UNIVERZITA PARDUBICE
Fakulta ekonomicko-správní

Využití data-miningových metod pro zpracování dat z oblasti sociální politiky

Bc. Michal Knížek

Diplomová práce

2011

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2010/2011**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Michal KNÍŽEK**
Osobní číslo:	**E090531**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Regionální a informační management**
Název tématu:	**Využití data-miningových metod pro zpracování dat z oblasti sociální politiky**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Práce je zaměřena na problematiku dobývání znalostí z databází. V práci budou analyzovány silné a slabé stránky vybraných metod pro danou oblast, specifikovány vhodné metody a navržena a ověřena jejich využitelnost.
Potřebné databáza autor získá z ČSU a pomocí dotazníkových šetření.
Zaměří se na vybraný region v ČR.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **BERKA, P. Dobývání znalostí z databází. 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.**
2. **ASUNCION, A., NEWMAN D.J. UCI Repository Of Machine Learning Databases and Domain Theories [online]. Irwine, USA. Dostupné z www: < http://archive.ics.uci.edu/ml >**
3. **KREBS, V. Sociální politika. 4. vyd. Praha : ASPI, 2007. 504s. ISBN 80-7357-276-1.**

Vedoucí diplomové práce: **Ing. Pavel Jirava, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **4. října 2010**

Termín odevzdání diplomové práce: **6. května 2011**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. října 2010

Prohlašuji:

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 28. 04. 2011

Bc. Michal Knížek

Poděkování

Rád bych poděkoval Ing. Pavlovi Jiravovi, Ph.D. za vedení diplomové práce, připomínky a cenné rady při jejím vypracovávání.

ANOTACE

Tato práce se zabývá využití data-miningových metod pro zpracování dat z oblasti sociální politiky. Jsou vybrány ty obce z regionu Královéhradeckého a Pardubického, které mají pověřený obecní úřad. Vybraná data těchto obcí jsou pomocí navržených modelů analyzována a jsou zde definovány silné a slabé stránky jednotlivých modelů.

KLÍČOVÁ SLOVA

Sociální politika, data mining, asociační pravidla, shluková analýza, rozhodovací stromy, vícerozměrná lineární regrese.

TITLE

Use data-mining methods for processing data from the social policy.

ANNOTATION

This work deals with the use of data-mining methods for processing data from the social policy. They selected the communities from the region of Hradec Kralove and Pardubice, which have a responsible local authority. Selected data from these communities are analyzed using the proposed model and are defined here as the strengths and weaknesses of individual models.

KEYWORDS

Social policy, data mining, association rules, clustering, decision trees, multivariate linear regression.

Obsah

Seznam obrázků

Seznam tabulek

Seznam rovnic

Úvod

Sociální politika je téma, které se bezpochyby dotýká každého občana. V současné době jsme svědky úsporných opatření za účelem snížení deficitu státního rozpočtu. Úsporná opatření vlády se dotýkají sociální politiky ve všech jejích oblastech, počínaje reformou zdravotnictví, reformou školství, důchodovou reformou. Pokud se podíváme na výdaje důchodů ze státního rozpočtu, zjistíme, že každým rokem dochází k jejich velikému nárůstu. V roce 2007 činily 282,6 miliardy korun, v roce 2008 dosáhly 304,9 miliardy, v roce 2009 narostly o dalších zhruba 26 miliard na 330,5 miliardy. V roce 2010 stát na důchodech vydal zhruba 337,5 miliardy korun.

Průzkum společnosti CVVM[1] z listopadu minulého roku ukázal, jak jsou občané spokojeni s výdaji na sociální politiku. Celých 48 % obyvatel České republiky považuje výdaje, které stát vynakládá na zabezpečení sociální politiky, za nízké. Za odpovídající považuje výdaje na sociální politiku 35 % občanů. Vysoké jsou podle desetiny dotázaných.

Pro modelování dat z oblasti sociální politiky za pomoci data-miningových metod z regionu Královéhradeckého a Pardubického byla vybrána data za všechny obce s pověřeným obecním úřadem v těchto dvou regionech. Celkem se jedná o 61 takovýchto obcí. Práce je rozdělena celkem do čtyř kapitol.

První kapitola se snaží o definici sociální politiky, definuje, co jsou objekty a subjekty sociální politiky a jaké jsou jednotlivé funkce sociální politiky. Je zde i definice hlavních typů a znaků sociální politiky. Dále je zde popsán pojem data-mining, metodologie CRISP-DM[2] a jsou rozebrány jednotlivé typy dat tak, jak na ně data-mining nahlíží. Poslední část kapitoly stručně charakterizuje systém Clementine, pomocí kterého bylo provedeno jednotlivé modelování dat.

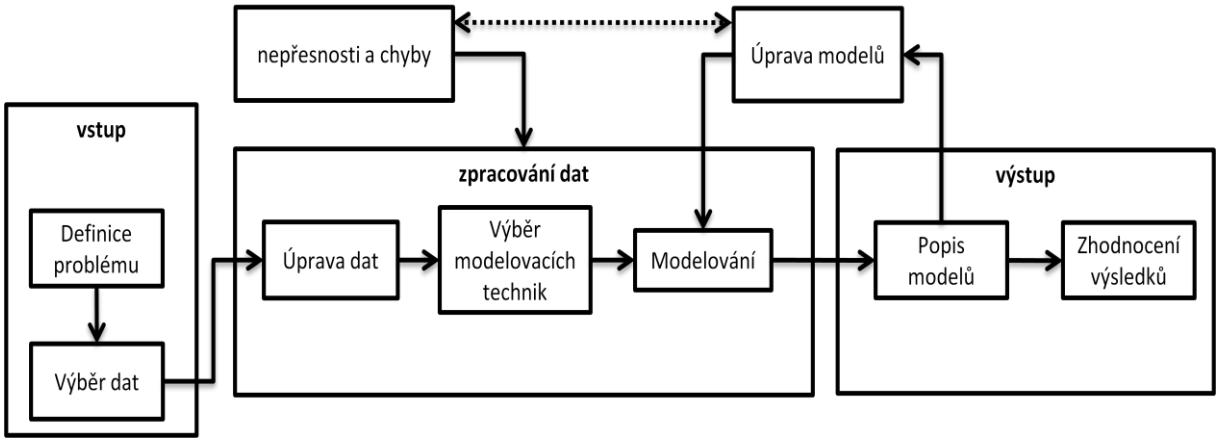
Druhá kapitola pojednává o jednotlivých vybraných modelovacích technikách, které jsou použity pro modelování dat z oblasti sociální politiky, jedná se o konkrétně asociační pravidla, rozhodovací stromy, shlukovou analýzu a vícerozměrnou lineární regresi.

[1] CVVM – centrum pro výzkum veřejného mínění
[2] CRISP-DM - Cross Industry Standard Process for Data Mining (meziprofesní standardní postup pro data mining)

Další kapitola, třetí, se zabývá výběrem vstupních dat, vstupní data (atributy) jsou vyjmenována a přesněji charakterizována. Analýza dat pomocí datového slovníku definuje jednotlivé typy dat, jejich rozsah a upravené názvy jednotlivých atributů. Je provedeno prvotní načtení dat do systému Clementine a provedena základní statistická analýza. Posledním tématem této kapitoly je posouzení kvality dat, kde jsou data zkontrolována, neobsahují li extrémní hodnoty.

Čtvrtá kapitola, stěžejní, se zabývá návrhem modelů. V této kapitole jsou popsány postupy tvorby jednotlivých modelů, jejich různé nastavení a jsou provedeny jejich analýzy na základě výsledných modelů. Pokud to daný model dovoluje, jsou zde zobrazeny u daného modelu i jeho konkrétní grafické výstupy. Jsou zde popsány silné a slabé stránky jednotlivých modelů a vždy je zobrazen výsledný návrh (stream) příslušného modelu.

Pro vypracování čtvrté kapitoly a tím je i hlavní cíl této práce, byl navržen obecný model systému pro modelování dat z oblasti sociální politiky, který je zobrazen na obrázku 1. Model je složen ze tří částí, které obsahují vstup, zpracování a výstup. Model zahrnuje zpětnou vazbu (úprava modelů) a vnější vlivy (nepřesnosti a chyby), mezi zpětnou vazbou a vnějšími vlivy je vzájemná interakce. Pokud bude zjištěna nepřesnost či chyba, pomocí úpravy modelů dojde k jejich odstranění.



Obrázek 1 - návrh obecného modelu. Zdroj [vlastní].

1. Definice hlavních oblastí

První kapitola se zabývá definicí sociální politiky, oblastí zájmu, objekty a subjekty sociální politiky, funkcemi a typy sociální politiky. Dále je definován pojem data mining, jeho jednotlivé fáze podle metodiky CRISP-DM.

1.1 Sociální politika

Sociální politika je velice široké téma, které zahrnuje mnoho oblastí. Následující text se zabývá definicí sociální politiky.

Hledáme-li odpověď na otázku, co je SP[3], je vhodné také vyjít od obecného vymezení politiky vůbec. Politiku lze obecně chápat jako specifickou společenskou činnost (projevující se zejména souborem různých opatření), konkrétní jednání různých subjektů na různých úrovních (tedy nejen státu), kterými je ovlivňována společenská realita v nejširším slova smyslu. Toto obecné vymezení politiky je možné aplikovat i na sociální politiku s tím, že ovlivňuje nikoli společenskou, ale sociální (v užším a nejužším slova smyslu) realitu [1].

Skutečnost, že sociální realita je složitá, že je různě chápána a je obtížné ji souhrnně postihnout, je příčinou toho, že neexistuje ani jednoznačná definice sociální politiky, ale naopak určitá libovůle v jejím chápání, a to jak v teorii, tak i v praxi. SP zpravidla zahrnuje politiku sociálního zabezpečení včetně osobních sociálních služeb, rodinnou politiku, politiku zaměstnanosti a vzdělávací politiku [1].

Sociální politika je snaha po změnách společenského zřízení prostředky společenskými tak, aby členské zájmy lidí ve společnosti byly uspokojovány způsobem trvale prospěšným celku. Sociální politika není obor nebo oddíl politiky, nýbrž způsob, směr nebo hledisko, které by mělo pronikat veškeru politiku, ať hospodářskou, kulturní, v politice hospodářské pak všechny její obory zemědělský, průmyslový, obchodní i dopravní (když už se tak obyčejně dělí), úpravu výroby i rozdělení statků, organizaci hospodářství soukromého i veřejného [2].

Sociální politika jako věda je disciplínou, která zkoumá politické procesy tvorby politik, jež se dotýkají sociálních podmínek života občanů. Předmětem sociální politiky jsou sociální problémy a kritické situace v životě jednotlivce (např. chudoba, mateřství, rodičovství,

[3] SP – sociální politika

nemoc, stáří, invalidita, nezaměstnanost). Následující témata mohou spadat do zájmu oblasti sociální politiky [3]:

- politika boje s chudobou a sociálním vyloučení,
- politika sociálního zabezpečení,
- politika zaměstnanosti,
- vzdělávací politika,
- bytová politika,
- zdravotní politika,
- rodinná politika.

Pod pojmem objekty sociální politiky rozumíme všechny obyvatele dané země, ať již jako jednotlivce, či určité sociální skupiny. Sociální skupinu lze vymezit jako skupinu osob, mezi nimiž existuje určitá interakce, a tyto osoby i okolí si tuto skutečnost uvědomují [1].

Subjekty jsou ti, kdo mají zájem, schopnosti, vůli, předpoklady, možnosti a prostředky k určité sociální činnosti či chování a kdo takové činnosti a chování může iniciovat a naplňovat. K subjektům sociální politiky patří [1]:

- stát a jeho orgány,
- zaměstnavatelé a firmy,
- zaměstnavatelské, zaměstnanecké a odborové orgány,
- regiony, místní komunity, obce, jejich orgány a instituce,
- občanské organizace a iniciativy,
- církve,
- občané, rodiny, domácnosti.

Sociální politika plní řadu funkcí, které mohou být různě členěny. Funkce spolu vzájemně souvisejí, působí komplexně na jedince či sociální skupiny a mají i jistý globální vliv na společnost jako celek. Nejčastěji se v sociální politice hovoří o funkci ochranné, rozdělovací a přerozdělovací, homogenizační, stimulační a preventivní. Funkcemi sociální politiky jsou [1]:

Ochranná funkce - je historicky nejstarší funkcí, tvoří tradiční a stabilní prvek SP. Řeší situace, kdy jedinec či sociální skupina (rodina) je znevýhodněna ve vztahu k ostatním, ať ekonomicky či sociálně. Jde o zmírnění nebo odstranění důsledků určitých sociálních událostí

(např. nezaměstnanost, škodlivé pracovní prostředí, stáří, příjmová situace vícedětných rodin, nemoc, osiření…).

Rozdělovací a přerozdělovací funkce - je nejsložitější a nejvýznamnější funkce, zaměřuje se nejen na rozdělování důchodů, ale i životních šancí. Určuje podíl jednotlivců (sociál. skupin) na společenském bohatství a místo člověka ve společnosti.

Homogenizační funkce - je relativně novou funkcí SP. Je spojena s předchozí funkcí neboť má zmírňovat sociální rozdíly v životních podmínkách jedinců a sociál. skupin (cestou poskytování stejných životních šancí) a odstraňovat (zmírňovat) neodůvodněné rozdíly mezi lidmi.

Stimulační funkce - cílem této funkce je podporovat, podněcovat, vyvolávat žádoucí sociální jednání jednotlivců a sociál. skupin jak v oblasti ekonomické, tak i mimo ni.

Preventivní funkce - je spojena se snahou zabránit zcela nebo alespoň v co největší míře nežádoucím sociálním situacím (chudoba, nezaměstnanost, zdravotní poškození…) a omezovat či vylučovat faktory, které brání integraci člověka do společnosti.

Budoucí podobu sociální politiky budou ovlivňovat tyto trendy: dynamičnost poznání, globalista, subjektivita, demografické poměry, nutnost propojování ekonomických a sociálních úvah a cílů, dosažený stav transformace společnosti.

Sociální politika v současnosti zdůrazňuje především ochranné aspekty. Pro budoucnost je potřeba podpořit široké vnímání sociální politiky, tj. kromě sociální ochrany zdůraznit i její aktivní prvky a fakt, že je i procesem kultivace člověka a přispívá k určitému morálnímu profilu společnosti [1].

Stát jako bezprostřední vykonavatel se v sociální oblasti stále příliš angažuje, ale jeho možnosti nést tíhu rozsáhlého zabezpečení jsou omezené. Větší roli v tomto směru musí v sociální politice sehrát nestátní subjekty. Role státu musí být posílena, pokud jde o koncepční, normotvorné a kontrolní aktivity v sociální politice [1].

Na solidaritu v sociální politice je nutné nahlížet z různých možných úhlů pohledu. Podpořit je nutno solidaritu uskutečňovanou na bázi nestátních subjektů. Celospolečenská solidarita nemá nutně kladné znaménko, může mít pozitivní i negativní důsledky. Pro budoucnost se jako žádoucí jeví položit větší důraz na tzv. aktivizující a integrační solidaritu (podpora

vzdělání a pracovních míst) a pečlivě zvažovat míru a důsledky tzv. solidarity pečovatelské [1].

Typy (modely) sociální politiky, které jsou rozlišovány v zemích OECD[4] [4]:

Redistributivní s dominantní rolí státu. Do svého působení zahrnuje celou populaci bez ohledu na to, zda je sociálně potřebná. Vyžaduje značný rozsah redistribuce a výrazně omezuje, někdy až ruší aktivity nestátních subjektů).

Výkonový či korporativní vychází z toho, že sociální potřeby mají být primárně uspokojovány na základě pracovního výkonu a zásluh. Je založen na širší kooperaci občanů a zpravidla také na aplikaci sociálního pojištění. Míra redistribuce je zde ve srovnání s prvním typem nižší a stát garantuje pouze základní společensky uznaná minima potřeb a vytváří prostor pro působení nestátních subjektů.

Reziduální se spoléhá téměř výhradně na trh a jeho instituce a na rodinu. Role státu jako subjektu sociální politiky je značně potlačena, míra redistribuce je zde ze všech typů nejnižší.

V České Republice je aplikován „mix“ ze všech výše uvedených typů sociální politiky. Jednotlivé typy a znaky sociální politiky jsou uvedeny v tabulce 1.

Tabulka 1 - typy a znaky sociální politiky. Zdroj [1].

TYPY SOCIÁLNIHO STÁTU CHARAKTERISTIKA	Reziduální liberální	Výkonový konzervativní	Institucionální sociálně - demokratický
Odpovědnost státu za uspokojování potřeb	minimální	optimální	úplná
Rozdělení podle potřeb	marginální	sekundární	primární
Rozsah povinně poskytovaných služeb	omezený	extenzivní	úplný
Populace pokrytá povinně poskytovanými službami	menšina	většina	všichni
Výše příspěvků	nízká	střední	vysoká
Část národního důchodu určená pro služby státu	nízká	střední	vysoká
Zkoumání potřebnosti	primární	sekundární	marginální
Charakter klientů	chudáci	občané	členové společnosti
Status klientů	nízký	střední	vysoký

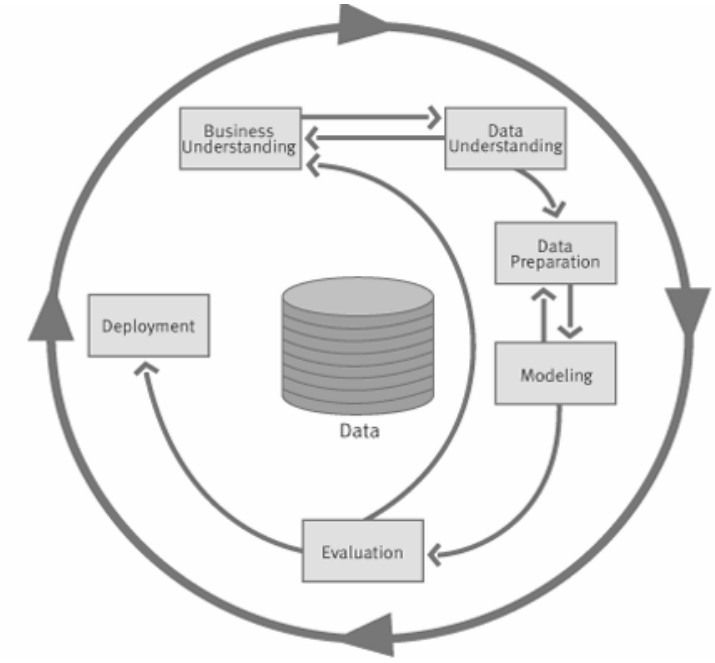
[4] Organizace pro hospodářskou spolupráci a rozvoj (zkráceně OECD z angl. Organisation for Economic Co-operation and Development)

1.2 Data mining

Data mining lze charakterizovat jako „proces extrakce relevantních, předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází“ [5]. Důležitou vlastností DM[5] je, že se jedná o analýzy odvozené z obsahu dat, nikoliv předem specifikované uživatelem nebo implementátorem. Jedná se především o odvozování prediktivních informací, nikoliv pouze deskriptivních. To znamená, že proces DM lze definovat jako netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat [5].

Metodika CRISP-DM vznikla v rámci Evropského výzkumného projektu. Cílem projektu bylo navrhnout univerzální postup (tzv. standardní model procesu dobývání znalostí z databází), který bude použitelný v nejrůznějších komerčních. Vytvoření takovéto metodiky umožní řešit rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady. Kromě návrhu standardního postupu má CRISP-DM nabízet „průvodce“ potenciálními problémy a řešeními, které se mohou vyskytnout v reálných aplikacích [6].

Metodika využívá šesti kroků, které jsou zobrazeny na obrázku 2:



Obrázek 2 - metodika CRISP-DM. Zdroj [6].

[5] DM – data mining

Porozumění problematice (business understanding) je úvodní fáze zaměřená na pochopení cílů projektu a požadavků na řešení formulovaných z manažerského hlediska. Tato manažerská formulace musí být převedena do zadání úlohy pro dobývání znalostí z databází [6].

Porozumění datům (data understanding) začíná prvotním sběrem dat. Následují činnosti, které umožní získat základní představu o datech, která jsou k dispozici (posouzení kvality dat, první „vhled" do dat, vytipování zajímavých podmnožin záznamů v databázi…). Obvykle se zjišťují různé deskriptivní charakteristiky dat (četnosti hodnot různých atributů, průměrné hodnoty, minima, maxima apod.), s výhodou se využívají i různé vizualizační techniky [6].

Příprava dat (data preparation) zahrnuje činnosti, které vedou k vytvoření datového souboru, který bude zpracováván jednotlivými analytickými metodami. Tato data by tedy měla obsahovat údaje relevantní k dané úloze, a mít podobu, která je vyžadována vlastními analytickými algoritmy [7].

