 2003).

### 2.4.6 Bayesovská klasifikace

Metody bayesovské klasifikace vychází z Bayesovské věty o podmíněných pravděpodobnostech. I přesto, že se jedná především o metody pravděpodobnostní, tak jsou v poslední době studovány v souvislosti se strojovým učením a uplatňují se i v data miningových systémech. Bayesova věta má tuto podobu:

$$P(G \mid T) = \frac{P(T \mid G) \times P(G)}{P(T)}$$

P(G|T) je podmíněná pravděpodobnost jevu G za předpokladu, že nastal jev T. Pro snazší pochopení Bayesovy věty bude uveden ilustrační příklad.

***Zadaní: Předpokládejme, že máme školu s 60 % chlapců a 40 % dívek. Všichni chlapci nosí kalhoty. Z dívek nosí kalhoty polovina. Pozorovatel vidí z dálky studenta v kalhotách. Jaká je pravděpodobnost, že tento student je dívka?***

Jako událost G označíme, že pozorovaný student je dívka. Jako událost T, že pozorovaný student nosí kalhoty. Pro výpočet podmíněné pravděpodobnosti P (G|T), že student v kalhotách je dívka, potřebujeme vědět:

P (T|G), neboli pravděpodobnost, že náhodně vybraná dívka nosí kalhoty. Vzhledem k tomu, že dívky nosí kalhoty a sukně stejně často, je tato pravděpodobnost 0,5.

P (G), neboli pravděpodobnost, že student je dívka. Vzhledem k tomu, že pozorovatel vidí náhodného studenta a zastoupení dívek je 40 %, je tato pravděpodobnost 0,4.

P (T), pravděpodobnost, že náhodně vybraný student bude nosit kalhoty. Víme přitom, že polovina dívek a všichni chlapci nosí kalhoty, takže to bude $0,4 \times 0,5 + 0,6 \times 1,0 = 0,8$.

Pak můžeme použit Bayesův vzorec a dostáváme:

$$P(G \mid T) = \frac{P(T \mid G) \times P(G)}{P(T)} = \frac{0,5 \times 0,4}{0.8} = 0.25$$

***Výsledkem tedy je, že pravděpodobnost, že pozorovatel vidí dívku v kalhotách, je 25 %.***

***Zdroj: (WIKISKRIPTA, 2014)***

#### Naivní Bayesovský klasifikátor

Naivní bayesovský klasifikátor je jeden z nejčastěji používaných algoritmů z této třídy DM technik. Vychází ze zjednodušujícího předpokladu, že jednotlivé evidence $T_1,\ldots T_k$ jsou podmíněně nezávislé při platnosti hypotézy G. Díky tomuto předpoklad pak lze spočítat pravděpodobnosti při platnosti všech evidencí. Obecný zápis tohoto klasifikátoru:

$$P(G \mid T1,\ldots,TK) = \frac{P(T1,\ldots,TK \mid G) \times P(G)}{P(T1,\ldots,TK)}$$

V případě klasifikace pomocí „Naive bayes“ se bude hledat hypotéza s největší aposteriozní pravděpodobností. Pro použití tohoto algoritmu je potřeba znát hodnoty P(G) a $P(G|T_1,\ldots,T_K)$. V kontextu DM lze tyto hodnoty získat během fáze učení. Evidence $T_k$ jsou hodnoty jednotlivých vstupních atributů a hypotézy G jsou pak jednotlivé cílové třídy. Na rozdíl od rozhodovacích stromů a rozhodovacích pravidel není třeba prohledávat celý prostor hypotéz, stačí pouze spočítat příslušné pravděpodobnosti na základě četnosti výskytů hodnot jednotlivých atributů (MICHIE et. al., 1994).

Přestože předpoklad podmíněné nezávislosti bývá splněn pouze málokdy, vykazuje Naivní bayesovský klasifikátor překvapivě dobré výsledky při klasifikaci, díky tomu je poměrně oblíbený. Další výhodou je velmi snadné aplikace, mírnou nevýhodou může být nižší srozumitelnost znalostí pomocí pravděpodobnosti (BERKA, 2003).

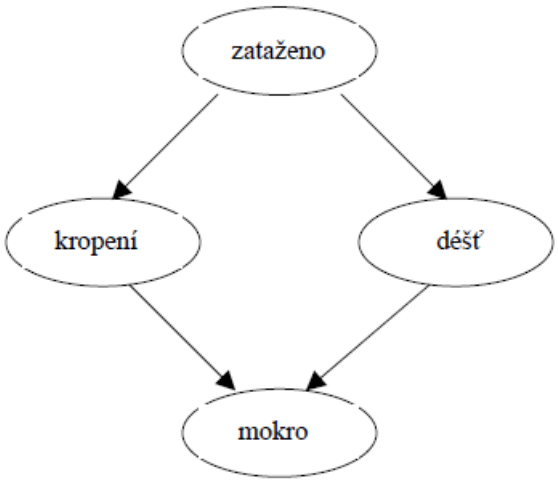
## Bayesovské sítě

Problém naivního bayesovského klasifikátoru s vzájemnou nezávislostí jednotlivých atributů lze řešit pomocí bayesovských sítí. Bayesovská síť je acyklický orientovaný graf, který zachycuje podmíněné závislosti mezi náhodnými veličinami pomocí hran. Základní pojem, jejž je třeba znát pro pochopení principu baeysovských sítí, je podmíněná nezávislost. Tento pojem lze zapsat takto:

***P(A,B|C) = P(A|C) P(B|C) a ekvivalentní vztahy P(A|B,C) = P(A|C) P(B|A,C) = P(B|C)***
***Případně: A | B|C – podmíněná nezávislost A a B při daném C***

Za předpokladu, že se sestaví a očísluje takový graf, že jsou všichni „rodiče“ před svými dětmi (mají nižší pořadové číslo), pak platí, že je podmíněně nezávislý na všech uzlech s nižším pořadovým číslem s výjimkou jeho rodičů. Toto lze vysvětlit na příkladu jednoduché bayesovské sítě na obrázku 14, mokro je podmíněně nezávislé na tom, zda bylo zataženo, ale závislé pouze na tom, zda se kropilo nebo byl déšť (BERKA, 2003).

***Obr. 14: Příklad bayesovské sítě***

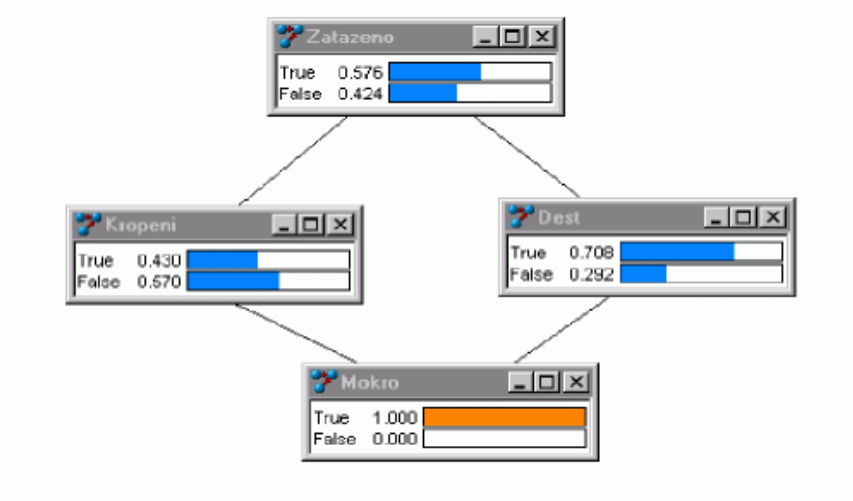



***Zdroj: (BERKA, 2003)***

Příklad využití je znázorněn na obrázku 15, kde pro síť na obrázku 14 byly vypočítány příslušné hodnoty pravděpodobností. Autory toho příkladu zajímalo, co a s jakou aposteriozní

pravděpodobností způsobilo, že bylo mokro. Vzhledem k faktu, že tento příklad slouží pouze jako názorná demonstrace, tak zde nebudou uvedeny jednotlivé hodnoty dílčích pravděpodobnostní, na základě nichž se dospělo k výsledným hodnotám v grafu na obrázku 15.

***Obr. 15: Příklad bayesovské sítě s aposteriozními pravděpodobnostmi***




***Zdroj: (BERKA, 2003)***

Bayesovské sítě tedy v sobě kombinují dva typy znalostí. Prvním typem je znalost struktury mezi atributy (hrany v grafu) a znalosti o pravděpodobnostech těchto atributů (ohodnocení uzlů v grafu). Při DZD jsou dvě následující možnosti, buďto se vychází ze známé struktury a z dat se odvozují pouze příslušné pravděpodobnosti. Tato struktura může být sestavena nějakým expertem v oboru. Druhá možnost je odvodit z dat jak strukturu, tak pravděpodobnosti. Na odvození struktury sítě existuje řada algoritmů, např. EM algoritmus (BERKA, 2003), (MITCHELL, 1997).

### 2.4.7 Techniky založené na analogii

Většina těchto technik je založena na pravidle, že v neznámé situaci lze použít to řešení, které se osvědčilo v podobné situaci. Mezi nejznámější techniky, které lze zařadit do této kategorie patří pravidlo nejbližšího souseda, případové usuzování (CBR), učení založené na instancích (IBL), líné učení, paměťové učení (MBL) a shlukování (clustering).

#### Podobnost mezi příklady

Podstatou těchto úloh je určení míry podobnosti mezi jednotlivými případy. K tomu se využívají různé metriky. Nejpoužívanější metriky jsou Euklidovská vzdálenost a Hammingova vzdálenost (známější pod pojmem „Manhattan“ nebo „city block“). Pro kategoriální atributy se používá často metrika „překrytí“ (overlap), pomocí něhož je počítán počet rozdílů v hodnotách stejných atributů (BERKA, 2003).

Uvedené metriky trpí dvěma zásadními nedostatky, prvním z nich je fakt, že se všechny vybrané atributy podílí na určení podobnosti stejně, tento problém lze odstranit využitím vah. Stanovení jednotlivých vah také není jednoduché, často je potřeba sehnat nějakého experta na danou problematiku. Druhým problémem je, že měření podobnosti kategoriálních atributů metodou overlap nedokáže často zachytit jemné nuance problému, protože se porovnává pouze shoda / neshoda hodnot.

Pro řešení tohoto problému se nabízí využít nějaké složitější metriky, např. „Value Difference Metric“, tato metrika bere do úvahy celkovou podobnost příkladů patřících do různých tříd v celé trénovací množině (BERKA, 2003).

Pomocí těchto metod lze řešit zajímavý dílčí problém a to podobnost mezi časovými řadami a sekvencemi. Nejjednodušší způsob jak měřit podobnost dvou časových řad stejné délky je pomocí euklidovské metriky, nicméně tento způsob je velmi citlivý na drobné posuny v časové ose, proto se jako lepší řešení naskýtá využít metody „dynamic time warping“, která provádí dynamickou deformaci časové osy (BERKA, 2003).

### Učení založené na instancích

Učení založené na instancích se skládá ze tří základních kroků. První je volba správné metriky pro porovnání jednotlivých instancí. Nejpoužívanější metriky byly popsány v předchozí podkapitole. Druhý krok je volba klíčových instancích, které se uloží do databáze a se kterými se při klasifikaci porovnávají další případy. Toto je klíčový krok celého algoritmu. Existuje několik řešení. Nejjednodušším řešením je uložit do databáze všechny příklady z trénovačích dat, toto řešení dobře pracuje i pro data zatížené šumem. Druhá možnost je, že během učení se pokusí systém každý příklad klasifikovat a do databáze následovně zařadí pouze příklady, které se nepodařilo klasifikovat správně (vzniká problém v případě šumu). Nejsložitější způsob je založen na kritériu správnosti klasifikace. Nejprve tedy uloží do databáze všechny příklady z trénovací množiny a ty používá pro klasifikaci nových případů a v DB zůstanou pouze ty, které měli největší úspěšnost klasifikace. Třetím krokem v celém procesu je pak samotná klasifikace, pro tu se nejčastěji využije algoritmu „k-nearest neighbour rule“. Problém této metody je, že při velké obsáhlosti dat sekvenční procházení uložených instancí při klasifikaci je velmi časově náročná. Proto je potřeba využít specifických indexovacích technik, pomocí nichž lze redukovat dobu hledání v databázích (tvorba stromů, kde jsou v uzlech atributy) (BERKA, 2003).

Specifickým příkladem je metoda nejbližšího souseda. U této metody se ve fázi učení neprovádí žádná generalizace a všechny příklady jsou chápány jako body v N-rozměrném prostoru. Prakticky tato metoda funguje tak, že si během učení zapamatuje systém veškeré příklady z trénovací množiny a pak se při klasifikaci vybere „K“ (parametr, jež lze nastavit) nejbližších bodů z prostoru a ty pak „hlasují“ o zařazení daného příkladu k dané třídě. Předpokladem pro tento algoritmus je ten, že cílový atribut je kategoriální, tedy klasifikuje se do určitého počtu tříd (MICHIE et. al., 1994).

## 2.5 SW nástroje

V této kapitole bude uvedeno několik příkladů SW společností, jež vyrábějí nástroje využívané pro data mining. Nástroje budou rozděleny do dvou kategorií a to volně dostupné a komerční. U každého nástroje budou stručně popsány jeho funkce, případně historie. Pro výběr zde zmíněných nástrojů bude primárně vycházeno z analýzy společnosti Gartner. Tato společnost je považována za nejlepší společnost v oblasti poradenství a výzkumu o informačních technologiích.

Každý rok sestavuje žebříček SW nástrojů v jednotlivých odvětvích IT, jeden z žebříčků se týká právě oblasti data miningu. Oficiálně tuto oblast pojmenovali „Advanced Analytics Platforms“. Jedním z hlavních výstupů analýz od Gartneru je Magic Quadrant. Je nutné podotknout, že tento graf byl sestaven jedním autorem na základě předem vybraných ukazatelů (GARTNER, 2015).

Dalšími analýzami, jež byly využity pro výběr SW nástrojů zmiňovaných v této kapitole, jsou dva články, které vyšly na renomovaném webu Predictive Analytics Today (PAT). Tento web se zabývá problematikou DZD, Big data, BI atd. První článek je zaměřen primárně pouze na volně dostupné DM nástroje, kde bylo vybráno autory této studie 40 nejlepších nástrojů. Druhý článek je zaměřen na volně dostupné i komerční nástroje, celkem se zmiňuje o 31 nástrojích (PAT, 2014), (PAT, 2013).

Na základě těchto tří analýz bude vybráno několik nejlepších společností / SW nástrojů z každé kategorie.

***Tab. 5: Komerční SW data miningové nástroje***

| Název společnosti | Popis | Zdroj |
|---|---|---|
| SAS | Dle analýzy společnosti Gartner patří SAS mezi špičku v oboru, dokonce byl vybrán jako nejlepší. Nástroj od společnosti SAS, který se zabývá primárně DM, se nazývá SAS Enterprise Miner. První verze tohoto nástroje byla vydána v roce 1999. SAS vyvíjí i několik dalších nástrojů, které se do určité míry zaobírají DM, případně ulehčují DM proces. Mezi tyto nástroje lze zařadit např. SAS Visual Analytics, který slouží primárně k počáteční datové analýze.<br>V nástroji SAS Enterprise Miner lze provést veškeré procedury spojené s DM procesem, od předzpracovaní dat přes modelování až po evaluaci výsledků. Co se týče podporovaných DM technik, tak tento nástroj podporuje většinu algoritmů, jež byly zmíněny v předchozí kapitole (rozhodovací stromy, neuronové sítě, clustering, analýza nákupního košíku, analýza časových řad a mnoho dalších).<br>Licence na tento nástroj stojí řádově desítky tisíc dolarů na rok. | (PAT, 2014) (SAS, 2014) |
| IBM | Podobně jako SAS je společnost IBM zařazena společností Gartner mezi leadery na trhu v oblasti DM nástrojů. Nástroj, jež primárně slouží pro DM se nazývá IBM SPSS Modeler. První verze tohoto nástroje vznikla v roce 1994 ještě pod názvem Clementine. Společnost IBM odkoupila tento nástroj v roce 2009.<br>Tento SW opět podporuje celý DM proces, primárně vychází z metodiky CRISP-DM. Lze zde nalézt širokou podporu jednotlivých algoritmů, např. bayesovské sítě, neuronové sítě, rozhodovací stromy (C5.0, CHAID), Apriori, …<br>Licence na tento nástroj v té nejzákladnější verzi stoji €15,000 / rok. Společnost IBM poskytuje trial verzi na vyzkoušení zdarma. | (IBM, 2015) |

| | | |
|---|---|---|
| SAP | Nástroj společnosti SAP pro DM se nazývá SAP Predictive Analysis. V roce 2013 firma SAP odkoupila firmu KXEN, která se zabývala podobně DM softwarem (primárně orientována na Big Data). Verze SAP Predictive Analysis, které vyšly od zmiňované akvizice, již obsahují funkce a moduly z KXENu.<br>SW nástroj podporuje celý DM proces a opět vychází primárně z metodiky CRISP-DM. Dle uživatelského manuálu neobsahuje tolik funkcí a nepodporuje tolik algoritmů jako SW nástroje od SASu nebo od IBM. Ze zde zmíněných funkcí se jedná o neuronové sítě, 1 typ rozhodovacího stromu a apriori algoritmus. Podobně jako IBM nabízí firma SAP tento nástroj ve třicetidenní trial verzi s omezenou funkčností. | (HOLST et. al., 2014) |
| Microsoft | Microsoft (MS) nevyvíjí speciální nástroj pro DM. Jejich DM řešení je pouze doplněk do nástroje, který se nazývá Microsoft SQL server. Tento doplněk se jmenuje Microsoft Analysis Services. MS zahájil vývoj DM nástrojů již v roce 1996, kdy vznikla platforma Plato, z které se postupem času stal MS Analysis Services (první verze rok 2005).<br>MS Analysis Services obsahuje pouze modelovací funkce, tudíž předzpracování a přípravu dat je potřeba udělat v jiném nástroji nebo modulu MS SQL serveru. Ze zde zmíněných algoritmů jsou v tomto nástroji podporovány mj. naivní bayesovský klasifikátor, asociační pravidla, rozhodovací stromy a neuronové sítě.<br>Pro využívání toho nástroje je potřeba si zakoupit rozšířenou licenci MS SQL Serveru, která vyjde na minimálně $3,700. | (MICROSOFT, 2015) |
| Oracle | Společnost Oracle vyvíjí DM nástroj, který se nazývá Oracle Data Mining. Zajímavostí je, že v roce 2014 byl hodnocen společností Gartner poměrně dobře (ve skupině vyzyvatelů leaderů na trhu (SAS a IBM), ale v roce 2015 nebyl již do analýzy vůbec zařazen. První verze toho nástroje byla vydána v roce 2002. Oracle vydává většinou novou verzi tohoto SW nástroje společně s novou verzí jejich DB. Oracle se primárně soustředí na vývoj databázových systémů, nicméně poslední verze tohoto nástroje je z roku 2009, přesto že v roce 2012 vydali novou verzi jejich databáze.<br>Podobně jako nástroje od SAPu nebo od IBM vychází opět z metodiky CRISP-DM, tudíž lze simulovat celý DM proces. Funkčnost je velmi podobná nástroji od SAPu. Lze zde opět najít základní klasifikační algoritmy (naive bayes, rozhodovací stromy), KMeans algoritmus, apriori atd.. ). | (ORACLE, 2009) |

| | | |
|---|---|---|
| | Oracle Data Mining je k dispozici v rámci OTN licence, tudíž lze tento nástroj za určitých předpokladů používat k nekomerčním účelům „zdarma“. | |

***Zdroj: uvedený v tabulce, vlastní zpracování***

***Tab. 6: Volně dostupné SW data miningové nástroje***

| Název společnosti | Popis | Zdroj |
|---|---|---|
| Orange | Orange SW byl na webu PAT vyhlášen jako nejlepší free SW nástroj. Vývoj Orange započal na fakultě biometriky na univerzitě v Ljublani. První verze tohoto nástroje vyšla v roce 1996, na vývoji se stále pracuje, poslední rozsáhlý update proběhl v roce 2013, kdy došlo k výrazné proměně grafického rozhraní.<br>Funkcionalita tohoto nástroje není na takové úrovni jako u nejlepších komerčních nástrojů, nicméně lze zde provést základní klasifikaci pomocí algoritmů naive bayes, neuronových sítí, knn algoritmu, několika typů rozhodovacích stromů, případně algoritmus SVM. Obsahuje i několik funkcí, pomocí nichž lze upravovat vstupní data, tudíž lze prohlásit, že lze provést celý DM proces v rámci tohoto nástroje.<br>Jedná se o OpenSource SW, tudíž kdokoliv si může stáhnout zdrojový kód a tento nástroj upravovat / vylepšovat v rámci GNU/GPL licence. | (ORANGE, 2013) |
| Weka | Nástroj Weka vznikl jako projekt na novozélandské univerzitě Waikato. NZ vláda tento projekt podporovala od roku 1993, první veřejná verze vyšla v roce 1996. Poslední verze byla vydána v prosinci roku 2014. Weka byla uvedena jako druhá v pořadí v analýze volně dostupných nástrojů na webu PAT.<br>Předzpracování dat v tomto SW nástroji je možné, nicméně z vlastní zkušenosti mohu potvrdit, že to není uživatelsky příjemné. Nástroj Weka je primárně určen pro klasifikaci a strojové učení, tudíž je zde velké množství klasifikační algoritmů (desítky rozhodovacích stromů, rozhodovací pravidla, bayesovské klasifikátory, …).<br>Tento nástroj je podobně jako Orange k dispozici v rámci GNU/GPL licence. Zajímavostí je, že nástroj je vyvíjen v Javě, tudíž lze importovat vlastní DM algoritmy jako třídy v Javě. | (WEKA, 2014) |
| Rapid Miner | RapidMiner je poměrně nový SW nástroj, protože jeho první verze vyšla až v roce 2006, což je řádově o 10 let později než většina zde zmíněných SW. Nástroj byl vyvíjen v rámci dortmundské univerzity. V roce 2007 dva vývojáři založili | (RAPID MINER, 2015) |

| | | |
|---|---|---|
| | společnost Rapid-I a částečně se tak distancovali od akademického prostředí. V roce 2014 a 2015 dokonce zařadila společnost Gartner tento nástroj mezi leadery na trhu. SW nástroj využilo již přes 3 milióny stažení a údajně ho využívá přes 200 tisíc uživatelů.<br>Tento nástroj opět podporuje všechny součásti DM procesu. Co se týče funkcionality, tak může konkurovat tento nástroj nejlepším komerčním nástrojům. Jsou zde desítky klasifikačních algoritmů (bayes, rozhodovací stromy a pravidla, neuronové sítě), asociační pravidla a mnoho dalších algoritmů.<br>Tento nástroj je OpenSource podobně jako Orange a Weka k dispozici v rámci GNU/GPL licence. Vývojáři dávají k dispozici vždy zdrojový kód všech verzí kromě té nejaktuálnější. Je možnost vytvářet si vlastní pluginy a rozšiřovat tak funkcionalitu, podobně jako u Weky. | |
| Knime | Knime se podobně jako RapidMiner v roce 2015 umístil mezi leadery na trhu v analýze společnosti Gartner. Vývoj tohoto nástroje začal v roce 2004. Od roku 2006 se používá při řade farmaceutických výzkumů. Poslední verze byla vydána v dubnu roku 2015. V současnosti má kolem 15 tisíc uživatelů.<br>Workflow tohoto nástroje je rozděleno na pět částí a přibližně tak odpovídá konceptu CRISP-DM. Funkcionalita je opět velmi podobná ostatním zmíněným nástrojům. Mezi algoritmy, které lze využít v tomto nástroji patří bayesovské klasifikátory, rozhodovací stromy, neuronové sítě, rozhodovací pravidla (fuzzy), asociační pravidla a další.<br>Tento nástroj je k dispozici v rámci GNU/GPL licence. Lze opět přidávat pomocí pluginů další funkcionalitu. | (KNIME , 2015) |
| LisP Miner | Lisp Miner je otevřený akademický systém pro podporu výzkumu a výuky v oblasti DZD. Vývoj započal v roce 1996 na VŠE v Praze. Od té doby byla na vývoj tohoto nástroje poskytnuta celá řada grantů. Nástroj se neustále vyvíjí a jsou přidávány nové moduly, případně opravovány různé chyby, poslední verze vyšla na konci června 2015.<br>Tento SW nástroj je primárně zaměřen na tvorbu asociačních pravidel (navazuje zde na metodu GUHA). Tímto se zásadně odlišuje od zbylých zde uvedených SW nástrojů.<br>Nástroj je volně dostupný k akademickým účelům. | (LISPMINER, 2015) |

***Zdroj: uvedený v tabulce, vlastní zpracování***

## Kapitola 3

# Statistické šetření

V této kapitole bude popsána problematika statistických šetření. Popsány jednotlivé typy, způsoby sběru dat a nakonec uvedeny některé příklady probíhajících šetření.

## 3.1 Základní pojmy

Zdrojem statistických údajů je statistické šetření. Prakticky se jedná o zjišťování neznámých dat o hodnotách určených statistických znaků jednotlivých jednotek zkoumaného statistického souboru. Náplní těchto šetření není pouze samotné získávání dat, ale jedná se také o teoretické a praktické postupy při tomto zjišťování. V případě rozsáhlých sociálně-ekonomických šetření se často lze setkat s velmi rozsáhlými základními soubory, proto je potřeba rozhodnout, zda se bude jednat o vyčerpávající (úplné) šetření anebo o šetření výběrové (neúplné). V případě výběrových šetření je potřeba zvolit nejvhodnější typ šetření, díky kterému lze získat co nejpřesnější (VALENTOVÁ, 2013).

Při přípravě každého statistického šetření nejprve přesně vymezit účel připraveného zkoumání (k čemu má sloužit a na jaké otázky má dát odpověď). Na základě tohoto účelu se posléze musí stanovit předmět zkoumání, neboli se vymezí statistický soubor, na němž se zkoumání provede. Dále je potřeba stanovit obsah tohoto zkoumání, neboli se určí statistické proměnné (znaky), a nakonec je potřeba zvolit metody analýzy, pomocí kterých se každé zkoumání završí (ČERMÁK et. al., 1998).

Základem každého statistického šetření jsou data. Data lze získat v zásadě dvěma způsoby. Buďto se jedná o tzv. sekundární data, což jsou data, která jsou převzata z různých zdrojů. U tohoto typu dat je potřeba nejprve prověřit důvěryhodnost vybraného zdroje dat. Druhý typ dat jsou tzv. primární data, což jsou data, která nejsou převzatá, jsou „originální“, získaná na základě vlastního zjišťování (VALENTOVÁ, 2013).