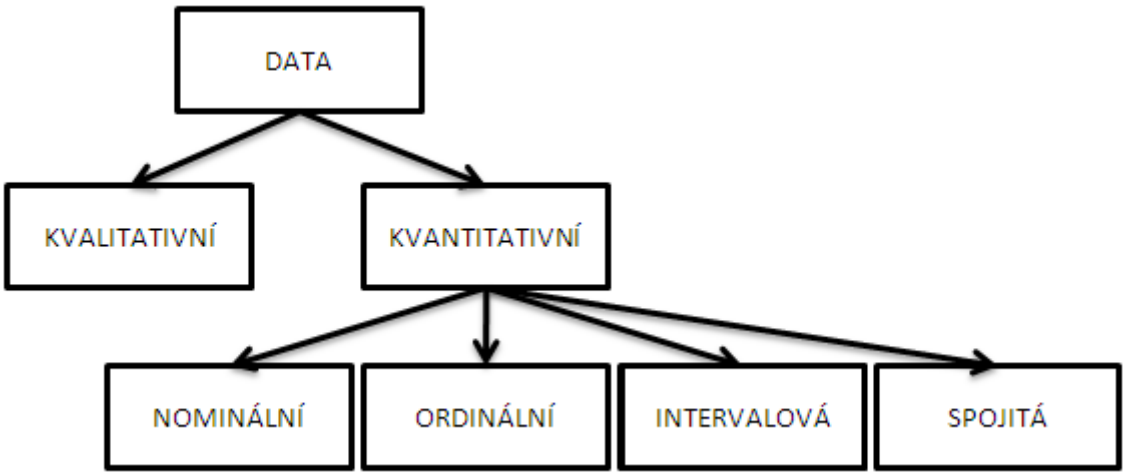
Modelování (modeling) používá analytické metody s algoritmy pro dobývání znalostí. Obvykle existuje řada různých metod pro řešení dané úlohy, je tedy třeba vybrat ty nejvhodnější (doporučuje se použít více různých metod a jejich výsledky kombinovat) a vhodně nastavit jejich parametry. Jde tedy opět o iterativní činnost (opakovaná aplikace algoritmů s různými parametry), navíc, použití analytických algoritmů může vést k potřebě modifikovat data a tedy k návratu k datovým transformacím z předcházející fáze [7].

Vyhodnocení výsledků (evaluation) se zabývá problémem, zda byly splněny cíle formulované na počátku projektu [6].

Využití výsledků (deployment). Vytvořením vhodného modelu celý projekt obecně nekončí. Dokonce i v případě, že řešenou úlohou byl „pouze" popis dat, získané znalosti je třeba upravit do podoby použitelné pro podporu rozhodování. Podle typu úlohy tedy využití (nasazení) výsledků může na jedné straně znamenat prosté sepsání závěrečné zprávy, na straně druhé pak zavedení (hardwarové, softwarové, organizační) systému pro automatickou klasifikaci nových případů [7].

Data mining pro práci s daty využívá nástroje jako jsou například asociační pravidla, shlukovou analýzu, rozhodovací stromy, statistickou analýzu, neuronové sítě a mnohé další metody pro dobývání znalostí z dat.

Data klasifikujeme do dvou základních skupin a to data kvantitativní a data kvalitativní. Obrázek 3 zobrazuje klasifikaci dat.



Obrázek 3 - klasifikace dat. Zdroj [vlastní].

V kvalitativních datech se odlišují proměnné pomocí popisných pojmů. Například pohlaví se všeobecně klasifikuje jako „M“ jako muž a „Ž“ jako žena. Kvalitativní data lze použít pro segmentaci a klasifikaci [5].

Kvantitativní data jsou charakteristická číselnými proměnnými. Kvantitativní data se používají k vytváření prediktivních modelů. Rozlišujeme čtyři typy kvantitativních dat [5]:

Nominální data jsou číselná data, která reprezentují kategorie neboli atributy. Důležitou vlastností nominálních dat je to, že nemají relativní význam.

Ordinální data jsou číselná data, která představují kategorie a mají relativní význam.

Intervalová data jsou číselná data, která mají relativní význam a nemají nulový bod.

Spojitá data jsou nejčastějším typem dat používaných při vytváření prediktivních modelů. Se spojitými daty lze provádět všechny základní aritmetické operace včetně sčítání, odečítání, násobení a dělení.

1.3 O programu Clementine

Systém Clementine vyvinula britská firma Integral Solutions Ltd. v polovině 90. let. K 1. lednu 1999 tuto firmu (a s ní i systém Clementine) převzal přední výrobce statistického software, firma SPSS. Clementine patří mezi přední komerční systémy pro dobývání znalostí. Systém důsledně vychází z metodologie CRISP-DM. Systém nabízí řadu metod pro klasifikační (predikční) i deskriptivní úlohy, mimo jiné standardní algoritmy C5.0 (rozhodovací stromy), apriori (asociační pravidla), vívevrstvý perceptron (neuronové sítě), metodu k-středů, nebo lineární regresi [13].

Clementine má velice propracovaný způsob ovládání, tzv. vizuální programování (vizual programming). Z nástrojů v jednotlivých paletách (tyto palety odpovídají jednotlivým krokům procesu dobývání znalostí; předzpracování, modelování, vizualizace a interpretace) se na pracovní ploše poskládá sekvence řešení úlohy (tzv. stream) [13].

Na Clementine se spoléhá mnoho komerčních organizací, vládních a akademických institucí na celém světě a to především proto, aby odkryli informace skryté v datech. To jim pomáhá nejen zlepšit výsledky ale také dosáhnout stanovených cílů [8].

Clementine pomáhá [8]:

- Bankovním a finančním institucím realizovat efektivněji marketingové kampaně, spolehlivěji hodnotit úvěrové riziko a odhalit podvodné aktivity,
- Pojišťovnám zlepšit marketingové úsilí a zefektivnit proces vyřizování pojistných, událostí díky identifikaci podvodných pojistných událostí,
- Telekomunikačním společnostem vytvářet těsnější vztah se zákazníky pro budování a posílení loajality a snižování odchodu zákazníků ke konkurenci,
- Obchodníkům zlepšovat plánování zásob a zefektivnit úsilí vynaložené na marketing a budování loajality zákazníků,
- Dodavatelům energií a veřejných služeb nabízet svým zákazníkům služby šité na míru. Analýzy hrají rovněž velkou roli při preventivních údržbách, čímž je zajištěna větší spolehlivost s nižšími náklady,
- Zdravotnickým zařízením proaktivně řídit své zdroje a zdokonalovat léčebné postupy tak, aby poskytovali pacientům lepší péči,
- Vysokým školám a univerzitám řídit celý životní cyklus studentů, od nabírání správného složení studentů po nabízení správných studijních a asistenčních programů.

2. Vybrané modelovací techniky

Druhá kapitola se zabývá popisem jednotlivých vybraných technik pro analýzu dat. Pro analýzu dat byla zvolena asociační pravidla, vícerozměrná lineární regresní analýza, shluková analýza a rozhodovací stromy.

2.1 Asociační pravidla

IF-THEN konstrukce nalezneme ve všech programovacích jazycích, používají se i v běžné mluvě (nebude-li pršet, nezmoknem). Není tedy divu, že pravidla s touto syntaxí patří společně s rozhodovacími stromy k nejčastěji používaným prostředkům pro reprezentaci znalostí, ať už získaných od expertů, nebo vytvořených automatizovaně z dat [6].

Termín asociační pravidla široce zpopularizoval počátkem 90. let Agrawal[6] v souvislosti s analýzou nákupního košíku. Při této analýze se zjišťuje, jaké druhy zboží si současně kupují zákazníci v supermarketech (např. pivo a párek). Jde tedy o hledání vzájemných vazeb (asociací) mezi různými položkami sortimentu prodejny. Přitom není upřednostňován žádný speciální druh zboží jako závěr pravidla [6].

U pravidel vytvořených z dat nás obvykle zajímá, kolik příkladů splňuje předpoklad (*Ant*) a kolik závěr (*Suc*) pravidla, kolik příkladů splňuje předpoklad i závěr současně, kolik příkladů splňuje předpoklad a nesplňuje závěr. Tedy, zajímá nás, jak pro pravidlo [9]:

Ant => Suc,

kde *Ant* (předpoklad, levá strana pravidla, antecedent) i *Suc* (závěr, pravá strana pravidla, sukcedent) jsou kombinace kategorií vypadá příslušná kontingenční tabulka. Zjednodušená kontingenční tabulka o rozměrech 3x3 se nazývá čtyřpolní tabulka (v tomto případě je doplněna ještě o řádkové i sloupcové sumy). Čtyřpolní tabulka je zobrazena v tabulce 2.

[6] Agrawal a kol, 1993

Tabulka 2 - čtyřpolní kontingenční tabulka. Zdroj [vlastní].

	závěr	¬závěr	∑
předpoklad	a	b	r = a + b
¬předpoklad	c	d	s = c + d
∑	k = a + c	l = b + d	n=a+b+c+d

Z tabulky 2 se počítají základní charakteristiky asociačních pravidel, které jsou podpora a spolehlivost. Podpora udává, jak často lze dané pravidlo použít a vypočítá se pomocí rovnice 1 [10]:

Rovnice 1 - podpora. Zdroj [10].

$$P(Ant \wedge Suc) = \frac{a}{a+b+c+d},$$

kde a znamená počet záznamů splňujících předpoklad i závěr současně a b znamená počet záznamů splňujících předpoklad a nesplňujících závěr. Spolehlivost udává, jak moc se na dané pravidlo můžeme spolehnout a vypočítá se pomocí rovnice 2 [10]:

Rovnice 2 - spolehlivost. Zdroj [10].

$$P(Suc \mid Ant) = \frac{a}{a+b}$$

Asociační pravidla se z dat získávají nejčastěji pomocí algoritmu apriori. Jeho vstupem je vhodná reprezentace datové struktury a parametry *minconf*[7] a *minsup*[8]. Výstupem jsou všechna asociační pravidla $X \rightarrow Y$ taková, že *conf (X→Y) minconf* a *sup (X→Y) minsup*. Algoritmus pracuje tak, že nejprve vyhledá všechny frekventované podmnožiny položek *k*, které se v datech vyskytují dostatečně často (tj. s podporou *minsup*) a z těchto sestaví asociační pravidla splňující *conf (X → Y) minconf* [11].

Algoritmus apriori podle [6]:

1. do L_1 přiřaď všechny kategorie, které dosahují alespoň požadované četnosti
2. polož $k = 2$
3. dokud $L_{k-1} \neq \emptyset$

   3.1 pomocí funkce apriori-gen vygeneruj na základě L_{k-1} množinu kandidátů C_k

   3.2 do L_k zařaď ty kombinace z C_k, které dosáhly alespoň požadovanou četnost

   3.3 zvětši počítadlo *k*

[7] minconf – označení pro minimální spolehlivost pravidla
[8] minsup – označení pro minimální podporu pravidla

Kde L_1 je frekventovaná množina o velikosti *1*, *k* označuje položky frekventované množiny, C_k označuje množinu kandidátů.

Funkce apriori-gen(L_{k-1}) podle [6]:

1. pro všechny dvojce kombinací $Comb_p$, $Comb_q$ z L_{k-1}
   1.1 pokud $Comb_p$ a $Comb_q$ se shodují v *k-2* kategoriích, přidej $Comb_p$ ^ $Comb_q$ do C_k
2. pro každou kombinaci *Comb* z C_k
   2.1 pokud některá z jejich podkombinací délky *k-1* není obsažena v L_{k-1} odstraň *Comb* z C_k

Kde $Comb_p$ a $Comb_q$ jsou kombinace předpokladu a závěru. Jádrem algoritmu je hledání často se opakujících množin položek (frequent itemsets). Jedná se o kombinace (konjunkce) kategorií, které dosahují předem zadané četnosti (podpory *minsup*) v datech. Při hledání kombinací délky *k*, které mají vysokou četnost, se využívá toho, že již známe kombinace délky *k-1*. Při vytváření kombinace délky *k*, spojujeme kombinace délky *k-1* [9].

2.2 Vícerozměrný model lineární regresní analýzy

Lineární regresní model je schopen změřit pouze lineární, přímkový vztah. Jsou-li body v kruhu, regresní analýza nebude detekovat lineární vztah. Nejdůležitějším kritériem linearity je Pearsonův korelační koeficient *r*. Blíží-li se jeho hodnota *+1* nebo *-1*, jde o přímkový vztah; blíží-li se však nule, nejde o lineární (přímkový) vztah. Perfektní přímka má *r* rovno *+1* (vzestupná přímka) nebo *r* rovno *-1* (sestupná přímka) [12].

Při vícerozměrném modelu lineární regresní analýzy se snažíme o odhad závislé proměnné Y_i pomocí následujících parametrů výběrové regresní rovnice 3[13]:

Rovnice 3 - výběrová regresní funkce. Zdroj [13].

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad \text{k = 1, 2, …, n}$$

kde člen *Y* je závislá proměnná, β_0 je průsečíkem přímky s osou *y*. Odhady β_k jsou směrnicemi přímky a parametry x_k jsou nezávislými parametry. Závislá proměnná Y_i se změní o tolik, o kolik jednotek se změní nezávislé parametry x_{ik}.

2.3 Shluková analýza

Pojem shluková analýza je souhrnným názvem pro celou řadu výpočetních postupů, jejichž cílem je rozklad daného souboru dat na několik relativně homogenních podmnožin, shluků. Rozklad množiny dat by měl být proveden takovým způsobem, aby si objekty uvnitř jednotlivých shluků byly co nejvíce podobné. Objekty patřící do různých shluků by si naopak měly být podobné co nejméně [14].

Podobnost se převádí na vzdálenost. Jedná se tedy o shlukování založené na vzdálenosti (existuje také shlukování založené na konceptech, objekty potom patří do stejného shluku, pokud shluk definuje koncept společný všem objektům) [15].

Základní vzdáleností u shlukové analýzy je Euklidovská vzdálenost, která se počítá podle rovnice 4 [16]:

Rovnice 4 - Euklidovská vzdálenost. Zdroj [16].

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2},$$

kde x_{ik} je hodnota k-tého pozorování na i-tém prvku a x_{jk} je pozorování na j-tém prvku.

Shlukování dělíme na shlukování:

- hierarchické,
- nehierarchické.

Bude popsáno pro účely této práce pouze shlukování nehierarchické. Skupina nehierarchických metod, na rozdíl od hierarchických, nevytváří hierarchickou strukturu, ale rozkládá výchozí množinu objektů do několika podmožin takovým způsobem, aby bylo splněno určité kritérium. Prvotní rozklad původní množiny objektů do několika podmnožin se potom v dalších krocích výpočtu mění. Cílem je dosažení optimální hodnoty jistého, pro danou metodu, specifického kritéria [14].

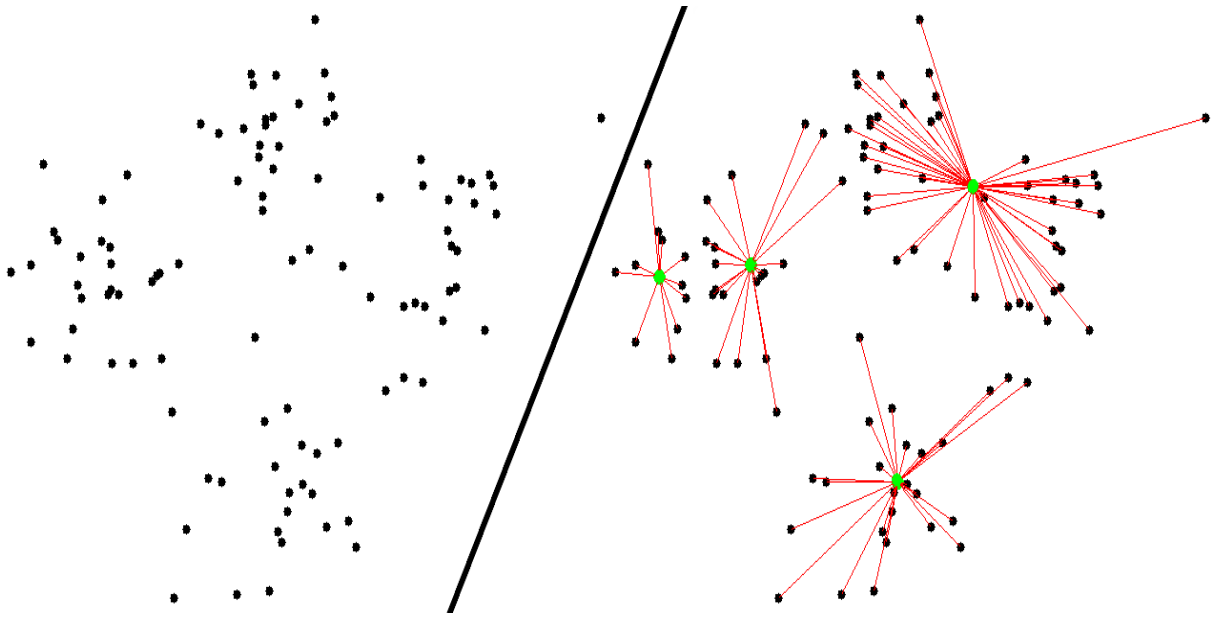
Pro nehierarchické shlukování se nejvíce využívá metoda K-MEANS (k-středů). Při shlukování metodou K-středů předpokládáme, že víme do kolika shluků je možno příklady rozdělit. Počet shluků se tedy během výpočtu nemění, mění se pouze zařazení příkladů k

těmto shlukům. Proto je tato metoda méně výpočetně náročná než hierarchické shlukování (a tudíž vhodnější pro větší datové soubory) [6].

Algoritmus metody k-means lze obecně popsat prostřednictvím čtyř následujících postupných kroků [6]:

1. náhodně zvol rozklad do K shluků,
2. urči centroidy pro všechny shluky v aktuálním rozkladu,
3. pro každý příklad x,
   3.1 urči vzdálenosti $d(x,c_k)$, $k=1,\ldots,k$, kde c_k je centroid k-tého shluku,
   3.2 nechť $d(x,c_k) = min_k\, d(x,c_k)$,
   3.3 není-li x součástí shluku l (k jehož centroidu c_l má nejblíže) přesuň x do shluku
4. došlo-li k nějakému přesunu potom, jdi na 2 jinak konec.

Kde K označujeme shluk, x označujeme jednotlivý případ, c_k označujeme centroid k-tého shluku, d označujeme vzdálenost mezi jednotlivým případem x a centroidem c_k k-tého shluku, min_k označujeme minimální vzdálenost k-tého shluku. Obrázek 4 zobrazuje výchozí rozmístění objektů na levé části obrázku a na pravé části obrázku jsou již konečné čtyři shluky. Každý shluk má svůj centroid, který je označen zelenou barvou. Červené čáry potom představují vzdálenosti od centroidu k danému objektu.



Obrázek 4 - příklad shlukování. Zdroj [17].

2.4 Rozhodovací stromy

Rozhodovací stromy jsou analytické nástroje sloužící k nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně. Cílem je určit takové proměnné, které dokážou záznamy rozdělit a snižují tak nejistotu. Problémem může být určení, na kolik „větví“ se má dělit každá proměnná. Pokud záznamy rozdělíme podle proměnné do příliš mnoha skupin, může nastat situace, kdy do každé z těchto skupin přísluší pouze několik málo záznamů a nelze tak vyvodit žádná rozhodovací pravidla. Rozhodovací stromy jsou vhodné pro úlohy, ve kterých má být provedena klasifikace nebo předpověď. Užitečné jsou v oblastech, ve kterých můžeme hodnoty proměnných rozdělit do relativně malého počtu skupin [18].

Vlastní klasifikace pomocí rozhodovacího stromu probíhá cestou záznamu od kořene stromu k jeho listu. V každém kroku je záznam otestován podle testu v aktuálním uzlu rozhodovacího stromu a dále pokračuje po větvi shodné s konkrétním výsledkem testu. Pokud takto záznam dojde až do listového uzlu, je oklasifikován třídou identifikovanou hodnotou příslušného listu rozhodovacího stromu [19].

Jedním ze základních algoritmů rozhodovacích stromů je algoritmus TDIDT [6]:

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.

Uvedený algoritmus bude fungovat pro kategoriální data (počet podmnožin-uzlů vytvářený v kroku 2 odpovídá počtu hodnot daného atributu), která nejsou zatížena šumem (růst stromu se podle bodu 3 zastaví v okamžiku, kdy všechny příklady v daném uzlu patří do téže třídy). Vhodný atribut se vybírá pomocí entropie, informačního zisku, poměrného informačního zisku, chí-kvadrátu, nebo Giniho indexu [6].

Rozhodovací stromy jsou založeny na mnoha dalších algoritmech [20]:

- CHAID - rychlý statistický víceúrovňový stromový algoritmus pro účinné zkoumání interakčních vztahů v datech,
- Úplný CHAID - kompletní statistický víceúrovňový stromový diagram pro komplexní prohledávání dat,
- Klasifikační a regresní strom (C&RT) - úplný binární stromový algoritmus pro postupné binární štěpení datového souboru a tvorbu homogenních podmnožin,
- QUEST - Statistický algoritmus pro selekci proměnných bez vychýlení; sestavuje přesné binární stromy rychle a účinně.

3. Výběr vstupních dat

Třetí kapitola popisuje vstupní data, je provedena analýza dat. Dále je zde zobrazen datový slovník a je posouzena kvalita vstupních dat.

3.1 Popis vstupních dat

Zdrojem vstupních dat posloužila MOS[9] a statistické ročenky Českého statistického úřadu pro Královéhradecký a Pardubický kraj, konkrétně byly vybrány všechny obce s pověřeným obecním úřadem z těchto dvou krajů. Celkem se jedná o 61 takovýchto obcí. Pro potřeby této práce bylo vybráno 22 atributů. Jedná se tedy o matici dat o 61 řádcích a 22 sloupcích. Datová matice v příloze 6. Popis jednotlivých atributů je následující:

- správní obvody obcí – názvy jednotlivých obcí,
- počet obyvatel do 14 let,
- počet obyvatel od 15 do 64 let,
- počet obyvatel nad 64 let,
- ekonomicky aktivní obyvatelstvo - ekonomicky aktivní obyvatelstvo, čili pracovní sílu, tvoří zaměstnaní a nezaměstnaní. Za zaměstnané jsou považovány všechny osoby starší 15 - ti let, které během referenčního období příslušely mezi placené zaměstnance, příslušníky armády nebo osoby zaměstnané ve vlastním podniku. Za nezaměstnané jsou považovány osoby 15leté a starší, které ve sledovaném období souběžně neměli placené zaměstnání ani sebezaměstnání, zaměstnání aktivně hledaly nebo byly připraveny k nástupu do práce (tj. nejpozději do 2 týdnů) [21],
- uchazeči o zaměstnání - uchazečem o zaměstnání je občan, který není v pracovním nebo obdobném vztahu ani nevykonává samostatnou výdělečnou činnost ani se nepřipravuje soustavně pro povolání a osobně se u úřadu práce uchází na základě písemné žádosti o zprostředkování vhodného zaměstnání,
- nezaměstnanost – za nezaměstnaného se považuje ten, kdo je starší patnácti let, aktivně vyhledává práci a je připraven k nástupu do práce nejpozději do čtrnácti dnů,
- ekonomické subjekty – ekonomické subjekty tvoří domácnosti, firmy, stát, zahraniční subjekty, neziskové subjekty,
- počet mateřských škol,

[9] MOS – městská a obecní statistika

- počet základních škol,
- počet středních škol,
- domovy s pečovatelskou službou – dům s pečovatelskou službou je určen pro staré občany, kteří dosáhli věku rozhodného pro přiznání starobního důchodu a pro občany, kteří jsou plné invalidní a jejich celkový zdravotní stav je takový, že nepotřebují komplexní péči, poskytuje se jim ubytování a základní péče [22],
- domovy důchodců - domov pro seniory je určen především pro staré občany, kteří dosáhli věku rozhodného pro přiznání starobního důchodu a kteří pro trvalé změny zdravotního stavu potřebují komplexní péči, jež jim nemůže být zajištěna členy jejich rodiny ani pečovatelskou službou nebo jinými službami sociální péče a dále pro staré občany, kteří toto umístění nezbytně potřebují z jiných vážných důvodů, poskytují svým klientům sociální, zdravotnické, stravovací, ubytovací, lékařské a mnohé další služby, včetně volnočasových aktivit, tyto instituce mohou být státní i soukromé [22],
- úřad práce - úřad práce je státní instituce, jejíž hlavní činností je poskytování informací z oblasti pracovního trhu nejen v České republice, ale i v Evropské unii, evidence uchazečů o zaměstnání a volných pracovních míst, spadá pod správu MPSV[10] České republiky, ale jeho působnost se dělí do menších administrativních celků, které působí na jednotlivých pobočkách po celé České republice. Na úřadu práce, který je místně příslušný bydlišti, je možné vyhledat aktuální pracovní nabídky, s příslušným referentem si společně vytvořit svůj kariérní profil, dále se prostřednictvím úřadu můžete přihlásit na nejrůznější rekvalifikační školení apod. [23],
- počet ordinací lékařů pro dospělé,
- počet ordinací lékařů pro děti a dorost,
- počet narozených dětí,
- počet zemřelých občanů,
- sňatečnost,
- rozvodovost,
- výstavba nových bytů.

Všechny výše uvedené atributy a jejich sledované hodnoty jsou za sledované období 2009.

[10] MPSV – ministerstvo práce a sociálních věcí

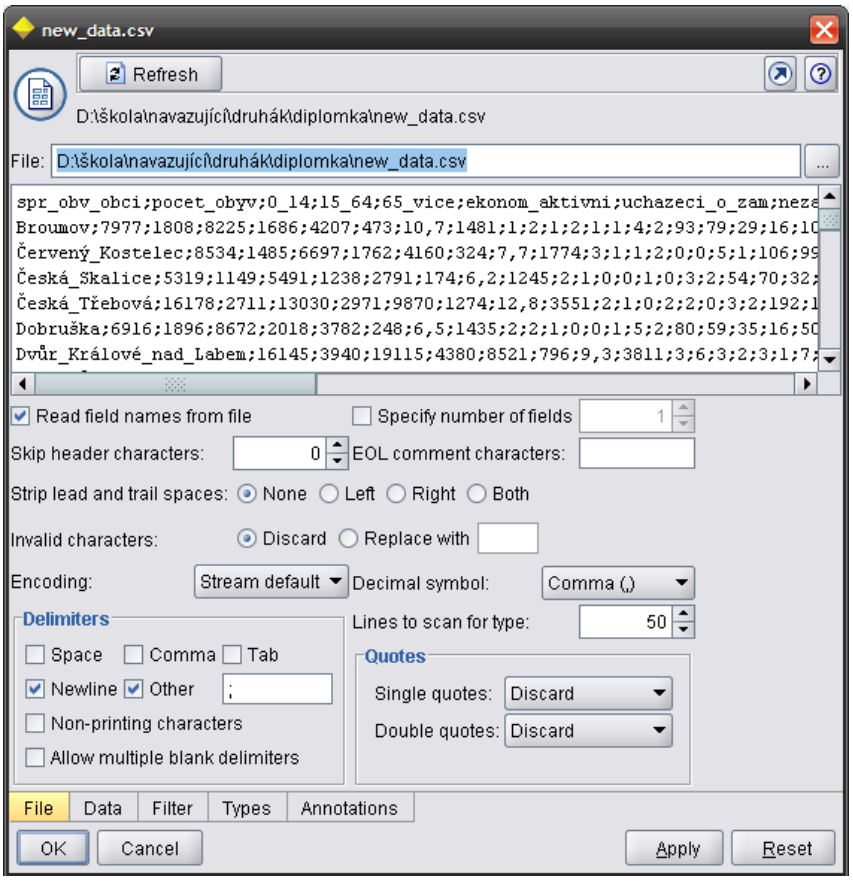
3.2 Analýza vstupních dat

Shromážděná data byla původně zpracovaná v MS Excel, kde byly upraveny názvy atributů, konkrétně byla odstraněna diakritika a mezery nahrazeny podtržítkem. Zde byl také vytvořen datový slovník, který je zobrazen v tabulce 3. Poté byl soubor uložen s koncovkou *.csv. Takto uložený soubor s touto koncovkou je již možné importovat do systému Clementine.

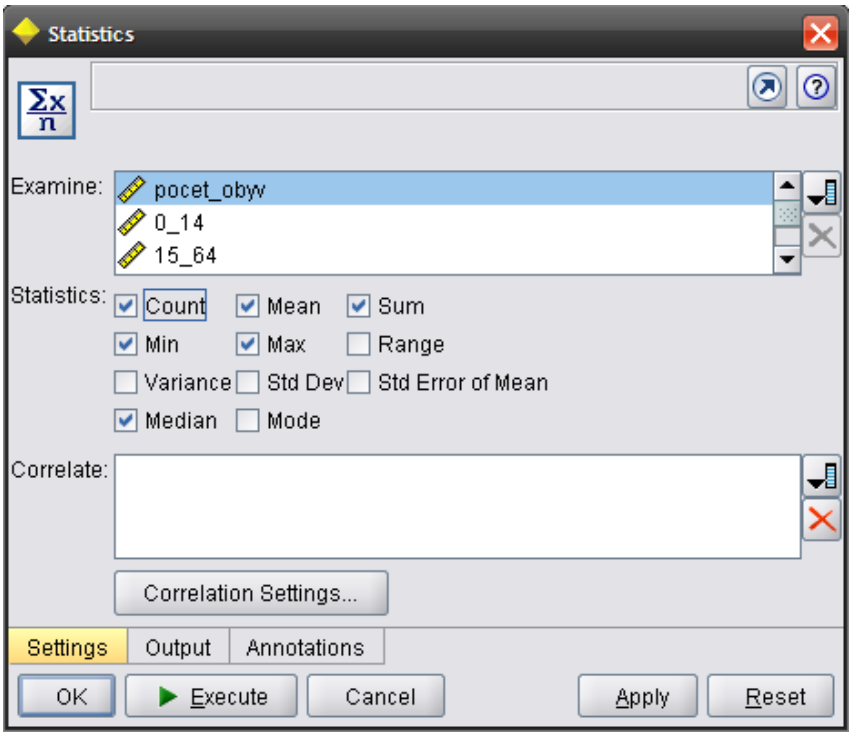
Tabulka 3 - datový slovník. Zdroj [vlastní].

atribut	typ proměnné	typ v Clementine	rozsah dat	původní název
spr_obv_obci	kategorická	set	Broumov,…	správní obvod obce
pocet_obyv	spojitá	range	[1651;94493]	počet obyvatel
0_14	spojitá	range	[462;15521]	0-14 let
15_64	spojitá	range	[2347;80705]	15-64 let
65_vice	spojitá	range	[412;20374]	více než 64 let
ekonom_aktivni	spojitá	range	[884;57913]	počet ekonomicky aktivních
uchazeci_o_zam	spojitá	range	[73;4370]	uchazeči o zaměstnání
nezam	spojitá	range	[5.7;17.6]	velikost nezaměstnanosti %
ekonom_subj	spojitá	range	[361;28714]	počet ekonomických subjektů
ms	spojitá	range	[1;29]	počet mateřských škol
zs	spojitá	range	[1;19]	počet základních škol
ss	spojitá	range	[0;17]	počet středních škol
dom_s_pec_sl	spojitá	range	[0;7]	počet domovů s pečovatelskou službou
dom_duchodcu	spojitá	range	[0;3]	počet domovů důchodců
urad_prace	spojitá	range	[0;1]	počet úřadů práce
lekar_dospeli	spojitá	range	[1;48]	počet ordinací lékařů pro dospělé
lekar_deti	spojitá	range	[0;27]	počet ordinací lékařů pro děti a dorost
narozeni	spojitá	range	[13;1027]	počet narozených
zemreli	spojitá	range	[11;970]	počet zemřelých
snatky	spojitá	range	[5;436]	počet sňatků
rozvody	spojitá	range	[1;265]	počet rozvodů
vystavba_bytu	spojitá	range	[3;516]	počet nových bytů

Data jsou načtena pomocí uzlu Var. File, a je provedena základní statistická analýza dat pomocí uzlu STATISTICS. Úprava tohoto uzlu je na obrázku 5. Je potřeba upravit oddělovače (delimiters) a to tak, že oddělovačemi budou Newline a Other, kde u Other je vložen oddělující znak středník, který pro oddělení jednotlivých atributů a případů používá právě středník. Základní statistika je potom zobrazena na obrázku 6.

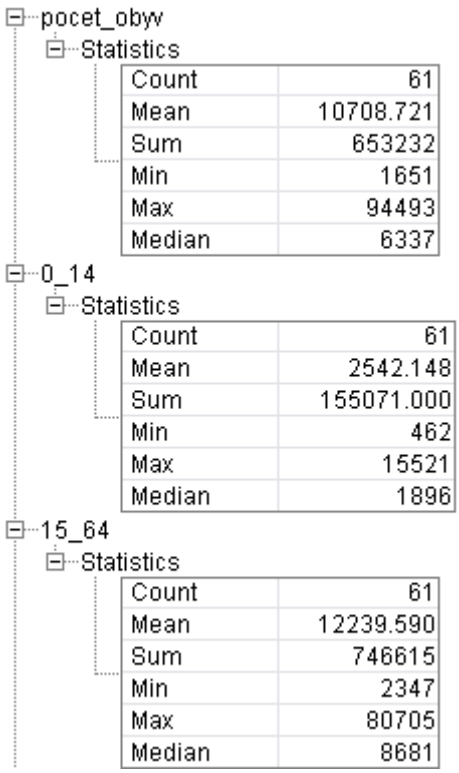


Obrázek 5 - nastavení uzlu Var. File. Zdroj [vlastní].



Obrázek 6 - základní statistika. Zdroj [vlastní].

Základní statistika byla provedena pomocí Count (počet), Min (minimální hodnota), Max (maximální hodnota), Mean (střední hodnota), Sum (suma), Median (median). Výstup je na obrázku 7. Kompletní statistika je v příloze 1.



Obrázek 7 - výstup statistiky. Zdroj [vlastní].

3.3 Posouzení kvality dat

Posouzení kvality dat se provádí z důvodu zjištění, kolik dat je validních, zda existují v datech nějaké extrémy, zda jsou data kompletní, zda obsahují prázdné případy. K této analýze slouží blok Data Audit, který má tu výhodu, že zároveň se základním posouzením jednotlivých atributů, zobrazuje i aktuální grafický výstup ke každému atributu podle jeho hodnot. Obrázek 8 zobrazuje kvalitu jednotlivých atributů.

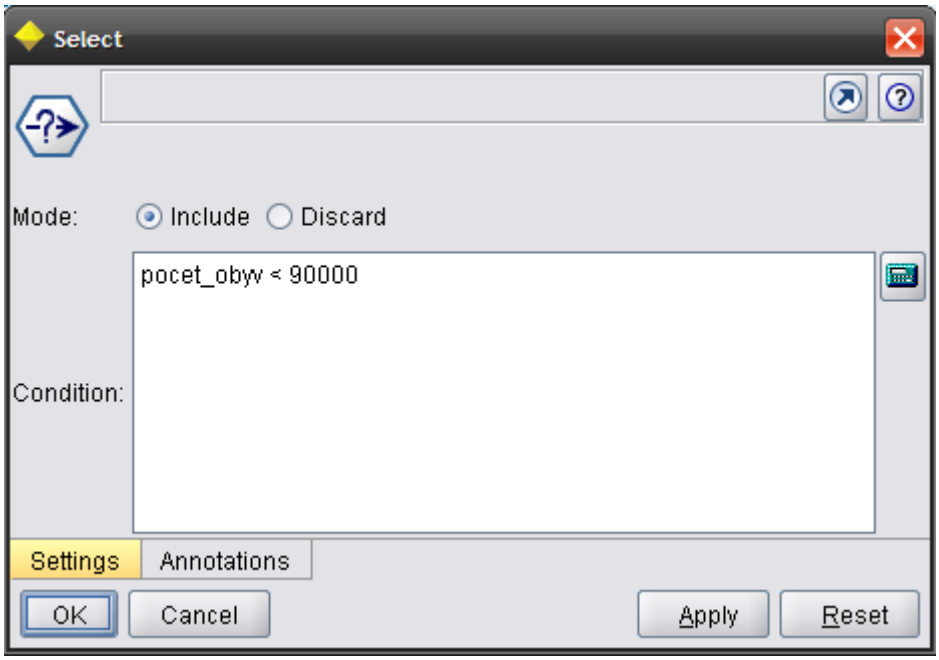
Field	Type	Outliers	Extremes	% Complete	Valid Records
pocet_obyv	Range	1	1	100	61
0_14	Range	2	0	100	61
15_64	Range	2	0	100	61
65_vice	Range	1	1	100	61
ekonom_akti...	Range	1	1	100	61
uchazeci_o_...	Range	2	0	100	61
nezam	Range	1	0	100	61
ekonom_subj	Range	1	1	100	61
ms	Range	1	1	100	61
zs	Range	1	1	100	61
ss	Range	1	1	100	61
dom_s_pec_...	Range	1	0	100	61
dom_duchod...	Range	1	0	100	61
urad_prace	Range	0	0	100	61
lekar_dospeli	Range	1	1	100	61
lekar_deti	Range	1	1	100	61
narozeni	Range	1	1	100	61
zemreli	Range	1	1	100	61
snatky	Range	1	1	100	61
rozvody	Range	1	1	100	61
vystavba_bytu	Range	1	1	100	61

Obrázek 8 - kvalita dat. Zdroj [vlastní].

Z obrázku 8 je patrné, že jsou data kompletní, datový záznam obsahuje u všech atributů shodně 61 záznamů. Vidíme, že se v datech vyskytují extrémy, ty budou odstraněny. Extrémy jsou způsobeny dvěma případy a to konkrétně městy Hradec Králové a Pardubice. Je to z důvodu velikosti měst a tím pádem i očekávaným velikým zvětšením nebo zmenšením hodnot u jednotlivých atributů. Tato města budou pro další analýzu vyřazena a nebudou zařazena do modelování právě z důvodu možnosti nepřesnosti výsledku jednotlivých modelů.

K odstranění těchto dvou měst nám poslouží blok Select, kde pomocí podmínky vyřadíme města Hradec Králové a Pardubice. Blok Select s danou podmínkou jsou zobrazeny na obrázku 9. Vidíme, že podmínka byla nastavena na kritérium, že budou vybrána města, která mají méně než 90000 tisíc obyvatel. Hradec Králové i Pardubice mají více než 90000 tisíc obyvatel, proto budou díky této podmínce odstraněna.

Následná kontrola, zda data stále obsahují extrémní hodnoty, zobrazuje obrázek 10.



Obrázek 9 - podmínka Select. Zdroj [vlastní].

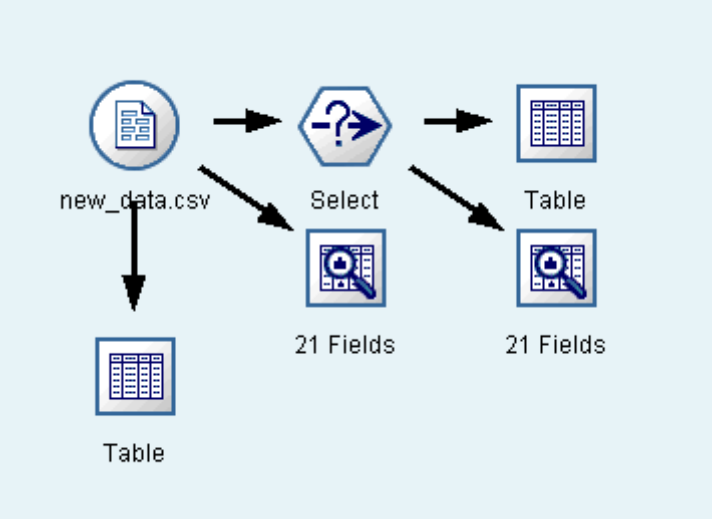
Field	Type	Outliers	Extremes	% Complete	Valid Records
pocet_obyv	Range	1	0	100	59
0_14	Range	1	0	100	59
15_64	Range	1	0	100	59
65_vice	Range	0	0	100	59
ekonom_akti...	Range	0	0	100	59
uchazeci_o_...	Range	1	0	100	59
nezam	Range	1	0	100	59
ekonom_subj	Range	1	0	100	59
ms	Range	0	0	100	59
zs	Range	3	0	100	59
ss	Range	2	0	100	59
dom_s_pec_...	Range	1	0	100	59
dom_duchod...	Range	1	0	100	59
urad_prace	Range	0	0	100	59
lekar_dospeli	Range	2	0	100	59
lekar_deti	Range	0	0	100	59
narozeni	Range	1	0	100	59
zemreli	Range	1	0	100	59
snatky	Range	1	0	100	59
rozvody	Range	1	0	100	59
vystavba_bytu	Range	1	0	100	59

Obrázek 10 - kontrola kvality dat. Zdroj [vlastní].

Z obrázku 10 je patrné, že extrémní hodnoty byly odstraněny a o dva záznamy ubylo validních záznamů.