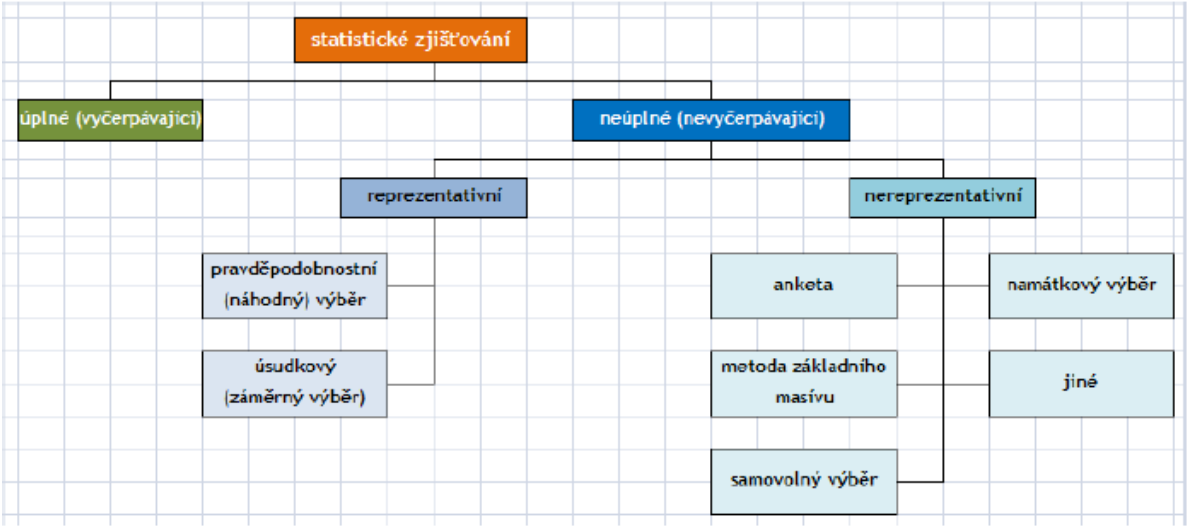
Každé statistické šetření tedy obsahuje statistické proměnné. Tyto proměnné lze rozdělit do těchto tří kategorií:

- Nominální – je to taková proměnná, o které lze pouze prohlásit, zda je stejná nebo různá. Příkladem takovéto proměnné může být škola, fakulta, obec atd. Hodnotami těchto proměnných mohou být texty, případně číselné kódy.
- Ordinální – tento typ proměnných se často nazývá pořadový. U tohoto typu lze na rozdíl od nominálního typu určit pořadí. Jedná se tak například o úroveň spokojenosti s nějakou službou, úroveň vzdělání. Lze pouze říct, která hodnota je vyšší, ale nelze říct o kolik nebo kolikrát.

- Kvantitativní – u tohoto typu proměnných už lze prohlásit o kolik nebo kolikrát jsou vyšší než druhá (jsou-li kladné). Mezi tyto proměnné lze např. zařadit měsíční příjem domácnosti, výšku, váhu atd. (ČERMÁK et. al., 1998).

V předchozí pasáži tohoto textu byla zmíněna problematika úplného a výběrového šetření. Úplné šetření oproti výběrovému několik zásadních výhod. Zejména se jedná o to, že poskytuje naprosto přesnou charakteristiku souboru, což umožňuje činit velmi široké závěry. Další výhodou je že úplné šetření poskytuje informace nejen o celém souboru, ale také o každém jednotlivém prvku. Většinou jsou tato šetření podložena zákonem, tudíž lze předpokládat, že se u respondentů setkají s větším pochopením. Často jsou doprovázena různými osvětovými kampaněmi, aby se dostali do lidského povědomí. Nicméně toto je spojeno s vysokými náklady na tato šetření, např. SLDB v roce 2011 stálo státní pokladnu řádově 2,5 miliardy korun. Tudíž zde vyvstává otázka, zda takto vysoké náklady přináší kýžený efekt v podobě získaných výsledků. Další nevýhodou úplných šetření je praktická neproveditelnost. Toto je spojeno zejména s časovou náročností a s tím, že by u části souboru mohlo dojít k neochotě a tím poskytnutí nepřesných údajů, případně naprosté odmítnutí účasti v daném šetření. Proto je nutné často vzít zavděk pouze výběrovými šetřeními. Na obrázku 16 je znázorněno rozdělení jednotlivých druhů šetření (ČERMÁK et. al., 1998).

***Obr. 16: Rozdělení druhů šetření***




***Zdroj: (ČERMÁK et. al., 1998)***

## 3.2 Výběrové šetření

Jak bylo naznačeno v předchozí kapitole, výběrové šetření je vhodné využít v případech, kdy by úplné zjišťování nešlo provést nebo bylo příliš neefektivní nebo nehospodárné. Zásadní výhodou jsou menší náklady a větší rychlost s jakými lze získat výsledky z výběrového šetření v porovnání s úplnými. V některých případech se výběrové šetření kombinují s úplnými, např. že u všech jedinců se zjišťují pouze hlavní údaje (základní rysy) daného šetření a navíc se každému dvacátému účastníkovi přidá ještě nějaký dodatečný dotazník s rozšiřujícími otázkami (ČERMÁK et. al., 1998).

Určitě je namístě zmínit zde i nevýhody výběrových šetření. Mezi ně lze zařadit např. to, že na základě výběrových šetření lze tvořit dostatečně přesné odhady za celý soubor (celý stát), ale již nelze tvořit odhady za menší jednotky (kraje, obce), přitom právě tyto odhady se často v praxi

požadují[5]. Další nevýhodou je i to, že již zmíněná úspora nákladů není přímo úměrná snížení počtu jednotek. Jedním z hlavních důvodů tohoto jevu jsou dodatečné „fixní“ náklady, které jsou způsobeny důkladnější organizační přípravou, kontrolou správnosti zjišťovaných údajů nebo na složitější výpočty při přípravě šetření a při vyhodnocování přesnosti údajů (ČERMÁK et. al., 1998).

Jak již bylo znázorněno na obrázku 16 výběrové šetření lze rozdělit do dvou kategorií a to na reprezentativní a nereprezentativní šetření. U nereprezentativních šetření zobecnění výsledků na základní soubor není možné, případně velmi problematické. Mezi nereprezentativní šetření patří zejména anketa a metoda základního masivu, případně lze zařadit i namátkový a samovolný výběr. U reprezentativních šetření již existuje možnost zobecnění výsledků šetření na základní soubor, protože soubor představuje velmi věrnou zmenšeninu souboru základního. Mezi tyto šetření se řadí pravděpodobnostní (náhodný) výběr a záměrný (úsudkový) výběr (VALENTOVÁ, 2013).

### 3.2.1 Nereprezentativní šetření

V této kapitole budou popsány vybrané typy nereprezentativních šetření a to anketa a metoda základního masivu.

#### Anketa

Anketa je způsob šetření, kdy je oslovena jen určitá část statistických jednotek (určitý okruh osob, podniků a institucí). Nejčastější formou jsou dotazníky. Většinou je založena na principu dobrovolnosti. Návratnost těchto dotazníků je velmi malá. (Valentová, 2013) V průměru je návratnost ankety pouze 33 %, proto nelze brát výsledky a vypočtené charakteristiky jako obecně platné. Dalším problémem u anket je velmi úzký vztah mezi odpovědí nebo odmítnutím na jedné straně a dotazovanou skutečností na straně druhé. Často se tak stává např. to, že dotazník o výši příjmů nevyplní a nevrátí lidé s vysokými příjmy nebo naopak s velmi nízkými příjmy (mohou se stydět). Pouze jen v některých případech lze zobecnit výsledky anket tak, aby šly odhadnout hodnoty charakteristik celého souboru (ČERMÁK et. al., 1998).

#### Metoda základního masivu

Tato metoda je vhodná za předpokladu, že soubor obsahuje několik velmi velkých jednotek a zároveň velký počet jednotek malých. Pokud probíhá zkoumaný jev (např. prodej nebo výroba nějaké komodity) převážně ve velkých jednotkách, tak stačí k získání hrubého odhadu o objemu nebo kvalitě této komodity prošetřit jen tyto velké jednotky. Malé lze za určitých předpokladů z tohoto šetření vynechat (VALENTOVÁ, 2013). Předností tohoto druhu zjišťování je to, že podchytí převážnou část zkoumaného jevu a zároveň ušetří velké množství práce (ČERMÁK et. al., 1998).

Podobně jako u anket ani tato metoda nedovoluje zobecňovat získané výsledky na celý soubor. Hlavním důvodem je fakt, že v neprošetřené části souboru (malé jednotky) mohou být jiné zákonitosti a tendence než u prošetřených (velkých) jednotek. Nicméně lze nalézt i nějaké výjimky, např. příklady, kdy mezi velikostí jednotky a zkoumaným znakem je prokazatelně nulová korelace či asociace (ČERMÁK et. al., 1998).

[5] tímto se v posledních letech zabývá poměrně nově vniknuvší úsek teorie výběrových šetření „odhadování za malá území“ – small area estimation (Šídlo, LS 2013/2014)

### 3.2.2 Reprezentativní šetření

V této kapitole budou popsány vybrané druhy reprezentativních šetření a to náhodný a záměrný výběr.

#### Záměrný výběr

Tento výběr provádí zkušený odborník na základě vlastního úsudku. Výběr je prováděn ze základního souboru vybrané statistické jednotky tak, aby výběrový soubor byl co nejvíce reprezentativní. Tím, že tento výběr provádí nějaký odborník, existuje zde velké riziko subjektivity (VALENTOVÁ, 2013). Počet jednotek bývá při tomto typu výběru předem dán finančními nebo pracovními možnostmi instituce, která je pověřena provedením daného statistického šetření (ČERMÁK et. al., 1998).

Velmi často se pak v praxi vybírají jednotky, tak aby se získané charakteristiky daly zobecnit. Tento požadavek se pak označuje jako požadavek reprezentativnosti. Zabezpečení tohoto požadavku je velmi obtížné. Nicméně ho lze dosáhnout následujícím způsobem. Jedná se o sestavení výběrového souboru takovým způsobem, že rozdělení četností u nějakého známého (pomocného) znaku odpovídá rozdělení v celém souboru. Často se volí více znaků a jedná se většinou o kategoriální proměnné. Prakticky se usiluje o shodu struktury výběrového a základního souboru. Způsoby, pomocí nichž tohoto stavu lze dosáhnout se nazývají kvótní výběr nebo tzv. metoda dokonalého průřezu. Druhou možností jak dosáhnout reprezentativního výběru je za pomocí tzv. typického výběru, při kterém se vybírají takové jednotky, které mají hodnoty zkoumaného znaku blízké průměru nebo s hodnotami modálními. Nicméně ani aplikace těchto metod nemusí zaručit, že výběr bude dobrým reprezentantem celého souboru (ČERMÁK et. al., 1998).

Navíc četné zkušenosti z praxe naznačují, že i ti nejzkušenější a nejobjektivnější znalci mají tendenci dělat ve svých výběrech odchylky od průměru. Dopouští se tak systematické chyby. Dalším závažným nedostatkem úsudkových výběrů je fakt, že nelze objektivně stanovit přesnost odhadů sestrojených na jeho základě, tj. vypočítat nějakou průměrnou nebo maximální chybu odhadu. Případně lze tyto údaje odhadnout, ale jsou opět tyto odhady provedené nějakým znalcem (většinou tím stejným, jež stanovoval pravidla výběru) (ČERMÁK et. al., 1998). Mezi další nedostatky lze zařadit zdánlivou neexistenci non-response při kvótních výběrech. Non-response je pouze skrytá za náborem respondentů (JEŘÁBKOVÁ, 2012).

#### Náhodný výběr

Výběrová data jsou v těchto šetřeních pořizována náhodným způsobem. Tato šetření jsou vždy reprezentativní. Reprezentativnost těchto šetření je zabezpečena působením zákonitosti náhody. Pomocí metod matematické statistiky lze pak výsledky těchto šetření zobecnit na základní soubor (VALENTOVÁ, 2013).

Tento způsob šetření se provádí tak, že se nejprve celý soubor rozdělí na výběrové jednotky. Tyto jednotky jsou zpravidla totožné s elementárními (statistickými) jednotkami, ale mohou to být také menší nebo větší skupiny. Každé této jednotce se následovně přiřadí pravděpodobnost zařazení do výběru. Tato pravděpodobnost nemusí být stejná, podstatné je, že je známá a vyčíslitelná. Následná selekce pak proběhne, tak aby o výběru rozhodovala pouze náhoda (ČERMÁK et. al., 1998).

U náhodných výběrů je možné sestavit takové odhady, které s rostoucím rozsahem výběru konvergují ke skutečné hodnotě. Jedná se o tzv. nevychýlené odhady, tedy při každém rozsahu výběru

nenadhodnocují / nepodhodnocují skutečnou hodnotu. Přesnost těchto výběrů lze při každém rozsahu objektivně změřit (určit střední velikost jejich chyb, případně stanovit tzv. intervalový odhad[6]). Odhady získané ze záměrných výběrů nemají ani jednu z těchto vlastností (ČERMÁK et. al., 1998).

Existuje celá řada metod, pomocí nichž lze realizovat náhodný výběr. Nejjednodušší z nich je prostý náhodný výběr, při němž má každá jednotka stejnou pravděpodobnost výběru. Lze ho realizovat buďto jako výběr s vracením (pravděpodobnost, že jedna jednotka bude vybrána vícekrát) nebo bez vracení (zde se pravděpodobnost výběru jednotky zvětšuje s každým tahem) (VALENTOVÁ, 2013).

Mezi složitější metody náhodného výběru lze zařadit stratifikovaný výběr, při němž se nejdříve základní soubor rozdělí na straty. Straty by měly být co nejvíce homogenní (obsahovat podobné jednotky z určitého hlediska). V dalším kroku je proveden v každé oblasti náhodný výběr daného počtu prvků. Nejčastěji se provádí proporcionální výběr, výběrové rozsahy v oblastech jsou úměrné velikostem oblastí. Další pokročilou metodou je dvoustupňový náhodný výběr, kde se nejprve základní soubor rozdělí do skupin. V prvním stupni se ze základního souboru vyberou skupiny jednotek tzv. klastrů. Ve druhém stupni se ve vybraných jednotkách náhodně vybírají statistické jednotky. Tento způsob výběru odstraňuje problém prostorové rozptýlenosti u stratifikovaného výběru. Nicméně tento způsob e má oproti stratifikovanému výběru nižší statistickou efektivitu. Zejména proto, že nejsou na konečném stupni výběru jednotky shlukovány zcela náhodně a mají tendence být v řade aspektů více „podobné“, tudíž existuje vnitřní homogenita klastrů. Čím vyšší tato homogenita je, tím je větší ztráta na efektivnosti výběru (VALENTOVÁ, 2013), (JEŘÁBKOVÁ, 2012).

## 3.3 Způsoby sběru dat

Ve většině průzkumech jsou nejčastěji zdrojem shromažďování údajů lidé. Buďto jsou sami výběrovými jednotkami nebo výběrové jednotky zastupují. Proto má většinou získávání údajů formu mezilidské komunikace. Nicméně s rozvojem informačních technologií a rozšířením internetu do domácností toto nemusí déle platit (ČERMÁK et. al., 1998). Existují čtyři základní způsoby sběru dat při šetření. Pro každý z těchto způsobů existuje čtyřpísmenná anglická zkratka:

- PAPI (Pen and Paper Interview) – dotazníkové šetření pomocí tužky a papíru, při tomto druhu šetření nehrozí technické problémy. Je vhodné zejména pro problematické lokality (lidé s nedostatečným technickým vzdělání, případně s určitým druhem postižení - mentální retardace, demence). Většinou tento druh šetření dělají externisté.
- CAPI (Computer Assisted Personal Interview) – dotazníkové šetření pomocí elektronického formuláře. Šetření je plně automatické, provádí dotazníkem a upozorňuje na chyby. I zde je potřeba externisty, aby obcházeli domácnosti, protože je prováděno osobně. Je jednodušší oproti PAPI co se týče metodiky a odpadají zde problémy s čitelností.
- CATI (Computer Assisted Telephone Interview) – dotazníkové asistované telefonické šetření. Jedná se prakticky o CAPI akorát s tím rozdílem, že interview je vedeno přes telefon a ne osobně.

[6] interval, v němž se téměř jistě nachází skutečná hodnota

- CAWI (Computer Assisted Web Interview) – jedná se o dotazníkové šetření pomocí webového formuláře. Realizátor výzkumu vytvoří webový formulář, na který respondent odpovídá sám pouze za pomocí klávesnice a myši (JEŘÁBKOVÁ, 2012).

Tyto metody lze zařadit mezi standardizované postupy, tyto postupy se využívají zejména při kvantitativním výzkumu, cílem těchto výzkumů je získat číselné odpovědi a odpověď na otázku „kolik“. Existují také nestandardizované (hloubkové) metody, které se využívají primárně při kvalitativních výzkumech, jejíž cílem je získat odpovědí na otázky „proč“ nebo „jak“. U těchto postupů bývá dotazník nahrazen volnějším scénářem rozhovoru. Takovýto rozhovor se často označuje jako hloubkové interview (případně se může jednat o skupinový rozhovor). Hlavním rozdílem oproti standardizovaným postupům je zde možnost přizpůsobování rozhovoru vzhledem k odpovědím dotázaných. Předpokladem pro tento typ šetření je velmi dobrá průprava osob, jež tyto rozhovory vedou. Dále taky velmi dobrý způsob záznamu a následného rozboru celého interview. Mezi nestandardizované postupy lze zařadit i pozorování. Zde je potřeba připravit jako podklad osnovu zahrnující typy akci a projevů, jež si má pozorovatel všimnout. Vzhledem k náročnosti se zpravidla těchto postupů využívá pouze na velmi malých výběrových souborech. Pro výzkum se vybírají průměrní jedinci v sociologickém nebo sociálně psychologickém slova smyslu (ČERMÁK et. al., 1998).

## 3.4 Chybějící údaje

Jedním se závažných problémů při zpracování dat ze statistických šetření jsou chybějící data. V této kapitole nejdříve budou popsány základní problémy, které chybějící data způsobují a dále zde budou uvedeny příklady, jak lze problém s chybějícími daty vyřešit.

Při statistickém zpracování dat vznikají dva druhy chyb. Jedna nazývá výběrová chyba a ta vzniká pouze ve výběrových šetření. Vzniká vlivem variability zkoumaných proměnných v populaci v důsledku skutečnosti, že se vždy prošetřuje pouze jeden ze všech možných souborů. Druhý typ chyb vzniká právě díky existenci chybějících údajů, jedná se o tzv. non-sampling error. Při zvětšování rozsahu výběru má tato chyba tendenci k růstu. Tato chyba vzniká jak v případě výběrových, tak v případě úplných šetření. Mezi odborníky existují názory, že tento typ chyby způsobuje při interpretaci dat z šetření větší zkreslení celkových výsledků než chyby výběrové (VALENTOVÁ, 2013).

### 3.4.1 Druhy chybějících údajů

Existují v zásadě dva druhy chybějících údajů. Prvním z nich jsou uživatele definované chybějící údaje. V praxi to znamená, že uživatel (zpracovatel dotazníku) určuje, co bude za chybějící údaj považováno. Může to být např. nezodpovězená otázka v dotazníku, toto může nastat z několika důvodů. Hlavní důvod je respondentova tzv. non-response (kdy se vědomě rozhodl, že na zadanou otázku neodpoví), dále se může jednat o nečitelnou nebo špatně označenou odpověď či odpověď „nevím“. Jako chybějící údaje lze v některých specifických případech označit málo zastoupené kategorie nebo kategorie pro sledování určitého problému nepodstatné. V některých případech se za chybějící údaje označují odlehlá pozorování, která mohou výrazně zkreslit hodnoty některých statistických charakteristik (VALENTOVÁ, 2013).

Non-response je velmi specifický typ chybějících údajů, protože může upozornit na nějaké systémové chyby při sestavování dotazníku. Může se stát, že respondent neporozuměl otázce nebo není schopen vybrat odpověď, protože dotazník neobsahuje jeho názory či pocity. Dále se může stát,

že respondent nemá dost času na vyplňování dotazníku nebo ztratí během vyplňování zájem, tento problém lze identifikovat např. tím, že u respondentů není vyplněno více otázek za sebou. Dalším problémem jsou respondenti, kteří sice poskytnou některé údaje, ale úmyslně poskytnou špatné údaje a tak zkreslují stav zkoumaného problému (toto se velmi špatně ověřuje, částečně se tím zabývá disciplína zvaná datová kvalita) (VALENTOVÁ, 2013), (SINGH et. al., 1996).

Druhým typem chybějících údajů jsou tzv. chybějící údaje systémové. Ty mohou vzniknout v zásadě dvěma způsoby. Buďto při samotném vstupu dat, tedy v případě, kdy nebyla zadána žádná hodnota nebo byla vložena hodnota nepřípustná. Druhým způsobem je, že vznikly na základě chybných výpočtů, které jsou z matematického hlediska neproveditelné (dělení nulou) (VALENTOVÁ, 2013).

### 3.4.2 Postupy při práci s chybějícími údaji

Při práci s chybějícím je potřeba zvolit optimální postup, který bude minimalizovat nevýběrovou chybu. Nejprve potřeba rozhodnout zda se chybějící údaje v souboru ponechají či nikoli. Ponechání chybějících dat v souboru může být problém hned z několika důvodů. Zejména se „zmenšuje“ rozsah souboru, což může vést k oslabení statistické síly jednotlivých analýz. Nejzávažnějším problémem je skutečnost, že zbylá data v souboru mohou být značně zkreslená (VALENTOVÁ, 2013).

V některých statistických SW nástrojích existují speciální metody na zpracování chybějících údajů, případně pro práci s nimi. Existují i různé metody vypouštění údajů. Např. metoda Listwise, kde za předpokladu chybí hodnota, tak je vypuštěn z analýzy celý řádek datové matice s alespoň jednou chybějící hodnotou. Tato metoda má smysl pouze za předpokladu, že podíl takových řádků je menší než 5 %. Alternativou k této metodě je Pairwise, kdy jsou z analýzy vypouštěné pouze řádky, které obsahují chybějící hodnoty a jsou využívány v právě probíhající analýze. Toto může být ve výsledku matoucí, protože se pro výpočet různých hodnot využívají soubory o různém rozsahu. Doporučuje se využít pro malé soubory nebo pro soubory s vysokým podílem chybějících hodnot (PETRÚŠEK, 2013), (VALENTOVÁ, 2013).

Pokud se rozhodnou zpracovatelé, že nahradí chybějící data konkrétními hodnotami, naskýtá se zde několik možností.

- Nahrazení chybějících hodnot aritmetickým průměrem – jedná se o jednoduchý způsob, kdy se chybějící hodnoty nahradí průměrnou hodnotou zbylých hodnot dané proměnné. Tato metoda má řadu omezení, nelze ji např. využít za předpoklad, že je vysoký podíl chybějících údajů, je vysoká variabilita údajů (existují extrémní pozorování).
- Nahrazení mediánem, modem, maximální nebo minimální hodnotou – způsob velmi podobný nahrazení průměrem. Lze využít pro nominální proměnné, např. místo minimální hodnoty se dosadí hodnota s nejnižší četností.
- Nahrazení skupinovým průměrem – nejprve je potřeba hodnoty proměnné s chybějícími údaji rozdělit do skupin podle hodnoty jiné proměnné, následně je v těchto skupinách vypočten průměr / modus / medián a touto hodnotou se nahradí chybějící hodnota. Klíčovým prvkem této metody je volba skupinové proměnné, ideální je zvolit takovou proměnou, aby vytvořené skupiny byly uvnitř co nejvíce homogenní.

- Nahrazení chybějících údajů podle vzoru – hodnoty vybraných proměnných u respondenta, u něhož chybí údaj, jsou porovnávány s hodnotami těchto proměnných u jiných respondentů. Pokud se podaří nalézt respondenta se stejnými hodnotami ostatních proměnných, tak se chybějící údaj nahradí podle něho, jinak lze vzít v úvahu jiný set proměnných a celou akci opakovat nebo vybrat respondenta náhodně.
- Nahrazení chybějících údajů na základě metod regresní analýzy – z existujících hodnot jsou odhadnuty parametry modelu, vysvětlujícího hodnoty určité proměnné na základě hodnot jiných proměnných. Lze využít pouze v případě, že se jedná o numerické proměnné.

Volba jedné z těchto metod závisí primárně na situaci a charakteru dat (VALENTOVÁ, 2013), (PETRÚŠEK, 2013).

## 3.5 Příklady statistických šetření

V této kapitole bude uvedeno několik statistických šetření, které probíhají / proběhli na českém území.

### 3.5.1 Výběrové šetření pracovních sil (VŠPS)

V prosinci roku 1992 bylo v ČR poprvé provedeno toto šetření. VŠPS je primárně zaměřeno na sledování informací o trhu práce. Je prováděno v domácnostech bydlících v náhodně vybraných bytech. Pomocí VŠPS lze kvalifikovaně odhadnout výši zaměstnanosti. Zjišťuje se struktura zaměstnanosti dle pohlaví, věku, kvalifikace odvětví a charakteru zaměstnání respondentů. Dále lze pomocí šetření zjistit informace o celkové nezaměstnanosti a jejím charakteru z hlediska profesního, sociálního a kvalifikačního. Po vstupu ČR do EU byla metodika přizpůsobena mezinárodním požadavkům tak, aby šlo porovnávat výsledky v rámci EU. Rozsah šetření a ukazatele zaměstnanosti a nezaměstnanosti plně odpovídají definicím Mezinárodní organizace práce a metodickým doporučení Eurostatu (ČSÚ, 2015).

Výběr respondentů je prováděn náhodným výběrem z Registru sčítacích obvodů. Předmětem šetření jsou osoby obvykle bydlící v domácnostech šetřených bytů. VŠPS se v současnosti šetří pomocí elektronického dotazníku, který obsahuje soubor otázek a možných odpovědí. Každá vybraná jednotka zůstává v souboru pět po sobě jdoucích čtvrtletí, tedy se každé čtvrtletí obmění 20 % souboru. Šetření probíhá v cca 26000 domácnostech na celém území ČR, ale nevztahuje se na osoby bydlící např. v hromadných ubytovacích jednotkách. Výsledky VŠPS jsou vždy průměrné údaje za čtvrtletí a jsou pravidelně prezentovány na stránkách ČSÚ (JEŘÁBKOVÁ, 2012), (ČSÚ, 2015).