Výsledný stream pro analýzu dat

Obrázek 11 zobrazuje výsledný stream, který byl použit pro posouzení kvality dat.



Obrázek 11 - stream kvality dat. Zdroj [vlastní].

4. Návrh modelů

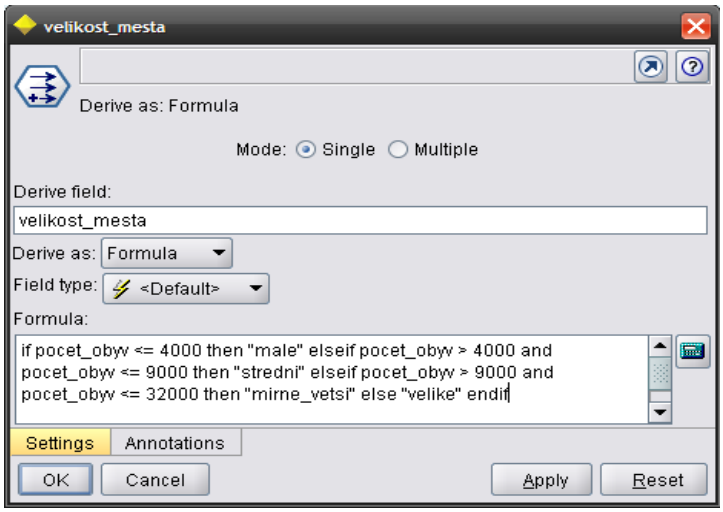
V této kapitole jsou popsány postupy tvorby jednotlivých modelů, jejich analýzy, pokud to jde, tak i grafické výstupy. Jsou zde popsány silné a slabé stránky jednotlivých modelů a vždy je zobrazen výsledný návrh příslušného modelu.

4.1 Asociační pravidla

Jak již bylo řečeno v kapitole 2.1, asociační pravidla jsou schopny generovat pravidla z dat. Pro naše účely budeme pomocí asociačních pravidel modelovat vybraná data z oblasti sociální politiky, kterými jsou:

- počet středních škol,
- počet ordinací lékaře pro dospělé,
- přítomnost domovu důchodců,
- počet nových bytů,
- velikost města.

Nejdříve musejí být upravena data do takového formátu, se kterým dokáže algoritmus APRIORI pracovat a vytvářet model. V tomto kroku budou data převedena z dat číselných do kategorických pomocí vhodně vybraných intervalů a modelovacího uzlu Derive, který opět za pomoci podmínky vytvoří nový atribut. Uzel Derive s danou podmínkou a nastavením je zobrazen na obrázku 12.



Obrázek 12 - uzel Derive, asociační pravidla. Zdroj [vlastní].

Z obrázku 12 je patrné, že vznikne nový kategorický atribut „**velikost_mesta**“ s podmínkou, která je v tabulce 4. Stejný postup musíme zopakovat ještě pro zbývající výše uvedené vstupní atributy.

Tabulka 4 - pocet_obyvatel. Zdroj [vlastní].

pocet_obyvatel	
podmínka (IF)	interval (THEN)
<= 4000	male
> 4000 and <= 9000	stredni
> 9000 and <= 32000	mirne_vetsi
> 32000	velike

Atribut „**pocet_ss**“ byl vygenerován s podmínkou:

```
if ss  = 0 then "zadna" elseif ss > 0  and ss  <= 2 then "nedostatecne" elseif ss > 2  and ss <= 5 then "dostatecne" else "hodne" endif
```

Atribut „**pocet_lekaru_dos**“ byl vygenerován s podmínkou:

```
if lekar_dospeli  <= 2 then "nedostatecne" elseif lekar_dospeli > 2  and lekar_dospeli  <= 10 then "dostatecne" else "hodne" endif
```

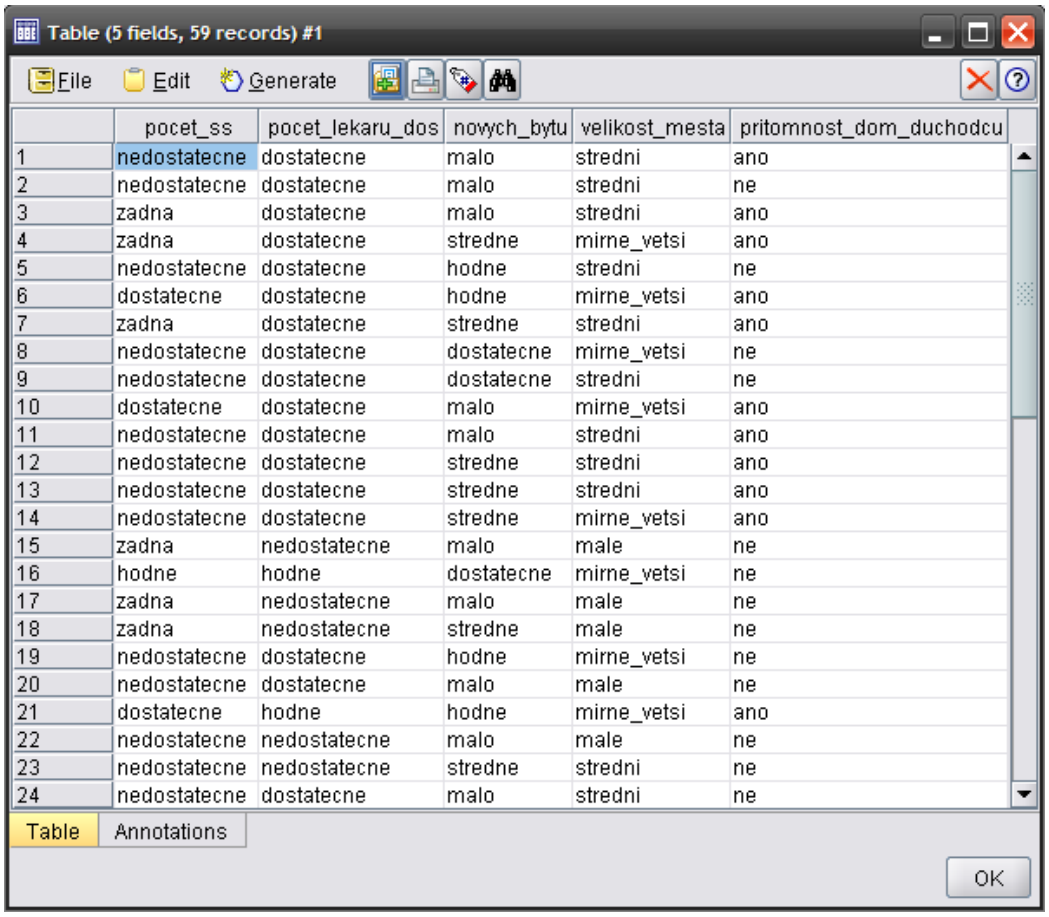
Atribut „**novych_bytu**“ byl vygenerován s podmínkou:

```
if vystavba_bytu <= 20 then "malo" elseif vystavba_bytu > 20 and vystavba_bytu <= 47 then "stredne" elseif vystavba_bytu > 47 and vystavba_bytu <= 80 then "hodne" else "dostatecne" endif
```

Atribut „**pritomnost_dom_duchodcu**“ byl vygenerován s podmínkou:

```
if dom_duchodcu = 0 then "ne" else "ano" endif
```

Výsledných pět nových atributů a jejich hodnoty jsou zobrazeny na obrázku 13. Je zde zobrazeno pět nových kategorických atributů s prvními 24 případy. Kompletní tabulka s vygenerovanými hodnotami je v příloze 2.



	pocet_ss	pocet_lekaru_dos	novych_bytu	velikost_mesta	pritomnost_dom_duchodcu
1	nedostatecne	dostatecne	malo	stredni	ano
2	nedostatecne	dostatecne	malo	stredni	ne
3	zadna	dostatecne	malo	stredni	ano
4	zadna	dostatecne	stredne	mirne_vetsi	ano
5	nedostatecne	dostatecne	hodne	stredni	ne
6	dostatecne	dostatecne	hodne	mirne_vetsi	ano
7	zadna	dostatecne	stredne	stredni	ano
8	nedostatecne	dostatecne	dostatecne	mirne_vetsi	ne
9	nedostatecne	dostatecne	dostatecne	stredni	ne
10	dostatecne	dostatecne	malo	mirne_vetsi	ano
11	nedostatecne	dostatecne	malo	stredni	ano
12	nedostatecne	dostatecne	stredne	stredni	ano
13	nedostatecne	dostatecne	stredne	stredni	ano
14	nedostatecne	dostatecne	stredne	mirne_vetsi	ano
15	zadna	nedostatecne	malo	male	ne
16	hodne	hodne	dostatecne	mirne_vetsi	ne
17	zadna	nedostatecne	malo	male	ne
18	zadna	nedostatecne	stredne	male	ne
19	nedostatecne	dostatecne	hodne	mirne_vetsi	ne
20	nedostatecne	dostatecne	malo	male	ne
21	dostatecne	hodne	hodne	mirne_vetsi	ano
22	nedostatecne	nedostatecne	malo	male	ne
23	nedostatecne	nedostatecne	stredne	stredni	ne
24	nedostatecne	dostatecne	malo	stredni	ne

Obrázek 13 - nové atributy Define. Zdroj [vlastní].

4.1.1 Nastavení uzlu APRIORI a GRI

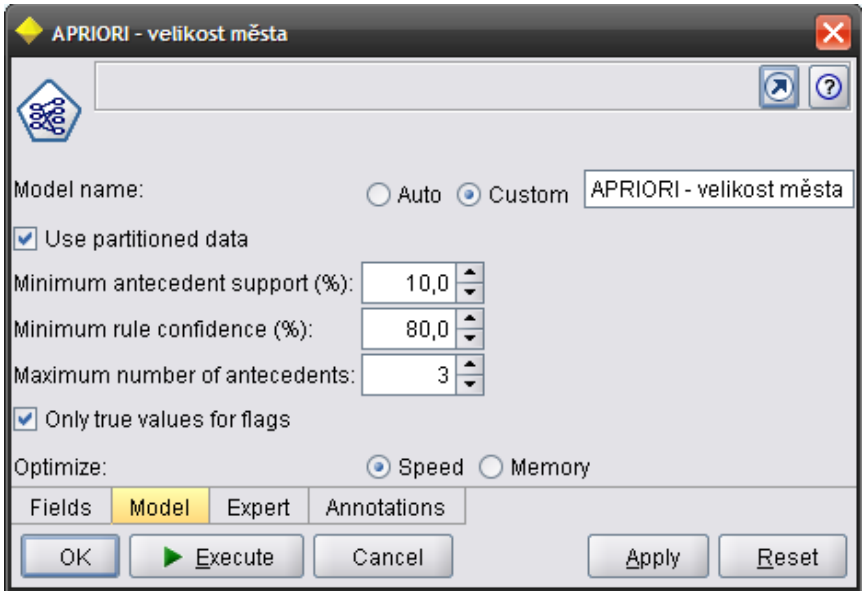
Ze základního nastavení uzlu APRIORI byly určeny vstupní atributy pro Consequents (závěry) a Antecedents (předpoklady). Pro vstupní pole předpokladů byly zvoleny atributy:

- novych_bytu,
- pocet_ss,
- pritomnost_dom_duchodcu,
- pocet_lekaru_dos.

Pro vstupní pole závěrů byl zvolen atribut:

- velikost_mesta.

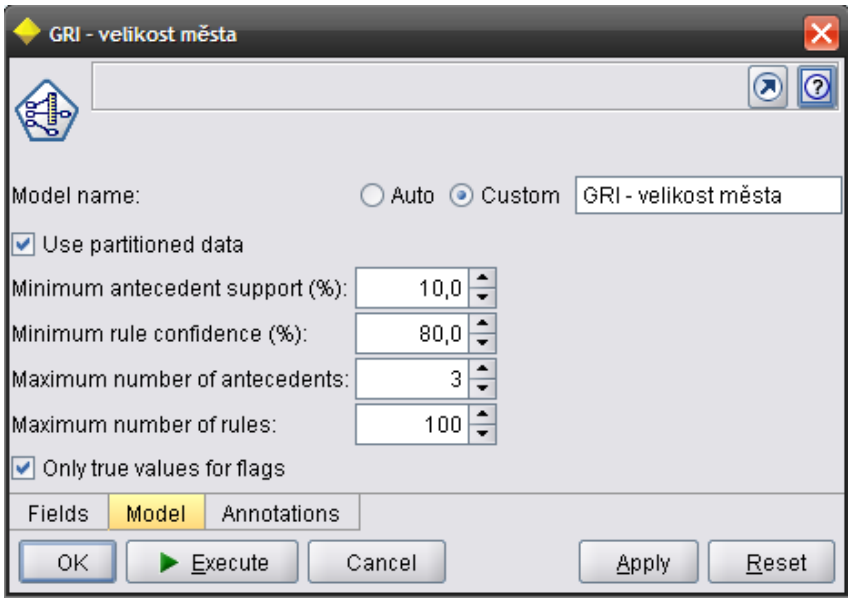
Další nastavení se týká nastavení modelu samotného a to konkrétně bylo nastaveno minimální podpora předpokladu a minimální podpora závěru. Ostatní nastavení zůstalo původní. Nastavení minimální podpory předpokladu a závěru je na obrázku 14.



Obrázek 14 - nastavení uzlu APRIORI. Zdroj [vlastní].

Uzel GRI[11] je další možností pro vygenerování asociačních pravidel. Asociační pravidla jsou zde vygenerována opět pomocí podmínky **if** *antecedent (s)* **then** *consequent (s)*. Nastavení uzlu pro vstupní pole předpokladů a závěrů je určeno stejně jako u uzlu APRIORI. Tím bude zajištěno za stejných podmínek vstupních dat a nastavení modelu se stejnými parametry jako je tomu opět u uzlu APRIORI vygenerování odlišných asociačních pravidel. Nastavení maximálního počtu předpokladů je u obou uzlů nastaveno na číslo 3. Obrázek 15 zobrazuje nastavení uzlu GRI. Maximální počet vygenerovaných pravidel je 100.

[11] GRI – generalized rule induction (zobecněné pravidlo indukce)



Obrázek 15 - nastavení GRI. Zdroj [vlastní].

4.1.3 Analýza pravidel APRIORI a GRI

Analýza vygenerovaných pravidel APRIORI ukazuje následující hodnoty:

- počet pravidel: 20,
- počet odpovídajících transakcí: 59,
- minimální podpora: 10,169%,
- maximální podpora: 30,508%,
- minimální spolehlivost: 80,0%,
- maximální spolehlivost: 100,0%.

První čtyři vygenerovaná pravidla uzlem APRIORI jsou v tabulce 5, všechna příslušná pravidla jsou v příloze 3.

Tabulka 5 - pravidla APRIORI. Zdroj [vlastní].

Consequent	Antecedent	Support %	Confidence %
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pritomnost_dom_duchodcu = ano	15,254	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and novych_bytu = malo	15,254	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and pocet_ss = zadna	20,339	100
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pritomnost_dom_duchodcu = ano and pocet_lekaru_dos = dostatecne	13,559	100

Interpretace prvního pravidla je následující:

JESTLIŽE počet středních škol je dostatečný (3 až 5 středních škol) A je zde přítomnost domova důchodců, POTOM je velikost města mírně větší (počet obyvatel je v rozmezí 9001 až 32000). Počet případů splňujících či nesplňujících předpoklad či závěr je v tabulce 6.

Tabulka 6 - čtyřpolní tabulka APRIORI. Zdroj [vlastní].

	závěr	¬závěr	∑
předpoklad	9	0	9
¬předpoklad	5	45	50
∑	14	45	59

Spolehlivost (confidence) = (9/9)*100 = **100%**, **podpora** (support) = (9/59)*100 = 15,254%.

Analýza vygenerovaných pravidel GRI ukazuje následující hodnoty:

- počet pravidel: 8,
- počet odpovídajících transakcí: 59,
- minimální podpora: 10,17%,
- maximální podpora: 30,51%,
- minimální spolehlivost: 80,0%,
- maximální spolehlivost: 100,0%.

Všechna pravidla jsou v tabulce 7. Počet případů splňujících či nesplňujících předpoklad či závěr je v tabulce 8.

Tabulka 7 - pravidla GRI. Zdroj [vlastní].

Consequent	Antecedent	Support %	Confidence %
velikost_mesta = male	pocet_ss = zadna and pocet_lekaru_dos = nedostatecne	20,34	100
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pritomnost_dom_duchodcu = ano	15,25	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and pritomnost_dom_duchodcu = ne	23,73	92,86
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne	20,34	91,67
velikost_mesta = stredni	pocet_ss = nedostatecne and pocet_lekaru_dos = dostatecne and novych_bytu = stredne	11,86	85,71
velikost_mesta = male	pocet_lekaru_dos = nedostatecne	30,51	83,33
velikost_mesta = stredni	pocet_ss = nedostatecne and novych_bytu = stredne and pritomnost_dom_duchodcu = ano	10,17	83,33

Consequent	Antecedent	Support %	Confidence %
velikost_mesta = stredni	pocet_ss = nedostatecne and pritomnost_dom_duchodcu = ano	16,95	80

Interpretace prvního pravidla je následující:

JESTLIŽE pocet_ss = zadna A pocet_lekar_dos = nedostatecne, POTOM velikost_mesta = male. Počet případů splňujících či nesplňujících předpoklad či závěr je v tabulce 8.

Tabulka 8 - čtyřpolní tabulka GRI. Zdroj [vlastní].

	závěr	¬závěr	∑
předpoklad	12	0	12
¬předpoklad	1	46	47
∑	13	46	59

Spolehlivost (confidence) = (12/12)*100 = **100%**, **podpora** (support) = (12/59)*100 = 20,339%.

4.1.5 Silné a slabé stránky použitých metod, výsledný stream

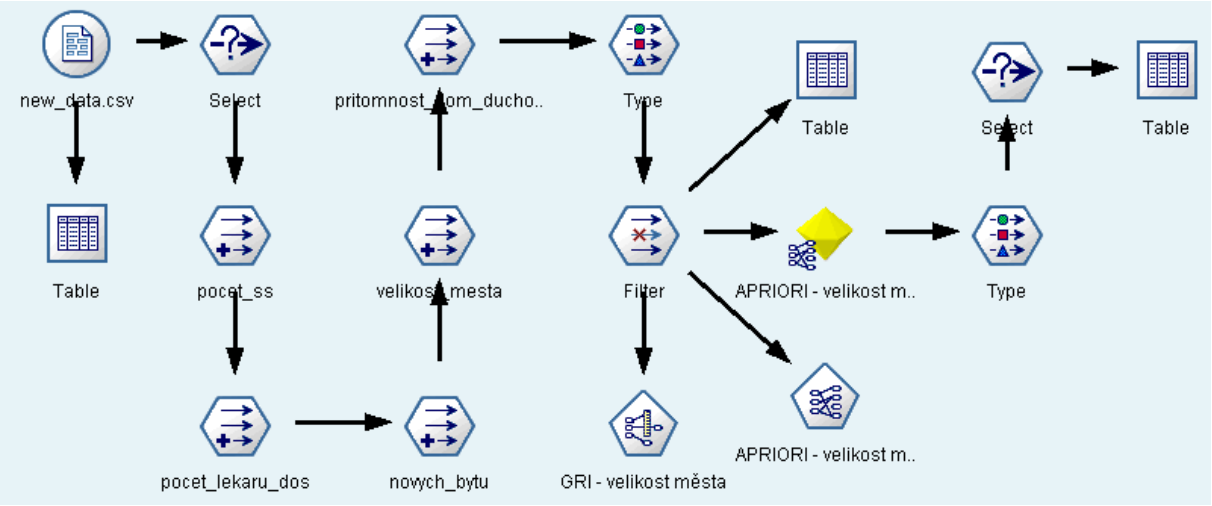
Silné stránky:

- možnost určení maximálního počtu předpokladů pro daný závěr,
- možnost určení počtu vygenerovaných pravidel,
- můžeme snadno odvodit, za jakých podmínek vstupní atributy (novych_bytu, pocet_ss, pritomnost_dom_duchodcu, pocet_lekar_dos) patří do příslušné velikosti města (velikost_mesta),
- každé pravidlo má přesně danou svoji spolehlivost a podporu.

Slabé stránky:

- nemožnost vytváření smyslupných grafických výstupů,
- v tomto případě u oboru algoritmů APRIORI a GRI nutnost převést spojitá data na data kategorická,
- s tím spojené riziko nevhodně nastavených intervalů (kategorií).

Výsledný stream vytvořený pro vygenerování asociačních pravidel je na obrázku 16.



Obrázek 16 - stream asociačních pravidel. Zdroj [vlastní].

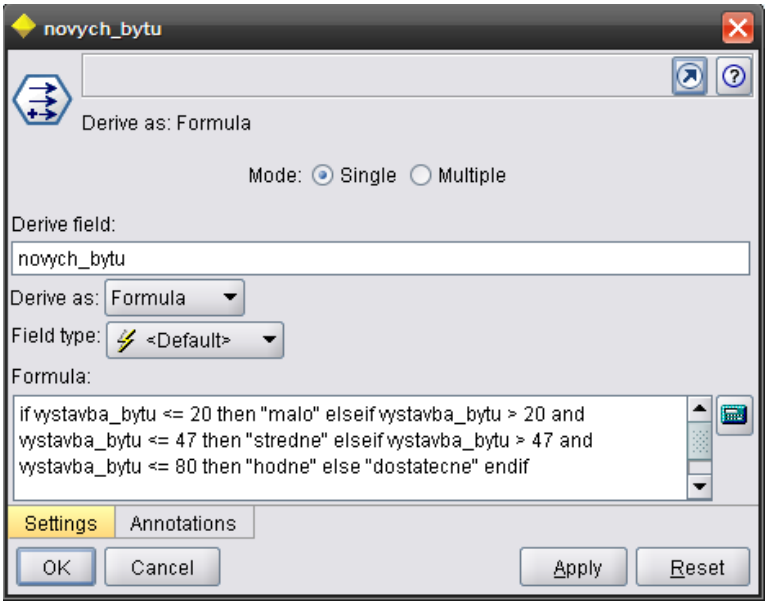
4.2 Shluková analýza

Pomocí shlukové analýzy budou modelována všechna spojitá data, která budou vstupními atributy uzlu K-Means. Vstupními atributy tedy budou:

- spr_obv_obci,
- pocet_obyv,
- 0_14,
- 15_64,
- 65_vice,
- ekonom_aktivni,
- uchazeci_o_zam,
- nezam,
- ekonom_subj,
- ms,
- zs,
- ss,
- dom_s_pec_sl,
- dom_duchodcu,
- urad_prace,
- lekar_dospeli,
- lekar_deti,

- narozeni,
- zemreli,
- snatky,
- rozvody,
- vystavba_bytu.

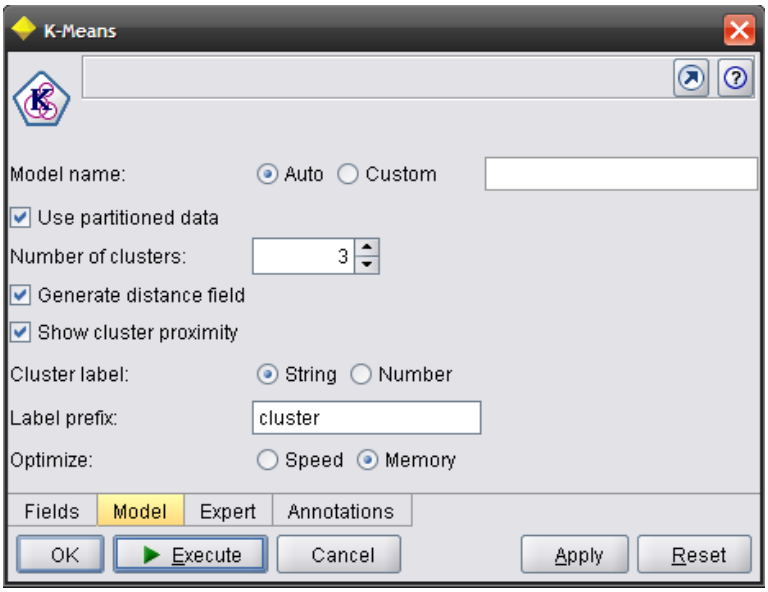
Z důvodu grafických výstupů budou atributy „pocet_obyvatel“ a „vystavba_bytu“ převedeny na atributy kategorické. Atribut „pocet_obyvatel“ použije stejnou podmínku i název jako tomu bylo u obrázku č. 11 v kapitole 4.1. Druhým novým atributem bude „novych_bytu“, který vznikne z atributu „vystavba_bytu“. Nastavení uzlu Derive i s podmínkou jsou zobrazeny na obrázku 17.



Obrázek 17 - uzel Derive, shluková analýza. Zdroj [vlastní].

4.2.1 Nastavení uzlu K-Means

Vstupními atributy budou výše zmíněné všechny atributy s numerickými hodnotami v části Fields. Samotné nastavení modelu v části Model bude provedeno podle obrázku 18. Zde je velice důležité se rozhodnout, do kolika počtu shluků se mají daná data roztřídit. Je zde možné i vlastní pojmenování nově vzniklých atributů, které budou reprezentovat jednotlivé shluky (clusters).



Obrázek 18 - nastavení uzlu K-Means. Zdroj [vlastní].

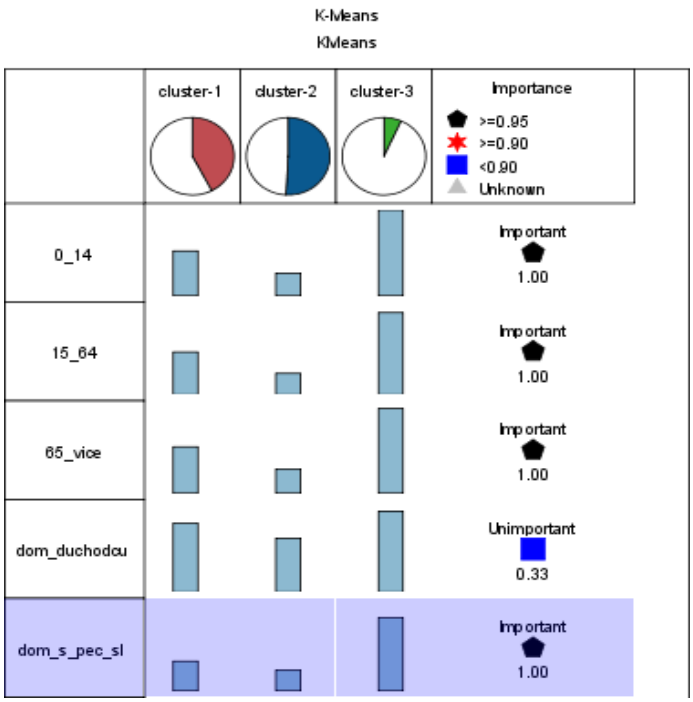
Z obrázku 18 je patrné, že bylo rozhodnuto o počtu shluků číslem tři. Tři shluky se zdají být odpovídající volbou na velikost datové matice, ze které vznikne model. Menší počet by mohl způsobit méně přesné zatřídění objektů a naopak větší počet je výhodnější pro velké objemy dat a to není tento případ.

4.2.2 Analýza vytvořeného K-Means modelu

Ze shrnutí modelu je patrné, že vznikly následující shluky:

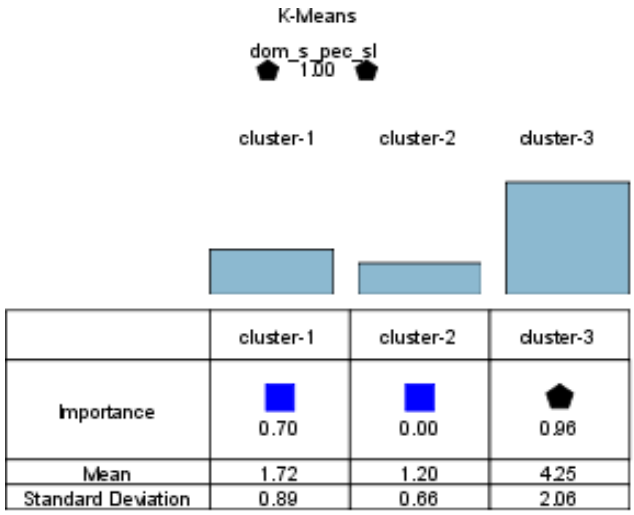
- cluster-1 o 25 záznamech,
- cluster-2 o 30 záznamech,
- cluster-3 o 4 záznamech.

Zobrazení jednotlivých shluků a důležitosti atributů je na obrázku 19.



Obrázek 19 - zobrazení klastrů. Zdroj [vlastní].

Na obrázku 19 vidíme jednotlivé klastry a prvních pět atributů s určenou důležitostí. Atribut „dom_duchodcu“ byl vyhodnocen jako nedůležitý s hodnotou 0.33. Naopak například atribut „dom_s_pec_sl“ byl vyhodnocen jako důležitý s hodnotou 1.00. Na každý atribut můžeme kliknout a tím tak získat podrobnější informace o daném atributu. Pokud klikneme na zvýrazněný atribut „dom_s_pec_sl“, uvidíme, jednotlivou důležitost pro každý shluk (cluster-1, 2, 3), střední hodnotu (Mean) daného atributu pro každý shluk a standardní odchylku. Obrázek 20 zobrazuje detailní pohled na atribut „dom_s_pec_sl“.



Obrázek 20 - detailní pohled K-Means. Zdroj [vlastní].

4.2.3 Výstupy pomocí bloku MATRIX, grafické výstupy

Blok matrix umožňuje přehledně zobrazovat jednotlivé hodnoty zadaných atributů v různých formátech, pro naše použití byly vyexportovány následující tabulky 9 a 10. Jedná se o obdobu kontingenční tabulky. V tabulce 9 jsou zobrazeny jednotlivé shluky a atribut „novych_bytu“.

Tabulka 9 - novych_bytu. Zdroj [vlastní].

	novych_bytu			
$KM-K-Means	**dostatecne**	**hodne**	**malo**	**stredne**
cluster-1	4	7	4	10
cluster-2	0	1	14	15
cluster-3	1	2	0	1

Z tabulky 9 je patrné, že do cluster-1 patří novych_bytu u kategorie „hodne“ sedm. Naopak u cluster-2 pouze jeden a u cluster-3 dva. Nejvíce nových bytů je tedy v kategorii hodně ve shluku cluster-1 a to sedm, nejméně jich je potom ve shluku cluster-2 počtem jeden.

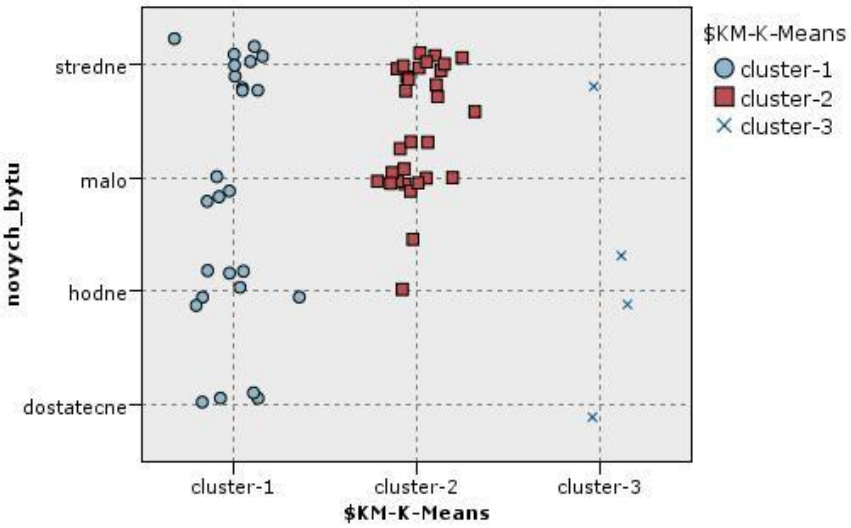
Tabulka 10 zobrazuje opět jednotlivé shluky, ale tentokrát v porovnání s atributem „velikost_mesta“.

Tabulka 10 - velikost_mesta. Zdroj [vlastní].

	velikost_mesta		
$KM-K-Means	**male**	**mirne_vetsi**	**stredni**
cluster-1	2	13	10
cluster-2	15	3	12
cluster-3	0	4	0

Z tabulky 10 je patrné, že podle atributu velikost_mesta, které je „stredni“, byly zaklasifikovány v cluster-1 deset měst, v cluster-2 dvanáct měst a v cluster-3 žádné město. Nejvíce podle počtu obyvatel je středních měst dvanást ve shluku cluster-2.

Grafické výstupy jsou realizovány pomocí bloku Plot. Vstupními atributy byly atributy „novych_bytu“ a „velikost_mesta“ a nově vzniklý atribut „$KM-K-Means“. Obrázek 21 zobrazuje pomocí grafu Plot atributy „$KM-K-Means“ a „novych_bytu“.



Obrázek 21 - shluky a nové byty. Zdroj [vlastní].

Na obrázku 21 zastupují osu x jednotlivé shluky a osu y jednotlivé kategorie atributu „novych_bytu“. Každý shluk má svůj tvar a barvu. Toto roztřídění je velice přehledné a má vysokou vypovídací schopnost. Jde v podstatě o grafický výstup pro tabulku č. 9. Naopak obrázek 22 je těžší na interpretaci. Jsou zde zobrazeny hned tři atributy, osu x představuje „velikost_mesta“ a osu y „novych_bytu“. Jednotlivé objekty, které patří k předchozím dvěma atributům, jsou obarveny a mají odlišnou velikost podle toho jednotlivých shluků, do kterých patří.

Obrázek 22 - velikost města, nových bytů. Zdroj [vlastní].

4.2.5 Silné a slabé stránky metody, výsledný stream

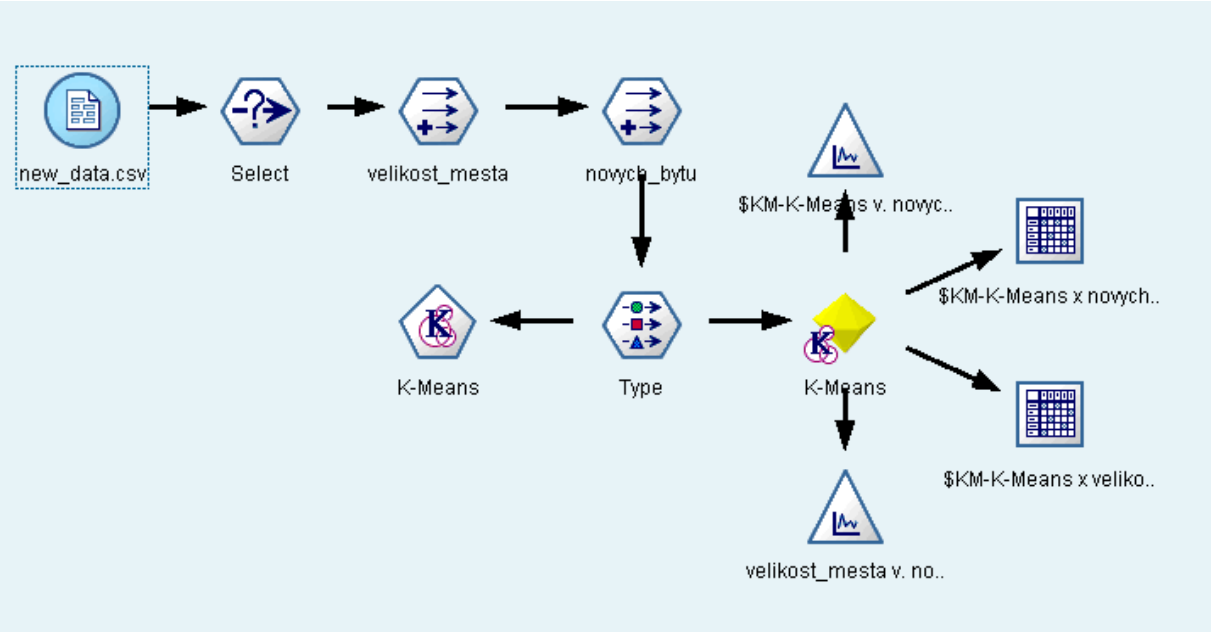
Silné stránky:

- spojitá vstupní data se nemusejí převádět na kategorická,
- možnost nastavení, do kolika shluků se mají data roztřídit,
- možnost grafických výstupů, které se snadno interpretují,
- pomocí uzlu MATRIX možnost přehledného tabulkového výstupu klasifikace určitého atributu do konkrétního shluku,
- u grafických výstupů je široká možnost použití zobrazení jednotlivých objektů, jako je tomu na obrázku 22, kdy hodnoty dvou atributů jsou zvýrazněny třetím atributem, v tomto případě jednotlivými shluky.

Slabé stránky:

- nutnost zvážení nastavení počtu shluků v závislosti na velikosti datové matice,
- s tím spojené riziko nevhodného nastavení počtu shluků,
- pro snadnější interpretaci grafických výstupů je zapotřebí převedení určených spojitých atributů (pocet_obyv, vystavba_bytu) na kategorické atributy (velikost_mesta, novych_bytu),

Výsledný stream vytvořený pro model shlukové analýzy K-Means je na obrázku 23.

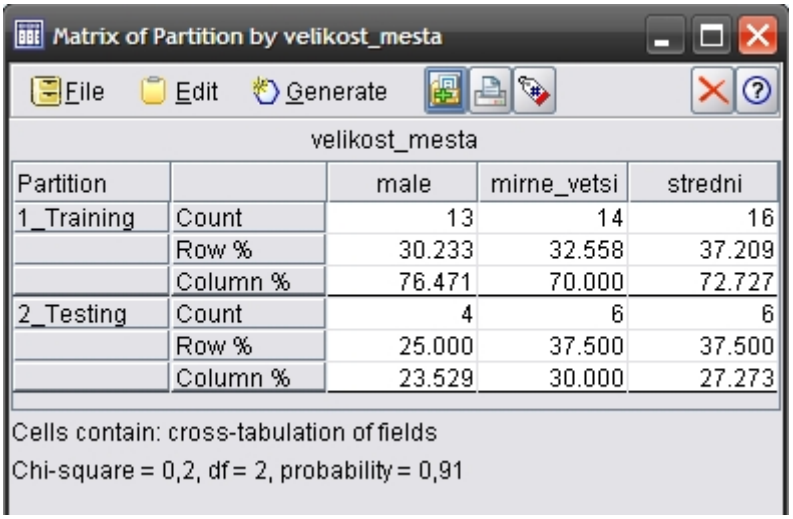


Obrázek 23 - stream shlukové analýzy. Zdroj [vlastní].

4.3 Rozhodovací stromy

U rozhodovacích stromů jsou nejdříve data rozdělena na množiny trénovací a testovací. Trénovací množina bude obsahovat 70% všech záznamů a testovací množina bude obsahovat 30% záznamů. Rozdělení se provede pomocí uzlu Partition. Dále je možné data rozdělit ještě na množinu validační, ale v našem případě toto dělat nebudeme, z důvodu toho, že nemáme tolik záznamů v datové matici.

Takovéto rozdělení dat se provede náhodně, pomocí algoritmu, který uzel Partition využívá a bude použitelné pro všechny následně vytvořené modely rozhodovacích stromů. V tomto případě budou vytvořeny binární rozhodovací stromy QUEST a C&RT. Na obrázku 24 je zobrazeno rozdělení záznamů na množinu testovací a trénovací, pomocí vstupních atributů „partition“ a „velikost_mesta“. Je zde i zobrazeno, kolik kterých záznamů v příslušné velikosti města patří do které množiny.



Obrázek 24 - rozdělení množin dat. Zdroj [vlastní].

4.3.1 Nastavení uzlu QUEST a C&RT

Ze základního nastavení pro uzly QUEST a C&RT byly určeny jako vstupní atributy následující atributy:

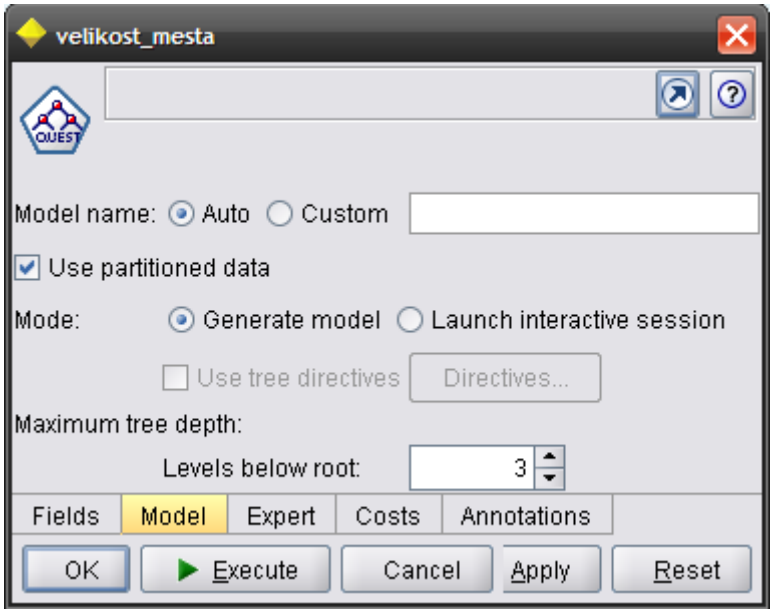
- novych_bytu,
- pocet_ss,

- pritomnost_dom_duchodcu,
- pocet_lekaru_dos.