### 3.5.2 Statistika rodinných účtů (SRÚ)

SRÚ poskytuje informace o výši vydání a struktuře spotřeby soukromých domácností. Pomocí tohoto šetření lze získat informace o odlišnostech spotřeby v domácnostech dle různých hledisek, případně rozličných faktorů např. pohyb cen, situace na trhu. Využití výsledků tohoto šetření slouží jako podklad pro kvalifikované rozhodování při realizaci sociální politiky státu, pro sociální a ekonomický výzkum, interní využití ČSÚ (tvorba spotřebitelských košů při revizi indexu spotřebitelských cen) (ČSÚ, 2012).

Domácnosti do tohoto šetření jsou vybírány kvótním výběrem. Domácnosti jsou ve zpravodajském soboru celý rok za předpokladu že se nezmění některá z jejích klíčových výběrových

charakteristik např. ekonomická aktivita osoby v čele. Jednotkou výběru a zpravodajskou jednotkou šetření je hospodařící domácnost, tedy soubor osob společně bydlících, které se společně podílejí na uhrazování základních výdajů (výživa, provoz, domácnosti, údržba bytu). Celkově je do tohoto šetření zapojeno 3000 domácnosti a jejich struktura je konstruována tak, aby jejich složení odpovídalo struktuře domácností v ČR. Existuje zde i doplňkový soubor, kde se nachází 400 domácností s minimálním příjmem. Oporou pro stanovení kvót jsou v tomto případě výsledky šetření Životní podmínky (EU-SILC) (JEŘÁBKOVÁ, 2012), (ČSÚ, 2012).

Prvotní údaje o peněžních příjmech, výdajích a spotřebě domácností se zjišťují tak, že vybraná domácnost denně zapisuje veškeré peněžní i naturální příjmy a výdaje za všechny členy domácnosti do „Deníku zpravodajské domácnosti“. V zájmu snížení zátěže respondentů a zefektivnění vynaložených nákladů vede každá domácnost od roku 2006 zápisy vydání pouze za potraviny a nealkoholické nápoje pouze dva měsíce v roce, v ostatních měsících sdělují pouze celkovou sumu za daný měsíc. Domácnostem je za řádně vyplněné a úplné záznamy vyplácí peněžitá odměna (ČSÚ, 2012).

### 3.5.3 Šetření životních podmínek (EU-SILC)

Primárním účelem tohoto šetření je zkoumání životních podmínek a příjmů domácností a jednotlivců. Výsledky tohoto šetření slouží jako podklad pro výpočet ukazatelů peněžní a materiální chudoby. V ČR je toto šetření povinné od roku 2005, je to dané nařízením EU. Toto šetření probíhá na území celé EU a v několika dalších státech (Norsko, Turecko, Island a Švýcarsko). Díky tomu existují srovnatelné údaje o sociální a ekonomické situaci domácností v EU (JEŘÁBKOVÁ, 2012).

Výběrovou jednotkou pro toto šetření je byt. Volba těchto jednotek probíhá náhodným dvoustupňovým výběrem, kdy jsou nejprve vybrány sčítací obvody a následně v každém z nich 10 bytů jednoznačně identifikovaných přesnou adresou a číslem bytu v domě. V šetření se používá rotační panel, což znamená, že vybrané domácnosti jsou navštěvovány v ročních intervalech po dobu 4 let. Každoročně se odmění asi čtvrtina domácností. Díky tomuto dlouhodobému pozorování jde pozorovat změny a vývoj sociální a ekonomické situace jednotlivých sledovaných jednotek. Pokud se některá sledovaná osoba přestěhuje, podléhá šetření na nové adrese. Výběrový soubor obsahuje 11000 bytů (ČSÚ, 2012).

Samotné zjišťování údajů probíhá formou osobního rozhovoru respondenta s tazatelem, který zjištěné údaje zaznamenává do připravených dotazníků (papírová i elektronická podoba). Šetření se skládá ze tří modulů (byt, osoby 16+ a domácnost), navíc je zde každý rok speciální modul na který je EU-SILC zaměřeno (2009 – sociální a materiální deprivace, 2010 – správa financí, 2011 – mezigenerační přenos znevýhodnění) (ČSÚ, 2012), (JEŘÁBKOVÁ, 2012).

### 3.5.4 Sčítání domů lidí a bytů (SDLB)

SDLB je vyčerpávající šetření prováděné na celém území státu. Toto šetření má v ČR dlouhou tradici, první sčítání proběhlo již v roce 1754. S postupem času se i zlepšovala kvalita sčítání. Za zlom, co se týče kvality sčítání lidu, je považován rok 1857, protože tehdy byl proveden první krok k modernímu sčítání lidu. Další důležitý mezník nastal o 12 let později, kdy bylo provedeno první sčítání lidu, kde byly podrobně definovány znaky zjišťované u obyvatelstva např. státní příslušnost, rodinný stav, náboženství, atd. Dále bylo stanoveno, že sčítání bude prováděno každých 10 let, což se s výjimkou válečného roku 1940 víceméně daří až do současnosti. Po vzniku ČSR (1918) a založení státního statistického úřadu (1919) byla touha pořádat sčítání každých 10 let v „desítkový“ rok, což se

ne vždy z různých důvodů podařilo (výjimky 1921, 1961 + novodobé sčítání 1991, 2001, 2011) (ČSÚ, 2012).

Poslední SDLB proběhlo v roce 2011, sčítání bylo provedeno ČSÚ a jeho smluvním partnerem byla Česká pošta, která zajišťovala terénní práce. Toto sčítání přineslo celou řadu novinek souvisejících především s rozvojem informačních technologií. Sčítací formuláře bylo možné vyplňovat elektronicky na internetu a odesílat je on-line nebo prostřednictvím datových schránek. Formuláře již neobsahovaly dotazy na vybavenost domácnosti (otázky vlastnictví osobního automobilu, pračky, telefonu a dalších věcí), jako tomu bylo v předchozích sčítáních. Z hlediska vybavenosti byla důležitá pouze otázka, zda má domácnost možnost využívat PC a má k dispozici internetovou přípojku. ČSÚ poprvé v historii přešel od tradiční metody sčítání ke kombinované metodě s využitím registrů (formuláře byly předvyplněny některými údaji). Byla zde i snaha započítat lidi bez přístřeší žijící v různých azylových domech. Nicméně podobně jako v roce 2001 stálo toto SDLB přes 2 miliardy korun a existuje zde řada kritiků tohoto šetření. Dokonce zde existuje možnost, že v roce 2021 již nebude SDLB provedeno jako vyčerpávající šetření (ČSÚ, 2014).

### 3.5.5 Další statistická šetření

Existuje celá řada statistických šetření, jež probíhají na našem území. Za zmínku stojí výběrové šetření o IT, jehož cílem je zjistit vybavenost domácností PC, internetem a případně zjistit, k čemu je využíván (internetové nakupování, práce, škola, apod.). Toto šetření je nasazeno jako modul ve 2. čtvrtletí v rámci VŠPS. Další pravidelné šetření, jež je realizováno čtvrtletně, je Výběrové šetření o cestovních ruchu, kde se zkoumá mj. participace členů domácnosti na cestovním ruchu, náklady na cestování a realizované cesty členů domácnosti. Další výběrové šetření pořádané na našem území je VŠ zdravotně postižených osob. Cílem tohoto šetření je získat základní charakteristiky zdravotně postižených (rodinný stav, druh bydlení, typ/délku/míru/důsledky ZP apod.). Toto šetření zde proběhlo v letech 2007 a 2013 (JEŘÁBKOVÁ, 2012).

Ústav zdravotnických informací a statistiky v ČR (ÚZIS) pořádá výběrové šetření o zdravotním stavu české populace (EHIS). Toto šetření zde proběhlo v letech 1993, 1996, 1999, 2002, 2008 a naposled v roce 2014, jako součást Evropského výběrového šetření o zdraví. Cílem tohoto šetření je získat obecnou představu o zdraví obyvatelstva ve zkoumaných zemích. Je složeno z celé řady komplexních otázek, mezi které patří např. frekvence návštěvy u lékaře, konzumace léků, kouření, dopravní nehody, zrak apod. (JEŘÁBKOVÁ, 2012).

## Kapitola 4

# Data miningová úloha

V této kapitole bude provedena reálná DM úloha. Jedná se o praktickou část této diplomové práce. Struktura této kapitoly bude kopírovat metodiku CRISP-DM popsanou v předchozí části práce. Tato metodika bude mírně upravena tak, aby odpovídala profilu dané úlohy. Kapitola bude modifikována tak, že v ní bude popsáno statistické šetření, ze kterého pochází testovaná data. Dále v této kapitole bude vymezeno zadání této úlohy.

V kapitole porozumění datům budou pomocí metod deskriptivní statistiky popsány základní charakteristiky vybraného datového souboru. V následující kapitole příprava dat budou vypsány veškeré změny, jež bylo nutné v daných provést, tak aby splňovaly požadavky jednotlivých DM algoritmů v modelovací fázi procesu. V kapitole modelování pak bude popsána tato fáze a veškeré modely a algoritmy pro zhotovení této úlohy. V kapitole hodnocení výsledků pak budou zmíněny zajímavé výsledky získané během DM procesu.

V této části práce budou využívány volně dostupné SW nástroje, které byli zmíněny v kapitole 2.5. Pro předpracování dat se bude jednat zejména o SW nástroj Knime. Tento nástroj jsem zvolil zejména proto, že zde nebyl žádný problém při práci s rozsáhlými datovými soubory (jako např. v nástroji Orange). V rámci modelování budou využit pro první část úlohy SW nástroj Weka, protože z volně dostupných nástrojů obsahuje nejvíce klasifikačních metod. Pro druhou část úlohy bude pak využit nástroj LispMiner, protože je specializovaný na daný typ DM analýz.

## 4.1 Porozumění problematice

V této kapitole bude zdůvodněn výběr dat pro testování, popsán jejich zdroj a také poskytnuty základní údaje. Dále bude v této kapitole upřesněno zadání jednotlivých data miningových úloh.

### 4.1.1 Zdroj dat

Jedním z cílů této práce je demonstrovat možnosti DM technik při zpracování / vyhodnocování dat ze statistických šetření, nejlépe ze šetření s „demografickou“ tematikou. Jako jeden z hlavních zdrojů dat pro demografii slouží SLDB, proto by bylo logické využít pro tuto demonstrativní úlohu data právě z tohoto šetření. Ideální data pro DM úlohu jsou data za jednotky, v případě SLDB by se jednalo o data za jednotlivé obyvatele. ČSÚ tato data z mnoha, asi celkem logických důvodů (ochrana osobních údajů), na svých stránkách neposkytuje. K dispozici jsou pouze většinou agregovaná v rámci některých publikací, které ČSÚ vydává po každém sčítání. Vzhledem k faktu, že mým cílem je v rámci této DP vytvořit pouze modelovou úlohu, má tato úloha předem

nejasný výsledek v tom smyslu, zda se podaří získat nějaké zajímavé výsledky na základě zde provedených analýz. Rozhodl jsem se tedy, že se před absolvováním složitého byrokratického procesu při snaze získat data z českého SLDB porozhlédnu, zda nelze získat podobná data z nějakého šetření, které proběhlo na území jiného státu.

Shodou okolností jsem narazil na projekt IPUMS (Integrated Public Use Microdata Series). V rámci tohoto projektu jsou k dispozici po registraci (která je zdarma) data ze sčítání lidu z více než 70 zemí z celého světa (jak za osoby, tak za domácnosti). Většinou se jedná o náhodné výběry (nejsou k dispozici úplná data). U každé země jsou tyto výběry rozdílně obsáhlé, např. data z Francie za rok 2006 obsahují informace o 20M osobách (řádově 33 % populace Francie), data za Čínu 1990 obsahují informace o 10M osobách (řádově 1 % populace Číny)[7]. Pro většinu zemí jsou k dispozici data z přelomu tohoto tisíciletí. Data za ČR nejsou v rámci tohoto projektu k dispozici (IPUMS, 2010).

Data z tohoto projektu jsou k dispozici zdarma, pokud potenciální uživatel splní následující podmínky.

> ***Za poskytnutí dat si nebude účtovat žádné poplatky.***
>
> ***Bude citovat IPUMS dle přesné citace uvedené na stánkách IPUMS.***
>
> ***Dá vědět autorům o tom, že využil data z tohoto projektu tak, že jim zašle výslednou publikaci, aby ji mohli přidat do interní bibliografie.***
>
> ***Nebude tato data využívat pro genealogický výzkum.***
>
> ***Bude data využívat ke konání dobra, nikoliv zla.***
>
> ***Pokud narazí na chyby v datech, dá autorům vědět na příslušnou mailovou adresu.***

***Zdroj: (IPUMS, 2010)***

Domnívám se tedy, že mi nic nebrání v tom využít data z tohoto projektu v rámci výzkumu v mé diplomové práci. Nyní je potřeba vybrat zemi, na které budou provedeny analýzy v další fázi práce. Mým cílem je analyzovat co nejaktuálnější data, z evropských zemí nejsou data starší než 10 let pouze u Irska. Relativně novější data jsou k dispozici z několika afrických zemí a jihoamerických zemí. Nicméně jako nejvhodnější kandidát mi připadá USA. IPUMS je projektem Univerzity v Minneapolis, tudíž předpokládám, že data za USA budou nejkvalitnější a nejpřesnější. Další výhodou je, že v USA probíhá obdoba SDLB každý rok od roku 2001. Tudíž jsou k dispozici data za více po sobě jdoucích let (u zbylých zemí se jedná pouze o data za maximálně 5 šetření s 5 nebo 10 letou periodou). Toto šetření se oficiálně nazývá American Community Survery (ACS) (IPUMS, 2010).

## American Community Survey

Toto výběrové šetření probíhá každoročně na území Spojených států amerických. Měsíčně je prošetřeno 250 tisíc adres (tedy ~3 miliony ročně). Souběžně s tímto šetřením probíhá v USA každých 10 let i vyčerpávající šetření obdobné českému SLDB (Decennial census). ACS slouží jako doplňkové šetření k desetiletým cenzům a poskytuje obdobné informace za 1 % obyvatelstva. Návratnost tohoto šetření je řádově 66 %, tedy 2 / 3 vybraných adres vrátí zpět dotazníky. Způsob výběrů jednotek do tohoto šetření probíhá pomocí stratifikovaného výběru. Co se týče metody sběru dat, jsou využívány

[7] Odkaz se seznamem veškerých datasetů v tomto projektu https://international.ipums.org/international/samples.shtml#br

tři možnosti. Nejprve je potřeba poslat vyplněný iniciační dotazník poštou (vyplnění bez asistence), poté následuje CATI a nebo CAPI (to provádí výzkumníci ze statistického úřadu) s cílem získat dodatečné informace. Díky tomuto výzkumu lze získat každoročně odhady jednotlivých charakteristik za větší územní celky (autoři této studie uvádí za územní celky s více než 65 tisíci obyvateli). Pro získání informací za menší celky by bylo potřeba kombinovat výsledky šetření za několik po sobě jdoucích let (tři roky za celky o velikost 20 – 65 tisíc obyvatel a 5let za celky s méně než 20 tisíci obyvateli) (CENSUS.GOV, 2014).

### 4.1.2 Zadání úlohy

Cílem této práce je demonstrace využití DM metod při zpracování dat ze statistických šetření. Proto jsem se rozhodl pro výběr dvou typických DM úloh. První z nich je klasifikační / predikční typ úlohy. A druhou úlohu jsem nazval hledání nuggetů, cílem této úlohy je nalézt zajímavé nebo nové znalosti v datech za pomoci asociačních pravidel.

#### Klasifikační / predikční úloha

Predikce znamená předpověď, jedná se o odhad budoucích hodnot na základě hodnot jiných, většinou minulých. Cílem tohoto typu úloh je nalézt znalosti, které pomáhají nová data klasifikovat (roztřídit) do určitých předem určených skupin. Podstatou tohoto typu úlohy je zvolení si nějakého atributu a následné sestavení modelu, jež znázorní vliv ostatních atributů na zvolený cílový atribut.

Dlouho jsem přemýšlel jaký atribut zvolit, tak aby úloha dávala smysl pro data z výběrového šetření typu SLDB. K dispozici je celá řada atributů, v rámci ACS se jedná o téměř 100 unikátních atributů za domácnosti (geografické, ekonomické, vybavení domácnosti, uspořádání domácnosti) a ještě více než dvojnásobek atributů za trvale bydlící obyvatelstvo, rovněž z několika kategorií, např. rodinný stav, základní demografické údaje, rasa a náboženství, vzdělání, práce, příjmy, zdravotní pojištění, postižení, status válečného veterána, dojezdy do práce, migrace a další. Seznam vybraných atributů do této úlohy se základní charakteristikou je k dispozici v kapitole 4.2.2.

Cílem je, aby tato úloha měla i nějaké praktičtější využití, aby to nebyla pouhá demonstrace možností a případné úspěšnosti jednotlivých klasifikátorů na zadaných datech. Jako vhodný typ úlohy se mi zdá využít klasifikační algoritmy pro predikci chybějících hodnot. Tímto problémem se zabýval (Pejčoch, 2011). Jde o to zjistit, zda a případně s jakou úspěšností by šlo klasifikovat / predikovat jednotlivé hodnoty chybějících atributů na datech z ACS. Jedná se o modifikaci postupu, který byl popsán v kapitole 3.4.2 o postupech při práci s chybějícími daty. Jedná se o nahrazování chybějících údajů na základě modelu sestaveného pomocí klasifikačních algoritmů. Mým cílem v rámci této úlohy je vybrat atributy, které mají relativně vysoký podíl chybějících hodnot a zjistit s jakou úspěšností lze hodnoty těchto atributů predikovat z hodnot ostatních atributů. Nutno podotknout, že v rámci IPUMS není přístup k raw datům, tedy prvotnímu zpracování jednotlivých dotazníků. Lze tedy předpokládat, že při zpracování dat před publikací dat v rámci IPUMS již mohlo dojít k nějakému procesu zlepšování datové kvality.

Klasickým příkladem této proměnné je výše platu. Jedná se o jednu z hodnot, která má obecně vyšší podíl nevyplnění. Lidé s vysokým platem se často bojí výši platu udat ze strachu (možnost krádeže), případně ze závisti okolí, naopak lidé s nižším platem si tyto hodnoty mohou upravovat, případně je neuvádět naopak ze studu. Myslím si, že je teoreticky možné odvodit výši platu na základě ostatních proměnných, které se sbírají v rámci tohoto výzkumu, např. dosažené vzdělání, typ průmyslu, ve kterém je zaměstnán, počet odpracovaných týdnů / hodin atd.

Dalšími dvěma proměnnými, jež jsem se rozhodl otestovat při klasifikaci v této práci, jsou: „typ domácnosti“, a zda se domácnost nachází v metropolitní oblasti. Obě tyto proměnné měly podíl chybějících údajů na úrovni 10 %. Klasifikační úloha bude rozdělena na tři podúlohy. V rámci každé podúlohy proběhne klasifikace jedné ze tří vybraných proměnných. Každá podúloha bude vyhodnocena nad speciálně upravenými datovými soubory (jejich popis bude proveden v kapitole 4.3.2). Klasifikace proběhne pro minimálně dva různé sety proměnných. Pro každou podúlohu bude vytvořen aspoň jeden speciálně upravený datový soubor, který bude obsahovat pouze proměnné, u kterých lze očekávat nějaký vliv na cílovou proměnnou. Jeden z datových souborů bude vždy obsahovat většinu vybraných proměnných do této práce (seznam vybraných proměnných se nachází v kapitole 4.2.2), aby se prokázalo, zda selektování proměnných zvýší úspěšnost klasifikace.

### Hledání nuggetů

Hledání nuggetů je obecně používaný název pro typ úlohy dobývání znalostí. Cílem této úlohy je nalézt nové, překvapivé znalosti, které nemusí pokrývat celý „koncept“ jako v případě klasifikační úlohy. Pro tento typ úloh jsou data z výběrového šetření ACS naprosto ideální, protože v těchto datech je hodně atributů (většinou kvalitativního typu s vysokým počtem možných hodnot) a vysoký počet pozorování. Existuje lze tedy velký předpoklad pro nalezení nějakých zajímavých hypotéz. Nejčastěji využívaná DM technika pro tento typ úlohy jsou asociační pravidla. Ideální dle mého názoru je využít metodu GUHA, která byla představena v kapitole 2.4.2.

Nejčastějším postupem při hledání zajímavých pravidel pomocí této metody je sestavení obecné analytické otázky. Obecná analytická otázka má následující podobu:

> ***?: M; $A \wedge B \approx C$***
>
> ***Jsou v matici dat M nějaké zajímavé vztahy mezi kombinacemi atributů ze skupin A a B na straně jedné a kombinacemi atributů ze skupiny C na straně druhé?***

Na tento typ analytických otázek se poté následně „hledá odpověď“ za pomoci 4FT kvantifikátorů (základní kvantifikátory byly popsány v kapitole 2.4.2). Obvykle na tvorbě těchto otázek spolupracuje analytik s autorem dat (zadavatelem úlohy) a snaží se najít odpověď pouze na nějaké klíčové problémy (RAUCH et. al., 2015).

Vzhledem k faktu, že zde v roli zadavatele úlohy působím já, jakožto autor práce, rozhodl jsem se sestavit několik analytických otázek. Jedná se o modelové otázky a mým cílem je sestavit tak, abych naznačil základní možnosti využití vybrané metody při zpracování dat ze statistických šetření. Bude provedeno celkem 6 různých analýz. Jejichž seznam včetně slovně definované analytické otázky je v následné tabulce.

***Tab. 7 : Seznam analytických otázek***

| Č. analýzy | Formulace analytické otázky |
|---|---|
| 1 | Je možné zařadit pracovníky v nějakém typu průmyslu (IND) a na určité pracovní pozici (OCC) do jedné nebo dvou platových skupin |
| 2 | Je možné určit počet odpracovaných hodin v nějakém typu průmyslu (IND) a na určité pracovní pozici (OCC). |

| | |
|---|---|
| 3 | Je možné nalézt nějaké společné rysy (pojištění, ekonomická aktivita, příjmová skupina sociálních dávek, hledání práce v posledních 5 letech, vzdělání) pro respondenty s různými typy zdravotních postižení. |
| 4 | Je možné určit respondentův dopravní prostředek, který využívá k přepravě na základě jeho rasy, státu, rodinného stavu nebo toho zda žije v centru nebo na okraji města |
| 5 | Zajímají nás státy, ve kterých je nadprůměrný výskyt respondentů nějaké rasy. |
| 6 | Zajímají nás státy, ve kterých je nadprůměrný výskyt respondentů na základě kombinace věku, rodinného stavu a typu domácnosti ve kterém žijí. |

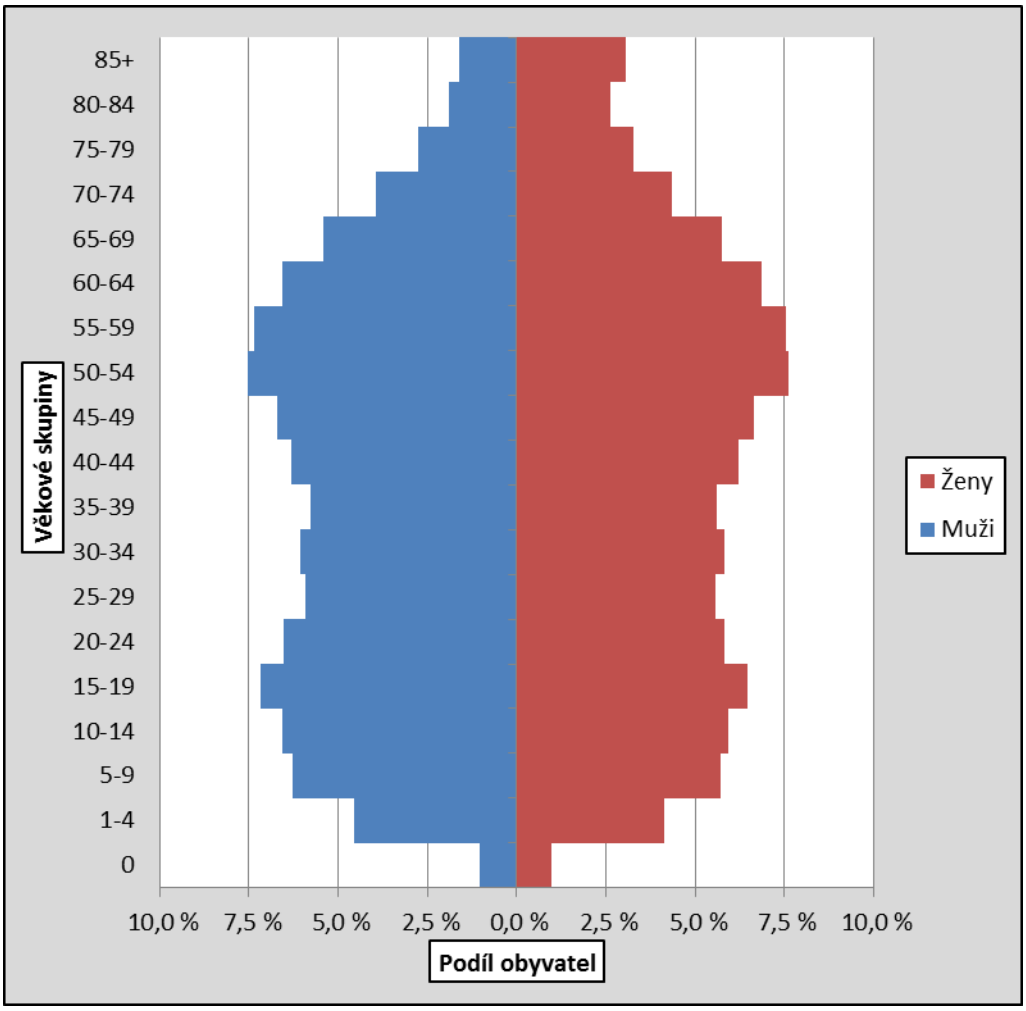
***Zdroj: vlastní zpracování***

## 4.2 Porozumění datům

V této kapitole bude nejprve za využití popsány základní demografické charakteristiky dat (věková struktura, vzdělanostní struktura, ekonomická aktivita a rodinný stav respondentů). Ve druhé části kapitoly budou popsány jednotlivé atributy, které byly vybrány do této úlohy.