Pro cílový atribut byl pro oba uzly zvolen atribut:

- velikost_mesta.

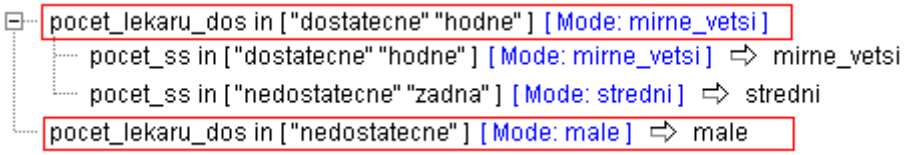
Shodné nastavení pro oba uzly zobrazuje obrázek 25. Zde je důležité poznamenat, že je zaškrtnuto pole „Use partitioned data“ a „Set random seed“. Dále je pro oba uzly určeno, do kolika podúrovní mají být stromy vygenerovány. Bylo určeno, že stromy budou generovány do maximálního počtu podúrovní čísla tři. Ostatní nastavení je ponecháno původnímu nastavení modelů.



Obrázek 25 - nastavení QUEST, C&RT. Zdroj [vlastní].

4.3.2 Analýza modelu QUEST a C&RT

Pro model QUEST byla vygenerována hloubka stromu o číslu dva pomocí pravidel, která jsou zobrazena na obrázku 26.



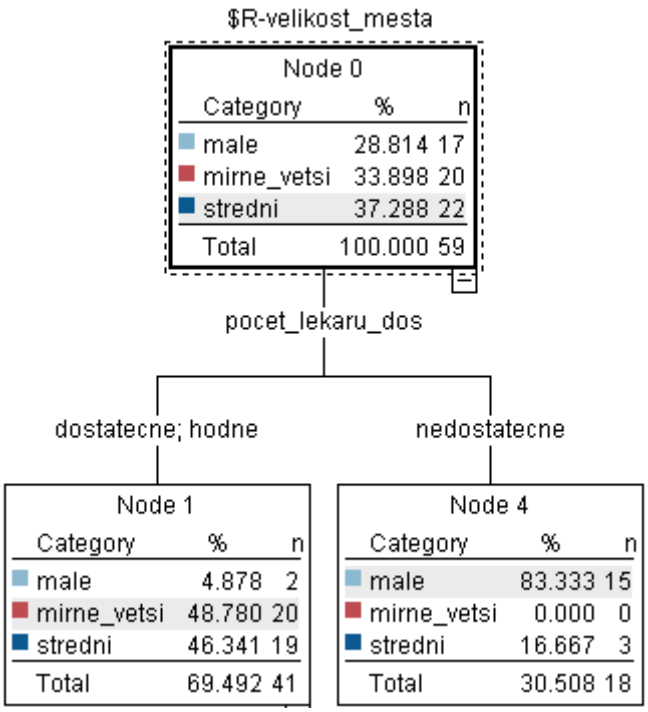
Obrázek 26 - quest pravidla. Zdroj [vlastní].

První dvě pravidla v první úrovni jsou interpretovatelná následujícím způsobem (pravidla v první úrovni jsou v červeném rámečku, pravidla v druhé úrovní jsou bez rámečku):

IF pocet_lekaru_dos = dostatecne AND pocet_lekaru_dos = hodne THEN velikost_mesta = mirne_vetsi

IF pocet_lekaru_dos = nedostatecne AND pocet_lekaru_dos = hodne THEN velikost_mesta = male

Stromová struktura rozhodovacího stromu QUEST je na obrázku 27.



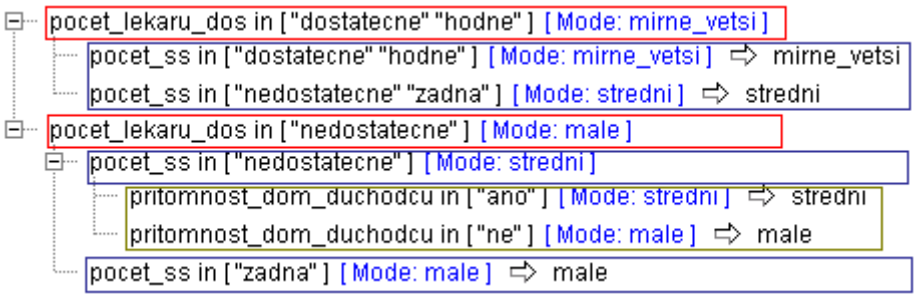
Obrázek 27 - quest strom. Zdroj [vlastní].

Zde je zobrazena pouze první úroveň stromu. Celý rozhodovací strom je poté v příloze 4. Vidíme kořenový atribut „velikost_mesta“, kde je zvolena velikost města „stredni“ díky tomu, že je nejvíce zastoupena počtem 22 a procentuálním zastoupením 37.288 %. Dále je strom

rozdělen do dvou podúrovní, tzv. listů s rozhodovacím kritériem „pocet_lekaru_dos“, které se dále dělí podle podkategorií tohoto atributu na „dostatecne; hodne“ a „nedostatecne“.

Zde také můžeme porovnat textový výstup uvedený v obrázku 26 daných pravidel s grafickým výstupem uvedeným na obrázku 27. Grafický výstup se zdá být pro první pohled přehlednější a tím pádem má i vyšší vypovídací schopnost.

Pro model C&RT byla vygenerována hloubka stromu o číslu tři pomocí pravidel, která jsou zobrazena na obrázku 28.



Obrázek 28 - c&rt pravidla. Zdroj [vlastní].

První úroveň je označena červeným rámečkem a obsahuje dvě stejná pravidla jako pravidla v první úrovni u QUEST rozhodovacího stromu.

Druhá úroveň je označena modrým rámečkem s počtem vygenerovaných pravidel čtyři a třetí úroveň je označena zeleným rámečkem s počtem vygenerovaných pravidel dva. Vygenerovaná pravidla v druhé úrovni rozhodovacího stromu jsou interpretovatelná následujícím způsobem:

IF pocet_ss = dostatecne AND pocet_ss = hodne THEN velikost_mesta = mirne_vetsi

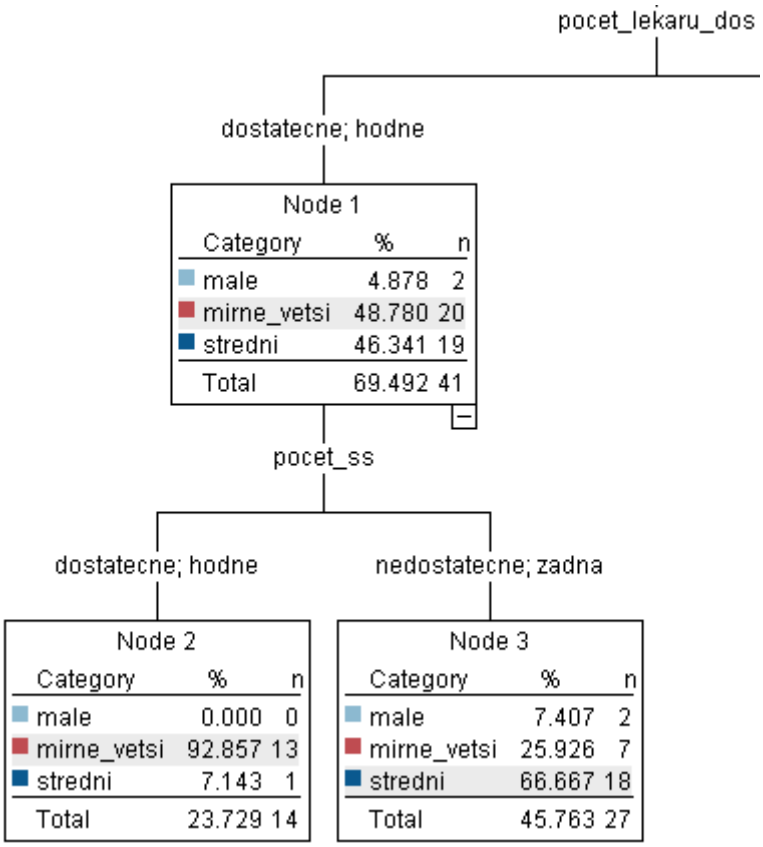
IF pocet_ss = nedostatecne AND pocet_ss = zadna THEN velikost_mesta = stredni

IF pocet_ss = nedostatecne THEN velikost_mesta = stredni

IF pocet_ss = zadna THEN velikost_mesta = male

Částečná struktura rozhodovacího stromu C&RT je na obrázku 29. Je zde zobrazena pouze první úroveň rozhodovacího atributu „pocet_lekaru_dos“ = „dostatecne; hodne“ a druhá

úroveň rozhodovacího stromu „pocet_ss“ = „dostatecne; hodne“ a „nedostatecne; zadna“. Celá stromová struktura rozhodovacích pravidel je zobrazena v příloze 5.



Obrázek 29 - část c&rt stromu. Zdroj [vlastní].

V první úrovní stromu vidíme vítěznou velikost města „mirne_vetsi“ s počtem zastoupení 20 a procentuálním vyjádřením 48.780 % na základě podmínky,

IF pocet_lekaru_dos = dostatecne AND pocet_lekaru_dos = hodne THEN velikost_mesta = mirne_vetsi

Dále je určeno rozhodovací pravidlo pro „pocet_ss“ = „dostatecne; hodne“, kterému byla určena velikost města „mirne_vetsi“ s počtem 13 a procentuálním vyjádřením 92.857 %.

Rozhodovací pravidlo „pocet_ss“ = „nedostatecne; zadna“ bylo vygenerováno pro „velikost_mesta“ = „stredni“ o počtu záznamů 18 a procentuálním vyjádřením 66.667%.

4.3.4 Zhodnocení modelů QUEST a C&RT

Pomocí uzlu Analysis hodnotíme, do jaké míry se hodnoty jednotlivých modelů shodly s hodnotami cílového atributu, v tomto případě atributu „velikost_mesta“. Na obrázku 30 je zobrazen výstup z uzlu Analysis.

Results for output field velikost_mesta

Individual Models

Comparing $R-velikost_mesta with velikost_mesta

'Partition'	1_Training		2_Testing	
Correct	34	79,07%	12	75%
Wrong	9	20,93%	4	25%
Total	43		16	

Comparing $R1-velikost_mesta with velikost_mesta

'Partition'	1_Training		2_Testing	
Correct	36	83,72%	12	75%
Wrong	7	16,28%	4	25%
Total	43		16	

Agreement between $R-velikost_mesta $R1-velikost_mesta

'Partition'	1_Training		2_Testing	
Agree	41	95,35%	16	100%
Disagree	2	4,65%	0	0%
Total	43		16	

Comparing Agreement with velikost_mesta

'Partition'	1_Training		2_Testing	
Correct	34	82,93%	12	75%
Wrong	7	17,07%	4	25%
Total	41		16	

Obrázek 30 - srovnání výsledků quest, c&rt. Zdroj [vlastní].

Červený rámeček zachycuje hodnoty modelu QUEST a C&RT. U modelu QUEST byla dosažena shoda trénovacích dat s daty cílového atributu („velikost_mesta“) v celkem 39 případech z celkových 43 případů, což činí 79.07%. Špatně bylo odhadnuto 9 případů. U dat testovacích došlo ke shodě v 12 případech z celkových 16, což činí 75%, špatně bylo odhadnuto 25% dat, což činí 4 případy.

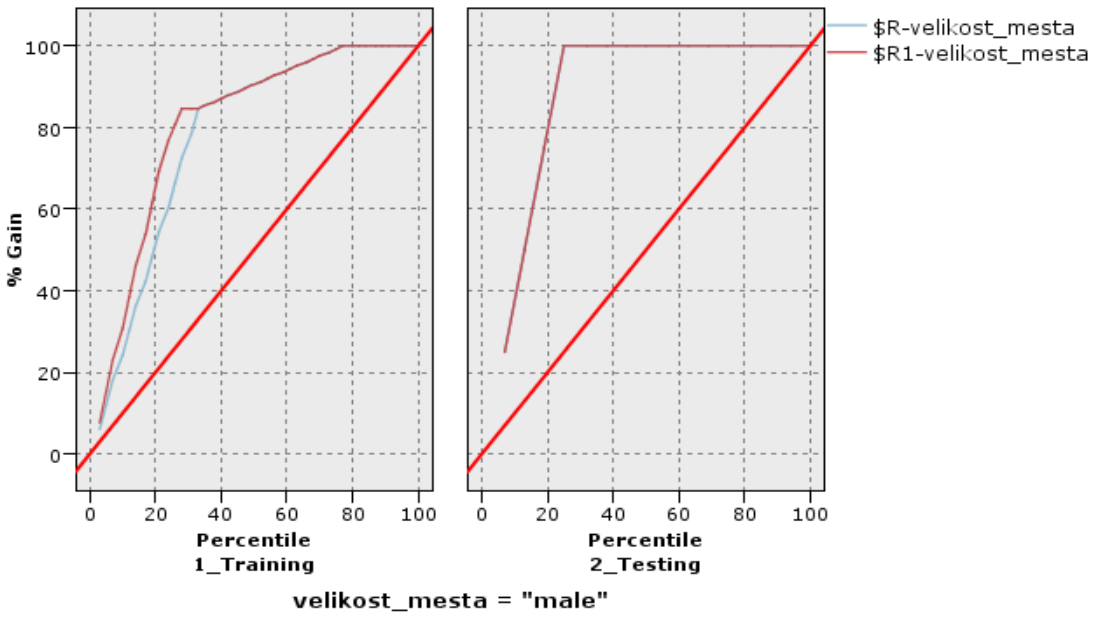
U modelu C&RT bylo dosaženo shody v trénovacích datech s cílovým atributem v 36 případech ze 43 celkem, špatně bylo přiřazeno 7 případů. U testovacích dat bylo dosaženo shody v 12 případech a data se neshodly ve 4 případech.

Při porovnávání shody s atributem „velikost_mesta“ metoda C&RT dopadla lépe s trénovacími daty než metoda QUEST. U testovacích dat dopadly obě metody shodně.

Černý rámeček zobrazuje porovnání výsledků metod QUEST a C&RT. U obou metod došlo u trénovacích dat ke shodě v 41 případech, což činí 95.35 %, k neshodě došlo ve 2 případech, což činí 4.65 %. U testovacích dat se modely ve všech případech a tudíž nedošlo k neshodě.

Poslední zelený rámeček zobrazuje porovnání shody s cílovým atributem, k němu došlo u trénovacích dat ve 34 případech a nedošlo v 7 případech ze 41 celkových případů. U testovacích dat tato shoda nastala ve 12 případech a nenastala ve 4 případech.

Pro grafické zobrazení byl vybrán uzel Evaluation, který zobrazuje proces učení dat tréninkových a testovacích. Výstup z uzlu Evaluation je na obrázku 31.



Obrázek 31 - evaluační graf pro quest, c&rt. Zdroj [vlastní].

Z obrázku 31 vidíme proces učení dat u tréninkových a testovacích dat pomocí cílového atributu „velikost_mesta“. Modrá čára označuje metodu QUEST a fialová čára metodu C&RT. U trénovacích dat došlo k naučení obou metod na 77 percentilu. U testovacích dat k naučení došlo podstatně dříve a to již na 25 percentilu.

4.3.5 Silné a slabé stránky metod, výsledný stream

Silné stránky:

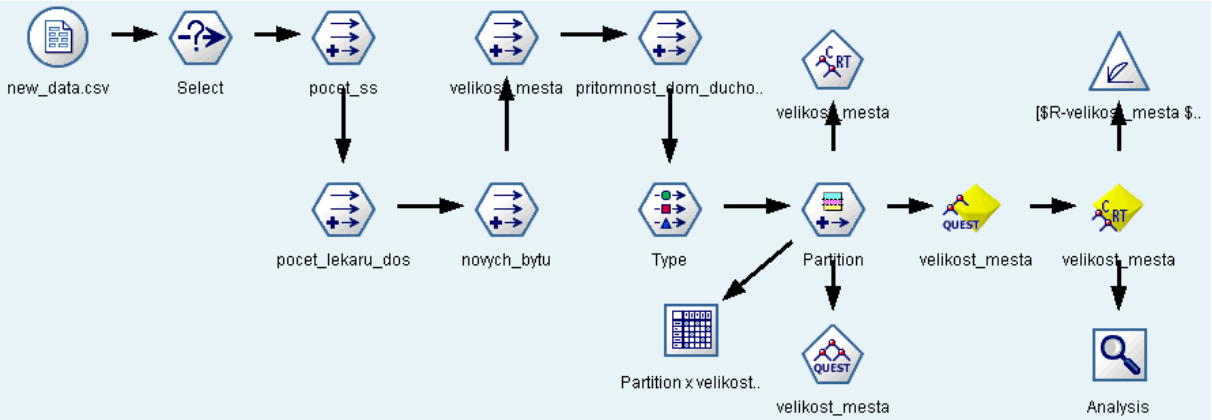
- pomocí stromové struktury snadná interpretace a vyhodnocení jednotlivých vygenerovaných pravidel,

- možnost zobrazení procesu učení jednotlivých modelů.

Slabé stránky:

- použité metody nedosáhly v trénovacích a testovacích datech v porovnání k cílovému atributu 100 % přesnosti.

Výsledný stream vytvořený pro model rozhodovacích stromů je na obrázku 32.



Obrázek 32 - stream rozhodovacích stromů. Zdroj [vlastní].

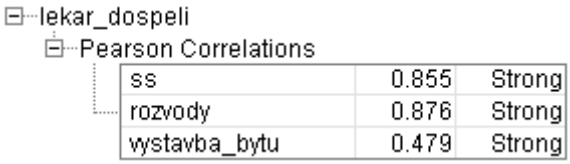
4.4 Vícerozměrná regresní analýza

Vícerozměrným modelem regresní analýzy budou odhadnuty pomocí nezávislých parametrů a hodnot regresních koeficientů závislé atributy. Nejdříve je nutné otestovat pomocí uzlu Statistics korelaci mezi závislým a jednotlivými nezávislými atributy pomocí Pearsonova korelačního koeficientu.

V prvním případě budeme predikovat počet lékařů pro dospělé pomocí nezávislých atributů:

- ekonom_subj,
- rozvody,
- vystavba_bytu.

Obrázek 33 zobrazuje výsledek korelace, všechny testované atributy mají silnou korelaci.



Obrázek 33 - Pearsonův korelační koeficient 1. Zdroj [vlastní].

Model je vytvořen pomocí uzlu Regression, kde je jako cílový atribut zadán atribut „lekar_dospeli“ a jako vstupní atributy výše uvedené nezávislé atributy, je vygenerována rovnice 5, která obsahuje následující parametry:

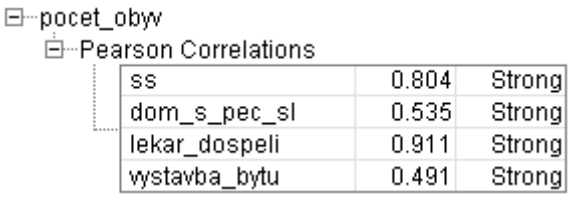
Rovnice 5 - rovnice pro výpočet počtu lékařů pro dospělé. Zdroj [vlastní].

*lekar_dospeli = 0,895 + ekonom_subj*0,0006737 + rozvody*0,0903 - vystavba_bytu*0,00286*

Ve druhém případě budeme predikovat počet obyvatel pomocí nezávislých atributů:

- ss,
- dom_s_pec_sl,
- lekar_dospeli,
- vystavba_bytu.

Obrázek 34 zobrazuje výsledek korelace, všechny testované atributy mají silnou korelaci.



Obrázek 34 - Pearsonův korelační koeficient 2. Zdroj [vlastní].

Model je vytvořen pomocí uzlu Regression, kde je jako cílový atribut zadán atribut „pocet_obyv“ a jako vstupní atributy výše uvedené nezávislé atributy, je vygenerována rovnice 6, která obsahuje následující parametry:

Rovnice 6 - rovnice pro výpočet počtu obyvatel. Zdroj [vlastní].