### 4.2.1 Obecné charakteristiky

***Graf 1: Věková struktura respondentů ACS 2013 - 5leté intervaly***



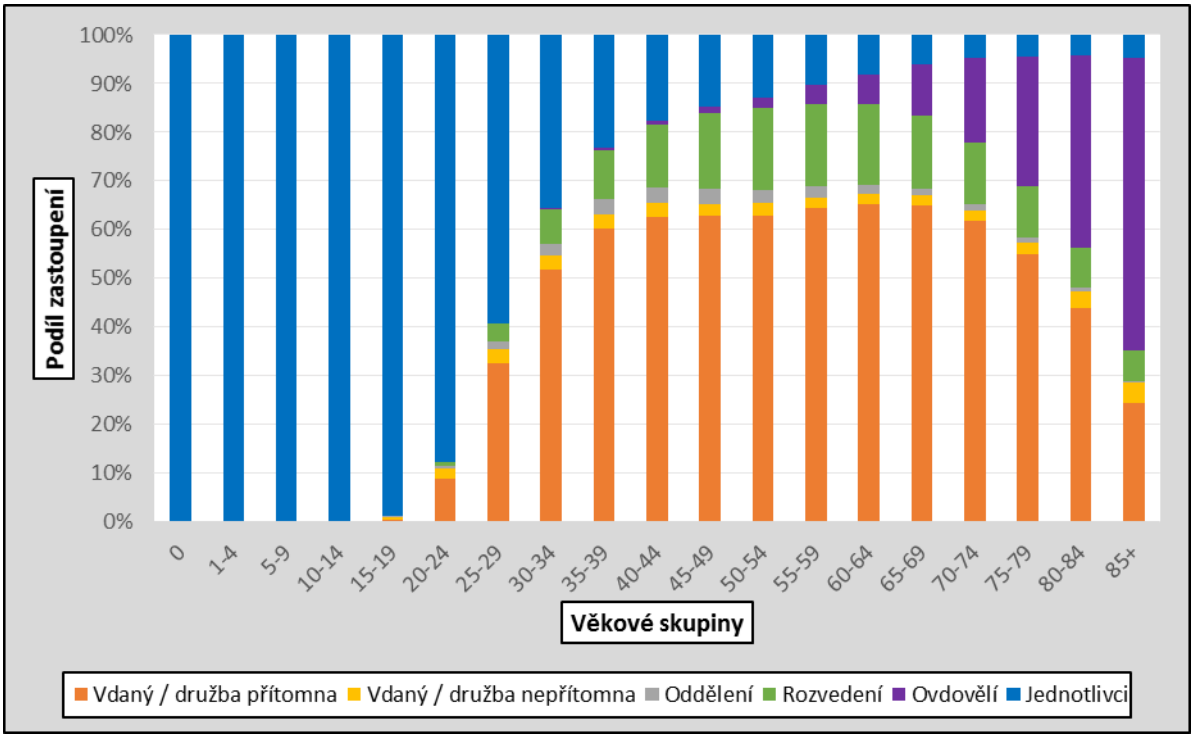

***Zdroj: (RUGGLES et. al., 2015), vlastní úpravy***

Na grafu 1 je znázorněna pomocí věkové pyramidy věková struktura účastníků průzkumu ACS v roce 2013. Lze si všimnout trendu stárnutí populace, nejčastěji zastoupená věková třída je 50–54let. V grafu si lze všimnout slabých populačních ročníků 1972–1979, kdy počet narozených v USA nedosáhl ani na hranici 3,5M narozených dětí za rok, což se od té doby neopakovalo až do současnosti. V současnosti se v USA rodí řádově 4M dětí ročně. Věkový průměr účastníků výzkumu byl 40,5 roku a medián je 41 let.

Na grafu 2 je podílový graf dle věku a rodinného stavu. Lze si všimnout obecného trendu ve vyspělých zemí s odkladem sňatků, ve skupině 20 – 24 letých je pouze 10 % obyvatel v manželském svazku. Věkový medián při prvním manželství je u mužů v USA 29,1 roku a u žen 26,9. Nikoho asi

nepřekvapí nárůst podílu ovdovělých ve skupinách s věkem 60 a více let. Zajímavostí je že v USA zaznamenávají obyvatele, jež žijí odděleně a se statusem rozvedený, vzájemný podíl těchto ukazatelů s věkem klesá, což by mohlo naznačovat nový trend, že mladší lidé radši žijí odděleně než, aby se rozváděli.

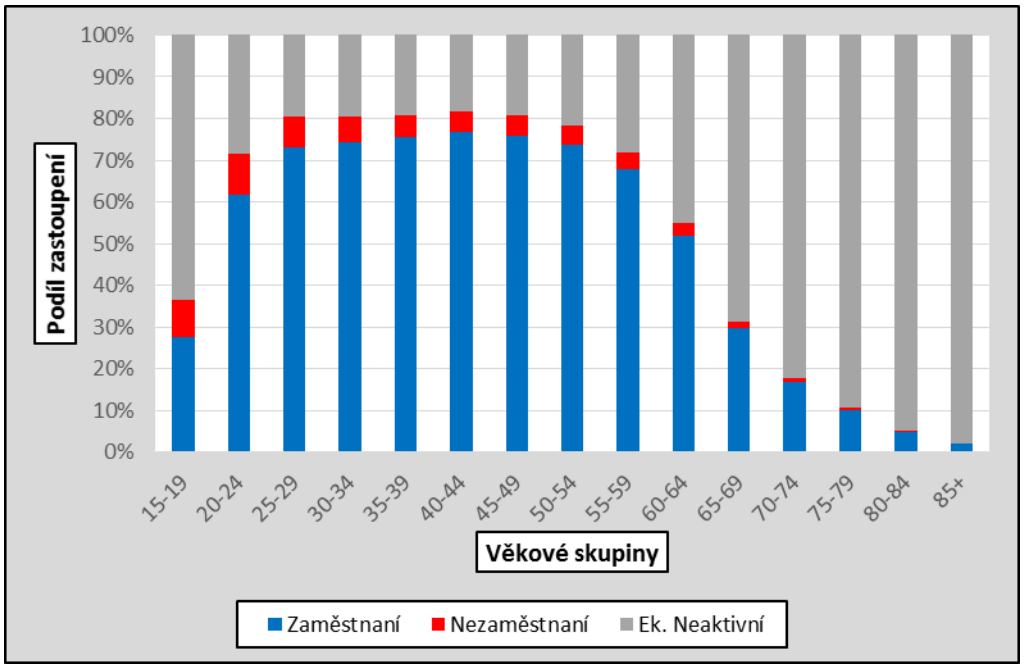
***Graf 2: Rodinný stav respondentů ACS 2013 dle věkových skupin***




***Zdroj: (RUGGLES et. al., 2015), vlastní úpravy***

Míra nezaměstnanosti byla v roce 2013 v USA 7,9 %. Na grafu 3 je znázorněn graf ekonomické aktivity. Mezi ekonomicky aktivní lze zařadit zaměstnané a nezaměstnané obyvatelstvo. Podíl ekonomicky neaktivních lidí ve skupině 15–65 let v roce 2013 činil 38 %. K ekonomické neaktivitě může vést řada důvodů, zejména se jedná o lidi v domácnosti, ve škole nebo obecně neschopné práce. V rámci ACS 2013 bohužel není k dispozici tato klasifikace (i přesto, že v číselníku jsou na to vymezeny speciální kódy u proměnné), podobně jako nelze klasifikovat nezaměstnané na absolventy a lidi, co již pracovali. Průměrný roční mzda skupiny „zaměstnaní“ činila v roce 2013 49 tisíc dolarů, medián byl 35 tisíc dolarů. Průměrný roční příjem nezaměstnaných byl 12 800 dolarů, medián příjmu byl 4 tisíce dolarů, do tohoto ukazatele mj. zahrnují sociální dávky, prodeje / pronájem nemovitostí apod.

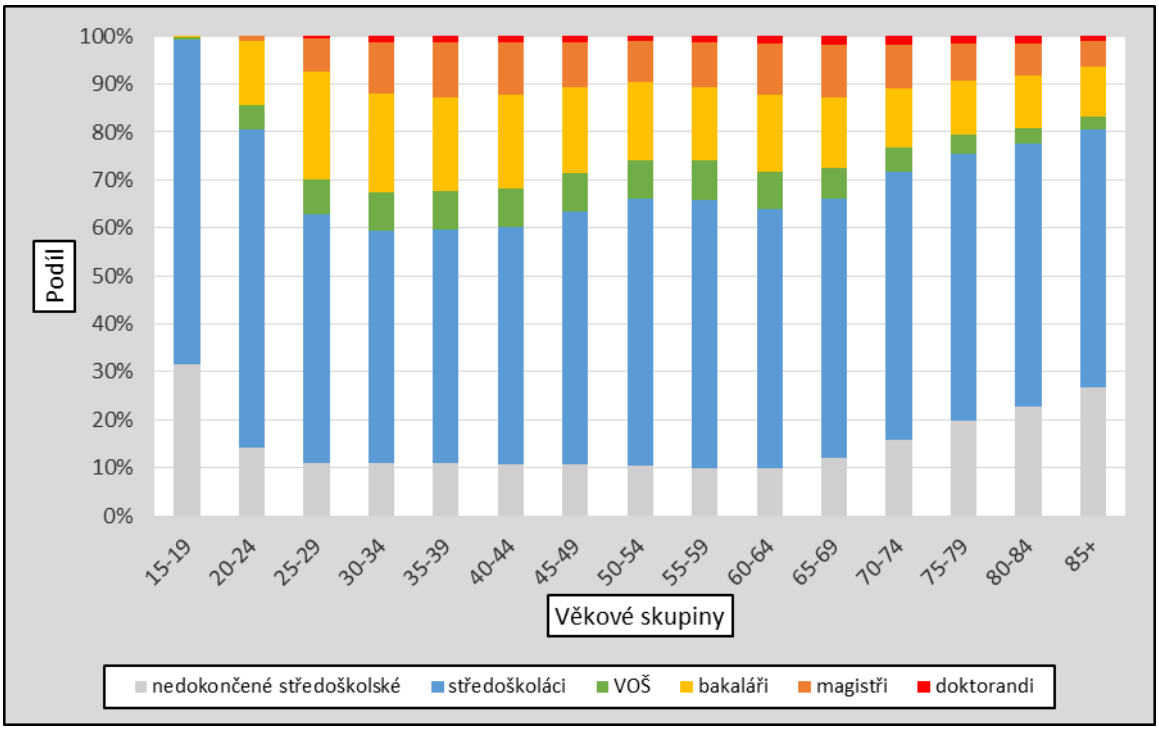
***Graf 3: Ekonomická aktivita respondentů ACS 2013 dle věkových skupin***




***Zdroj: (RUGGLES et. al., 2015), vlastní úpravy***

Na grafu 4 je znázorněn graf úrovně vzdělanosti dle věkových skupin. V grafu jsou zahrnuty pouze osoby, které na otázku v šetření týkající se aktuálního studia odpověděli „ne“. V USA mají odlišný vzdělávací systém od toho našeho. Otázka týkající se dosaženého vzdělání byla velmi podrobná, obsahuje odpovědi na nejvyšší vystudovaný rok. Skupina „nedokončené středoškolské“ vzdělání obsahuje všechny osoby, jež uvedli jinou odpověď než, že vystudovali s diplomem střední školu (tedy všechny ty co odpadli v nějakém od 1. do 12. ročníku základního / středoškolského studia. Do skupiny „středoškoláci“ byli zahrnuti všichni ti, co uvedli dokončené středoškolské vzdělání a že odstudovali 1 nebo více let na VŠ bez dosaženého diplomu. Ve skupině „VOŠ“ jsou absolventi komunitních a většinou dvouletých více specializovaných oborů na tzv. „junior college“, tyto školy považuji za ekvivalent českých vyšších odborných škol. Na grafu si lze všimnout, že podíl lidí s nedokončeným středoškolským vzdělání se zvyšuje s přibývajícím věkem, což naznačuje, že se začal klást větší důraz na vzdělání. Také se zvýšil podíl lidí s vyšším dokončeným vzdělání než středoškolským, tento rozdíl lze zaznamenat u všech věkových skupin 25–44 let, jedná se zejména o nárůst podílu lidí s bakalářským titulem.

***Graf 4: Ukončené studium respondentů ACS 2013 dle věkových skupin***




***Zdroj: (RUGGLES et. al., 2015), vlastní úpravy***

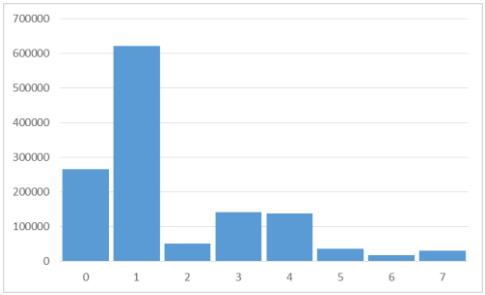
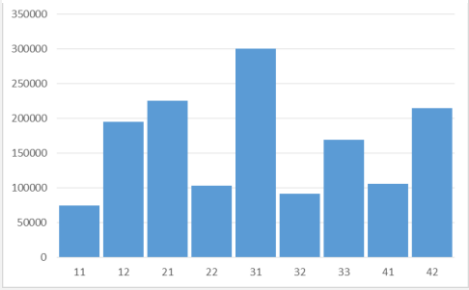
### 4.2.2 Technické parametry

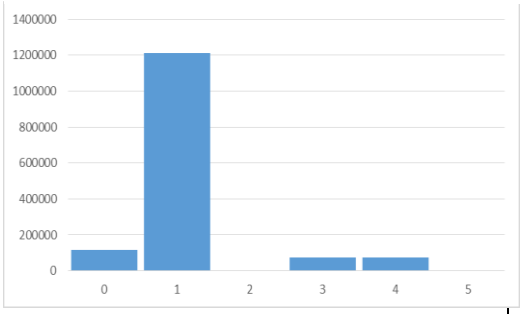
V této kapitole budou popsány veškeré atributy zahrnuté do testování. Byly vybrány hodnoty, kterého dle mého názoru jsou zajímavé pro vybraný typ úloh. Bude zde uveden název, typ atributu, popis, případně uveden obor hodnot (pokud bude u proměnné méně než 10 unikátních hodnot bude poskytnut jejich seznam a případný graf rozdělení hodnot u kvantitativních proměnných bude k dispozici univerzální histogram). Na vertikální ose u těchto grafů bude vždy uveden počet pozorování. Atributy budou rozděleny do dvou skupin (atributy za domácnosti a atributy za jednotlivé osoby). V rámci úspory místa v tabulce nebude u každého grafu uveden zdroj. Zdroj pro všechny tyto grafy, popis dat a hodnoty atributů je (RUGGLES et. al., 2015).

Vzhledem k faktu, že autoři IPUMS využívají kódových číselníků, odpadá zde problém zkoumání datové kvality, případně špatně zadaných údajů. V tabulkách 8 a bude popsáno celkem 48 proměnných, které jsem vybral z celkového počtu 214 proměnných, které jsou k dispozici v rámci IPUMS.

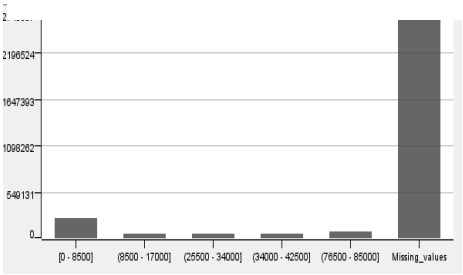
***Tab. 8: Popis vybraných atributů týkajících se domácností***

| Název proměnné | Typ proměnné | Popis |
|---|---|---|
| HHTYPE | Nominální | Typ domácnosti ve smyslu zda se jedná o rodinnou / nerodinnou domácnost, případně jaká osoba je v čele domácnosti |

<table>
<tr><td>Obor hodnot:</td><td colspan="2">[0] údaj není k dispozici<br>Rodinná domácnost<br>[1] Manželský pár<br>[2] Pouze muž + děti<br>[3] Pouze žena + děti<br>Nerodinná domácnost<br>[4] Muž žijící sám<br>[5] Muž v čele nežijící sám<br>[6] Žena žijící sama<br>[7] Žena v čele nežijící sama<br>Graf 5: [HHTYPE] - četnosti<br>  </td></tr>
<tr><td>REGION</td><td>Nominální</td><td>Region, ve kterém je domácnost umístěna. USA je v rámci tohoto výzkumu rozděleno celkem na 4 hlavní regiony a 9 oblastí</td></tr>
<tr><td>Obor hodnot:</td><td colspan="2">Severovýchodní Region<br>[11] Nová Anglie<br>[12] Středoatlantická oblast<br>Středozápad<br>[21] Severovýchodní část<br>[22] Severozápadní část<br>Jižní region<br>[31] Jižní atlantická oblast<br>[32] Východní jižní oblast<br>[33] Západní jižní oblast<br>Západní region<br>[41] Horská oblast<br>[42] Pacifická oblast<br>Graf 6: [REGION] - četnosti<br>  </td></tr>
<tr><td>STATEICP</td><td>Nominální</td><td>Hodnotou této proměnné je dvoučíselný kód určující stát, do kterého patří domácnost. Tyto kódy byly poprvé použity v rámci výzkumu Inter-University Consortium for Political and Social Research (ICPSR).</td></tr>
<tr><td>Obor hodnot:</td><td colspan="2">50 unikátních hodnot pro každý stát + 1 pro federální distrikt (District of Colombia)</td></tr>
<tr><td>METRO</td><td>Nominální</td><td>Hodnota této proměnné udává, zda se daná domácnost nachází v metropolitní oblasti a pokud ano zda je v centru nebo na okraji této oblasti.</td></tr>
</table>

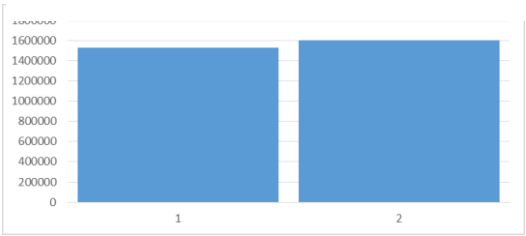
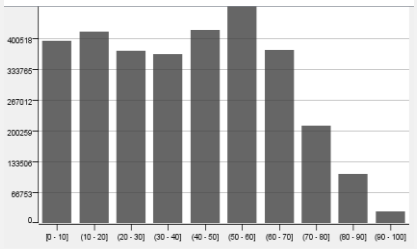
| | | |
|---|---|---|
| Obor hodnot: | [0] nelze identifikovat<br>[1] Nemetropolitní oblast<br>**Metropolitní oblast**<br>[2] Centrum metropole<br>[3] Okraj metropole<br>[4] Nelze určit | ***Graf 7: [METRO] - četnosti***<br>  |
| GQ | Nominální | Určuje skupinu, do které zařadit domácnost. Buďto se jedná rekreační jednotku, domácnost nebo skupiny domácnosti. Do skupinových domácností lze zařadit různé pečovatelské instituce, z těchto institucí jsou vybíráni pouze jedinci, tudíž nelze propojit všechny obyvatele z tohoto typu zařízení. |
| Obor hodnot: | [0] rekreační jednotka<br>**Domácnost**<br>[1] definice 1970<br>[2] definice 1990<br>**Skupinová domácnost**<br>[4] instituce<br>[5] ostatní<br>[6] další domácnost def 2000 | ***Graf 8: [GQ] - četnosti***<br>  |
| PUMA | Nominální | Jedná se 5 číselný kód označující oblast, ve které se nachází daná domácnost. Jedná se o speciální označení, které bylo poprvé použito při SDLB v roce 1990. Každá jednotka obsahuje minimálně 100 tisíc obyvatel, jednotky se vzájemně nepřekrývají a nezasahují do více než jednoho státu. |
| Obor hodnot: | 2378 unikátních PUMA jednotek bylo použito v šetření ACS 2013.[8] | |
| CITYPOP | Kvantitativní | Jedná se o přibližný počet obyvatel ve městě, ve kterém se nachází vybraná domácnost. Hodnoty jsou uvedeny v tisících. |

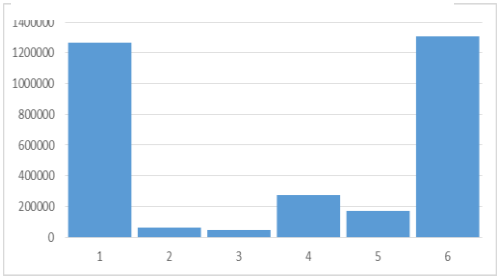
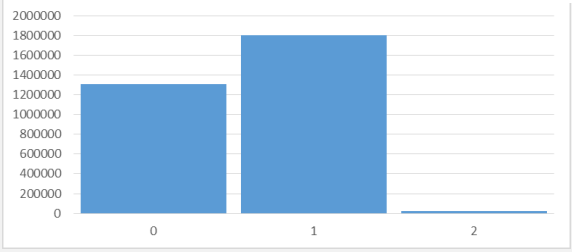
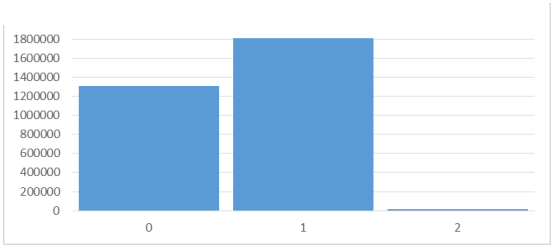
[8] https://usa.ipums.org/usa/volii/2010PUMAS.shtml na této adrese je dostupná mapa jednotek

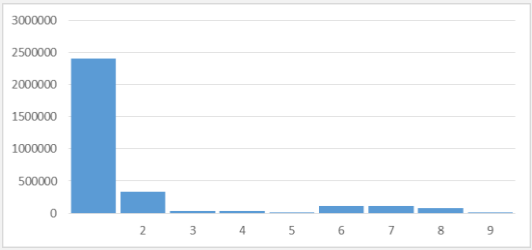
| | |
|---|---|
| Obor hodnot: | Podíl chyb. hodnot – 87 %<br>Maximum – 84 058<br>Minimum - 971<br>Průměr – 24 122<br>***Graf 10: [CITYPOP] - histogram***<br>  |

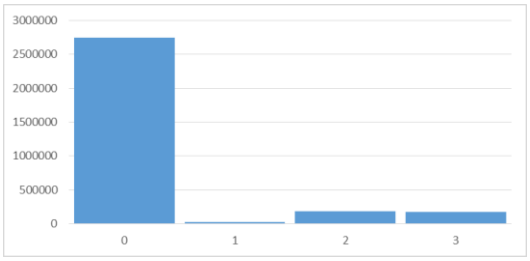
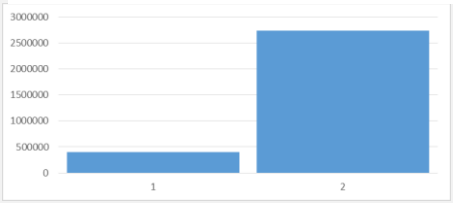
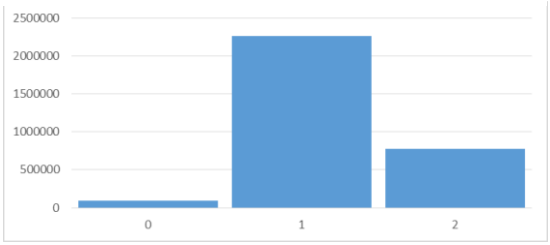
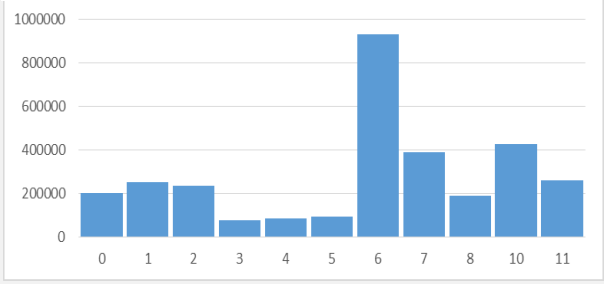
***Zdroj: (RUGGLES et. al., 2015), úprava vlastní***

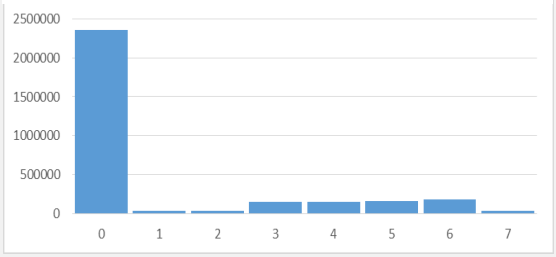
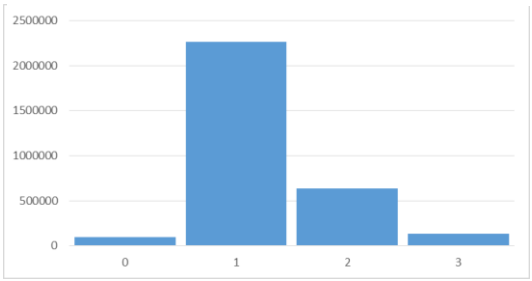
***Tab. 9: Popis vybraných atributů týkajících se respondentů***

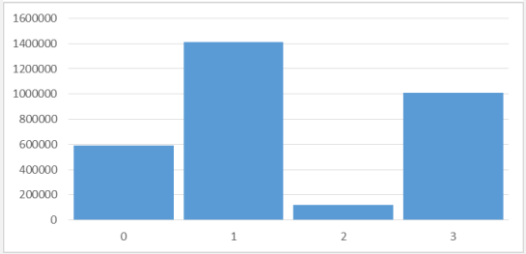
| Název proměnné | Typ proměnné | Popis |
|---|---|---|
| SEX | Nominální | Pohlaví respondenta |
| Obor hodnot: | [1] muž<br>[2] žena<br>***Graf 11: [SEX] - četnosti***<br>  | |
| AGE | Kvantitativní | Věk respondenta |
| Obor hodnot: | Podíl chyb. hodnot – 0 %<br>Maximum – 95<br>Minimum - 0<br>Průměr – 40,45<br>Medián - 41<br>Věk jednotlivých respondentů je lépe znázorněn na grafu 1 v kapitole 4.2.1<br>***Graf 12: [AGE] - histogram***<br>  | |
| MARST | Nominální | Rodinný stav respondenta. V rámci ACS se zjišťují údaje o tom, zda je v době sčítání manžela / manžel přítomen v domácnosti, což |

| | | |
|---|---|---|
| | | je případ [1], pokud je někde na služební cestě / ve válce apod. jedná se o případ [2]. |
| Obor hodnot: | [1] Vdaní, družba přítomna<br>[2] Vdaní, družba nepřítomna<br>[3] Oddělení<br>[4] Rozvedení<br>[5] Ovdovělí<br>[6] Single<br>***Graf 13:[MARST] - četnosti***<br>  <br> | |
| DIVINYR | Nominální | Charakteristika pomocí níž lze určit respondenty, jež se rozvedli v předchozím roce. Hodnota 0 [NA] je zaznamenána u lidí, jež se nemohli rozvést (tedy těch mimo rodinný stav „vdaný“) |
| Obor hodnot: | [0] N/A<br>[1] Ne<br>[2] Ano<br>***Graf 14: [DIVINYR] - četnosti***<br>  <br> | |
| WIDINYR | Nominální | Charakteristika pomocí níž lze určit respondenty, jež ovdověli v předchozím roce. Hodnota 0 [NA] je zaznamenána u lidí, jež nemohli ovdovět (tedy těch mimo rodinný stav „vdaný“) |
| Obor hodnot: | [0] N/A<br>[1] Ne<br>[2] Ano<br>***Graf 15: [WIDINYR] - četnosti***<br>  <br> | |
| RACE | Nominální | Rasa respondenta v nedetailní verzi, rozdělena do 9 základních skupin. Součástí europoidní (bílé) rasy jsou i hispánci. |