*pocet_obyv = 115 + lekar_dospeli * 1435,9 + dom_s_pec_sl * 289,0 + ss * 281,6 + vystavba_bytu * 14,05*

4.4.1 Analýza výsledků

V prvním případě se rovnice 5 zabývala výpočtem odhadu počtu lékařů pro dospělé. Výsledek je uveden v tabulce 11, tabulka 11 je tvořena třemi řádky. První řádek „rozdíl“ uvádí, jak velký byl rozdíl výsledku rovnice 5 v porovnání se skutečným počtem lékařů pro dospělé. Jsou zde hodnoty v rozmezí 0 – 5, kde hodnota 0 znamená, že výsledek rovnice 5 byl stejný v porovnání se skutečným počtem lékařů. Naopak hodnota 5 udává, že rovnice 5 vypočítala hodnotu počtu lékařů v rozdílu 5. Druhý řádek „počet“ udává počet případů k danému rozdílu a třetí řádek „%“ je procentuální vyjádření nastalých případů.

Tabulka 11 - hodnoty lékařů. Zdroj [vlastní].

rozdíl	0	1	2	3	4	5
počet	18	24	14	1	1	1
%	30,51	40,68	23,73	1,69	1,69	1,69

Vidíme, že v 18 případech se výsledek shoduje se skutečným počtem lékařů pro dospělé (rozdíl nula), což činí 30,51 %. Ve 24 případech se počet lékařů liší o jednoho, což činí 40,68 %. Ve 14 případech se počet lékařů liší o dva, což činí 23,73 %. Rozdíly počtu lékařů tři, čtyři a pět nastaly shodně v 1 případě, což činí pro každý případ 1,69%.

Ve druhém případě se rovnice 6 zabývala výpočtem odhadu počtu obyvatel. Ani v jednom případě se vypočítaná hodnota nerovná skutečnému stavu obyvatel. Kategorie rozdílů mezi odhadnutými a skutečnými hodnotami počtu obyvatel jsou zobrazeny v tabulce 12.

Tabulka 12 - odhad počtu obyvatel. Zdroj [vlastní].

rozdíl výsledku (%)	0,2-15	16-35	36-50	51-100	>100
počet	22	25	6	4	2

V tabulce je pět kategorií, do kterých byly výsledky zařazeny. První kategorie obsahuje 22 případů, kdy rozdíl dosáhl oproti skutečnému stavu počtu obyvatel 0,2 - 15 %. Druhá

kategorie obsahuje 22 případů s rozdílem 16 - 35 %. Třetí kategorie rozdílu 36 - 50 % obsahuje 6 případů. Čtvrtá kategorie obsahuje 4 případy s rozdílem výsledku 51 - 100 %. Pátá kategorie obsahuje 2 případy, kdy rozdíl dosáhl vyšší hodnoty než 100 %. Kompletní výpočet odhadu počtu obyvatel pro všechny obce je zobrazen v příloze 7.

4.4.2 Silné a slabé stránky modelu, výsledný stream

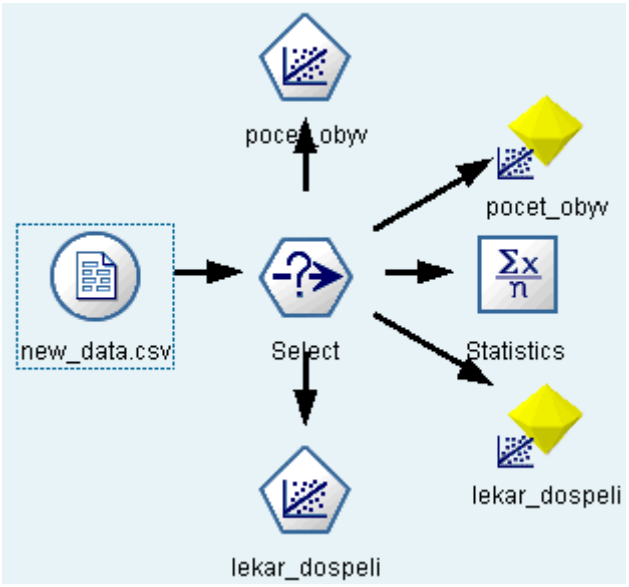
Silné stránky:

- rovnice 5, která se zabývala odhadem počtu lékařů pro dospělé, dokázala v 18 případech se 100 % přesností odhadnout počet lékařů pro dospělé v porovnání se skutečným stavem těchto lékařů v dané obci,
- rovnice 6, zabývající se odhadem počtu obyvatel, dokázala ve dvou případech odhadnout počet obyvatel s 0,2 % rozdílem oproti skutečnému stavu počtu obyvatel, což lze považovat při daných velikostech obcí (myšleno podle počtu obyvatel) za minimální rozdíl.

Slabé stránky:

- zároveň rovnice 5 odhadla ve dvou případech počet lékařů pro dospělé s rozdílem pěti těchto lékařů oproti skutečnému stavu,
- rovnice 6 odhadla počet obyvatel s rozdílem více než 100 % oproti skutečnému stavu ve dvou případech.

Výsledný stream vytvořený pro model regrese je na obrázku 35.



Obrázek 35 - model regrese. Zdroj [vlastní].

Závěr

Cílem této práce bylo využít data-miningové metody pro zpracování dat z oblasti sociální politiky. Pro potřeby vypracování této práce tak byly navrženy modely využívající asociační pravidla, shlukovou analýzu, rozhodovací stromy a vícerozměrnou lineární regresi. Byla využita data z databáze MOS a ze statistických ročenek Českého statistického úřadu pro Královéhradecký a Pardubický kraj. Tato data obsahovala všechna potřebná data a z tohoto důvodu bylo upuštěno od dotazníkového šetření.

V této práci jsou charakterizována vstupní data pro region Královéhradecký a Pardubický a je zde popsána jejich vstupní analýza. Konkrétně byla modelována data za všechny obce s pověřeným obecním úřadem v těchto dvou regionech. Celkem se jednalo o 61 takovýchto obcí. Pro modelování byly vybrány asociační pravidla, shluková analýza, rozhodovací stromy a vícerozměrná regresní analýza.

Asociační pravidla pracovala konkrétně s modely APRIORI a GRI. Jako vstupní atributy pro část pravidla „antecedent“ (předpoklad) byly vybrány atributy, které obsahují počet nových bytů, počet středních škol, přítomnost domovu důchodců a počet lékařů pro dospělé. Pro pole „consequent“ (závěr) byl vybrán atribut velikost města. Všechny výše uvedené atributy musely být nejprve pro použití v daných modelech převedeny z dat spojitých na data kategorická. Model APRIORI vygeneroval 20 různých pravidel a model GRI 8 různých pravidel. Tato jednotlivá pravidla ukázala, že určitá skladba předpokladů v pravidlech ovlivňuje výsledný závěr pravidla, tedy velikost města. Tato kapitola obsahuje i silné a slabé stránky výše zmíněných metod.

Shluková analýza pracovala s modelem K-Means, který používal pro svůj model všechna data spojitá. Pomocí tohoto modelu byly vytvořeny tři shluky. První shluk obsahoval 25 záznamů, druhý shluk 30 záznamů a třetí shluk 4 záznamy. Jednotlivé shluky a k nim příslušné atributy byly zobrazeny pomocí uzlu MATRIX, obdoby kontingenční tabulky, která tvoří přehlednou strukturu jednotlivých shluků a zvolených atributů „novych_bytu“ a „velikost_mesta“. Jsou zde provedeny dva grafické výstupy a definovány silné a slabé stránky metody K-Means.

Rozhodovací stromy pracovaly s kategorickými hodnotami určených atributů, kterými byly stejně jako u asociačních pravidel počet nových bytů, počet středních škol, přítomnost domovu důchodců a počet lékařů pro dospělé, tyto atributy byly zvoleny jako vstupní atributy. Cílový atribut byl zvolen stejně jako u asociačních pravidel atribut velikost města. Byly

vybrány binární rozhodovací stromy QUEST a C&RT, které pracují s kategorickými proměnnými. Model QUEST vytvořil 4 rozhodovací pravidla ve dvou podúrovních daného rozhodovacího stromu. Model C&RT vytvořil 6 rozhodovacích pravidel ve třech podúrovních daného rozhodovacího stromu. Pomocí grafických výstupů jsou tyto rozhodovací stromy zobrazeny a je zobrazen i proces učení trénovacích a testovacích dat. Závěr kapitoly obsahuje definované silné a slabé stránky použitých metod.

Pro potřeby vícerozměrného modelu regresní analýzy byly vytvořeny dva modely. První model predikuje počet obyvatel na základě nezávislých atributů počet lékařů pro dospělé, počet domovů s pečovatelskou službou, počet středních škol, počet výstavby nových bytů pomocí vygenerované regresní rovnice. Ve druhém případě je vygenerována rovnice pro predikci počtu lékařů pro dospělé opět na základě regresní rovnice, kde nezávislými atributy jsou počet ekonomických subjektů, rozvodovost a počet výstavby nových bytů. V závěru kapitoly jsou definovány silné a slabé stránky modelu.

Cílem této práce bylo získání dat z oblasti sociální politiky a tato data pomocí data-miningových metod zpracovat. Data byla získána a pomocí asociačních pravidel, shlukové analýzy, rozhodovacích stromů a vícerozměrné lineární regrese modelována a pro každý model byly definovány jeho silné a slabé stránky. Cíl této práce byl tímto splněn.

Seznam zdrojů

[1] KREBS, V. *Sociální politika*. 4. vyd. Praha : ASPI, 2007. 504s. ISBN 80-7357-276-1.

[2] *Sociální politika* [online].Vysoká škola ekonomická, 2007 [cit. 2011-03-07]. Co je sociální politika?. Dostupné z WWW: <http://ciks.vse.cz/Edice/Macek/socialnipolitika.aspx?autoredirect=asp>.

[3] *Sociální politika* [online]. 2008 [cit. 2011-03-07]. Sociální politika. Dostupné z WWW: <http://www.verejna-politika.cz/index.php?option=com_content&view=article&id=69&Itemid=81>.

[4] *Sociální politika* [online]. 2010 [cit. 2011-03-08]. Dostupné z WWW: <http://granty.vsers.cz/dokument/Socialni%20politika.pdf>.

[5] PETR, *P. Data Mining : Díl I.* Pardubice : Univerzita Pardubice, 2006. 144 s. ISBN 80-7194-886-1.

[6] BERKA, P. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[7] *NEJPOUŽÍVANĚJŠÍ METODOLOGIE* [online]. 2006 [cit. 2011-03-08]. NEJPOUŽÍVANĚJŠÍ METODOLOGIE. Dostupné z WWW: <http://www1.osu.cz/studium/dozna/crispdm.htm>.

[8] *IBM SPSS Modeler Professional - data mining, prediktivní analýzy, prediktivní modelování* [online]. c2010 [cit. 2011-03-10]. IBM SPSS Modeler Professional. Dostupné z WWW: <http://www.spss.cz/ibmspss_modeler.htm>.

[9] BERKA, Petr; RAUCH, Jan; ŠIMŮNEK, Milan. LISp-Miner: systém pro získávání znalostí z dat. In *LISp-Miner: systém pro získávání znalostí z dat* [online]. Praha : Vysoká škola ekonomická, 2007 [cit. 2011-03-10]. Dostupné z WWW: <http://sorry.vse.cz/~berka/docs/4iz450/LISp-Miner_popis.pdf>.

[10] *EAMOS - výukový systém* [online]. c2011 [cit. 2011-03-10]. EAMOS. Dostupné z WWW: <http://eamos.pf.jcu.cz/amos/kat_inf/externi/kat_inf_21586/files/studijni_texty/asociacni_pravidla.pdf>.

[11] RYDZI, Daniel; RAUCH, Jan. Aplikace asociačních pravidel ve společnosti Zinest s.r.o.. *Systémová integrace* [online]. 2008, 4/2008, [cit. 2011-03-10]. Dostupný z WWW: <http://www.cssi.cz/cssi/aplikace-asociacnich-pravidel-ve-spolecnosti-zinest-s-r-o>.

[12] MELOUN, Milan; MILITKÝ, Jiří. *KOMPENDIUM STATISTICKÉHO ZPRACOVÁNÍ DAT : Metody a řešené úlohy včetně CD*. Praha : Academia, 2002. 766 s., ISBN 80-200-1008-4.

[13] *Systémy pro dobývání znalostí z databází* [online]. 2002 [cit. 2011-03-12]. Dobývání znalostí z dat o hypertenzi. Dostupné z WWW: <http://euromise.vse.cz/kdd/index.php?page=kdd#Cl-krok4>.

[14] Metody shlukové analýzy. In *Metody shlukové analýzy* [online]. Praha : Jihočeská Univerzita, Zemědělská fakulta, 2006 [cit. 2011-03-12]. Dostupné z WWW: <http://www2.zf.jcu.cz/public/departments/kmi/MSMT_05/metody%20shlukove%20analyzy.pdf>.

[15] POŠÍK, Pavel. Část IV. Segmentace a shlukování. In *Data mining* [online]. Praha : České Vysoké Učení Technické, Katedra Kybernetiky, 2005 [cit. 2011-03-12]. Dostupné z WWW: <http://cyber.felk.cvut.cz/gerstner/teaching/zbd/DataMining4-hout.pdf>.

[16] KUBANOVÁ, Jana. *Statistické metody pro ekonomickou a technickou praxi*. 3. vydání. Bratislava: STATIS, 2008. 247 s. ISBN 978-80-85659-47-4.

[17] *K-means clustering. Data Points* [online]. 2007 [cit. 2011-03-12]. K-means clustering. Dostupné z WWW: <http://shabal.in/visuals/kmeans/1.html>.

[18] *Rozhodovací stromy* [online]. 2007 [cit. 2011-03-13]. Rozhodovací stromy. Dostupné z WWW: <http://datamining.xf.cz/view.php?cisloclanku=2002102802>.

[19] RYCHLÝ, M. *Klasifikace a predikce* [online]. 2006 [cit. 2011-03-12]. Dostupné z WWW: <http://www.fit.vutbr.cz/~rychly/docs/classification-and-prediction/classification-andprediction.pdf >.

[20] *AnswerTree - klasifikační a rozhodovací stromy* [online]. 2010 [cit. 2011-03-13]. AnswerTree. Dostupné z WWW: <http://www.spss.cz/sw_answertree.htm>.

[21] *Ekonomicky aktivní obyvatelstvo | Manažerský slovník | ELSE AZ s.r.o.* [online]. 2011 [cit. 2011-03-15]. Ekonomicky aktivní obyvatelstvo. Dostupné z WWW: <http://www.elseaz.cz/slovnik/ekonomicky-aktivni-obyvatelstvo/>.

[22] *Domov pro seniory* [online]. 2009 [cit. 2011-03-15]. Domov pro seniory. Dostupné z WWW: <http://www.mu-mohelnice.cz/domovy-duchodcu/domov.html>.

[23] *Úřad práce - wikipedie* [online]. 2006 [cit. 2011-03-15]. Úřad práce. Dostupné z WWW: <http://cs.wikipedia.org/wiki/%C3%9A%C5%99ad_pr%C3%A1ce>.

Přílohy

Příloha 1

Tabulka 13 - úplná základní statistika atributů. Zdroj [vlastní].

Statistics	pocet_obyv	0_14	15_64	65_vice	ekonom_aktivni	uchazeci_o_zam	nezam	ekonom_subj	ms	zs	ss
Count	61	61	61	61	61	61	61	61	61	61	61
Mean	10708.721	2542.148	12239.590	2771.230	7248.262	682.180	9.615	3223.426	3.426	2.492	2.049
Sum	653232	155071.000	746615	169045.000	442144.000	41613.000	586.500	196629.000	209.000	152.000	125.000
Min	1651	462	2347	412	884	73	5.700	361	1	1	0
Max	94493	15521	80705	20374	57913	4370	17.600	28714	29	19	17
Median	6337	1896	8681	2018	4593	423	9.100	1893	2	2	1

Statistics	dom_s_pec_sl	dom_duchodcu	urad_prace	lekar_dospeli	lekar_deti	narozeni	zemreli	snatky	rozvody	vystavba_bytu
Count	61	61	61	61	61	61	61	61	61	61
Mean	1.738	0.590	0.508	5.836	2.770	116.852	110.049	47.246	30.361	51.279
Sum	106	36	31.000	356.000	169.000	7128.000	6713.000	2882.000	1852.000	3128.000
Min	0	0	0	1	0	13	11	5	1	3
Max	7	3	1	48	27	1027	970	436	265	516
Median	2	0	1	4	2	80	70	30	18	30

Příloha 2

Tabulka 14 - úplné kategorie. Zdroj [vlastní].

	pocet_ss	pocet_lekaru_dos	novych_bytu	velikost_mesta	pritomnost_dom_duchodcu
1	nedostatecne	dostatecne	malo	stredni	ano
2	nedostatecne	dostatecne	malo	stredni	ne
3	zadna	dostatecne	malo	stredni	ano
4	zadna	dostatecne	stredne	mirne_vetsi	ano
5	nedostatecne	dostatecne	hodne	stredni	ne
6	dostatecne	dostatecne	hodne	mirne_vetsi	ano
7	zadna	dostatecne	stredne	stredni	ano
8	nedostatecne	dostatecne	dostatecne	mirne_vetsi	ne
9	nedostatecne	dostatecne	dostatecne	stredni	ne
10	dostatecne	dostatecne	malo	mirne_vetsi	ano
11	nedostatecne	dostatecne	malo	stredni	ano
12	nedostatecne	dostatecne	stredne	stredni	ano
13	nedostatecne	dostatecne	stredne	stredni	ano
14	nedostatecne	dostatecne	stredne	mirne_vetsi	ano
15	zadna	nedostatecne	malo	male	ne
16	hodne	hodne	dostatecne	mirne_vetsi	ne
17	zadna	nedostatecne	malo	male	ne
18	zadna	nedostatecne	stredne	male	ne
19	nedostatecne	dostatecne	hodne	mirne_vetsi	ne
20	nedostatecne	dostatecne	malo	male	ne
21	dostatecne	hodne	hodne	mirne_vetsi	ano
22	nedostatecne	nedostatecne	malo	male	ne
23	nedostatecne	nedostatecne	stredne	stredni	ne
24	nedostatecne	dostatecne	malo	stredni	ne
25	dostatecne	dostatecne	dostatecne	mirne_vetsi	ne
26	zadna	nedostatecne	malo	male	ne
27	nedostatecne	nedostatecne	stredne	male	ne
28	dostatecne	dostatecne	stredne	stredni	ne
29	dostatecne	dostatecne	hodne	mirne_vetsi	ano
30	dostatecne	dostatecne	stredne	mirne_vetsi	ano
31	dostatecne	dostatecne	stredne	mirne_vetsi	ano
32	zadna	nedostatecne	malo	male	ne
33	zadna	nedostatecne	malo	male	ano
34	dostatecne	dostatecne	stredne	mirne_vetsi	ano
35	nedostatecne	dostatecne	malo	mirne_vetsi	ne
36	nedostatecne	dostatecne	stredne	stredni	ne
37	zadna	dostatecne	stredne	male	ano
38	zadna	dostatecne	stredne	stredni	ano
39	nedostatecne	dostatecne	hodne	stredni	ano
40	nedostatecne	nedostatecne	stredne	stredni	ano
41	zadna	nedostatecne	malo	male	ne
42	nedostatecne	dostatecne	hodne	mirne_vetsi	ne
43	nedostatecne	dostatecne	stredne	stredni	ano

	pocet_ss	pocet_lekaru_dos	novych_bytu	velikost_mesta	pritomnost_dom_duchodcu
44	zadna	nedostatecne	stredne	male	ne
45	zadna	nedostatecne	malo	male	ne
46	dostatecne	dostatecne	hodne	mirne_vetsi	ne
47	zadna	nedostatecne	stredne	male	ne
48	zadna	nedostatecne	malo	male	ano
49	hodne	hodne	hodne	mirne_vetsi	ano
50	nedostatecne	nedostatecne	stredne	stredni	ano
51	nedostatecne	nedostatecne	stredne	male	ne
52	zadna	dostatecne	hodne	stredni	ano
53	nedostatecne	dostatecne	stredne	stredni	ne
54	dostatecne	dostatecne	stredne	mirne_vetsi	ano
55	zadna	dostatecne	malo	stredni	ne
56	nedostatecne	dostatecne	dostatecne	mirne_vetsi	ano
57	dostatecne	dostatecne	stredne	mirne_vetsi	ano
58	zadna	nedostatecne	stredne	male	ne
59	nedostatecne	dostatecne	stredne	stredni	ne

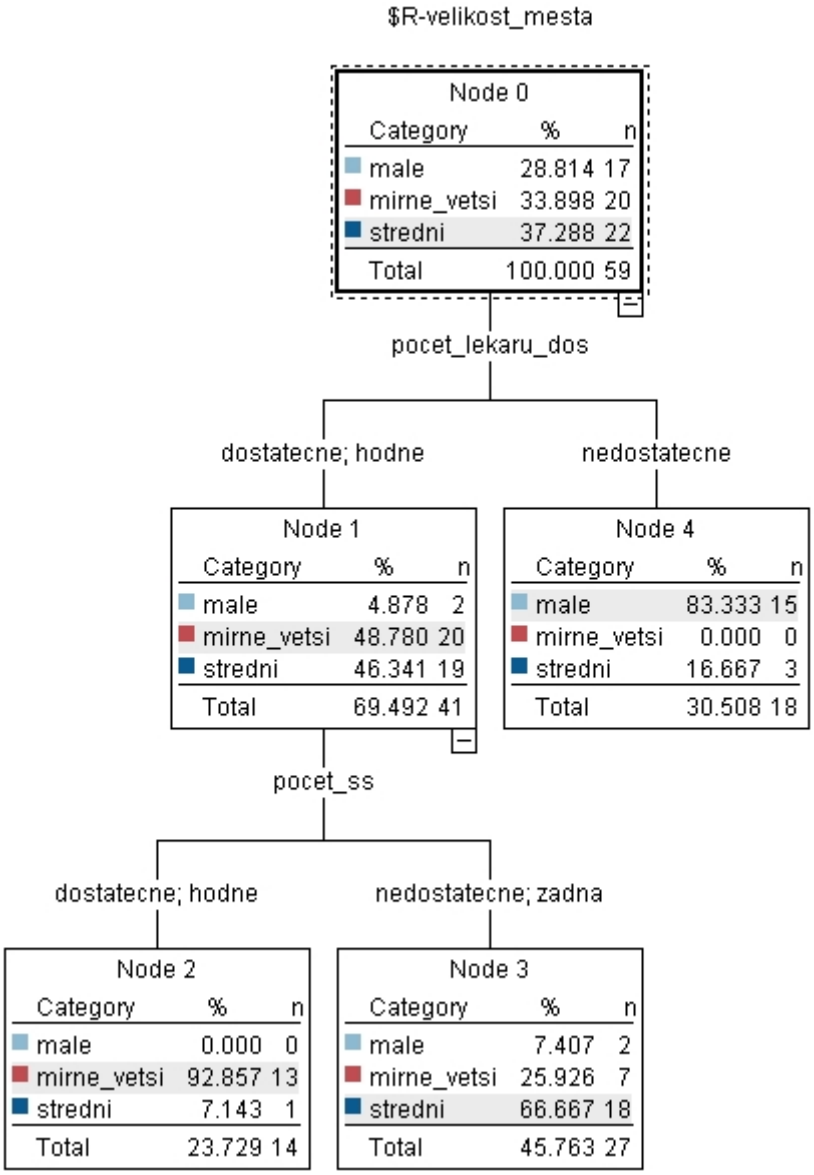
Příloha 3

Tabulka 15 - úplná pravidla apriori. Zdroj [vlastní].