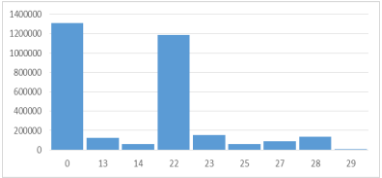
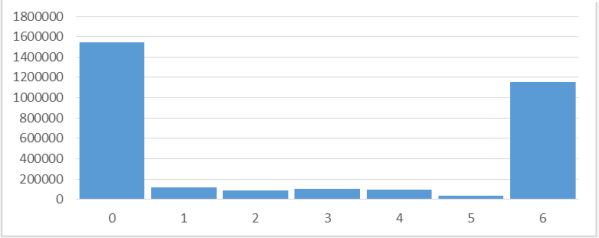
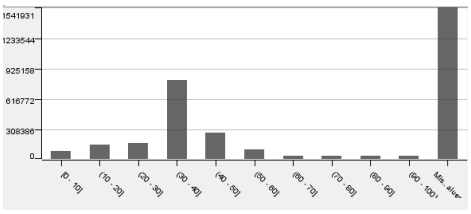
| | | |
|---|---|---|
| Obor hodnot: | [1] europoidní<br>[2] negroidní<br>[3] americký indián / původní obyvatel Aljašky<br>[4] čínská<br>[5] japonská<br>[6] Ostatní Asiati<br>[7] Ostatní rasy<br>[8] míšenec dvou ras<br>[9] míšenec tři nebo více ras<br>***Graf 16: [RACE] - četnosti***<br>  | |
| RACED | Nominální | Detailnější rozdělení ras (RACE). Pro každou uvedenou skupinu existuje ještě několik podskupin, např. původní obyvatelé jsou dotazováni na jejich kmeny. Bohužel v rámci tohoto výzkumu nebyli europoidní obyvatelé dotazováni na jejich původ. |
| Obor hodnot: | V číselníku je 258 unikátních skupin, do kterých lze klasifikovat jednotlivé obyvatele. V rámci tohoto šetření bylo zastoupeno 138 rasových skupin. | |
| BPL | Nominální | Místo narození. Jedná se tříčíselný kód, jež identifikuje stát, ve kterém se respondent narodil. První číslice vždy identifikuje světadíl. |
| Obor hodnot: | V průzkumu jsou respondenti pocházejících ze 128 území.<br>Vysvětlení pro jednotlivé kódy<br>0XX – americké státy<br>1XX – ostatní území severní Ameriky<br>2XX – státy střední Ameriky<br>3XX – jihoamerické státy<br>4XX – evropské státy<br>5XX – asijské státy<br>6XX – africké státy<br>7XX - Oceánie | |
| CITIZEN | Nominální | Statní příslušnost. Pomocí této charakteristiky lze kategorizovat respondenty do třech kategorií. Kategorie N/A obsahuje rozené Američany. |

| | | |
|---|---|---|
| Obor hodnot: | [0] N/A<br>[1] Narozen v cizině **Americkým rodičům**<br>[2] Naturalizovaný občan<br>[3] Není občan USA | ***Graf 17: [CITIZEN] - četnosti***<br>  |
| HCOVANY | Nominální | Charakteristika, pomocí níž se zjišťuje, zda je respondent pojištěn u zdravotní pojišťovny. |
| Obor hodnot: | [1] nepojištěn<br>[2] pojištěn | ***Graf 18: [HCOVANY] - četnosti***<br>  |
| SCHOOL | Nominální | Školní docházka. Zjišťuje se, zda respondent navštěvoval v určité době nějakou školu. |
| Obor hodnot: | [0] N/A<br>[1] Ne<br>[2] Ano | ***Graf 19: [SCHOOL] - četnosti***<br>  |
| EDUC | Ordinální | Nejvyšší úroveň dosaženého vzdělání respondenta. |
| Obor hodnot: | [0] N/A žádné dosažené vzdělání<br>[1] Předškolní - 4. úroveň<br>[2] Úrovně 5 - 8<br>[3] Úroveň 9<br>[4] Úroveň 10<br>[5] Úroveň 11<br>[6] Úroveň 12<br>[7] 1 rok VŠ<br>[8] 2 roky VŠ<br>[10] 4 roky VŠ<br>[11] 5 a více let VŠ | ***Graf 20: [EDUC] - četnosti***<br>  |

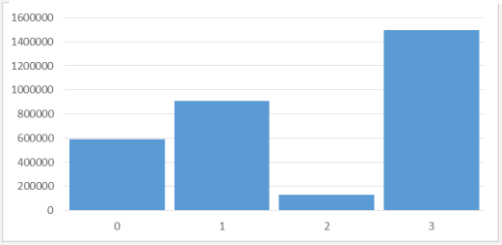
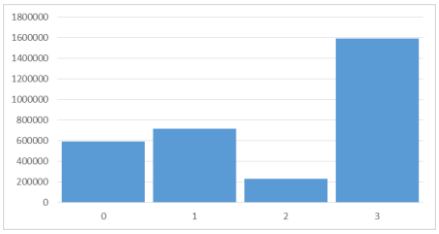
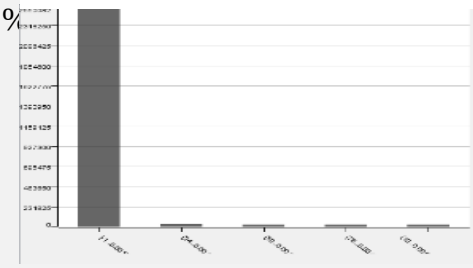
<table>
<tr><td>EDUCD</td><td>Nominální</td><td>Detailnější náhled na vzdělání, pouze pomocí této charakteristiky lze identifikovat přesný počet absolventů jednotlivých stupňů studia, případně zjistit dosažené tituly. Jedná se o tříčíselný kód, jehož první číslice je pro danou délku studia totožná s hodnotou EDUC.</td></tr>
<tr><td>Obor hodnot:</td><td colspan="2">V číselníku je k dispozici celkem 47 možných klasifikací, v rámci tohoto dotazníku je využito 33.</td></tr>
<tr><td>GRADEATT</td><td>Ordinální</td><td>Aktuální úroveň studia. Hodnota této proměnné se zjišťovala pouze u těch respondentů, jež vyplnili hodnotu „Ano“ u charakteristiky SCHOOL. Hodnota N/A tedy znamená, že dotiční nenavštěvují žádnou školu.</td></tr>
<tr><td>Obor hodnot:</td><td colspan="2">[0] N/A<br>[1] jesle<br>[2] školka<br>[3] 1. stupeň ZŠ<br>[4] 2. stupeň ZŠ<br>[5] střední škola<br>[6] VŠ bez titulu<br>[7] Navazující VŠ studium<br><br>Graf 21: [GRADEATT] - četnosti<br> 
</td></tr>
<tr><td>SCHLTYPE</td><td>Nominální</td><td>Typ školního zařízení ve smyslu zda se jedná o veřejné nebo soukromé zařízení. Hodnota N/A je uvedena u respondentů, jež nenavštěvovali žádnou školu. Zjišťováno pouze o respondentů, kteří vyplnili u charakteristiky SCHOOL možnost „ano“.</td></tr>
<tr><td>Obor hodnot:</td><td colspan="2">[0] N/A<br>[1] nevyplněn<br>[2] Veřejná škola<br>[3] Soukromá škola<br><br>Graf 22: [SCHLTYPE] - četnosti<br> 
</td></tr>
</table>

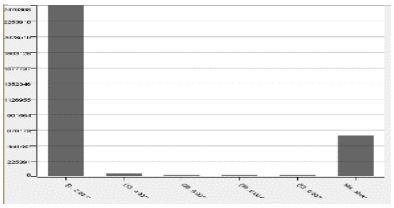
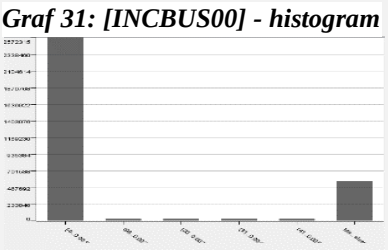
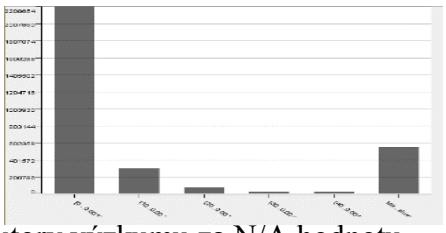
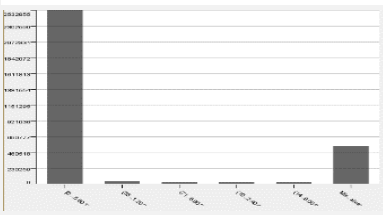
| | | |
|---|---|---|
| DEGFIELD / DEGFIELD2 | Nominální | Zjišťuje se u absolventů bakalářského typu studia, na kterém oboru ho získali. Jedná se o dvoučíselný kód. Zjišťují se dvě školy (pokud někdo vystudoval více oborů). |
| Obor hodnot: | V číselníku je k dispozici celkem 40 možných oborů | |
| DEGFIELDD / DEGFIELDD2 | Nominální | Zjišťuje se u absolventů bakalářského typu studia, na kterém oboru ho získali. Obory jsou tříděny detailněji. Jedná se o čtyřčíselný kód, kde vždy první dvě číslice jsou totožná s oborem jako DEGFIELD a následující dvě číslice identifikují podobor. Zjišťují se dvě školy (pokud někdo vystudoval více oborů). |
| Obor hodnot: | V číselníku je k dispozici celkem 190 možných oborů | |
| EMPSTAT | Nominální | Ekonomická aktivita. V rámci této charakteristiky jsou ještě ekonomicky aktivní rozdělení na zaměstnané a na lidi aktivně hledající práci. |
| Obor hodnot: | [0] N/A<br>[1] zaměstnaní<br>[2] nezaměstnaní<br>[3] ekonomicky neaktivní | ***Graf 23: [EMPSTAT] - četnosti***<br>  |
| OCC | Nominální | Druh povolání. Jedná se o 4 číselný kód určující typ povolání. Pokud respondent má více povolání, tak je uvedena primární, ta kde má vyšší plat. |
| Obor hodnot: | 0000 – N/A člověk 16 let a méně, který nikdy nepracoval<br>V číselníku pro rok 2013 se nacházelo 553 druhů povolání[9]. | |
| IND | Nominální | Odvětví. Jedná se o 4 číselný kód určující odvětví, ve kterém je respondent zaměstnán. Pokud respondent pracuje ve více odvětvích, tak je uvedeno primární, to kde má vyšší plat. |
| Obor hodnot: | 0000 – N/A člověk 16 let a méně, který nikdy nepracoval | |

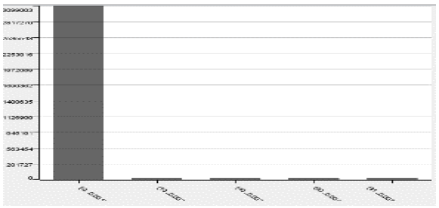
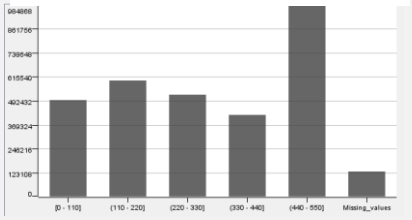
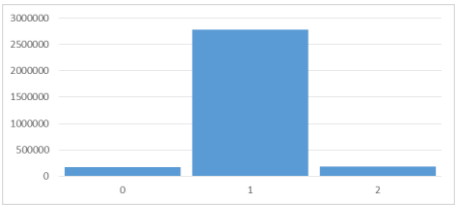
[9] Kompletní seznam k dispozici zde: https://usa.ipums.org/usa/volii/c2ssoccup.shtml

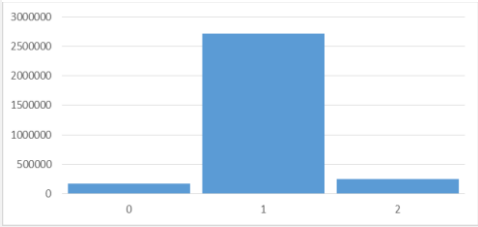
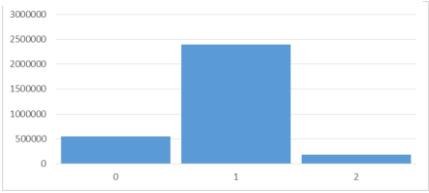
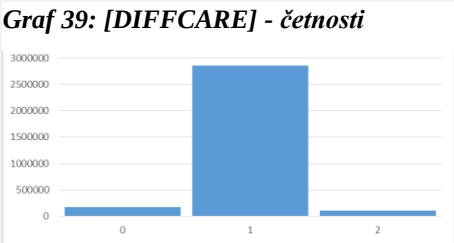
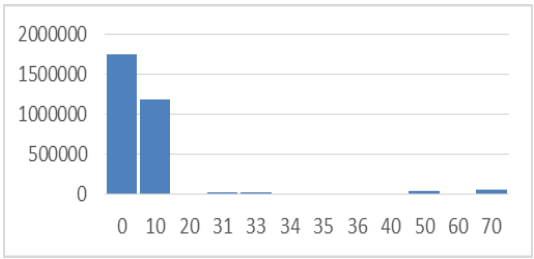
| | | |
|---|---|---|
| | V číselníku pro rok 2013 se nacházelo 260 druhů odvětví[10]. | |
| CLASSWRKD | Nominální | Typ zaměstnání. Jedná se o charakteristiku toho, zda respondent pracuje, pokud ano, zda je zaměstnán jako soukromí podnikatel nebo pracuje „za plat“. V tomto průzkumu je evidováno celkem 9 skupin v této kategorii. |
| Obor hodnot: | [0] N/A<br>[13] soukromý podnikatel netvořící akciovou společnost<br>[14] soukromý podnikatel tvořící akciovou společnost<br>[22] zaměstnanec v soukromé sféře<br>[23] zaměstnanec v neziskové sféře<br>[25] zaměstnanec federální vlády<br>[27] státní zaměstnanec<br>[28] zaměstnanec místní vlády<br>[29] neplacený zaměstnanec v rodinném podniku | ***Graf 24:[CLASSWRKD] - četnosti***<br>  |
| WKSWORK2 | Ordinální | Počet odpracovaných týdnů. Jedná se o intervalovou proměnnou. |
| Obor hodnot: | [0] N/A nebo chyb hodnota<br>[1] 1-13 týdnů<br>[2] 14-26 týdnů<br>[3] 27-39 týdnů<br>[4] 40-47 týdnů<br>[5] 48-49 týdnů<br>[6] 50-52 týdnů | ***Graf 25: [WKSWORK] - četnosti***<br>  |
| UHRSWORK | Kvantitativní | Počet odpracovaných hodin týdně, N/A jsou hodnoty u lidí, kteří jsou ekonomicky neaktivní. |
| Obor hodnot: | Podíl chyb. hodnot – 0,08 %<br>Maximum – 98<br>Minimum - 1<br>Průměr – 37,85<br>Medián - 40<br><br>U této charakteristiky je speciální hodnota 99, která má označit chybějící hodnotu. | ***Graf 26: [UHRSWORK] - histogram***<br>  |

[10] Kompletní seznam k dispozici zde: https://usa.ipums.org/usa/volii/13indus.shtml

| | | |
|---|---|---|
| LOOKING | Nominální | Hledání práce. Zjišťuje se, zda respondent v minulém měsíci poohlížel po nové práci. Tato charakteristika se zjišťuje u všech respondentů (včetně těch co pracují). |
| Obor hodnot: | [0] N/A<br>[1] Hledal práci<br>[2] Nehledal práci<br>[3] Nevyplněno | ***Graf 27: [LOOKING] - četnosti***<br>  |
| WORKEDYR | Nominální | Zjišťuje se, zda respondent pracoval v předchozích letech. Pokud ne, tak se dál zjišťuje, zda pracoval aspoň v předchozích pěti letech. |
| Obor hodnot: | [0] N/A<br>[1] Ne, nepracoval v předchozích pěti letech<br>[2] Ne, ale pracoval 1-5 let zpátky<br>[3] Yes | ***Graf 28: [WORKEDYR] - četnosti***<br>  |
| INCTOTAL | Kvantitativní | Jedná se o celkové příjmy nebo ztrátu, které měl respondent před zdaněním za předchozí rok. |
| Obor hodnot: | Podíl N/A hodnot – 17,5 %<br>Podíl nulových hodnot – 10,9 %<br>Maximum – 1,281,000<br>Minimum - -13,600<br>Průměr – 35,747<br>Medián – 20,800<br>Hodnota 9999999 je považována autory výzkumu za N/A hodnotu. Počet respondentů s touto hodnotou odpovídá přesnému počtu dětí mladších než 14 let. Vysoký podíl lidí, kteří vyplnili velmi nízké příjmy nebo příjmy nulové. | ***Graf 29: [INCTOTAL] - histogram***<br>  |

| | | |
|---|---|---|
| INCWAGE | Kvantitativní | Jedná se o celkové příjmy platové složky, které respondent obdržel jako zaměstnanec za předchozí rok. |
| Obor hodnot: | Podíl N/A hodnot – 18,9 %<br>Podíl nulových hodnot – 33,4 %<br>Maximum – 660,000<br>Minimum - 0<br>Průměr – 26,061<br>Medián – 6,600<br><br>Hodnota 999999 je považována autory výzkumu za N/A hodnotu. | ***Graf 30: [INCWAGE] - histogram***<br>  |
| INCBUS00 | Kvantitativní | Jedná se o celkové příjmy z vlastního podnikání a provozování farmy za předchozí rok. |
| Obor hodnot: | Podíl N/A hodnot – 17,5 %<br>Podíl nulových hodnot – 77,5 %<br>Maximum – 525,000<br>Minimum - -9,000<br>Průměr – 15,457<br>Medián – 0<br>Hodnota 999999 je považována autory výzkumu za N/A hodnotu. | ***Graf 31: [INCBUS00] - histogram***<br>  |
| INCSS | Kvantitativní | Jedná se o celkové příjmy, které respondent obdržel na sociálních dávkách za předchozí rok. |
| Obor hodnot: | Podíl N/A hodnot – 17,5 %<br>Podíl nulových hodnot – 63,5 %<br>Maximum – 50,000<br>Minimum - 0<br>Průměr – 2,955<br>Medián – 0<br>Hodnota 99999 je považována autory výzkumu za N/A hodnotu. | ***Graf 32: [INCSS] - histogram***<br>  |
| INCRETIR | Kvantitativní | Jedná se o celkové příjmy, které respondent obdržel jako penze nebo v rámci různých odškodnění za předchozí rok (nepočítají se do toho sociální dávky od státu). |
| Obor hodnot: | Podíl N/A hodnot – 17,5 %<br>Podíl nulových hodnot – 72,9 %<br>Maximum – 178,000<br>Minimum - 0<br>Průměr – 2,351<br>Medián – 0 | ***Graf 33: [INCRETIR] - histogram***<br>  |

| | | |
|---|---|---|
| | Hodnota 999999 je považována autory výzkumu za N/A hodnotu | |
| INCEARN | Kvantitativní | Jedná se o celkový „plat“, který respondent měl za předchozí rok, tedy INCWAGE + INCBUS00 |
| Obor hodnot: | Podíl N/A hodnot – 0 %<br>Podíl nulových hodnot – 69,2 %<br>Maximum – 1,019,000<br>Minimum - -9,000<br>Průměr – 22,601<br>Medián – 450<br>Hodnota 9999999 je považována autory výzkumu za N/A hodnotu. | ***Graf 34: [INCEARN] - histogram***<br>  |
| POVERTY | Kvantitativní | Míra chudoby. Jedná se číslo v procentech. Toto číslo vyjadřuje podíl celkových rodinných příjmů ku nějaké standardizované prahové hodnotě. Tato hodnota je ovlivněna počtem členů a dětí v jednotlivých domácnostech. Hraniční hodnoty jsou 1 % a 501 % |
| Obor hodnot: | Podíl N/A hodnot – 4 %<br>Maximum – 501<br>Minimum - 1<br>Průměr – 304,841<br>Medián – 308 | ***Graf 35:[POVERTY] - histogram***<br>  |
| DIFFREM | Nominální | Kognitivní poruchy. Zjišťuje se, zda respondent trpí nějakou mentální poruchou případně, zda má problémy s učením, koncentrací, pamětí nebo rozhodováním. |
| Obor hodnot: | [0] N/A<br>[1] netrpí kognitivní poruchou<br>[2] trpí kognitivní poruchou | ***Graf 36: [DIFFREM] - četnosti***<br>  |
| DIFFPHYS | Nominální | Fyzické postižení. Zjišťuje se, zda respondent je omezen v činnostech jako chození, chůze do schodů, zvedání věci, nošení věcí apod. |

| | | |
|---|---|---|
| Obor hodnot: | [0] N/A<br>[1] netrpí fyzickou poruchou<br>[2] trpí fyzickou poruchou | ***Graf 37: [DIFFPHYS] - četnosti***<br>  |
| DIFFMOB | Nominální | Zjišťuje se, zda respondent je schopen provozovat běžné aktivity mimo domov osamocen ať už z fyzických, mentálních nebo emociálních důvodů. Do této skupiny se neřadí dočasné problémy jako zlomené kosti / těhotenství. |
| Obor hodnot: | [0] N/A<br>[1] netrpí poruchou mobility<br>[2] trpí poruchou mobility | ***Graf 38: [DIFFMOB] - četnosti***<br>  |
| DIFFCARE | Nominální | Zjišťuje se, zda se je respondent postarat sám o sebe (koupání, oblékání nebo pohybování po domě). |
| Obor hodnot: | [0] N/A<br>[1] Je schopen<br>[2] Není schopen | ***Graf 39: [DIFFCARE] - četnosti***<br>  |
| TRANWORK | Nominální | Způsob dopravy do práce. |
| Obor hodnot: | [0] N/A<br>[10] Auto, dodávka<br>[20] Motorka<br>[31] Autobus / trolejbus<br>[33] Metro<br>[34] Vlak<br>[35] Taxi<br>[36] Trajekt<br>[40] Kolo<br>[50] Pěšky<br>[60] Ostatní<br>[70] Práce doma | ***Graf 40:[TRANWORK] - četnosti***<br>  |

***Zdroj:*** (RUGGLES et. al., 2015)***, úprava vlastní***

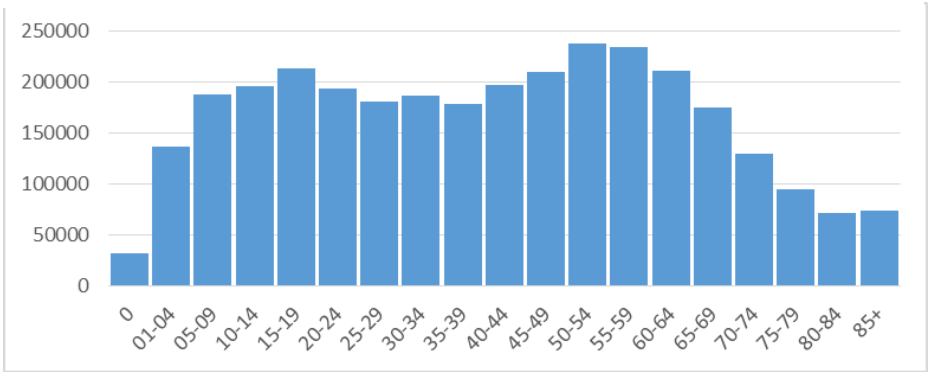
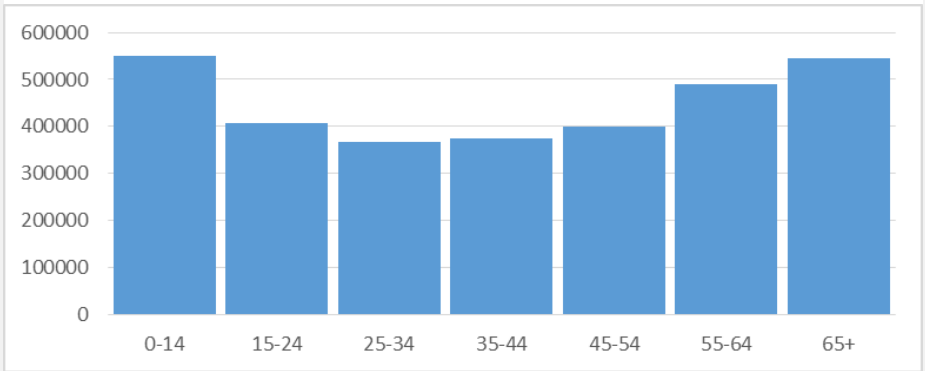
## 4.3 Příprava dat

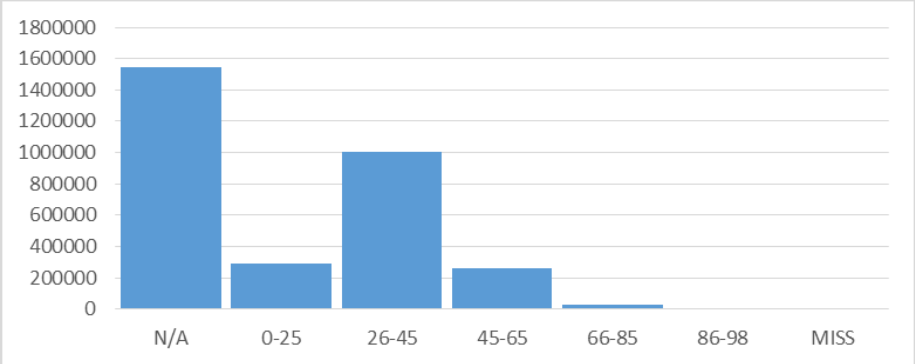
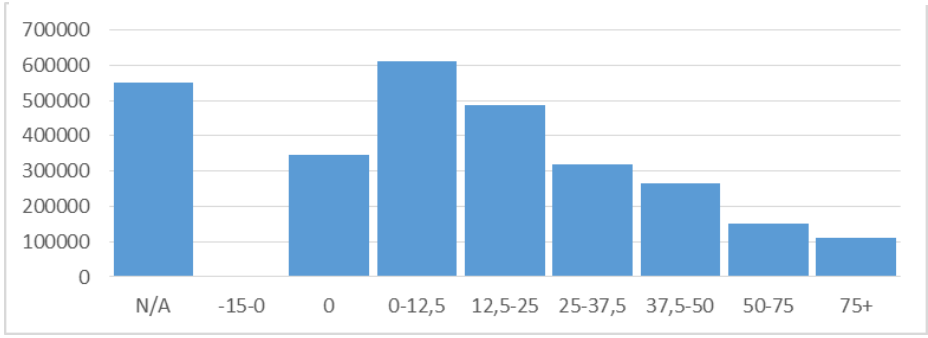
V této kapitole budou popsány veškeré úpravy, které budou provedeny nad zdrojovými daty před tím, než proběhne samotné modelování.

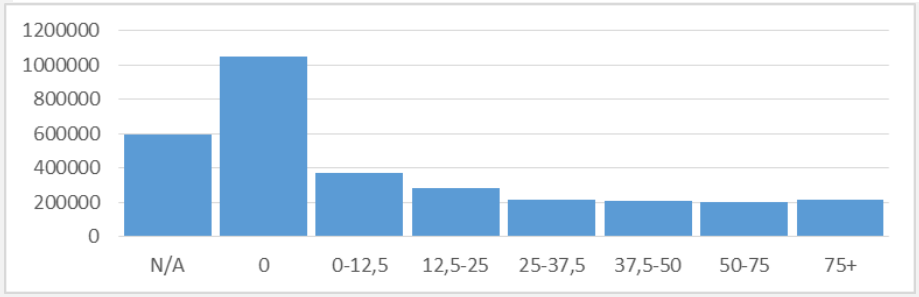
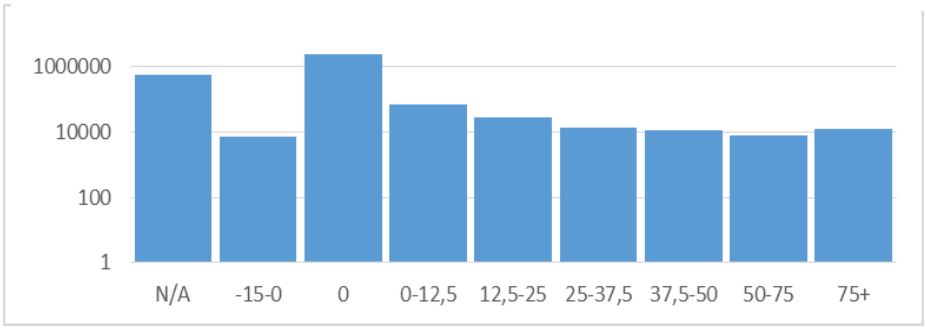
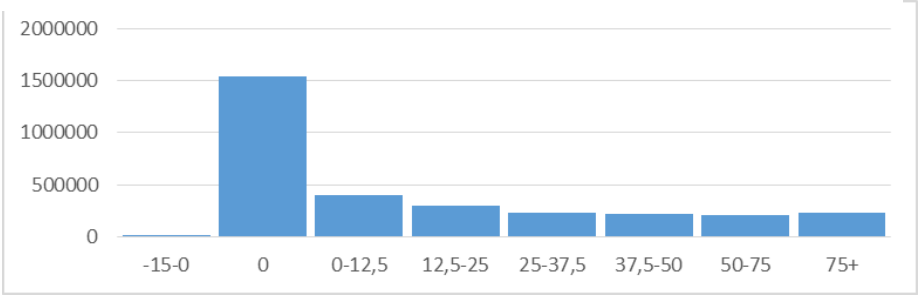
### 4.3.1 Překódování proměnných

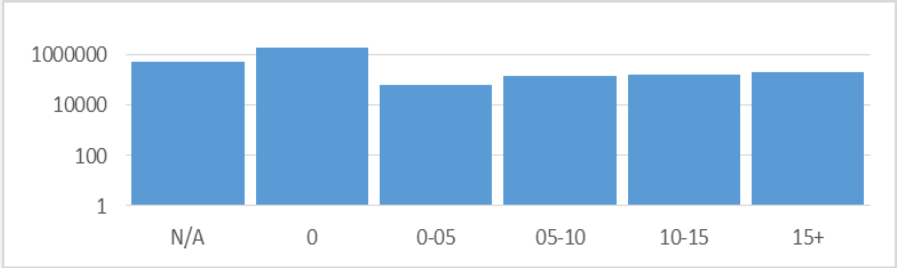
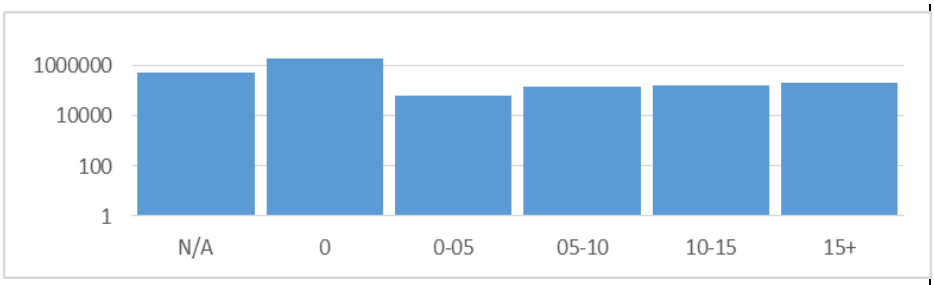
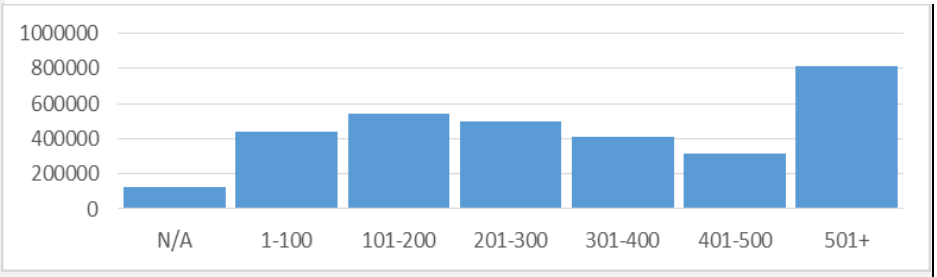
Vzhledem k faktu, že některé algoritmy mají problém s prací s kvantitativní proměnnými je potřeba je překódovat na kategoriální. Dále budou některé proměnné překódovány z důvodu lepší srozumitelnosti, např. vzdělání. Veškeré nové proměnné budou zapsány v podobné tabulce, jako byla ta v kapitole s technickými parametry. Bude tedy uveden název nové i staré proměnné, popsány intervaly a uveden graf četností.

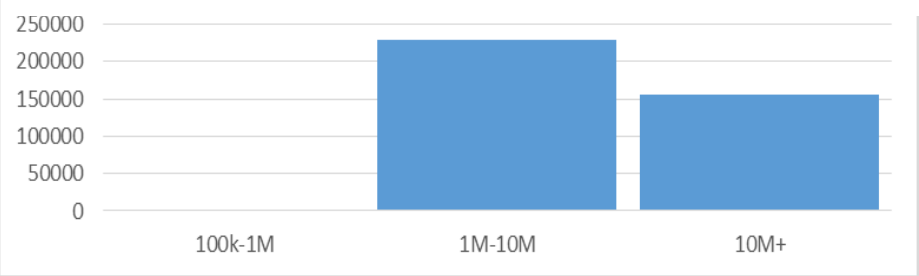
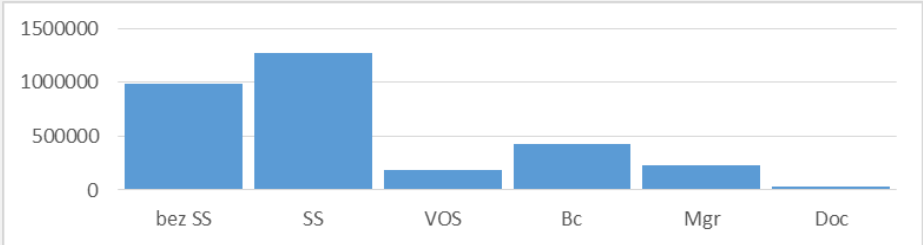
***Tab. 10: Seznam překódovaných proměnných***

<table>
<tr><th>Nová proměnná</th><th>Stará proměnná</th><th>Popis</th></tr>
<tr><td>AGE_INT5</td><td>AGE</td><td>Jedná se o překódování věku do 5letých intervalů s tím, že nula letí a lidé ve věku 85 a budou ve speciálním intervalu.</td></tr>
<tr><td>Graf četností:</td><td colspan="2">Graf 41: [AGE_INT5] - četnosti<br> 
</td></tr>
<tr><td>AGE_EA10</td><td>AGE</td><td>Z pragmatických důvodů jsem se rozhodl překódovat proměnnou věk ještě do 10 letých intervalů pro lidi v ekonomicky aktivním věku (15-65 let). Respondenti 0-14 let a 65+ budou mít vlastní interval.</td></tr>
<tr><td>Graf četností</td><td colspan="2">Graf 42: [AGE_EA10] - četnosti<br> 
</td></tr>
</table>

| | | |
|---|---|---|
| UHRWORK_INT20 | UHRWORK | Počet odpracovaných hodin týdně. Rozhodl jsem se pro interval, tak aby se dal vyjádřit typ pracovního úvazku. 0-25 by mohl být teoreticky poloviční úvazek, 26 - 45 plný pracovní úvazek, 46-65 jeden plný pracovník úvazek + jeden poloviční, 66-85 – dva plné pracovní úvazky a 86 hodin víc workoholik, u této proměnné N/A znázorňuje ekonomicky neaktivní respondenty a hodnota MISS lidí co nevyplnili tento atribut (původní kód „99“). |
| Graf četností | ***Graf 43: [UHRWORK_INT20] - četnosti***<br>  | |
| INCTOT_INT125 | INCTOT | V oficiální statistice USA se používá interval délky 25 tisíc dolarů na domácnost, já se rozhodl pro interval poloviční délky, protože jsou zde data za osoby. Do speciální skupiny jsem se rozhodl dát nulové hodnoty (respondenti co nezaznamenali příjem) a záporné hodnoty (roční ztráta). Zdvojnásobil jsem délku předposledního intervalu, protože v intervalu 62,5 – 75 by již byla malá skupina respondentů. Do skupiny N/A patří děti do 14 let. |
| Graf četností | ***Graf 44: [INCTOT_INT125] – četnosti***<br>  | |

| | | |
|---|---|---|
| INCWAGE_INT125 | INCWAGE | Vzhledem k faktu, že se rozdělení hodnot u této charakteristiky zásadně neodlišuje od INCTOT rozhodl jsem se pro zachování intervalů. Jediným rozdílem je to, že nelze mít záporný plat, tudíž zde lze vynechat jednu kategorii |
| Graf četností | ***Graf 45: [INCWAGE_INT125] - četnosti***<br>  | |
| INCBUS100_INT125 | INCBUS100 | Vzhledem k faktu, že se rozdělení hodnot u této charakteristiky zásadně neodlišuje od INCTOT rozhodl jsem se pro zachování intervalů bez rozdílu. |
| Graf četností (využito logaritmické měřítko, kvůli vysokému podílu nulových hodnot) | ***Graf 46: [INCBUS100_INT125] - četnosti***<br>  | |
| INCEARN_INT125 | INCEARN | Vzhledem k faktu, že se rozdělení hodnot u této charakteristiky zásadně neodlišuje od INCTOT rozhodl jsem se pro zachování intervalů. Jediným rozdílem je to, že nelze mít záporný plat, tudíž zde lze vynechat jednu kategorii |
| Graf četností | ***Graf 47: [INCEARN_INT125] - četnosti***<br>  | |

| | | |
|---|---|---|
| INCSS_INT5 | INCSS | Rozložení hodnot u této charakteristiky se odlišuje od INCTOTAL. Příjmy ze sociálních dávek nejsou tak vysoké jako ostatní příjmy (z podnikání, mzda apod.), proto jsem se rozhodl pro interval 5 tisíc. |
| Graf četností (využito logaritmické měřítko, kvůli vysokému podílu nulových hodnot) | ***Graf 48: [INCSS_INT5] - četnosti***<br>  | |
| INCRETIR_INT5 | INCRETIR | Rozložení hodnot u této charakteristiky je podobné jako u INCSS, proto jsem se rozhodl zachovat intervaly. |
| Graf četností (využito logaritmické měřítko, kvůli vysokému podílu nulových hodnot) | ***Graf 49: [INCRETIR_INT5] - četnosti***<br>  | |
| POVERTY_INT100 | POVERTY | Vzhledem k faktu, že možné hodnoty jsou od 1 do 501, rozhodl jsem se pro délku intervalu 100. V intervalu 1-100 budou tedy všichni, kteří nedosahují na prahovou hodnotu a dalo by se říct, že jsou ohroženi chudobou. |
| Graf četností | ***Graf 50: [POVERTY_INT100] - četnosti***<br>  | |
| CITYPOP_INT | CITYPOP | Vzhledem k faktu, že se jedná o velmi specifickou proměnnou s vysokým podílem chybějících hodnot. Rozhodl jsem se pro klasifikaci po řádech tedy od 100k do 1M, od 1M do 10M a 10M a výš. Tato |

| | | |
|---|---|---|
| | | charakteristika není vyplněna pro menší města a obce. |
| Graf četností | ***Graf 51: [CITYPOP_INT] - četnosti***<br>  <br>Podíl chybějících hodnot je u tohoto atributu 87 %. | |
| EDUC_REWORK | EDUCD | Překódování charakteristiky vzdělání, aby byla adekvátní českému školskému systéme. Ta z USA byla až moc podrobná, protože se evidoval každý rok školní docházky, ve kterém někdo odpadl. Proměnná EDUCD byla málo podrobná, nebyli zvýrazněny jednotlivé milníky v podobě dokončeného určitého stupně studia. |
| Graf četností | ***Graf 52: [EDUC_REWORK] - četnosti***<br>  | |
| INCEARN_4INT | INCEARN | Pro klasifikační část úlohy bude vytvořena ještě jedna speciální proměnná. Bude se jednat o příjem respondentů, který bude klasifikován do 4 automatických skupin, tak aby každá skupina obsahovala stejný počet případů, tedy rozdělí soubor na kvartily. Cílem je zjistit, zda se úspěšnost klasifikace změní, pokud se změní cílový atribut. Toto bude provedeno až po vybrání zaměstnaných respondentů (nebudou zahrnuti respondenti s příjmem 0). |
| Automatické hranice | 16,500; 34, 000; 59,900 | |

***Zdroj: vlastní zpracování***

### 4.3.2 Příprava datových souborů

Vzhledem k faktu, že ne pro každou část mnou navržených DM úloh se hodí mít celý datový soubor pohromadě, rozhodl jsem se vygenerovat několik speciálních souborů a na těch speciálně provádět různé druhy testů. V následujících tabulkách budou jednotlivé datové soubory popsány a bude v nich soupis proměnných.

***Tab. 11: Seznam datových souborů využitých při klasifikaci platových skupin***

| Název souboru | Omezení | Popis |
|---|---|---|
| DATASET1 | EMPSTAT = 1<br>15 % výběrový soubor | Cílem této části úlohy je klasifikovat respondenty do platových tříd. Proto jsem se u tohoto prvního setu rozhodl vybrat pouze proměnné, které dle mého názoru do určité míry výši platu a zjistit zda bude zásadní rozdíl v úspěšnosti klasifikace, pokud se bude klasifikovat pouze na základě těchto proměnných.<br>Do souboru jsem se rozhodl nezařadit proměnné INCTOTAL, INCBUS100 a INCWAGE, protože cílem úlohy je zjistit využití klasifikátorů při práci s chybějícími daty a pokud člověk neuvede svůj plat tak dle mého ve většině případů neuvede ani tyto další proměnné (některé tyto proměnné jsou kalkulovány právě na základě proměnné výše platu). Navíc je mezi těmito proměnnými silná závislost, což by pak výrazně ovlivnilo výsledky klasifikace (klasifikátor by např. klasifikoval pouze na základě těchto proměnných).<br>Dále jsem základní soubor omezil výběr pouze na zaměstnané lidi. |
| Vybrané charakteristiky | OCC, IND, EDUC_REWORK, AGE_EA10, UHRSWORK_INT20, WKSWORK2, SEX, POVERTY_INT100 | |
| DATASET2 | EMPSTAT = 1<br>15 % výběrový soubor | V druhém datovém souboru budou všechny charakteristiky zařazené do tohoto výzkumu, vyjma těch vysoce korelovaných, tedy veškerých proměnných obsahující nějaký typ příjmu. |
| Vynechané charakteristiky | INCTOT_INT125, INCSS_INT5, INCWAGE_INT125, INCBUS100_INT125, INTRETIR_INT5 | |
| DATASET3 | EMPSTAT = 1<br>15 % výběrový soubor | Do posledního datasetu k této podúloze byly zvoleny vysoce korelované proměnné, zde se očekává vysoká úspěšnost klasifikace, vynechám INC WAGE a INCBUS100, protože jejich součtem je právě INC_EARN. |

| | |
|---|---|
| Vybrané charakteristiky | INCTOT_INT125, INCSS_INT5, , INTRETIR_INT5 |

***Zdroj: vlastní zpracování***

***Tab. 12: Seznam datových souborů využitých při klasifikaci typu domácnosti***

| Název souboru | Omezení | Popis |
|---|---|---|
| DATASET4 | HHTYPE <> 0<br>10 % výběrový soubor | Při klasifikaci typu domácnosti vypustím ty respondenty, u nichž se nedal typ domácnosti identifikovat, resp. byl nevyplněn (HHTYPE=0).<br>Do tohoto datasetu budou zařazeny pouze ty proměnné, ze kterých se dá typ domácnosti částečně odvodit, jedná se o manželský stav a pohlaví. Dalšími charakteristikami jsou PUMA (tedy speciální územní celek, který se využívá při těchto cenzech) GQ a METRO, kde se dá předpokládat, že rodiny s dětmi budou více na okraji města apod. A v neposlední řadě rasa, protože lze předpokládat, že některé minoritní rasy žijí spíše pohromadě. |
| Vybrané charakteristiky | SEX, PUMA, MARST, RACE, GQ, METRO | |
| DATASET5 | HHTYPE <> 0<br>10 % výběrový soubor | Z datového souboru se všemi proměnnými není potřeba dle mého názoru žádnou vyřadit ze seznamu. Soubor neobsahuje žádnou silně korelované hodnoty jako v případě proměnné z předchozí podúlohy s platem. |

***Zdroj: vlastní zpracování***

***Tab. 13: Seznam datových souborů využitých při klasifikaci typu lokality***

| Název souboru | Omezení | Popis |
|---|---|---|
| DATASET6 | 10 % výběrový soubor | V tomto výběrovém datovém souboru bylo vybráno celkem 5 proměnných. Dvě proměnné určující geografickou polohu, STATEICP a PUMA. Další zajímavou proměnnou je typ dopravního prostředku, který respondent používá k dojezdu do práce. Poslední dvě proměnné jsou typ domácnosti a rasa, protože např. rodiny s dětmi se koncentrují na předměstí, kdežto minoritní rasy se mohou |

| | | |
|---|---|---|
| | | koncentrovat v centru metropole (čínské čtvrti, ghetta apod.). |
| Vybrané charakteristiky | HHTYPE, TRANSWORK, RACE, PUMA STATEICP | |
| DATASET7 | 10 % výběrový soubor | Podobně jako u klasifikace typu domácnosti není potřeba vyřadit některé charakteristiky a pro porovnání lze klasifikovat se všemi proměnnými v datovém souboru. |

***Zdroj: vlastní zpracování***

***Tab. 14: Datový soubor využitý v úloze hledání nuggetů***

| Název souboru | Omezení | Popis |
|---|---|---|
| DATASET8 | 40 % výběrový soubor | Pro tuto úloho není potřebné vytvářet nějaký speciální datový soubor. Pouze byla omezen jeho rozsah, zejména kvůli časové náročnosti jednotlivých procedur. |

***Zdroj: vlastní zpracování***

## 4.4 Modelování

V této kapitole budou provedeny analýzy za využití vybraných DM technik popsaných v kapitole 2.4. Kapitola bude rozdělena na dvě podkapitoly. V první podkapitole bude provedena klasifikační část úlohy a ve druhé podkapitole část zaměřena na hledání zajímavých souvislostech v datech. Veškeré použité algoritmy budou vždy zmíněny v úvodu každé podkapitoly.

### 4.4.1 Klasifikace

Většina proměnných ve vybraném šetření je kvalitativního typu, případně byly na tento typ transformovány. Díky tomuto faktu nelze využít neuronové sítě, které pro klasifikaci využívají primárně kvantitativní typ proměnných. Nejlépe s kvalitativním proměnnými při klasifikaci pracují algoritmy založeny na rozhodovacích stromech a bayesovské klasifikaci, případně lze využít i rozhodovací pravidla.

V této kapitole nejdříve bude v tabulce uveden seznam, u každého algoritmu bude uvedena skupina DM technik, do které patří a případně nějaký dodatečný popis / princip funkčnosti.

***Tab. 15: Použité klasifikační algoritmy***

| Název algoritmu | DM technika | Popis |
|---|---|---|
| ZeroR | Rozhodovací pravidla (kap. 2.4.3) | Jedná se o nejjednodušší možnou metodu klasifikace. Pomocí algoritmu se vybere hlavní kategorie (ve většině případů ta s největší četností) a pak jsou všechny případy klasifikovány jako tato třída (SAYAD, 2010). |

| | | |
|---|---|---|
| OneR | Rozhodovací pravidla (kap. 2.4.3) | Jedná se o další poměrně jednoduchou metodu klasifikace. V rámci modelování se pro každou vstupní hodnotu všech proměnných sestaví jedno pravidlo pro klasifikaci. Následně se vybere pouze jedna proměnná, která bude použitá pro následnou predikci hodnot (vybírá se proměnná s nejnižší chybovostí) (SAYAD, 2010). |
| NaiveBayes | Bayesovská klasifikace (kap. 2.4.6) | Princip funkcionality těchto dvou klasifikátorů byl detailněji popsán v kapitole 2.4.6. |
| BayesNet | | |
| J48 | Rozhodovací stromy (kap. 2.4.1) | Jedná se o OpenSource implementaci algoritmu C4.5. Vytváří se strom od shora dolů. Postup je následující, v každém uzlu se zvolí atribut, jež nejlépe charakterizuje data (nejlepší entropie a informační zisk[11]). Algoritmus končí buďto správným zařazením všech datových položek nebo použitím všech atributů jako uzlů stromu. Po sestavení takovéhoto stromu pak proběhne prořezávání stromu, tedy odstranění všech „zbytečných“ uzlů. Tedy těch, jejichž odstraněním se nesníží úspěšnost klasifikace (MRÁZOVÁ, 2011). |
| HoeffdingTree | Rozhodovací stromy (kap. 2.4.1) | Konstrukce tohoto stromu využívá principu tzv. Hoefdingovi meze. Při konstrukci tohoto stromu se vychází z předpokladu, že i malý vzorek dat stačí ke správné klasifikaci dat. Velikost tohoto vzorku je vypočítána pomocí zmíněného matematického teorému (WEKA, 2013). |
| RandomTree | Rozhodovací stromy (kap. 2.4.1) | Jedná se opět o rozhodovací strom sestavovaný na shora dolů podobně jako J48. Zásadní rozdíly oproti J48 jsou, že do každého uzlu je vybrán „K“ počet náhodných atributů a dále se neprovádí po sestavení tohoto stromu prořezávání atributů (OPENCV, 2011). |
| DesicisionStump | Rozhodovací stromy (kap. 2.4.1) | Algoritmus konstrukce tohoto stromu je velmi podobný algoritmu OneR. Výsledný strom vypadá tak, že má pouze jeden uzel a poté následují listy (počet listů je roven počtu klasifikačních tříd) (IBRA et. al., 1992). |

***Zdroj: uveden v tabulce, vlastní zpracování***

Pro klasifikační úlohu bylo vybráno několik různorodých algoritmů. Výběr jednotlivých klasifikátorů byl omezen možnostmi SW nástroje Weka, který byl k této úloze využit. Dalším faktorem, jež ovlivnil výběr jsou typy proměnných v datovém souboru a rozsah tohoto souboru.

[11] Informační zisk – očekávaná redukce entropie po rozdělení dat podle uvažovaného atributu

Některé klasifikátory mají problém si poradit s rozsáhlými soubory (200 tisíc a více záznamů) a velkým množstvím vstupních proměnných (30 a více). Případně by délka klasifikace trvala několik hodin.

V následující části této kapitoly budou všechny tyto klasifikátory aplikovány na datové soubory definované v kapitole v 4.3.2. Princip klasifikace je následující. Nejprve se datový soubor rozdělí v určitém poměru na trénovací a testovací data. Tento poměr jsem zvolil 66 % trénovací data a 34 % testovací. Poté je sestaven klasifikační model na trénovacích datech a ten se pak testuje na testovací množině. Pro každou podúlohu bude sestavena specifická tabulka, kde bude pro každý klasifikátor uvedena úspěšnost predikce hodnot na trénovací množině dat v jednotlivých datových souborech.

***Tab. 16: Výsledky klasifikace platových tříd (INCEARN_INT125)***

| Název algoritmu | DATASET1 | DATASET2 | DATASET3 |
|---|---|---|---|
| ZeroR | 19,44 % | 19,44 % | 19,44 % |
| OneR | 41,93 % | 41,93 % | 90,46 % |
| NaiveBayes | 52,94 % | 48,34 % | 90,46 % |
| BayesNet | 52,82 % | 48,3 % | 90,46 % |
| J48 | 52,86 % | 54,6 % | 93,36 % |
| HoeffdingTree | 52,99 % | 20,7 % | 92,84 % |
| RandomTree | 48,72 % | 45,35 % | 93,35 % |
| DesicisionStump | 31,90 % | 31,91 % | 35,44 % |

***Zdroj: vlastní zpracování***

Během této podúlohy probíhala klasifikace celkem do šesti platových tříd. DATASET3 zde byl pouze pro ukázku toho, že kdyby byla k dispozici nějaká charakteristika, která je velmi silně závislá na cílové proměnné tak bude klasifikace velmi úspěšná. Což se v tomto případě potvrdilo, u většiny algoritmů hodnoty úspěšnosti okolo 90 %. Při porovnání výběrového datasetu (DATASET1) s datasetem do kterého byli zařazeny všechny proměnné si lze všimnout, že úspěšnost klasifikace byla vyjma algoritmu J48 lepší u výběrového datasetu. Bylo to pouze o 3-4%. Výjimkou je HoeffDingTree, kde se ukázalo, že si tento klasifikátor neumí poradit s vysokým počtem proměnných.