Consequent	Antecedent	Support %	Confidence %
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pritomnost_dom_duchodcu = ano	15,254	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and novych_bytu = malo	15,254	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and pocet_ss = zadna	20,339	100
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pritomnost_dom_duchodcu = ano and pocet_lekaru_dos = dostatecne	13,559	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and novych_bytu = malo and pocet_ss = zadna	13,559	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and novych_bytu = malo and pritomnost_dom_duchodcu = ne	11,864	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and pocet_ss = zadna and pritomnost_dom_duchodcu = ne	16,949	100
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and pritomnost_dom_duchodcu = ne	23,729	92,857
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne	20,339	91,667
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and pocet_lekaru_dos = dostatecne	18,644	90,909
velikost_mesta = male	pocet_ss = zadna and pritomnost_dom_duchodcu = ne	18,644	90,909
velikost_mesta = male	pocet_lekaru_dos = nedostatecne and novych_bytu = stredne and pritomnost_dom_duchodcu = ne	11,864	85,714
velikost_mesta = male	novych_bytu = malo and pocet_ss = zadna and pritomnost_dom_duchodcu = ne	11,864	85,714
velikost_mesta = stredni	novych_bytu = stredne and pocet_ss = nedostatecne and pocet_lekaru_dos = dostatecne	11,864	85,714
velikost_mesta = male	pocet_lekaru_dos = nedostatecne	30,508	83,333
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and novych_bytu = stredne	10,169	83,333
velikost_mesta = mirne_vetsi	pocet_ss = dostatecne and novych_bytu = stredne and pocet_lekaru_dos = dostatecne	10,169	83,333
velikost_mesta = stredni	novych_bytu = stredne and pocet_ss = nedostatecne and pritomnost_dom_duchodcu = ano	10,169	83,333
velikost_mesta = male	novych_bytu = malo and pocet_ss = zadna	16,949	80
velikost_mesta = stredni	pocet_ss = nedostatecne and pritomnost_dom_duchodcu = ano	16,949	80

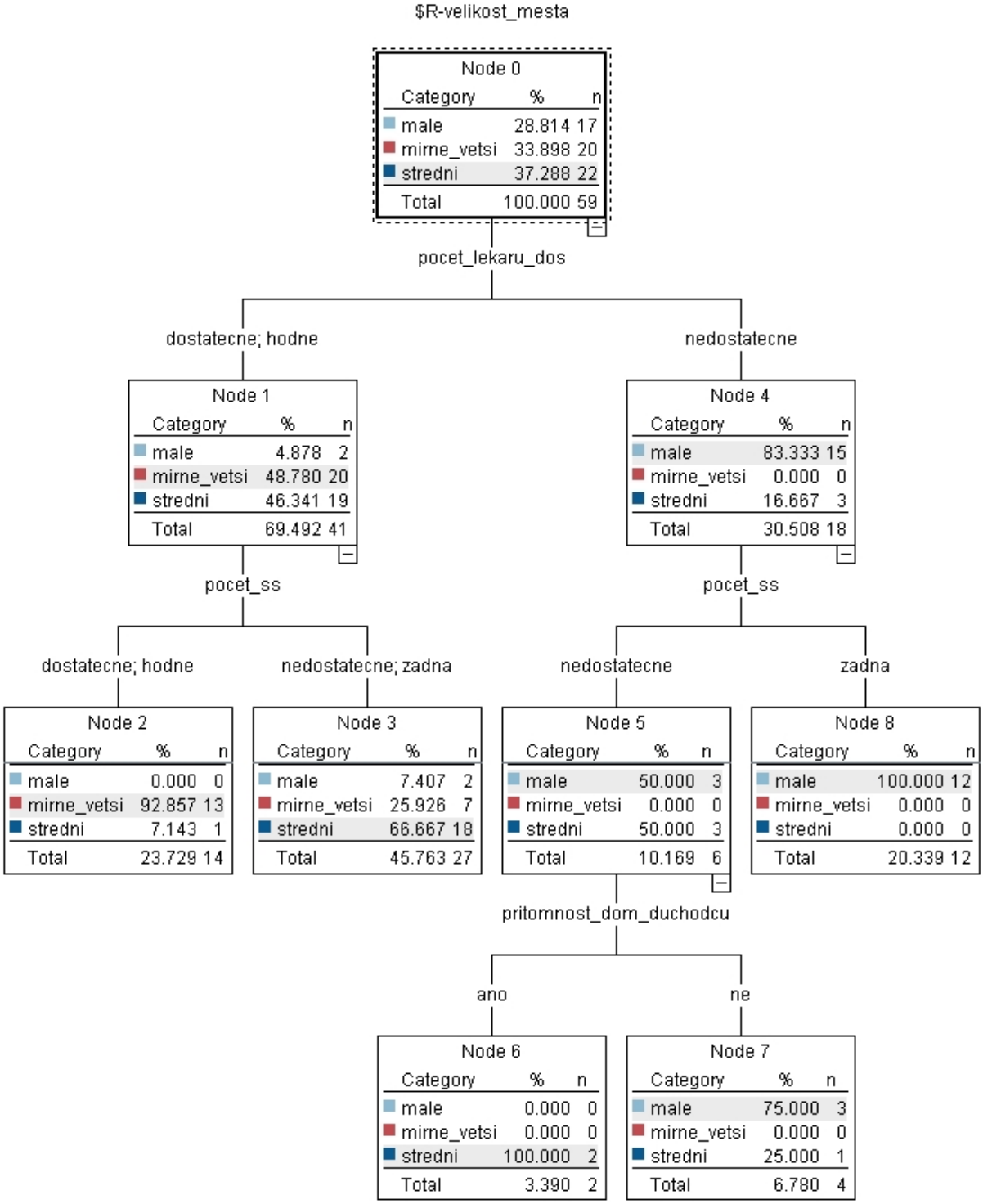
Příloha 4

Obrázek 36 - úplný strom quest. Zdroj [vlastní].



Příloha 5

Obrázek 37 - úplný strom c&rt. Zdroj [vlastní].



Příloha 6

Tabulka 16 - datová matice. Zdroj [vlastní].

spr_obv_obci	poc_ob	0_14	15_64	65_vic	ekon_akt	uch_o_zam	nezam	ekon_sbj	ms	zs	ss	dom_p_sl	dom_duch	ur_prace	lekar_dosp	lekar_deti	narozeni	zemreli	snatky	rozvody	vyst_bytu
Broumov	7977	1808	8225	1686	4207	473	10,7	1481	1	2	1	2	1	1	4	2	93	79	29	16	10
Červ_Kostelec	8534	1485	6697	1762	4160	324	7,7	1774	3	1	1	2	0	0	5	1	106	99	38	20	18
Česká_Skalice	5319	1149	5491	1238	2791	174	6,2	1245	2	1	0	0	1	0	3	2	54	70	32	11	12
Česká_Třebová	16178	2711	13030	2971	9870	1274	12,8	3551	2	1	0	2	2	0	3	2	192	169	71	41	33
Dobruška	6916	1896	8672	2018	3782	248	6,5	1435	2	2	1	0	0	1	5	2	80	59	35	16	50
Dvůr_Kr_n_Lab	16145	3940	19115	4380	8521	796	9,3	3811	3	6	3	2	3	1	7	3	183	192	68	43	62
Heřmanův_Městec	4992	1477	6920	1602	4741	509	10,7	1948	2	1	0	2	2	0	3	2	47	85	29	7	35
Hlinsko	10205	3136	14811	3558	10423	1176	11,2	4500	3	3	1	2	0	1	7	4	92	93	39	26	96
Holice	6498	2711	11965	2469	7745	607	7,8	3591	3	2	2	2	0	1	4	2	80	60	21	22	99
Hořice	9053	2759	13035	3004	4675	424	9	2348	3	3	5	2	1	1	6	4	100	92	31	20	20
Hostinné	4723	1243	5738	1098	2605	227	8,6	928	1	1	2	1	1	0	3	2	42	52	15	12	10
Hradec_Králové	94493	15500	79793	20374	50630	3010	5,7	26407	23	19	17	5	1	1	46	27	1027	970	408	256	395
Hronov	6220	1528	6728	1875	3215	225	6,8	1429	4	1	2	2	2	0	5	2	60	84	25	18	26
Chlumec_nad_Cidl	5401	1312	6321	1471	2707	296	10,5	1052	2	1	1	0	1	1	3	2	80	70	32	11	25
Choceň	9025	2157	9627	2242	6970	750	10,6	2963	4	2	2	1	1	0	4	2	88	109	30	29	39
Chrast	3202	2009	8681	1961	6121	808	13,2	2296	1	1	0	1	0	0	2	1	34	39	15	7	20
Chrudim	23323	5158	25284	5345	17835	1714	9,6	8259	7	6	7	4	0	1	14	5	250	215	105	71	102
Chvaletice	3240	828	4345	940	3161	286	8,9	1285	1	1	0	2	0	0	2	1	29	29	11	8	17
Jablonné_nad_Orl	3272	1187	5388	1121	3704	325	8,6	1510	2	1	0	2	0	0	2	2	32	21	24	10	35
Jaroměř	12770	3070	13415	2881	6499	663	9,9	2715	1	3	2	2	0	1	8	4	154	127	62	46	58
Jevíčko	2891	999	4704	1023	3238	544	16,5	1099	1	1	1	2	0	1	3	2	40	16	8	7	11
Jičín	16646	4301	21980	4755	8855	694	7,8	4481	5	4	4	3	1	1	12	4	169	166	65	45	55
Kopidlno	2240	1021	4931	1070	1093	121	11,1	432	1	1	1	1	0	0	2	1	20	21	8	6	8
Kostelec_nad_Orl	6237	2014	8861	2053	3166	273	8,6	1473	2	2	2	2	0	0	2	2	89	57	34	25	28
Králíky	4576	1341	6416	1323	4593	579	12,5	1893	4	2	1	2	0	1	3	1	46	35	17	8	17
Lanškroun	10196	3601	16200	3116	11243	1253	10,9	4346	6	3	3	3	0	1	5	2	93	89	36	35	123
Lázně_Bělohrad	3727	682	3272	867	1951	170	8,6	832	2	1	0	2	0	0	2	1	37	40	12	9	11
Lázně_Bohdaneč	3392	1544	7331	1610	4941	423	8,4	2175	1	1	1	0	0	0	2	1	36	22	21	10	37
Letohrad	6337	1220	5870	1151	3920	301	7,6	1565	6	3	3	3	0	1	5	2	66	49	30	19	40
Litomyšl	10275	4139	18649	3988	12836	1340	10,3	5397	3	3	5	2	1	1	7	4	95	89	39	28	59
Moravská_Třeb	10910	2971	14448	2993	10584	1892	17,6	3471	2	3	3	0	1	1	7	0	97	135	53	28	25

spr_obv_obci	poc_ob	0_14	15_64	65_vic	ekon_akt	uch_o_zam	nezam	ekon_sbj	ms	zs	ss	dom_p_sl	dom_duch	ur_prace	lekar_dosp	lekar_deti	narozeni	zemreli	snatky	rozvody	vyst_bytu
Náchod	20760	3769	17544	4227	10720	831	7,6	5399	7	3	5	1	1	1	10	5	239	225	94	59	37
Nasavrky	1651	462	2347	643	1522	219	14,4	674	1	1	0	1	0	0	1	1	16	12	7	3	10
Nechanice	2312	752	3648	775	1095	87	7,7	468	1	1	0	0	1	0	1	1	24	23	9	8	15
Nová Paka	9372	1847	9310	2279	4660	407	8,6	2252	2	2	3	2	1	1	4	3	101	101	33	20	26
Nov_Měs_n_Metu	9878	2075	9870	2446	5141	374	7,1	2364	3	3	1	1	0	0	6	3	91	111	44	26	19
Nový_Bydžov	7177	2493	12102	2752	3530	340	9,6	1602	2	2	2	2	0	1	4	3	87	76	34	8	32
Opočno	3128	1169	5131	1166	1481	98	6,6	638	1	1	0	1	1	0	4	2	22	34	19	10	24
Pardubice	90077	15521	80705	19249	57913	4370	7,5	28714	29	15	13	5	1	1	48	15	998	875	436	265	516
Police_nad_Metu	4287	1202	5438	1376	2210	160	7,2	922	1	1	0	1	1	0	3	1	61	55	23	13	35
Polička	8877	2961	13546	3021	9460	848	8,7	3545	5	2	2	1	1	1	5	3	83	97	42	26	71
Přelouč	8751	2398	12898	2967	9205	847	9,1	4087	2	2	1	2	1	1	2	0	83	100	32	34	41
Rokyt_v_Orl_h	2316	524	2390	412	1211	106	8,8	466	1	1	0	1	0	0	1	1	24	18	9	9	7
Rychn_nad_Kněžn	11466	3510	16571	3275	6184	380	6,1	2781	6	3	2	1	0	1	7	4	132	108	47	50	76
Skuteč	5300	1954	9183	2094	6264	760	12,1	2500	3	2	1	1	1	1	3	1	58	50	10	4	21
Smiřice	3050	1210	5772	1118	1609	164	9,5	638	1	1	0	0	0	0	2	1	45	29	10	6	45
Sobotka	2487	621	3083	801	1159	73	6,1	437	1	1	0	1	0	0	2	1	22	17	11	4	15
Svitavy	17067	4754	22445	4567	16202	2288	13,8	5957	7	5	3	2	0	1	8	4	198	176	75	48	53
Svoboda_nad_Úp	2137	609	3538	685	1218	111	9	649	1	1	0	1	0	0	1	1	13	11	8	8	24
Teplice_nad_Metu	1763	698	3836	823	884	92	10,4	361	1	1	0	1	1	0	1	0	19	19	5	1	3
Trutnov	31005	5972	28778	5700	16705	1643	9,6	7556	1	6	7	4	1	1	15	6	363	303	155	108	62
Třebech_pod_O	5848	1236	5278	1104	2715	232	8,4	1269	1	1	1	1	1	0	2	1	66	58	27	15	30
Třemošnice	3160	1058	5471	1285	3865	501	12,9	1413	1	1	1	1	0	1	2	1	32	32	10	5	21
Týniště_nad_Orlicí	6364	1661	8370	2167	3191	239	7,5	1304	2	1	0	1	1	0	3	0	63	50	33	21	48
Úpice	5957	2102	9735	2452	2916	336	11,4	1122	1	2	2	2	0	0	3	2	66	67	30	17	29
Ústí_nad_Orlicí	14565	4025	18539	4067	13583	1532	11	5689	8	3	4	7	2	1	8	4	160	165	62	40	30
Vamberk	4698	1064	5101	1213	2393	239	10	1017	2	1	0	1	0	0	3	1	52	46	23	18	10
Vrchlabí	12710	2836	14210	2970	6873	539	7,8	3434	7	2	1	2	1	1	6	4	148	197	63	32	168
Vysoké_Mýto	12578	2928	12921	2738	9167	1025	11	3952	4	2	4	3	1	1	9	4	153	118	55	39	24
Žacléř	3553	753	3486	853	1683	227	13,3	658	2	1	0	2	0	0	1	1	40	44	16	18	23
Žamberk	6025	2010	9426	1875	6603	646	9,5	3069	2	2	1	1	0	0	5	3	58	63	17	27	37

Příloha 7

Tabulka 17 - celkový odhad počtu obyvatel. Zdroj [vlastní].

spr_obv_obci	pocet_obyv	odhad_pocet_obyv	rozdíl %
Broumov	7977	6859	16,3
Červený_Kostelec	8534	8407	1,5
Česká_Skalice	5319	4591	15,9
Česká_Třebová	16178	5464	196,1
Dobruška	6916	8279	19,7
Dvůr_Králové_nad_Labem	16145	12460	29,6
Heřmanův_Městec	4992	5492	10,0
Hlinsko	10205	12375	21,3
Holice	6498	8391	29,1
Hořice	9053	10997	21,5
Hostinné	4723	5415	14,7
Hronov	6220	8801	41,5
Chlumec_nad_Cidlinou	5401	5055	6,8
Choceň	9025	7259	24,3
Chrast	3202	3557	11,1
Chrudim	23323	24778	6,2
Chvaletice	3240	3804	17,4
Jablonné_nad_Orlicí	3272	4056	24,0
Jaroměř	12770	13558	6,2
Jevíčko	2891	5437	88,1
Jičín	16646	20112	20,8
Kopidlno	2240	3670	63,8
Kostelec_nad_Orlicí	6237	4521	38,0
Králíky	4576	5521	20,7
Lanškroun	10196	10734	5,3
Lázně_Bělohrad	3727	3719	0,2
Lázně_Bohdaneč	3392	3788	11,7
Letohrad	6337	9568	51,0
Litomyšl	10275	12981	26,3
Moravská_Třebová	10910	11362	4,1

spr_obv_obci	pocet_obyv	odhad_pocet_obyv	rozdíl %
Náchod	20760	16691	24,4
Nasavrky	1651	1980	19,9
Nechanice	2312	1762	31,2
Nová Paka	9372	7647	22,6
Nové_Město_nad_Metují	9878	9568	3,2
Nový_Bydžov	7177	7449	3,8
Opočno	3128	6485	107,3
Police_nad_Metují	4287	5203	21,4
Polička	8877	9144	3,0
Přelouč	8751	4422	97,9
Rokytnice_v_Orl_h	2316	1938	19,5
Rychnov_nad_Kněžnou	11466	12086	5,4
Skuteč	5300	5288	0,2
Smiřice	3050	3619	18,7
Sobotka	2487	3486	40,2
Svitavy	17067	13770	23,9
Svoboda_nad_Úpou	2137	2177	1,9
Teplice_nad_Metují	1763	1882	6,7
Trutnov	31005	25652	20,9
Třebechovice_pod_O	5848	3979	47,0
Třemošnice	3160	3852	21,9
Týniště_nad_Orlicí	6364	5386	18,2
Úpice	5957	5971	0,2
Ústí_nad_Orlicí	14565	15173	4,2
Vamberk	4698	4852	3,3
Vrchlabí	12710	11950	6,4
Vysoké_Mýto	12578	15369	22,2
Žacléř	3553	2452	44,9
Žamberk	6025	8385	39,2