Jako nejúspěšnější klasifikátor se v této podúloze jeví rozhodovací strom J48. Je potřeba podotknout, že tento algoritmus je velmi náročný na využití paměti a výkon procesoru v porovnání s bayesovskými algoritmy. Vytvořit klasifikační model nad všemi záznamy výběrového šetření ACS pomocí algoritmu J48 by bylo prakticky nemožné na běžných PC sestavách. Z tohoto důvodu se musí limitovat buďto počet záznamů nebo vstupních proměnných, případně překódovat proměnné a snížit počet hodnot, kterých můžou vybrané proměnné nabývat. Bayesovské algoritmy takovéto problémy nemají a v případě klasifikace platových tříd dosáhli obdobných výsledků. U DATASETu1 byli dokonce algoritmus NaiveBayes o 0,08% lepší při klasifikaci než algoritmus J48.

***Tab. 17: Výsledky klasifikace platových tříd (INCEARN_4INT)***

| Název algoritmu | DATASET1 | DATASET2 | DATASET3 |
|---|---|---|---|
| ZeroR | 25,72 % | 25,72 % | 25,72 % |
| OneR | 53,8 % | 53,81 % | 79,34 % |
| NaiveBayes | 65,2 % | 60,51 % | 80,66 % |
| BayesNet | 65,1 % | 60,47 % | 80,66 % |
| J48 | 65,05 % | 67,2 % | 82,63 % |
| HoeffdingTree | 64,06 % | 32,68 % | 82,60 % |
| RandomTree | 61,33 % | 52,52 % | 82,61 % |
| DesicisionStump | 42,65 % | 42,65 % | 41,96 % |

***Zdroj: vlastní zpracování***

Tato podúloha je pouze modifikací úlohy předchozí. Pouze se upravila klasifikační třída. Místo platových tříd s intervalem 12,5 tisíce dolarů byly vygenerovány 4 ekvifrekvenční intervaly. Osobně jsem předpokládal, že úspěšnost klasifikace v této podúloze bude vyšší, zejména kvůli snížení počtu klasifikačních tříd. Tento předpoklad se potvrdil. Úspěšnost klasifikace se zvýšila v případě DATASETu1 a DATASETU2 řádově o 11-13% na 65 %, což je dle mého názoru již celkem solidní výsledek. I zde se potvrdilo, že je mírně úspěšnější klasifikace výběrového datasetu, vyjma algoritmu J48.

Naopak u DATASETU3 došlo ke zhoršení úspěšnosti klasifikace o 10-11%. To má velmi jednoduché vysvětlení. Byla přeškálována pouze proměnná s celkovým platem, ale proměnná s celkovým příjmem zůstala na původních hodnotách, kde byly použity podobné intervaly jako u cílové třídy INCEARN_INT125.

***Tab. 18: Výsledky klasifikace typu domácnosti (HH_TYPE)***

| Název algoritmu | DATASET4 | DATASET5 |
|---|---|---|
| ZeroR | 63,82 % | 63,82 % |
| OneR | 68,48 % | 68,48 % |
| NaiveBayes | 71,63 % | 62,22 % |
| BayesNet | 71,73 % | 59,55 % |
| J48 | 72,18 % | 76,62 % |
| HoeffdingTree | 72,16 % | 63,87 % |
| RandomTree | 72,17 % | 70,35 % |
| DesicisionStump | 63,82 % | 63,82 % |

***Zdroj: vlastní zpracování***

Typ domácnosti je velmi specifická proměnná, protože hned 61 % hodnot je typu rodinná domácnost. I přesto si s tím vybrané klasifikátory poradily celkem solidně. U výběrového datového souboru (DATASET4) se úspěšnost klasifikace pohybovala kolem 72 %. Tradičně nejúspěšnější byl klasifikátor J48, který měl úspěšnost klasifikace u DATASETU5 76,6 %. V porovnání s první

podúlohou si lze všimnout zlepšení klasifikace téměř o 300 % pomocí algoritmu ZeroR, má to právě souvislost s tím, že je velká podíl hodnot v jedné kategorii. Dokonce i HoeffdingTree a DesicionStump u datového souboru se všemi proměnnými mají výrazně lepší úspěšnost než u předchozí podúlohy. Zajímavostí je poměrně nízká úspěšnost pomocí BayesNet u DATASETU5, ve všech předchozích podúlohách dosahoval přibližně stejné úspěšnosti jako NaïveBayes. V tomto případě je o 2,7% horší. U DATASETU5 jsou obecně bayesovské klasifikátory horší než klasifikátory založené na rozhodovacích stromech a pravidlech

***Tab. 19: Výsledky klasifikace typu lokality (METRO)***

| Název algoritmu | DATASET6 | DATASET7 |
|---|---|---|
| ZeroR | 45,95 % | 45,95 % |
| OneR | 57,44 % | 57,44 % |
| NaiveBayes | 58,95 % | 57,06 % |
| BayesNet | 58,84 % | 58,19 % |
| J48 | 59,29 % | N/A |
| HoeffdingTree | 58,75 % | 46,03 % |
| RandomTree | 59,15 % | 49,36 % |
| DesicisionStump | 48,12 % | 51,67 % |

***Zdroj: vlastní zpracování***

Úspěšnost klasifikace této podúlohy byla nejvyšší. U těch nejlepších algoritmů dosahovala až k hranici 80 %. Zajímavostí je, že i přesto, že DATASET7 byl velmi podobný DATASETu5 z předchozí podúlohy, tak se nepodařilo provést klasifikaci pomocí algoritmu J48. Na tuto úlohu alokováno přes 4GB fyzické paměti a přesto to nestačilo. SW nástroj Weka, ve kterém byly veškeré analýzy provedeny, byl vždy ukončen s chybnou hláškou o nedostatku fyzické paměti. Jedná se o volně dostupný nástroj a dle mého názoru má celkem problémy se správou a využitím fyzické paměti. Toto je důvod, proč není v příslušném poli tabulky uvedena hodnota.

Zajímavostí této podúlohy vysoká úspěšnost bayesovských klasifikátorů a to jak v případě výběrových dat, tak i v případě DATASETu7, ve kterém byly zahrnuty všechny vybrané proměnné. Vysoká úspěšnost klasifikace v této podúloze je dle mého názoru způsobena silnou závislostí proměnné METRO na proměnné PUMA. PUMA je vlastně geografická oblast, která slučuje minimálně 100 tisíc obyvatel, tudíž dle mého názoru se tyto jednotky nachází s velkou pravděpodobností v jednom typu lokality.

### 4.4.2 Hledání nugetů

***Tab. 20: Použité 4FT miner kvantifikátory***

| Název | Zápis | Interpretace |
|---|---|---|
| Fundovaná implikace (FUI) | $\varphi \Rightarrow_{p,Base} \psi$<br>$\frac{a}{a+b} \geq p \wedge a \geq Base$ | Zajímá nás, zda platnost nějaké kombinace φ znamená s vysokou pravděpodobností i platnost nějaké jiné kombinace ψ (RAUCH et. al., 2015). |

| | | |
|---|---|---|
| AA kvantifikátor | $\varphi \Rightarrow^{+}_{p,Base} \psi$<br>$\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge$<br>$\wedge\, a \geq Base$ | Zajímá nás, zda platnost nějaké kombinace znamená výrazné zvýšení relativní četnosti nějaké jiné kombinace (RAUCH et. al., 2015). |

***Vysvětlení vzorců: parametry „a“, „b“, „c“, „d“ jsou hodnoty ze čtyřpolní tabulky (obrázek 6).***

***P a Base jsou dva parametry pomocí níž lze omezit počet nalezených hypotéz***

***Zdroj: uveden v tabulce, vlastní zpracování***

V tabulce 20 jsou popsány 2 základní 4FT kvantifikátory, které budou v následující části práce použity k nalezení zajímavých hypotéz. Jak bylo popsáno v kapitole 2.4.2 v části o generování kombinací asociačních pravidel, jedná se o velmi výpočetně složitý proces. Vzhledem k faktu, že bylo pro tuto analýzu vybráno 51 proměnných a každá proměnná může nabývat vyššího počtu hodnot (v extrémních případech těch hodnot je přes 200 – OCC a IND). Je tedy zřejmé, že celkový počet možných kombinací je velmi vysoké číslo. Tudíž je potřeba tedy počet testovaných hypotéz nějak omezit. Toto lze v zásadě dvěma způsoby. Prvním způsobem je omezit proměnné, které jsou zařazeny do antecedentu (předpokladu) a succedentu (závěru). Druhou možností je omezit určitým způsobem hodnoty, jednotlivých atributů, které vstupují do analýzy. K tomu je v rámci procedury 4 FT-Miner několik možností. Tyto možnosti jsou popsány v tabulce 21. V praxi se nejčastěji používá kombinace obou těchto možností. Tímto lze omezit počet testovaných hypotéz. Poté je ještě možnost upravit počet výsledných (zajímavých) hypotéz a to pomocí parametrů „base“ a „p“.

***Tab. 21: Možnosti omezení počtu generovaných hypotéz - parametry literálů***

| | |
|---|---|
| Subset | Podmnožiny kategorií zadané velikosti. Při zadané hodnotě 1 je možná u každého atributu pouze jedna hodnota. Pokud by tedy byl hypotetický příklad, kde jeden atribut může nabývat pěti různých hodnot, budou sestaveny hypotézy pro všech 5 hodnot tohoto atributu.<br>Při subset = 2 se sestavují hypotézy, kde se uvažují maximálně dvě hodnoty pro daný atribut. Existuje tedy 15 různých kombinací pro atribut v tomto hypotetickém příkladě.<br>Tedy čím nižší hodnota subset tím méně se testuje hypotéz. |
| One category | Sestavují se hypotézy pouze pro jednu kategorii pro daný atribut. |
| Sequence | Lze využít pro ordinální proměnné. Nastavuje se minimální a maximální délka intervalu. Lze využít např. u věku, že se budou sestavovat pro 5leté intervaly následujícím způsobem: 0-4, 1-5, 2-6 atd. |
| Cyclical sequence | Podobné jako sequence akorát s tím rozdílem, že lze využít tzv. zacyklení, což se hodí třeba u dnů v týdnu, že pokud by byl příklad, že budou sestavovány hypotézy interval délky 3 |

| | |
|---|---|
| | pro atribut den v týdnu tak by poslední nebyl literál „pátek – sobota neděle“, ale „neděle, pondělí, úterý“. |
| Cuts | Levý i pravý řez. Lze využít pro analýzu extrémních hodnot nějakého atributu. Lze zadat počet extrémních hodnot, který se bude testovat. V případě cuts se berou hodnoty „z obou stran“ soubor, tedy jak maximální tak i minimální.<br>Lze modifikovat pomocí tzv. „Left Cuts“ a „Right cuts“ tudíž se budou testovat pouze minimální nebo maximální hodnoty. |

***Zdroj: (RAUCH et. al., 2015), (LISPMINER, 2015)***

V následující části kapitoly budou popsány jednotlivé provedené jednotlivé testy. Testy budou popsány následující formou, vždy bude uveden nějaký předpoklad (neboli analytická otázka), bude následovat použitý 4FT kvantifikátor a jeho parametry. Následovat bude obecný zápis analytické otázky (použité atributy a koeficienty) a poté seznam nalezených hypotéz včetně jejich interpretace.

## Analýza 1

***Tab. 22: Analýza 1 - protokol***

| | |
|---|---|
| Analytická otázka | Je možné zařadit pracovníky v nějakém typu průmyslu (IND) a na určité pracovní pozici (OCC) do jedné nebo dvou platových skupin. |
| Parametry | Kvantifikátor: fundovaná implikace<br>p = 0,9<br>Base = 30 |
| Obecný zápis | ?: ,Base; OCC(subset=1) $\wedge$ IND(subset=1) $\Rightarrow_{p,Base}$ INCEARN_INT125(subset=2) |
| Počet testovaných kombinací | 3 594 416 |
| Počet nalezených hypotéz | 94 |
| Čas analýzy | 34 min 38sec |
| Příklady nalezených hypotéz | • Ind(Unemployed, last worked 5 years ago) & Occ(Unemployed, with No Work) >÷< Incwage_int125(0)<br>• Ind(Motion pictures and video industries) & Occ(Counter Attendant, Cafeteria, Food Concession, and Coffee Shop) >÷< Incwage_int125(0, 0-12,5)<br>• Ind(Recreational vehicle parks and camps, and rooming and boarding houses) & Occ(Recreation and Fitness Workers) >÷< Incwage_int125(0, 0-12,5)<br>• Ind(Religious organizations) & Occ(Childcare Workers) >÷< Incwage_int125(0, 0-12,5) |

| | |
|---|---|
| | • Ind(Other general government and support) & Occ(Miscellaneous office and administrative support workers including desktop publishers) >÷< Incwage_int125(0, 0-12,5)<br>• Ind(Electronic component and product manufacturing) & Occ(Engineering Managers) >÷< Incwage_int125(50-75, 75+)<br>• Ind(Electric power generation, transmission and distribution) & Occ(General and Operations Managers) >÷< Incwage_int125(50-75, 75+)<br>• Ind(Colleges, universities, and professional schools, including junior colleges) & Occ(Residential Advisors) >÷< Incwage_int125(0, 0-12,5)<br>• Occ(Counter Attendant, Cafeteria, Food Concession, and Coffee Shop) >÷< Incwage_int125(0, 0-12,5) |

***Zdroj: vlastní zpracování***

Výsledné hypotézy lze interpretovat následovně: Minimálně 90 % respondentů pracujících v daném oboru a vykonávající danou funkci patří do daných platových tříd.

Většina výsledků této analýzy pravděpodobně moc lidí nepřekvapí. Nicméně pro mě je celkem překvapivé, že byl nalezen tak malý počet výsledných hypotéz. Očekával jsem, že příjmová skupina půjde lépe klasifikovat na základě oboru a pozice.

## Analýza 2

***Tab. 23: Analýza 2 - protokol***

| | |
|---|---|
| Analytická otázka | Je možné určit počet odpracovaných hodin v nějakém typu průmyslu (IND) a na určité pracovní pozici (OCC). |
| Parametry | Kvantifikátor: fundovaná implikace<br>p = 0,9<br>Base = 30 |
| Obecný zápis | ?: ,Base; OCC(subset=1) ∧ DATASET8 (IND(subset=1)) $\Rightarrow_{p,Base}$ UhrsWORK_INT20(subset=1) |
| Počet testovaných kombinací | 641 860 |
| Počet nalezených hypotéz | 66 |
| Čas analýzy | 8 min 4sec |
| Příklady nalezených hypotéz | • Ind(Religious organizations) & Occ(Childcare Workers) >÷< Uhrswork_int20(0-25)<br>• Occ(Tire Builders) >÷< Uhrswork_int20(26-45) |

| | |
|---|---|
| | • Ind(Justice, public order, and safety activities) & Occ(Crossing Guards) >÷< Uhrswork_int20(0-25)<br>• Ind(Hospitals) & Occ(Billing and Posting Clerks) >÷< Uhrswork_int20(26-45)<br>• Ind(Justice, public order, and safety activities) & Occ(Secretaries and Administrative Assistants) >÷< Uhrswork_int20(26-45)<br>• Ind(Banking and related activities) & Occ(Secretaries and Administrative Assistants) >÷< Uhrswork_int20(26-45) |

***Zdroj: vlastní zpracování***

Výsledné hypotézy lze interpretovat následovně: Minimálně 90 % respondentů pracujících v daném oboru a vykonávající danou pracuje na daný úvazek.

Valná většina z 66 hypotéz byla kombinace průmyslu a zastávané funkce a počtu odpracovaných hodin z intervalu 26–45, což odpovídá jednomu úvazku (považuji za relativně nezajímavé hypotézy). Byly nalezeny také dvě hypotézy, které se týkaly částečného úvazku (interval 0–25).

## Analýza 3

***Tab.: 24: Analýza 3 - protokol***

| | |
|---|---|
| Analytická otázka | Je možné určit respondentův dopravní prostředek, který využívá k přepravě na základě jeho rasy, státu, rodinného stavu nebo toho zda žije v centru nebo na okraji města |
| Parametry | Kvantifikátor: fundovaná implikace<br>p = 0,95<br>Base = 500 |
| Obecný zápis | ?: DATASET8; MARST(subset=1) ∧ STATEICP(subset=1) ∧ (RACE(subset=1) ∧ METRO(subset=1) $\Rightarrow_{p,\ Base}$ TRANWORK(subset=1) |
| Počet testovaných kombinací | 24516 |
| Počet nalezených hypotéz | 21 |
| Čas analýzy | 2 min 47sec |
| Příklady nalezených hypotéz | • Stateicp(Alabama) & Race(negroidní) & Marst(ženatý/vdaná) >÷< Tranwork(auto)<br>• Stateicp(Alabama) & Race(negroidní) >÷< Tranwork(Auto) |

| | |
|---|---|
| | • Metro(okraj metropole) & Stateicp(Louisiana) & Race(europoidní) & Marst(ženatý/vdaná) >÷< Tranwork(auto)<br>• Stateicp(Louisiana) & Marst(ženatý/vdaná) >÷< Tranwork(Auto)<br>• Metro(nemetropilitní oblast) & Stateicp(Kentucky) & Marst(ženatý/vdaná) >÷< Tranwork(auto)<br>• Stateicp(South Carolina) & Race(negroidní) & Marst(vdaní) >÷< Tranwork(auto) |

***Zdroj: vlastní zpracování***

Výsledné hypotézy lze interpretovat následovně: Minimálně 95 % respondentů daného rodinného stavu, dané rasy, bydlícího v daném státě v na nemetropolitní oblasti, okraji města nebo v centru jezdí do práce daným dopravním prostředkem.

Tedy první zde zmíněné pravidlo by se dalo interpretovat takto. Vdaní/ženatí zástupci negroidní rasy z Alabamy jezdí do práce autem. Toto pravidlo jde pomocí dalšího pravidla zobecnit pouze na negroidní obyvatele z Alabamy.

## Analýza 4

***Tab.: 25: Analýza 4 - protokol***

| | |
|---|---|
| Analytická otázka | Je možné nalézt nějaké společné rysy (pojištění, ekonomická aktivita, příjmová skupina sociálních dávek, hledání práce v posledních 5 letech, vzdělání) pro respondenty s různými typy zdravotních postižení. |
| Parametry | Kvantifikátor: fundovaná implikace<br>p = 0,90<br>Base = 30 |
| Obecný zápis | ?: DATASET8; DIFFCARE(onecategory=2) ∧ DIFFMOB(onecategory=2) ∧ DIFFPHYS(onecategory=2) ∧ DIFFREM(onecategory=2) $\Rightarrow_{p,Base}$ METRO(subset=1) ∧ INCSS_INT5(subset=1) ∧ LOOKING(subset=1) ∧ HCOVANY(subset=1) ∧ EMPSTAT(subset=1) ∧ EDUCD_CZCONV(subset=1) |
| Počet testovaných kombinací | 114 304 |
| Počet nalezených hypotéz | 37 |
| Čas analýzy | 5 min 53sec |

| | |
|---|---|
| Příklady nalezených hypotéz | • Diffcare(Není se o sebe schopen postarat) & Diffmob(trpí poruchou mobility) & Diffphys(trpí fyzickou poruchou) & Diffrem(trpí kognitivní poruchou) >÷< Hcovany(Je pojištěn)<br>• Diffcare(Není se o sebe schopen postarat) & Diffmob(trpí poruchou mobility) & Diffphys(trpí fyzickou poruchou) & Diffrem(trpí kognitivní poruchou) >÷< Empstat(ekonomicky neaktivní) |

***Zdroj: vlastní zpracování***

Z 37 hypotéz jsem vybral pouze tyto dvě, protože zbytek jsou víceméně pouze „deriváty".

Dvě zmíněné hypotézy asi nikoho nepřekvapí, že je pravidlem, že lidé s určitým postižením jsou většinou ekonomicky neaktivní a jsou také pojištění u nějaké pojišťovny. Zajímavé ovšem je, že se nejedná o 100 % lidí, ale pouze o 96,8 %. Tedy pouze 96,8 %, kteří trpí zároveň všemi těmito druhy postižení, jsou někde pojištěni. Kdybych se to snažil nějak rozklíčovat, tak je pojištěno 95,8 % lidí, kteří se mají problém starat sami o sebe, 94,6 % lidí s pohybovými problémy, 93,4 % s fyzickými problémy a pouze 90,4 % lidí s kognitivní poruchou.

Dále nebyla nalezena žádná další pravidla, která by signalizovala něco o tom, že by postižení trpící těmito poruchami hledali práci nebo se nacházeli v podobných příjmových skupinách sociálních dávek apod.

## Analýza 5

***Tab.: 26: Analýza 5 - protokol***

| | |
|---|---|
| Analytická otázka | Zajímají nás státy, ve kterých je nadprůměrný výskyt respondentů nějaké rasy. |
| Parametry | Kvantifikátor: AA<br>p = 1,50<br>Base = 30 |
| Obecný zápis | ?: DATASET8; RACE(subset=1) $\Rightarrow^{+}_{p,Base}$ B STATEICP(subset=1) |
| Počet testovaných kombinací | 510 |
| Počet nalezených hypotéz | 23 |
| Čas analýzy | 1 min 3sec |
| Příklady nalezených hypotéz | • Race(japonská) >÷< Stateicp(Hawaii) – 50,958<br>• Race(míšenec tři nebo více ras) >÷< Stateicp(Hawaii) – 36,107<br>• Race(americký indián / původní obyvatel Aljašky) >÷< Stateicp(Alaska) – 28,914 |

| | |
|---|---|
| | • Race(americký indián / původní obyvatel Aljašky) >÷< Stateicp(New Mexico) - 14,109<br>• Race(americký indián / původní obyvatel Aljašky) >÷< Stateicp(South Dakota) – 7,9<br>• Race(americký indián / původní obyvatel Aljašky) >÷< Stateicp(Oklahoma) - 7,8<br>• Race(čínská) >÷< Stateicp(Hawaii) - 7,3<br>• Race(negroidní) >÷< Stateicp(District of Columbia) – 3,5 |

***Zdroj: vlastní zpracování***

Pomocí AA kvantifikátoru lze nalézt hodnoty, které se výrazně odlišují od průměru. Toto je nejjednodušší možný příklad úlohy s využitím tohoto kvantifikátoru. Výsledky tedy pravděpodobně nikoho se základní znalostí zeměpisu zásadně nepřekvapí.

Nicméně je možné interpretovat takto. Na Havaji žije 51krát více Japonců než je celostátní průměr. Podobně lze prohlásit, že nadprůměrný počet původních obyvatel žije ve státech Aljaška, v Novém Mexiku, Jižní Dakotě a v Oklahomě.

## Analýza 6

***Tab.: 27: Analýza 6 - protokol***

| | |
|---|---|
| Analytická otázka | Zajímají nás státy, ve kterých je nadprůměrný výskyt respondentů na základě kombinace věku, rodinného stavu a typu domácnosti ve kterém žijí. |
| Parametry | Kvantifikátor: AA<br>p = 1,50<br>Base = 30 |
| Obecný zápis | ?: DATASET8; RACE (subset=1) ∧ HHTYPE(subset=1) ∧ MARST (subset=1) $\Rightarrow^{+}_{p,Base}$ B STATEICP(subset=1) |
| Počet testovaných kombinací | 45 135 |
| Počet nalezených hypotéz | 37 |
| Čas analýzy | 2 min 7sec |
| Příklady nalezených hypotéz | • Hhtype(Žena v čele nežijící sama) & Age_int5(25-29) >÷< Stateicp(District of Colombia) – 5,6<br>• Hhtype(Žena žijící sama) & Age_int5(30-34) >÷< Stateicp(District of Colombia) – 4,6<br>• Marst(Vdané / ženatí) & Age_int5(20-24) >÷< Stateicp(Utah) – 2,7 |

| | |
|---|---|
| | • Marst(Vdané / ženatí, družba nepřítomna během sčítání) & Hhtype(Manželský pár) >÷< Stateicp(Hawaii) - 2,17<br>• Hhtype(Muž v čele nežijící sám) >÷< Stateicp(District of Colombia) – 1,59<br>• Hhtype(Muž v čele nežijící sám) & Age_int5(80-84) >÷< Stateicp(Florida) – 1,56 |

***Zdroj: vlastní zpracování***

Z nalezených 37 hypotéz se hned polovina z nich týkala státu District of Colombia. Dle mého názoru je to tím, že se jedná o "městský" stát tudíž tam panují odlišné zvyklosti, co se týče bydlení.

Z těch zajímavějších výsledků stojí za zmínku, že nadprůměr respondentů vdaných / ženatých ve věku 20–24 let se nachází v Utahu. Na Floridě je nadprůměrný výskyt domácností s muži v čele (co nežijí sami) ve kterých se vyskytují respondent ve věku 80–84 let.

## 4.5 Hodnocení výsledků

Dle obecného postupu dle metodiky CRISP-DM by nyní měly následovat dvě podkapitoly zaměřené na interpretaci výsledků a zanesení výsledků do praxe. Dle mého názoru lze dané kapitoly v rámci této modelové úlohy sloučit do jedné. Zejména kvůli faktu, že jsem zadavatelem úlohy i autorem analýzy, mi přijde zbytečné dělat speciální výstupy (Powerpoint prezentace, detailnější reporty) pro zadavatele úlohy, což je obvyklá náplň poslední fáze dle metodiky CRISP-DM. Kapitola bude rozdělena opět na dvě části, a to na zhodnocení výsledků klasifikační úlohy a následné zhodnocení výsledků úlohy hledání nuggetů.

### 4.5.1 Klasifikace

Byly provedeny celkem 4 podúlohy s využitím 8 druhů klasifikační algoritmů. Úspěšnost odhadu jednotlivých algoritmů se lišila atribut od atributu. Nicméně nenastal případ, že by byla úspěšnost nižší než 50 % pro minimálně jeden klasifikační algoritmus v rámci každé podúlohy. Nejnižší úspešnost měla klasifikace do platových tříd s intervalem 12,5 tisíce. Zde se pohybovala úspěšnost klasifikace 52 %. Vzhledem k faktu, že se jednalo o klasifikaci celkem do 6 tříd s víceméně rovnoměrně rozloženými hodnotami v rámci jednotlivých klasifikačních tříd, tak považuji tento výsledek za celkem kvalitní. Při konstrukci modelu nad DATASETEM2 (který obsahoval veškeré vybrané atributy v rámci této práce) se za pomocí algoritmu J46 dosáhlo dokonce úspěšnosti 55 %.

Při mírné modifikaci úlohy klasifikace platových tříd (klasifikace do 4 ekvifrevenčních tříd s automaticky generovanými hranicemi) bylo dosaženo úspěšnosti klasifikace 65–67 % u nejlepších klasifikátorů. Což je řádově o 10–13 % více než při klasifikaci do 6 tříd. Pokud by byl k dispozici např. z nějakého registru celkový příjem dané osoby rozdělený do podobných intervalů, jako má cílová klasifikační proměnná s platovými třídami, lze dosáhnout úspěšnosti klasifikace platové třídy 93 %.

Úspěšnost klasifikace u proměnné „typ domácnosti" se pohybovala na úrovni 70–76 % u těch nejlepších klasifikačních algoritmů. „Typ domácnosti" je velmi specifická proměnná, protože existuje 7 klasifikačních tříd, kde jedna z těchto tříd obsahuje vysoký podíl hodnot (třída manželský pár obsahuje 61 % hodnot). Kdyby se použila nejprimitivnější metoda nahrazení chybějících údajů uvedená v kapitole 3.4.2 (nahrazení pomocí modální hodnoty, což se přímo nabízí na podobných

datech) lze předpokládat úspěšnost klasifikace řádově 61 %. Na podobném principu je založen algoritmus ZeroR, pomocí něhož bylo dosaženo úspěšnosti klasifikace 61,8 %. Nejvyšší úspěšnost v této podúloze zaznamenal opět algoritmus J46, a to 76 %, což je tedy o 14% více, než pokud by byly chybějící hodnoty nahrazeny modální hodnotou.

Poslední úloha byla zaměřena na klasifikaci typu lokality, ve které se nachází domácnost. V této podúloze byla úspěšnost klasifikace vůbec nejvyšší, bylo dosaženo úspěšnosti 80 %. Zde to lze přisuzovat patrně tomu, že jedna z proměnných byla územní jednotka PUMA, která obsahovala údaje vždy za 100 tisíc obyvatel. Lze předpokládat, že v rámci jedné jednotky jsou většinou domácnosti ve stejném typu lokality, nicméně asi to není pravidlem, protože jinak by byla úspěšnost klasifikace ještě vyšší. Bohužel se v rámci této podúlohy z technických důvodů nepodařila klasifikace pomocí algoritmu J46. Na základě výsledků z předchozích úloh by bylo možné předpokládat, že by byla úspěšnost někde kolem 82–84 %.

Pokud bych měl vybrat obecně nejúspěšnější klasifikátor, pak by se jednalo o algoritmus J46, minimálně z těch, které byly zařazeny do této úlohy. Tento algoritmus byl většinou nejúspěšnější ve všech datových souborech s jednou výjimkou, a tou byl DATASET1, kde byly lepší bayesovské algoritmy. Problémem tohoto algoritmu je vysoká HW náročnost, zejména náročnost na paměť. V SW nástroji Weka, který byl použit pro klasifikační úlohu se zdál jako strop 40 vstupních atributů a 250 tisíc testovaných pozorování. Určitě také záleží na počtu hodnot, kterých může atribut nabývat, čím méně hodnot, tím menší spotřeba paměti (která byla v SW nástroji Weka tím hlavním limitem), tím větší rychlost generování modelu. Jako další v pořadí co do úspěšnosti lze zařadit bayesovské klasifikátory, které dosahovaly srovnatelných výsledků s J46 (většinou horší řádově 1–2%). Naopak nejméně úspěšné klasifikátory byly oba algoritmy fungující na principu rozhodovacích pravidel (OneR, ZeroR) a rozhodovací strom DecisionStump.

Odpověď na otázku, zda se při klasifikační úloze vyplatí provádět výběrový datový soubor pro klasifikaci, není jednoznačná. Oba bayesovské klasifikátory a všechny algoritmy založené na principech rozhodovacích stromů (vyjma J46) dosahovaly lepších výsledků nad výběrovým datovým souborem, a to řádově o 15–25 %. Pro rozhodovací pravidla byla úspěšnost totožná, je to díky tomu, jak jsou tyto algoritmy konstruovány. Naopak při klasifikaci pomocí algoritmu J46 se vyplatí mít zde co nejvíc atributů, protože zde platí přímá úměra, že čím víc atributů, tím je úspěšnost vyšší. Je to tím, že se jedná o zoptimalizovaný rozhodovací strom, kde se díky prořezávání odstraní všechny větve (případně listy) tohoto stromu, které nezvyšují úspěšnost klasifikace.

Pokud bych měl na závěr sdělit svůj názor na využití těchto metod v praxi, tak si myslím, že pro určitý typ atributů (kategoriálních) se nejedná o špatnou metodu a že za pomoci těchto metod lze dosáhnout relativně vysoké úspěšnosti za poměrně „nízkých nákladů“ v podobě času (vyexportovat a připravit data pro klasifikaci je otázka několika málo hodin, pro zkušené uživatele i minut). Bohužel v datech z IPUMS nebyl dostatek kvalitativních atributů k otestování dalších klasifikátorů, jako např. neuronových sítí, algoritmů založených na analogii apod. Nicméně je nutné položit si před využitím těchto metod otázku, zda 55–80 % úspěšnost klasifikace chybějících hodnot je dostatečná pro potřeby daného výzkumu.

### 4.5.2 Hledání nuggetů

V rámci těchto analýz bylo otestováno několik milionů potenciálně zajímavých kombinací. Na základě těchto testů byly nalezeny desítky pravidel. V dané fázi procesu u těchto typů úloh by měla přijít diskuze s odborníkem (zadavatelem úlohy), aby odpověděl na klíčovou otázku, zda se jedná o zajímavá pravidla neboli nové dosud nezjištěné vlastnosti, a nebo se jedná o obecně známá fakta. Já jako student s téměř nulovou praxí a pouze obecnými znalostmi o USA se za takového odborníka prohlásit netroufám.

Z nalezených hypotéz by se za „ověřená“ fakta dalo považovat např. to, že na Havaji žije nadprůměrný počet lidí mongoloidní rasy (původem z Japonska) nebo že nadprůměrný počet původních obyvatel a indiánských kmenů žije na území států Aljašky, Nového Mexika a Jižní Dakoty. To, že postižení lidé s různými poruchami mají zdravotní pojištění a jedná se o ekonomicky neaktivní lidi, také asi nikoho nepřekvapí, ovšem fakt, že se mezi takovými jedinci vyskytuje podíl lidí bez zdravotního pojištění, mě osobně překvapil.

Dle mého názoru bylo nalezeno několik potenciálně zajímavých hypotéz. Mezi ně lze zařadit např. to, že reprezentanti negroidní rasy v Alabamě jezdí do práce autem. Zajímavý, dle mého názoru, je fakt, že zmíněné bylo prokázáno pro tento typ lidí pouze v Alabamě, V žádném jiném státě a pouze pro negroidní část obyvatelstva. Zde by se dal např. přizvat ke konzultaci sociolog a dalo by se zkoumat, proč tomu tak je (je za tím rasismus, že se negroidní obyvatelé bojí jezdit autobusem nebo jsou tak bohatí, že si všichni mohou dovolit jezdit do práce autem apod.). To je příklad, že i ze zdánlivě pro někoho nezajímavé hypotézy lze vydedukovat určité zajímavé závěry.

Mezi další zajímavé hypotézy lze zařadit to, že ve státu Utah žije nadprůměrný počet sezdaných lidí ve věku 20–24 let. Dále také to, že na Floridě žije nadprůměrný počet mužů „na hromádce“ ve věku 80–84 let. Další poměrně zajímavou věcí je, že v manželských domácnostech na Havaji nebyl v nadprůměrném počtu případů partner přítomen během sčítání. Zde je zase možné vysvětlení, že byl na moři z pracovních důvodů.

Mezi další, dle mého názoru, zajímavé zjištění patří to, že v USA nelze u většiny povolání pouze za pomoci oblasti průmyslu a pozice klasifikovat s více než 90 % platovou třídu a to ani za předpokladu, že je možné klasifikovat do dvou tříd zároveň. Ze 3,6 milionů testovaných verifikací se podařilo klasifikovat pouze v 97 případech. Podobně se nepodařilo klasifikovat na stejném základě počet odpracovaných hodin, z 641 tisíc verifikací bylo nalezeno 66 hypotéz.

Doufám, že se mi podařilo pomocí těchto analýz dokázat, že existuje využití daných metod v praxi při zpracování dat z výzkumů. U hledání nuggetů je komunikace mezi zadavatelem a analytikem velmi důležitá, protože je potřeba omezit a specifikovat oblast, ve které se mají zajímavé znalosti vyhledávat (pomocí nastavení hodnot parametrů). Já jsem například vyhledával hypotézy, ve kterých je předpoklad i závěr splněn ve více než 90 %, což je velmi vysoké číslo. V některých praktických úlohách se např. považuje za úspěch, pokud je nalezena hypotéza s spolehlivostí okolo 75 %.

## Kapitola 5

# Závěr

Cílem této práce bylo představit jednotlivé data miningové metody a poté realizovat DM úlohu za využití obecné metodiky procesu DZD na reálných datech. Pro tuto úlohu byla zvolena data z amerického výběrového šetření ACS, což je obdoba SLDB v ČR. Tato data jsou každoročně publikována jako součást projektu IPUMS. V rámci praktické úlohy byly se získanými daty provedeny dva typy DM úloh. Prvním typem úlohy byla klasifikační úloha, jejímž cílem bylo za pomoci sestavení modelu klasifikovat hodnoty vybraných atributů. Tento způsob lze teoreticky využít při odhadu chybějících hodnot při zpracování dat ze statistických šetření. Druhým typem úlohy bylo hledání zajímavých charakteristik nad zdrojovými daty za využití asociačních pravidel a metody GUHA.

Při klasifikační úloze byla analyzována úspěšnost klasifikace čtyř vybraných atributů s poměrem chybějících hodnot vyšším než 10 %. Jednalo se o platovou třídu (intervaly 12,5 tisíc dolarů), platovou třídu (4 ekvifrekvenční intervaly), typ lokality a typ domácnosti. Pro každou klasifikační úlohu byl sestaven modelový datový soubor s vybranými atributy, u nichž se dá předpokládat, že jejich hodnoty ovlivňují cílovou (klasifikační) třídu. Posléze byly výsledky porovnány s klasifikací s využitím všech atributů vybraných v rámci tohoto výzkumu. Každý datový soubor byl klasifikován pomocí 8 vybraných klasifikačních algoritmů. Čtyři algoritmy byly ze skupiny rozhodovacích stromů, 2 ze skupiny rozhodovacích pravidel a 2 ze skupiny bayesovských klasifikátorů.

Úspěšnost klasifikace nejúspěšnějších klasifikátorů se pohybovala v intervalu 55–80 %, což znamená, že v tolika procentech případů byla správně odhadnuta hodnota klasifikovaného atributu. Nejúspěšnější klasifikátor byl rozhodovací strom J46. Jednalo se také o jediný klasifikátor, který měl vyšší úspěšnost klasifikace nad datovým souborem se všemi atributy. V ostatních případech byla úspěšnost klasifikace nad modelovým datovým souborem s vybranými atributy vyšší o 8–15 %. Nutno podotknout, že J46 je velmi náročný algoritmus na využití operační paměti při vysokém počtu atributů s mnoha hodnotami. Na pomyslném druhém místě, co se týče úspěšnosti klasifikace, se umístily oba bayesovské klasifikátory, pomocí nichž bylo dosaženo velmi podobné úspěšnosti klasifikace nad datovým soubor s omezeným počtem atributů jako v případě algoritmu J46. Nad kompletním datovým souborem byly výsledky těchto klasifikátorů horší o 15–25 %.

Pro druhou úlohu bylo sestaveno 6 analytických otázek, které byly řešeny pomocí 2 základních 4-FT kvantifikátorů (fundovaná implikace a AA kvantifikátor). Celkem 4 otázky byly řešeny pomocí fundované implikace a dvě pomocí AA kvantifikátoru. Bylo nalezeno několik desítek potenciálně zajímavých hypotéz. Obecně lze rozdělit nalezené hypotézy do dvou kategorií. První kategorií jsou nové, překvapivé a dosud neznámé znalosti. Toto jsou hypotézy, které se v rámci daných

úloh primárně vyhledávají. Druhým typem jsou určité obecně známé hypotézy, jedná se o pouhé potvrzení faktů.

Dle mého názoru se v rámci této úlohy podařilo najít oba dva typy hypotéz. Za obecně nalezené fakty lze považovat hypotézy nalezené pomocí AA kvantifikátoru, např. na Havaji žije nadprůměrný počet obyvatel japonského původu nebo nadprůměrný počet původních obyvatel žije ve státech Aljaška, Nové Mexiko a v Jižní Dakotě. Podařilo se také najít potenciálně zajímavé hypotézy, např. v Utahu žije nadprůměrný počet sezdaných lidí ve věku 20–24 let nebo na Floridě žije „na hromádce" nadprůměrný počet mužů ve věku 80–84 let. Několik zajímavých pravidel se povedlo nalézt i pomocí kvantifikátoru fundované implikace. Mezi ně lze zařadit např. minimálně 90 % účastníků výzkumu negroidní rasy pocházejících z Alabamy jezdí do práce autem.

Co se týče návaznosti na tuto práci, existuje, dle mého názoru, velké množství možností, protože se jedná o stále poměrně neprobádanou oblast, zejména co se týče propojení data miningu a demografie, především zkoumání vybraných demografických jevů pomocí DM metod. Co se týká návaznosti na problematiku DM a demografických šetření, zde lze také do určité míry navázat. Jako příklad mě napadá klasifikovat primární data z konkrétního reálného šetření a využít přitom všech atributů. Na základě této analýzy pak je možné zjistit, zda úspěšnost klasifikace všech atributů bude stále na úrovni 55–80 %. Případně následně může být provedena statistická analýza špatně klasifikovaných jednotek u vybraných atributů. Poté se mohou porovnat špatně klasifikované vzorky pro více atributů a odhalit tak pozorování, která jsou špatně klasifikována pro všechny vybrané atributy.

# Citovaná literatura

A Brief History of Data Mining, 2007. *Business Intelligence Wiki* [online] [cit. 2015-06-10]. Dostupné z: https://sites.google.com/site/fsubiwiki/home/data-mining/history

AFFENDEY, L. S. a N. MUSTAPHA, 2010. Ranking of Influencing Factors in Predicting Students Academic Performance. *Information Technology Journal*, č 9. Dostupné také z: http://www.researchgate.net/publication/47366475_Ranking_of_Influencing_Factors_in_Predicting_Students_Academic_Performance

BERKA, P., 2003. *Dobývání znalostí z databází.* Praha: Academia. ISBN 80-200-1062-9. Dostupné také z: http://sorry.vse.cz/~berka/4IZ450/

BROŽOVÁ, H., 2007. *Rozhodovací modely a znalostní management*. Praha: Česká zemědělská univerzita. ISBN 978-80-213-1633-1. Dostupné také z: http://etext.czu.cz/php/skripta/kapitola.php?titul_key=78&idkapitola=14

CARBUREANU, M., 2007. The Divorce Rate Prediction using Data Mining Techniques. *Universitatii - Gaze din Ploiesti*, č 2. Dostupné také z: http://bmif.unde.ro/docs/20072/Buletin_UPG_MIF_Nr2_2007-05.pdf

CENSUS.GOV, 2014. Census.gov. *American Community Survey (ACS)* [online] [cit. 2015-07-29]. Dostupné z: http://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html

ČERMÁK, V. a M. VRABEC, 1998. *Teorie výběrových šetření*. 2. díl. Praha : VŠE v Praze: Nakladatelství Oeconomica. ISBN 80-7079-609-X.

ČSÚ, 2012. Český statistický úřad. *Historie sčítání lidu České republiky I.* [online]. 18. 01. 2012 [cit. 2015-07-09]. Dostupné z: https://www.czso.cz/csu/czso/historie_scitani_lidu_na_uzemi_ceske_republiky_i_

ČSÚ, 2012. Český statistický úřad. *Statistika rodinných účtů* [online]. 06. 06. 2012 [cit. 2015-07-09]. Dostupné z: https://www.czso.cz/csu/vykazy/rodinne_ucty

ČSÚ, 2012. Český statistický úřad. *Výběrové šetření příjmů a životních podmínek domácností* [online]. 27. 11. 2012 [cit. 2015-09-07]. Dostupné z: https://www.czso.cz/csu/vykazy/vyberove_setreni_prijmu_a_zivotnich_podminek_domacnosti

ČSÚ, 2014. Český statistický úřad. *Historie sčítání lidu na území České republiky II.* [online]. 13. 11. 2014 [cit. 2015-07-09]. Dostupné z: https://www.czso.cz/csu/czso/historie_scitani_lidu_na_uzemi_ceske_republiky_ii_

ČSÚ, 2015. Český statistický úřad. *Výběrové šetření pracovních sil - VŠPS* [online]. 15. 01. 2015 [cit. 2015-07-08]. Dostupné z: https://www.czso.cz/csu/xa/vyberove_setreni_pracovnich_sil_vsps

FAYYAD, U., G. PIATETSKY-SHAPIRO a R. UTHURUSAMY, 1996. *Advances in Knowledge Discovery and Data Mining*. Massachusetts: AAAI Press/MIT Press. ISBN 0-262-56097-6.

GAMS, M. a J. KRIVEC, 2008. Demographic Analysis of Fertility. *Informatica*, č 32. Dostupné také z: http://www.informatica.si/index.php/informatica/article/viewFile/187/183

GARTNER, 2015. Gartner. *Magic Quadrant for Advanced Analytics Platforms* [online]. 19. 02. 2015 [cit. 2015-07-01]. Dostupné z: http://www.gartner.com/technology/reprints.do?id=1-2A881DN&ct=150219&st=sb

GOODWIN, L. K. et al., 2001. Data Mining Methods Find Demographic Predictors of Preterm Birth. *Nursing Research*, č 50.

GROVES, R. M. et al., 2009. *Survey Methodology*. New Jersey: John Wiley & Sons. ISBN 9781118211342.

HÁJEK, P.. H. T., 1978. *Mechanizing Hypothesis Formation*. Berlin: Springer-Velag. ISBN 3-540-08738-9.

HOLST, K. a A. MANGA, 2014. SAP. *SAP Predictive Analysis* [online]. 13. 01. 2014 [cit. 2017-07-01]. Dostupné z: http://www.sdn.sap.com/irj/scn/go/portal/prtroot/docs/library/uuid/a003a83f-4e54-3110-bcab-cdd8e75732b0?QuickLink=index&overridelayout=true&59017145622288

HUDEČEK, R., 2006. GYNEKOLOGICKO – PORODNICKÁ KLINIKA LÉKAŘSKÉ FAKULTY MASARYKOVY UNIVERZITY A FAKULTNÍ NEMOCNICE BRNO. In: *Ovariální hyperstimulační syndrom v programu asistované reprodukce - analýza rizikových faktorů …* [online]. Brno:2006 [cit. 2015-07-28]. Dostupné z: https://is.muni.cz/th/77349/lf_d/disertace_text.txt

CHALUPNÍK, V., 2012. Root. In: *Biologické algoritmy (4) - Neuronové sítě* [online]. 25. 04. 2012 [cit. 2015-07-05]. Dostupné z: http://www.root.cz/clanky/biologicke-algoritmy-4-neuronove-site/

CHALUPNIK, V., 2012. Root.cz. In: *Biologické algoritmy (1) - Evoluční algoritmy* [online]. 04. 04. 2012 [cit. 2015-06-20]. Dostupné z: http://www.root.cz/clanky/biologicke-algoritmy-1-evolucni-algoritmy

IBM, 2015. IBM. *IBM SPSS Modeler* [online] [cit. 2014-07-01]. Dostupné z: http://www-01.ibm.com/software/analytics/spss/products/modeler/features.html

IBRA, W. a P. LANGLEY, 1992. *Introduction of One-Level Decision Trees* [online]. Aberdeen: Morgan Kaufmann Publishers [cit. 2015-07-23]. Dostupné z: http://lyonesse.stanford.edu/~langley/papers/stump.ml92.pdf

IPUMS, 2010. IPUMS USA. *FAQ* [online] [cit. 2015-07-18]. Dostupné z: https://usa.ipums.org/usa-action/faq

JEŘÁBKOVÁ, V., 2012. In: *Přednášky v rámci kurzu 4ES405 - Sociální statistika* [online]. Praha:2012, verze 18.5.2012 [cit. 2012. Dostupné z: https://isis.vse.cz/auth/dok_server/slozka.pl?id=115980;lang=cz

KANTARDZIC, M., 2003. Data Minning. *UNC* [online] [cit. 2015-06-10]. Dostupné z: http://www.unc.edu/~xluan/258/datamining.html

KNIME, 2015. Knime. *Features* [online] [cit. 2015-07-02]. Dostupné z: https://www.knime.org/introduction/features

LISPMINER, 2015. LispMiner. *The official site of the LISp-Miner project* [online] [cit. 2015-07-02]. Dostupné z: http://lispminer.vse.cz/index.html

MICROSOFT, 2015. MSDN. *Data Mining Tools* [online] [cit. 2015-07-01]. Dostupné z: https://msdn.microsoft.com/en-us/library/ms174467.aspx

MICHIE, , et al., 1994. *Machine learning, neural and statistical classification*. Ellis Horwood Upper Saddle Rive: NJ, USA. ISBN:0-13-106360-X.

MITCHELL, T. M., 1997. *Machine Learning*. McGraw-Hill Science/Engineer. ISBN: 0070428077.

MRÁZOVÁ, I., 2010. Katedra teoretické informatiky MFF UK v Praze. In: *Dobývání znalostí - Asociační pravidla* [online].2010 [cit. 2015-07-24]. Dostupné z: http://ksvi.mff.cuni.cz/~mraz/datamining/lecture/Dobyvani_Znalosti_Prednaska_Asociacni_pravidla.pdf

MRÁZOVÁ, I., 2011. Katedra teoretické informatiky MFF UK v Praze. In: *Dobývání znalostí - rozhodovací stromy* [online].2011 [cit. 2015-07-23]. Dostupné z: http://ksvi.mff.cuni.cz/~mraz/datamining/lecture/Dobyvani_Znalosti_Prednaska_Rozhodovaci_stromy.pdf

OPENCV, 2011. OpenCV 2.4.11.0 documentation. *Random Trees* [online] [cit. 2015-07-23]. Dostupné z: http://docs.opencv.org/modules/ml/doc/random_trees.html

ORACLE, 2009. Oracle. *Data Mining Concepts* [online] [cit. 2015-07-01]. Dostupné z: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm

ORANGE, 2013. Orange. *Orange Documentation* [online] [cit. 2015-07-02]. Dostupné z: http://docs.orange.biolab.si/

PAT, 2013. Predictive Analytics Today. *Top 31 Predictive analytics SW* [online] [cit. 2015-07-01]. Dostupné z: http://www.predictiveanalyticstoday.com/top-predictive-analytics-software

PAT, 2014. Predictive Analytics Today. *40 top free Data Mining software* [online] [cit. 2015-07-01]. Dostupné z: http://www.predictiveanalyticstoday.com/top-free-data-mining-software

PEJČOCH, D., 2011. Metody řešení problematiky neúplných dat. *Forum Statisticum Slovacum*, roč. VII. ISSN 1336-7420. Dostupné také z: http://www.ssds.sk/casopis/archiv/2011/fss0711.pdf

PETRÚŠEK, I., 2013. FSV UK. *Chybějící hodnoty (item nonresponse)* [online] [cit. 2015-07-15]. Dostupné z: http://bit.ly/1MpCG7F

PIATETSKY, G., 2014. KDnuggets. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects* [online]. 08. 11. 2014 [cit. 2015-07-27]. Dostupné z: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

RAPIDMINER, 2015. Rapidminer. *Documentation* [online] [cit. 2015-07-02]. Dostupné z: http://docs.rapidminer.com/

RAUCH, J. a M. ŠIMŮNEK, 2015. *Dobývání znalostí z databází, LISp-Miner a GUHA*. PRAHA:VŠE: Oeconomica. ISBN 978-80-245-2033-9.

RITSCHARD, G., 2004. Universite de Géneve. In: *Data mining methods for longitudinal data* [online].2004 [cit. 2015-07-28]. Dostupné z: https://cigev.unige.ch/files/1514/2555/5587/sem20041209.pdf

RODRIGUES, M. D. F., C. RAMOS a P. R. HENRIQUES, 1999. Intelligent system to study demographic evolution. *Data Mining and Knowledge Discovery: Theory, Tools, and Technology*, č 3695. Dostupné také z: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=986454

RUGGLES, S. et al., 2015. *Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database]* [Data]. Minneapolis [cit. 2015-07-11].

RYCHLÝ, M., 2008. FIT VUTBR. In: *Klasifikace a Predikce* [online].2008 [cit. 2015-06-11]. Dostupné z: http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.pdf

SAS, 2014. SAS. *SAS Enterprise Miner* [online] [cit. 2015-07-01]. Dostupné z: http://www.sas.com/en_in/software/analytics/enterprise-miner.html

SAYAD, S., 2010. n Introduction to Data Mining. *Classification* [online] [cit. 2015-07-23]. Dostupné z: http://www.saedsayad.com/classification.htm

SINGH, R. a N. S. MANGAT, 1996. *Elements of Survey Sampling*. New York: Springer Science & Business Media. ISBN:9780792340454.

SONG, H. S., K. J. KIM a S. H. KIM, 2001. Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, č 21. Dostupné také z: http://www.sciencedirect.com/science/article/pii/S0957417401000379

SQL Data Mining, 2009. *History of Data mining* [online] [cit. 2015-06-10]. Dostupné z: http://www.sqldatamining.com/category/data-mining-basics

STATSOFT, 2013. StatSoft. *Úvod do data miningu* [online] [cit. 2015-06-10]. Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf

UMAYAPARVATHI, V. a K. IYAKUTTI, 2012. Applications of Data Mining Techniques in Telecom. *International Journal of Computer*, 42 (20). Dostupné také z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.1368&rep=rep1&type=pdf

VALENTOVÁ, V. H., 2013. MultiEdu TUL. In: *E učebnice* [online]. 18. 02. 2013 [cit. 2015-07-08]. Dostupné z: http://multiedu.tul.cz/~vladimira.valentova/multiedu/Statisticky_rozbor_dat_z_dotaznikovych_setreni/E_ucebnice.pdf

WEKA, 2013. Weka documentation. *Class HoeffdingTree* [online] [cit. 2015-07-23]. Dostupné z: http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/HoeffdingTree.html

WEKA, 2014. Weka. *Documentation* [online] [cit. 2015-07-02]. Dostupné z: http://www.cs.waikato.ac.nz/ml/weka/documentation.html

WIKISKRIPTA, 2014. Wikiskripta. In: *Bayesova věta* [online]. 02. 12. 2014 [cit. 2015-06-22]. Dostupné z: http://www.wikiskripta.eu/index.php/Bayesova_v%C4%9Bta

WIRTH, R. a J. HIPP, 2001. In: *Crisp DM: Towards a Standard Process Model for Data Mining* [online].2001 [cit. 2015-06-15]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf