Algoritmy v dataminingu

Algorithms in datamining

Mgr. Leona Štablová

Bakalářská práce
2010

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2009/2010

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Mgr. Leona ŠTÁBLOVÁ**
Osobní číslo:	**A06011**
Studijní program:	**B 3902 Inženýrská informatika**
Studijní obor:	**Informační a řídicí technologie**
Téma práce:	**Algoritmy v dataminigu**

Zásady pro vypracování:

1. **Nastudujte základní metody používáných pro analýzu dat.**
2. **Zvolte metody pro vyhodnocení sociálních dat.**
3. **Připravte data a statistické analýzy.**
4. **Vytvořte model.**
5. **Otestujte vytvořený model.**
6. **Srovnejte získané informace s praktickými znalostmi.**

Rozsah bakalářské práce:

Rozsah příloh:

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

Seznam odborné literatury:

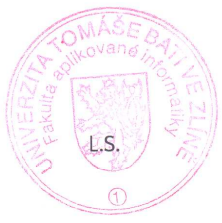
1. **COOPER, Donald R. Business research methods. 10th ed. Boston: McGraw-Hill/Irwin, c2008. 746 s. ISBN 978-0-07-340175-1.**
2. **DUBITZKY, Werner. Fundamentals of data mining in genomics and proteomics. New York: Springer, c2007. 281 s. ISBN 0-387-47508-7.**
3. **KLÍMEK, Petr. Získávání znalostí z podnikových dat (data mining) = Knowledge discovery in company data (data mining) : teze disertační práce. Zlín: Univerzita Tomáše Bati ve Zlíně, 2005. 35 s. ISBN 8073182416.**
4. **RUD, Olivia Parr. Data mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha : Computer Press, 2001. 329 s. ISBN 8072265776.**
5. **CONOLLY, Thomas; BEGG, Carolyn; HOLOWCZAK, Richard. /Mistrovství-Databáze/., Computer Press, 2009. 584 s. ISBN 978-80-251-2328-7.**

Vedoucí bakalářské práce: **Ing. Michal Procházka**
Ústav aplikované informatiky

Datum zadání bakalářské práce: **5. března 2010**

Termín odevzdání bakalářské práce: **1. června 2010**

Ve Zlíně dne 5. března 2010



prof. Ing. Vladimír Vašek, CSc.
děkan

doc. Ing. Ivan Zelinka, Ph.D.
ředitel ústavu

ABSTRAKT

Práce se zabývá zpracováním sociologických dat z Úřadu práce ve Zlíně pomocí dataminingových nástrojů. Cílem je vytvoření modelu a nalezení možných souvislostí mezi jednotlivými atributy.

Klíčová slova: Datamining, Dobývání znalostí z databází,

ABSTRACT

This work is solving the sociological data processing from Labour Office in Zlin, due to datamining´s tools. The goal is to create a model and to find possible relationship between attributes.

Keywords: Datamining, Knowledge Discovery in Database

Děkuji vedoucímu práce Ing. Michalovi Procházkovi za jeho připomínky a pomoc při řešení problémů.

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na bakalářské práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně 1.6.2010

…….
podpis diplomanta

OBSAH

ÚVOD

V současné době každý subjekt, ať to komerční nebo nekomerční shromažďuje velké množství dat ve svých databázích. Tyto data mohou obsahovat významné informace důležité pro další rozhodování v oblasti řízení procesů a zároveň mohou vést k získání konkurenční výhody na trhu. Datamining jako proces aplikace metod pro vyhledávání zajímavých vztahů v datech je v současné době velmi rozvíjející se oblast. Požadavek firem na pravidelné zpracovávání firemních dat roste. Velké množství dat se uchovává a následně zpracovává i v sektoru veřejné správy. Například data úřadů práce jsou statisticky zpracovávány a jsou jedním z prvků při stanovení politiky zaměstnanosti ČR ministerstvem práce a sociálních věcí a následném stanovení způsobu financování této politiky na několik let dopředu. Velké množství dat uchovávají o svých klientech i jednotlivé úřady práce a výsledky jednotlivých statistik jsou součástí hodnocení hospodářské situace dané oblasti. Proto nalezení nových informací v této oblasti by mohlo vést k podpoření stávajících nebo nalezení nových možností v oblasti trhu práce.

Práce je zaměřena na využití datamaningových nástrojů na sociologická data získaná z Úřadu práce ve Zlíně. Cílem této práce je nalezení souvislostí mezi atributem způsob ukončení evidence (nalezl či nenalezl si práci) a ostatními vstupními atributy. Při návrhu vstupů byly brány v úvahu atributy, které mohou ovlivnit uplatnění na trhu práce a které jsou sledovány jako významné statistické ukazatele Ministerstvem práce a sociálních věcí. Pokud by se podařilo popsat pomocí vstupních atributů jednotlivé skupiny (skupina osob, které si dokáží nalézt práci sami, skupina osob, které se dokáží zaměstnat s pomocí úřadu práce a skupina osob, která zůstává v evidenci), bylo by možné lépe a efektivněji postupovat při práci s těmito lidmi.

Data byla zpracována několika algoritmy, aby byly výsledky dostatečně podložené. Graficky znázorněné výsledky jsou součástí této práce. Na závěr této práce je vyhodnocení daných výsledků a diskuse s možnými závěry.

I. TEORETICKÁ ČÁST

1 DATA MINING A JEHO POZICE V KDD

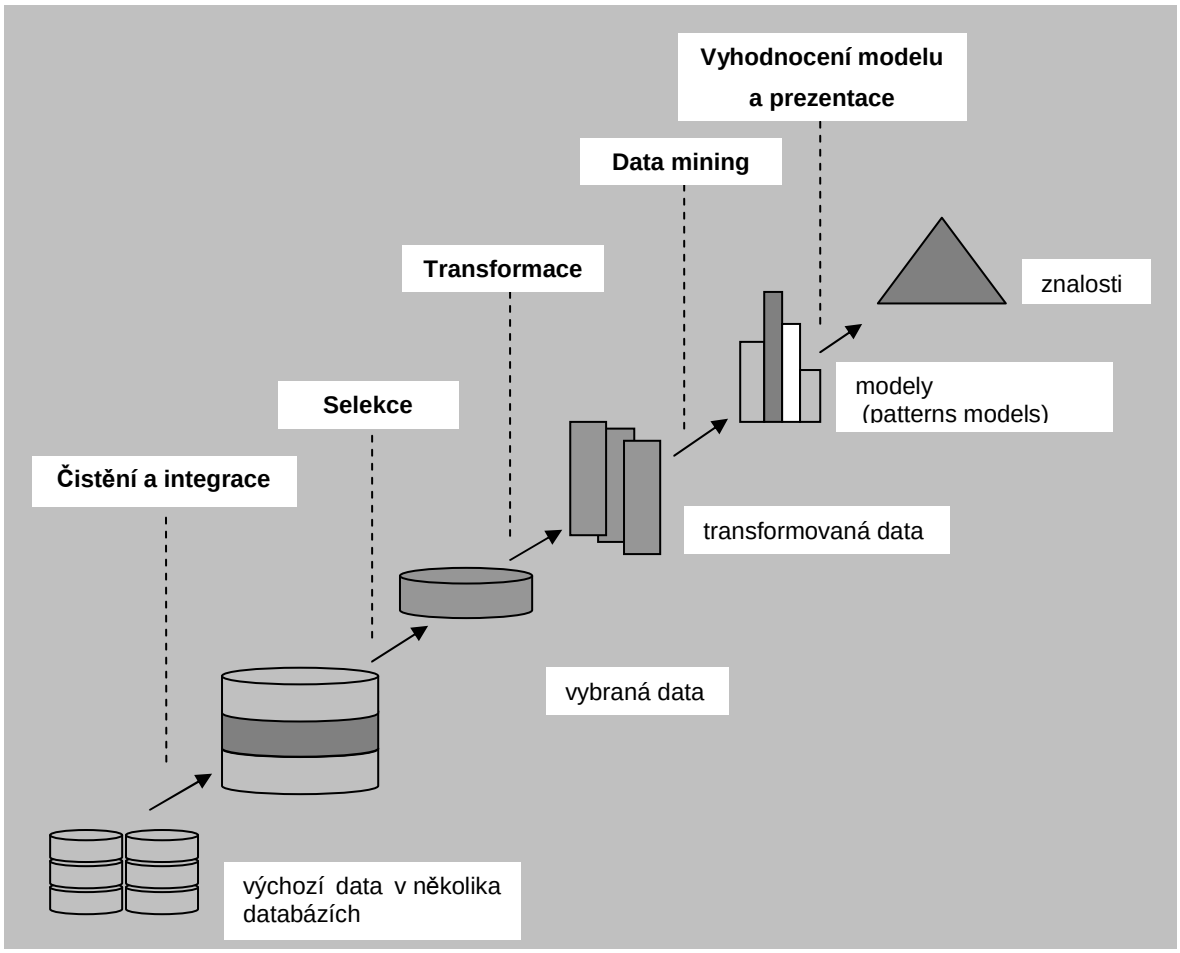
1.1 Získávání znalostí z databází

Velké množství dat, které se dnes uchovávají v databázích přesahuje možnosti člověka analyzovat tyto data bez použití sofistikovaných počítačových technik a získat tak hodnotné informace potřebné pro potenciálního uživatele.

Získávání znalostí z databází Knowledge Discovery in Database (KDD) je obor zabývající se právě problémy, které vznikají při práci s komplexními a objemnými daty. KDD je definováno jako: „Netriviální získávání neznámých a potenciálně užitečných informací z dat“.

KDD je proces, skládající se z několika kroků:

1. Čistění dat (Cleaning) – odstranění šumu a nekonzistentnosti dat, doplnění chybějících dat.
2. Integrace dat (Data integration) – získání a příprava potřebného množství dat z různých zdrojů.
3. Selekce dat (Selection) – výběr relevantních dat pro analýzu.
4. Transformace dat (Transformation) – úprava dat do formy vhodné pro vybrané metody dolování dat (např. aplikace sumarizačních nebo agregačních operací).
5. Datamining – základní proces, kde jsou aplikovány inteligentní metody za účelem vyhledávání zajímavých vztahů v datech.
6. Hodnocení modelu (Evaluation) – identifikování opravdu zajímavých datových modelů z množství výsledků, které byly získány na základě aplikace jednotlivých dataminingových metod.
7. Prezentace získaných znalostí (Presentation) – prezentace vytěžených znalostí uživateli.



Obrázek 1 - Proces dobývání znalostí z databází

1.2 Typy úloh řešené pomocí KDD

Hlavním impulsem pro KDD je stanovení reálného problému a cílem tohoto procesu je získání co nejvíce relevantních informací pro řešení stanoveného problému. V rámci procesu KDD lze řešit několik obecných úloh:

- Klasifikace
- Predikce
- Deskripce
- Hledání „nugetů“ [9]

1.2.1 Klasifikace

Cílem klasifikace je nalezení znalostí, které lze použít pro hodnocení (klasifikaci) nových případů. Vytváří určité skupiny a požaduje, aby získané znalosti co nejvíce odpovídaly konkrétní skupině.

1.2.2 Predikce

Cílem predikce je zjistit budoucí vývoj na základě analýzy předcházejících hodnot. U predikce je velmi důležitým faktorem čas. Jde o postup, kdy na základě analýzy známé množiny vstupních a jim odpovídajících známých výstupních hodnot se hledá nejpravděpodobnější vazba mezi vstupy a výstupy. Nalezením těchto vazeb lze při nových vstupech odvodit pravděpodobnou hodnotu výstupu.

1.2.3 Deskripce

Při deskripci (popisu) je cílem nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech. Požadujeme srozumitelné znalosti pokrývající daný koncept; dáváme tedy přednost menšímu množství méně přesných znalostí.[9]

1.2.4 Hledání nugetů

Hledáme-li nugety, požadujeme zajímavé (nové, překvapivé) znalosti, které nemusí plně pokrývat daný koncept. [9]

2 DATAMINING

Pohled na datamining může být širší a užší. V širším pohledu je DM chápán jako synonymum k procesu KDD. V užším podhledu je datamining chápán jako jeden z kroků procesu KDD (Montreal, 1995).

Datamining, jako krok v procesu KDD zahrnuje výběry a aplikace metod pro vyhledávání zajímavých vztahů v datech. Obvykle jsou jednotlivé metody aplikovány vícekrát. Zpravidla se nejedná o aplikaci pouze jedné metody, ale jednotlivé typy metod se navzájem kombinují na základě dílčích výsledků předcházející metody. Tyto odhalené vztahy mezi daty jsou použity jako důležitý faktor při rozhodování a plánování obchodních postupů.

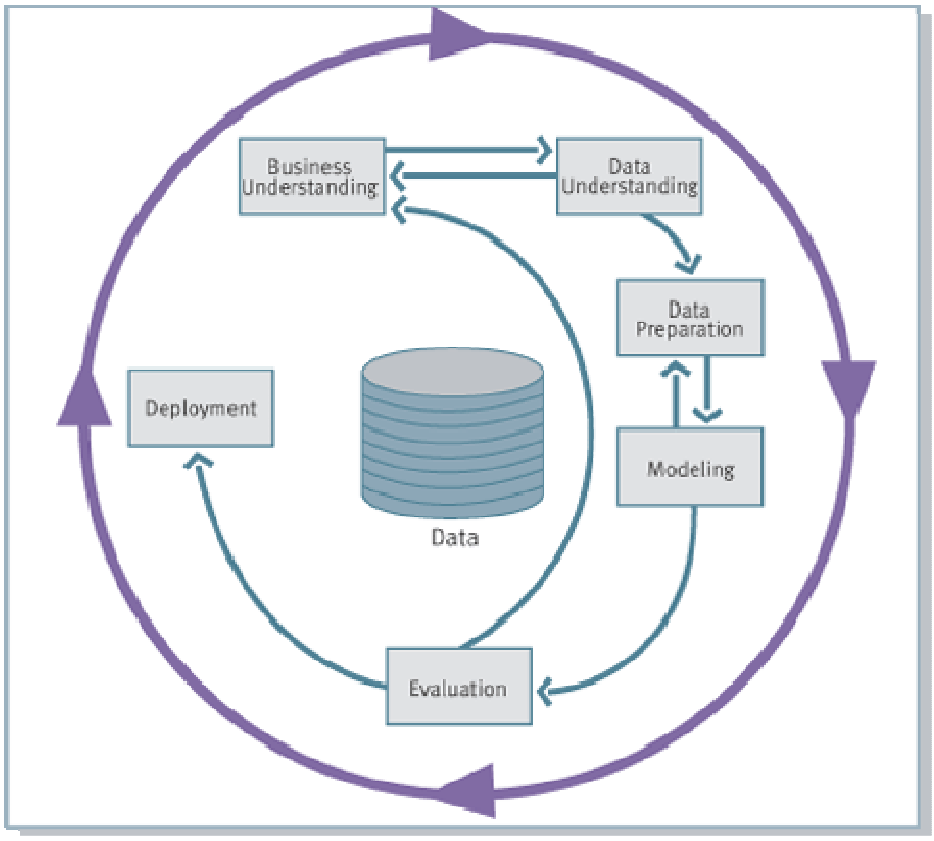
2.1 Postupy v DM

Postupným vývojem tohoto odvětví vznikaly i různé metodiky, které měly za úkol poskytnout uživatelům jednotný rámec pro řešení různých úloh. Tyto metodiky umožnily přenášení zkušeností z jednotlivých úspěšných projektů. Část z nich vznikla v souvislosti s vývojem některých softwarových aplikací (metoda „5A“ firmy SPSS nebo metoda „SEMMA“ firmy SAS). Další část těchto metodik vznikla ve spolupráci výzkumných a komerčních institucí, jako např. CRISP-DM.

2.2 Metodologie CRoss-Industry Standard Process for Data Minig

CRoss-Industry Standard Process for Data Minig (CRIPS-DM) metodologie popisuje životní cyklus projektu těžby dat a zároveň jednotlivé fáze projektu, jejich úkoly, a vztahy mezi těmito úkoly.

Životní cyklus projektu tvoří šest fází. Pořadí jednotlivých fází není pevně dáno. Je závislé na volbě a výsledcích předcházející fáze. Vnější kruh znázorňuje cyklickou povahu procesu.



Obrázek 2 - Fáze CRISP-DM modelu

[6]

2.2.1 Porozumění problematice, definování cílů (Business Understanding)

Jde o vstupní, ale velmi důležitou část tohoto procesu. Tato fáze je zaměřena na definování cílů projektu a požadavků z obchodního hlediska. Úkolem je zjistit požadavky klienta. V této fázi by měly být známy požadované vstupy, a to časové, finanční, hmotné, lidské (obchodní experti, specialisté na KDD, DM odborníci) a datové zdroje (přístup k datovým skladům atd.). Zde dochází také ke stanovení dataminingových cílů a kritérií, pro hodnocení výsledků dataminingu. Na závěr této fáze by měl být sestaven plán projektu, ve kterém je popsán celý proces dosažení cílů DM a popis jednotlivých kroků atd.

2.2.2 Porozumění datům (Data Understanding)

Fáze porozumění datům začíná prvotním sběrem dat a pokračuje činnostmi, které směřují k seznámení se s daty: identifikace problémů kvality dat, vytipování různých podmnožin záznamů v databázi, zjištění různých charakteristik dat (např. četnost různých hodnot, průměrné hodnoty, maxima, formát dat atd.).

2.2.3 Příprava dat (Data preparation)

Tato fáze zahrnuje všechny činnosti, které směřují k vytvoření finálnímu datovému souboru, na který budou aplikovány analytické metody [2.2.4]. Proto je považována za jednu z nejsložitějších v procesu DM.

Nejdříve je třeba rozhodnout, která data budou použita pro analýzu. Tento výběr je prováděn dle stanovených cílů DM [2.2.1].V této rozhodovací části může vzniknout požadavek na omezení objemu dat, nebo typů dat.

Dále je třeba upravit data dle požadavků vybraných analytických technik. Některé techniky vyžadují určitou kvalitu dat. V tomto případě je zapotřebí, aby vybraná data prošla tzv. čištěním.

Informace potřebné pro analýzu, mohou být uloženy v několika databázích, případně tabulkách. Většina datamingových nástrojů s nimi neumí pracovat. Proto je potřeba sloučit data do jedné tabulky, případně provést na daných datech agregace (výpočet nové hodnoty) nebo sumarizace. Při úpravách vstupních dat, je vždy třeba zvažovat, zda-li tyto úpravy nebudou mít vliv na výsledky analýz.

Poslední částí přípravy dat je naformátování dat, dle požadavků zvoleného algoritmu. Některé algoritmy požadují atributy v určitém pořadí, někdy jsou vyžadovány i změny pořadí záznamů, jiné mohou, však mohou vyžadovat, aby záznamy byly uspořádány náhodně (neuronové sítě). Další úpravy, které se provádí v této části jsou například syntaktické zněny (odstranění čárek a háčků z textu, vymazání různých oddělovačů), případně převedení např. textových hodnoty na hodnoty numerické.

2.2.4 Modelování (Modeling)

V této fázi je třeba si vybrat vhodný algoritmus, který bude dále použit pro analýzu. Z důvodu ověření výsledků je vhodné zvolit více algoritmů.

Před vytvořením modelu, je třeba stanovit, jakým způsobem budeme testovat kvalitu a správnost daného modelu. Datovou sadu rozdělujeme do dvou částí, učící se a testovací sadu dat. Model se vytváří na učící se části a kvalita modelu se určuje na testovací části dat (např. procentuálním vyjádřením chybných klasifikacích u klasifikačního algoritmu).

Během modelování jsou vytvářeny jeden nebo více modelů. Jak už bylo zmíněno v předchozím bodu, u některých algoritmů záleží na pořadí vstupů. Jednotlivé modely

mohou vznikat testováním optimálního pořadí nebo různých dalších změn v nastavovaných hodnotách.

Posledním krokem této fáze je ohodnocení modelů. Modely jsou ohodnoceny převážně podle kritérií přesnosti, které byly definovány v první fázi.

2.2.5 Vyhodnocení (Evaluation)

Jde o vyhodnocení, zda-li daný model splňuje obchodní cíle, které byly stanoveny v první fázi. Jednou z možností, jak vyhodnotit platnost a přesnost modelu je užitím modelu v reálném životě. Toto je však velmi náročné na čas.

V této fázi je důležitým krokem provedení revize použitých dat a celého postupu. Dále je nutné zajištění potřebných kvalitních a správných dat pro budoucí analýzy.

Na základě vyhodnocených výsledků, se rozhodne, zdali se přejde do závěrečné fáze „Implementace“, nebo zdali budou některé kroky opakovány, nebo bude proveden celý DM projekt znovu.

2.2.6 Implementace (Deployment)

Jde o vyvození strategií a implementací pro obchodní činnost, na základě výsledků vyhodnocení. Součástí zavádění modelů do praxe, by mělo být i stanovení kontrolních procesů a pro udržitelnost tohoto modelu v praxi.

2.3 DM úlohy

Pomocí Dataminingu je možné řešit velké množství úloh. Dle povahy je možné vymezit 4 základní skupiny úloh, na které je poté možné aplikovat většinu definovaných problémů:

2.3.1 Klasifikace

Podstatou klasifikačních úloh lze vstupní data s danými atributy, do výstupních, předem zvolených tříd. Typickou klasifikační úlohou je například stanovení zda-li je možné zákazníku poskytnout úvěr, na základně vstupních atributů: věk, příjem, oblast ve které bydlí, aj..

2.3.2 Regrese

Regresní model předpovídá převážně budoucí hodnotu na základě dosavadních zkušeností. Jde o statistickou metodu, která popisuje důležitost vstupních proměnných vzhledem k výstupu. Tyto metody mají schopnost odhadovat chyby modelu nebo možnost hledat závislosti na různých kombinacích vstupních proměnných. Například odhad ceny domu, vzhledem k lokalitě, velikosti domu, atd.

2.3.3 Analýza vztahů

Při analýze vztahů jsou pomocí asociačního algoritmu získávána pravidla, tj. implikace typu IF *fakt* - THEN *fakt*. Typickým příkladem analýzy vztahů je analýza nákupního košíku. Jde o analýzu produktů s cílem zjistit, které produkty se prodávají společně. Hlavním cílem je nalezení asociačních pravidel (jestliže si zákazník koupí produkt X, potom s 70% pravděpodobností koupí zároveň i produkt Y).

2.3.4 Segmentace (shlukování)

Jde o nejstarší nástroj DM. Podstatou je rozdělení objektů do skupin (shluků). Tyto skupiny nejsou předem stanoveny. Objekty uvnitř shluku jsou si podobné, rozdíly mezi jednotlivými shluky však musí být maximální. Vzhledem k tomu, že jednotlivé shluky nejsou předem známy, ne vždy se podaří zjistit jejich význam. Každý nový objekt je pak zařazen do některého shluku podle charakteristických vztahů, či vlastností, které daných shluk definují.

2.3.5 Predikce

Predikce je metoda, která na základě známých vstupních hodnot a jim odpovídajících výstupních hodnot, hledá nejpravděpodobnější hodnotu nového výstupu, pro danou kombinaci vstupních hodnot. Typickým příkladem je například vytvoření prediktivního modelu zákazníků banky, kdy na základě předcházejících zkušeností stanovíme prediktivní model dlužníka. Na základě tohoto modelu je každému nové příchozímu žadateli o úvěr vyhodnocena míra rizika dlužníka.

2.3.6 Detekce odchylek

Umožňuje nalézt velmi neobvyklé jevy, např. odchylky chování zákazníka od chování ostatních. Používají se při odhalování podvodů aj.

3 ALGORITMY

Podle jednotlivých úloh se stanovují metody řešení.

Tab. 1- rozdělení algoritmů DM dle úloh [7]

Úlohy	**Algoritmy**
Klasifikace	Diskriminační analýza Logistická regresní analýza Klasifikační (rozhodovací) stromy Neuronové sítě
Regrese	Lineární regresní analýza Nelineární regresní analýza Neuronové sítě
Analýza vztahů	Asociační algoritmus
Segmentace (shlukování)	Clustering (shluková analýza) Genetické algoritmy Neuronové shlukování (Kohenenovy mapy)
Predikce	Lineární regresní analýza Nelineární regresní analýza Neuronové sítě (RBF -- "radial basis function")
Detekce odchylek	Vizualizace Statistické postupy

3.1 Clustering (shluková analýza)

Shluková analýza je tvořena řadou metod. Hlavním cílem těchto metod je roztřídění *n* objektů, z nichž je každý popsán *p* rozměrným vektorem pozorování, do několika homogenních shluků. Přesný počet shluků většinou stanovuje uživatel. Další vlastností jednotlivých shluků je maximální podobnost objektů uvnitř shluku a co nejmenší podobnost objektů mezi různými shluky.

3.1.1 Podobnost objektů

Podobnost objektů se posuzuje podle různých kritérií. Pro intervalové proměnné se nejčastěji používá tzv. euklidovská vzdálenost. Tato vzdálenost je vypočtena pro všechny dvojce objektů a vytvořena matice vzdálenosti.

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{pmatrix} \quad (1)$$

Platí:

d(1,2)>0 (je nezáporné číslo);

d(i,j) je vzdálenost mezi objekty i a j.

d(i,j)=d(j,i) a zároveň d(i,i)=0;

$d(i,j) \leq d(i,h)+d(h,j)$

3.1.2 Hierarchické shlukování

Principem této procedury je postupné shlukování objektů. V prvních krocích se shlukují objekty, které mají nejmenší vzdálenost, v dalších krocích se do shluku začleňují objekty s větší vzdáleností.

Hierarchické metody:

- Aglomerativní – na počátku je každý objekt shlukem a postupně se sdružují
- Divisivní – na počátku jsou všechny objekty jeden shluk a postupně se rozkládají

Aglomerativní algoritmus

1. krok: Každý objekt je považován za samostatný shluk.
2. krok: Výpočet matice vzdáleností objektů. (1)
3. krok: Na základě porovnání vzdáleností jednotlivých shluků, nalezneme shluky s minimální vzdáleností v dané hladině.
4. krok: Tyto dva shluky spojíme do nového, shluku hierarchicky vyššího.
5. krok: Přepočítáme matici vzdáleností. Její řád se sníží o 1. Opakujeme kroky 3 - 5.

3.1.3 Nehierarchické metody

- K-means – každý shluk je reprezentován střední hodnotou objektů ve shluku

- K-medoids – každý shluk je reprezentován jedním z objektů ležícím blízko středu shluku
- Neuronové sítě

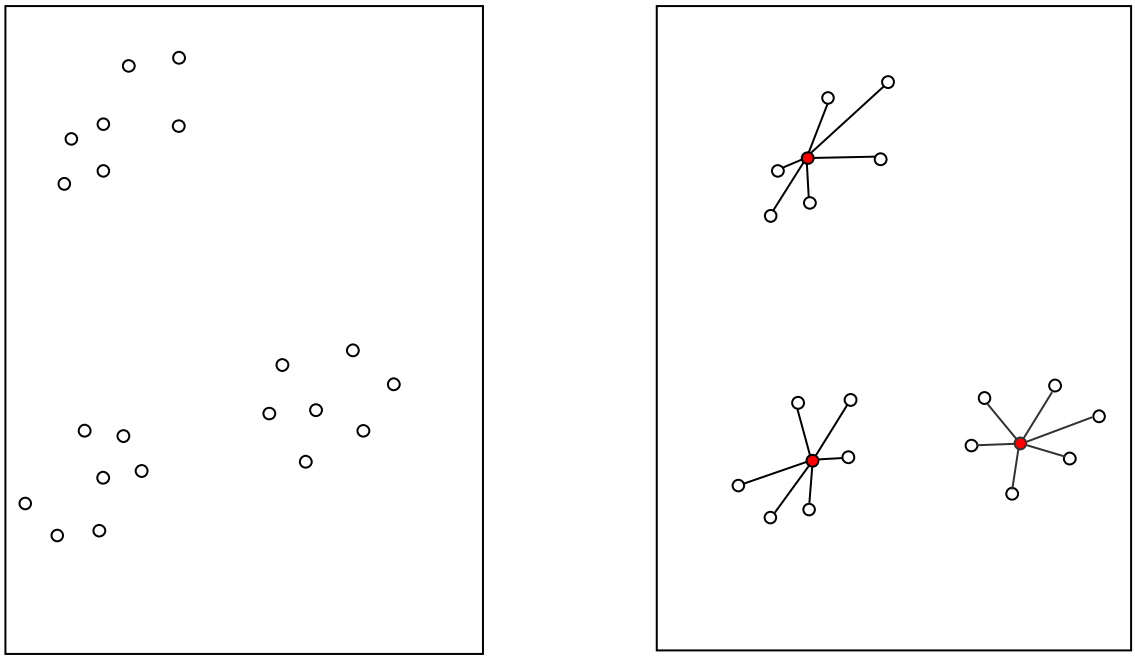
Vzdálenosti mezi objekty lze počítat jako:

1. vzdálenost nejbližšího souseda (vyhledávají se objekty s nejmenší vzdáleností)
2. vzdálenost nejvzdálenějšího souseda (vyhledávají se objekty s největší vzdáleností)
3. vzdálenost metodou průměrné vazby (vyhledávají se objekty, jejichž vzdálenost je průměrem všech vzdáleností mezi objekty).
4. vzdálenost mezi reprezentativními prvky (medoidy)
5. vzdálenost mezi středy shluků

K-means algoritmus

Jde o jeden z nejjednodušších algoritmů shlukování. Hlavním principem tohoto algoritmu je definovat tzv. centroidy, pro každý klastr jeden. Tyto centroidy musí být umístěny co nejdále od sebe. V dalším kroku se každý objekt spojuje s nejbližším centroidem.

1. krok: Rozdělení počáteční množiny *n* objektů do *k* shluků.
2. krok: Určení centroidů v aktuálních shlucích.
3. krok: Přiřazení každého bodu k nejbližšímu centroidu a jemu odpovídajícímu shluku.
4. krok: Vypočtení těžiště každého shluku.
5. krok: Definování nových centroidů ve vypočítaných těžištích.
6. krok: Pokud došlo ke změně v přiřazení bodů shlukům, opakují se od body 2-6.



Obrázek 3 - 3 shluky, K-středová metoda, 3 centroidy

3.2 Rozhodovací stromy

Metoda rozhodovacích stromů patří k nejznámějším algoritmům z oblasti metod symbolických metod strojového učení. Při tvorbě rozhodovacích stromů se postupuje metodou „rozděl a panuj“. Trénovací data se postupně rozdělují do menších a menších podmnožin (uzlů) tak, aby v těchto podmnožinách převládaly příklady jedné třídy. Na začátku tvoří trénovací data jednu množinu, na konci zůstanou podmnožiny tvořené z téže třídy. Postupuje se stromem shora dolů. Začínáme jedním uzlem (kořen stromu). Cílem je nalézt strom s trénovacími daty.

Je několik algoritmů pro tvorbu rozhodovacích stromů. Jedním z nich je i TDIDT algoritmus.

TDIDT algoritmus

1. krok: Zvolení jednoho atributu jako kořen dílčího stromu.
2. krok: Rozdělení dat v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a předání uzlu pro každou podmnožinu.

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

Nevýhodou rozhodovacích stromů je požadavek na data nezatížená šumem. [11]

3.3 Asociační pravidla

Asociační pravidla jsou vytvářena na základě pravidla IF-THEN. Vzhledem k tomu, že používání tento vztah používáme i v běžném životě, staly se tato pravidla prvními, spolu s rozhodovacími stromy které se začaly využívat při získávání znalostí z databází prostřednictvím strojového učení.

3.3.1 Základní charakteristiky pravidel

U asociačních pravidel nás zajímá, kolik příkladů splňuje předpoklad a kolik závěr pravidla, kolik příkladů splňuje předpoklad i závěr současně, kolik příkladů splňuje předpoklad a nesplňuje závěr. Cílem je tedy vytvoření pravidla tvaru:

$$\text{Ant} \Rightarrow \text{Suc},$$

kde Ant (předpoklad, levá strana pravidla, antecedent) i Suc (závěr, pravá strana pravidla, sukcedent).

Základními charakteristikami asociačních pravidel v podpora (support) a spolehlivost (confidence). Podpora je (absolutní resp. relativní4) počet objektů, splňujících předpoklad i závěr. Spolehlivost je vlastně podmíněná pravděpodobností závěru pokud platí předpoklad.

Základem všech algoritmů pro hledání asociačních pravidel je generování kombinací (konjunkcí) hodnot atributů. Při generování vlastně procházíme (prohledáváme) prostor všech přípustných konjunkcí. Metod je několik:

- Do šířky
- Do hloubky
- Heuristicky

Generování *do šířky* - kombinace se generují tak, že se nejprve vygenerují všechny kombinace délky jedna, pak všechny kombinace délky dvě, atd. Jde tedy o generování kombinací *podle délek*

Generování *do hloubky* - vyjde se od první kombinace délky jedna a ta se pak prodlužuje (vždy o první kategorii dalšího atributu) dokud to lze[10]. Nelze-li kombinaci prodloužit, změní se kategorie „posledního“ atributu. Nelze-li provést ani to (vyčerpaly se kategorie posledního atributu), kombinace se zkrátí a současně se změní poslední kategorie.

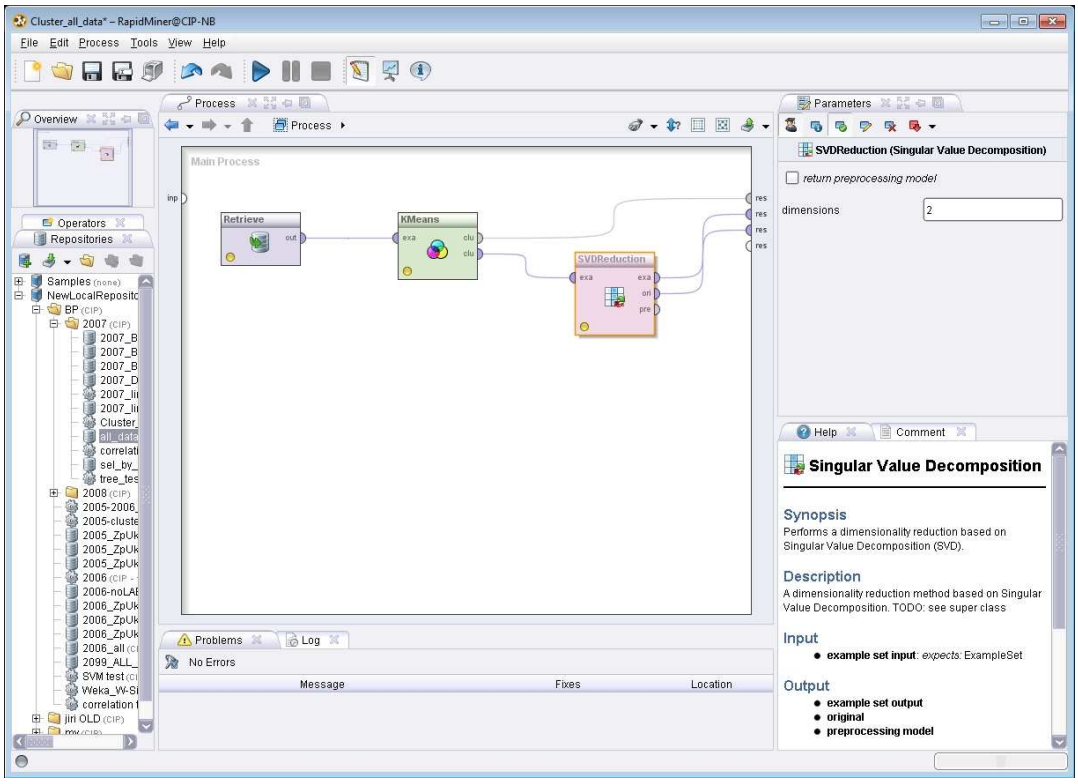
Genenerování *podle četností* - vytváří kombinace v pořadí podle jejich výskytu v datech. Jedná se o příklad heuristického prohledávání prostoru kombinací, kde heuristikou je „uvažuj kombinaci s nejvyšší četností". Při tomto způsobu generování se kombinace s nulovou četnosti objeví až na konci seznamu. [12]

4 POUŽÍTÝ SOFTWARE

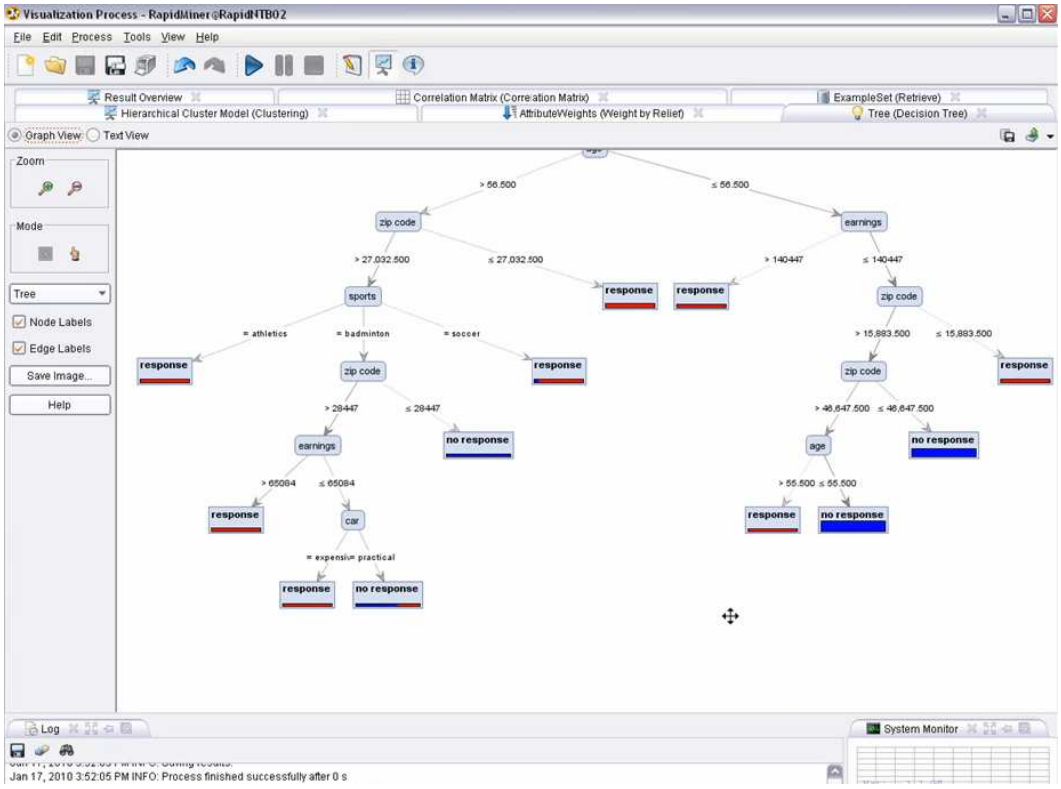
Systémů, které se zabývají programovým zpracováním dat v současné době roste. Touto oblastí vývoje se zabývají jak soukromé firmy se svými komerčními programy, tak i akademické volně šířené.

4.1 Rapid Miner 5

Rapid Miner nabízí softwerové řešení a služby v oblasti prediktivní analýzy dat a data miningu. Tento nástroj je zaměřen na sofistikovanou analýzu velkého objemu dat jako jsou databázové systémy, nestrukturovaná data a texty. Rapid Miner nabízí mnoho nástrojů na zpracování dataminingového modelu a pro jeho vyhodnocení. Následně nabízí také nástroje pro vizualizaci dat, modelů a dalších výsledků. Další významnou oblastí tohoto programu je i hodnocení a odhadování výkonnosti.



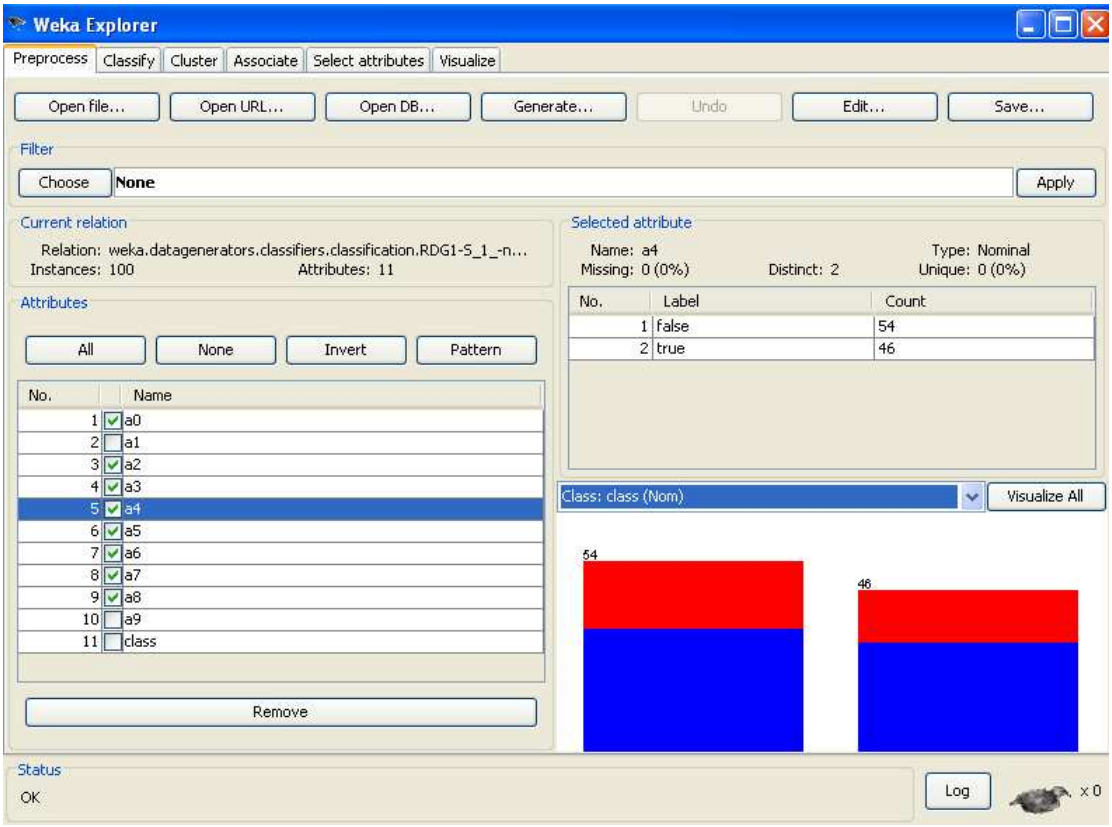
Obrázek 4 – modelování procesu



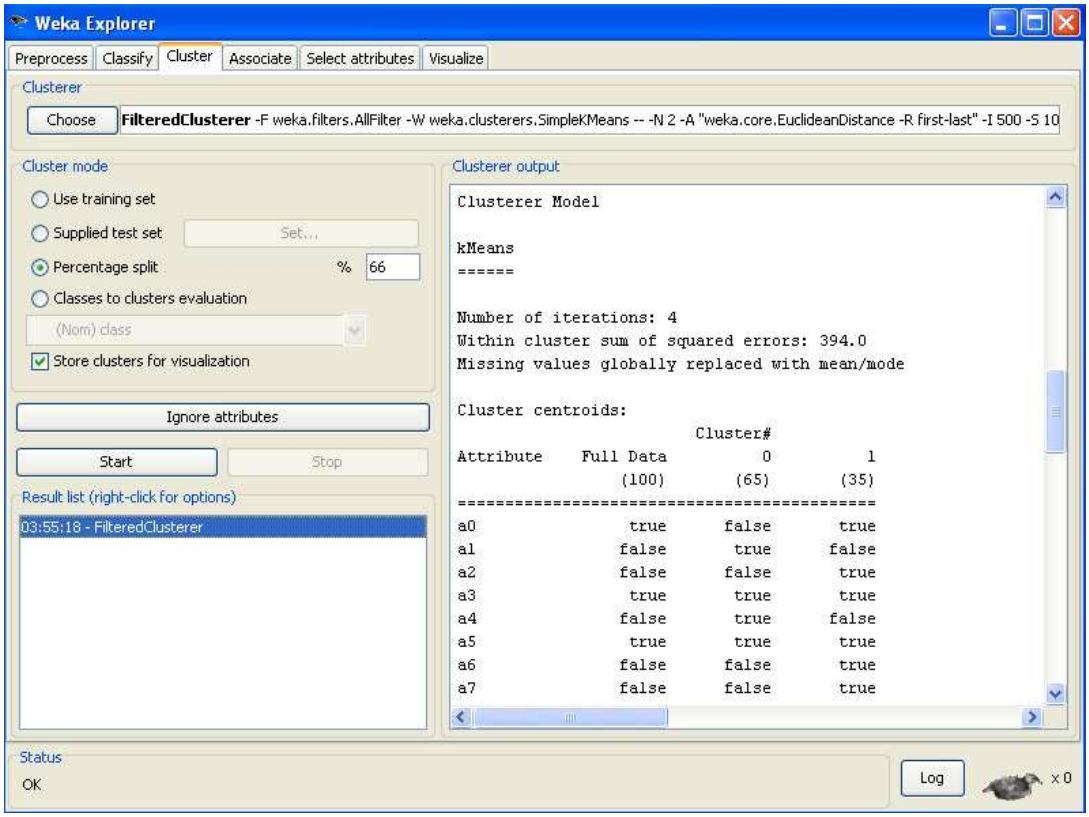
Obrázek 5 – vizualizace dat

4.2 Weka

Weka je volně šířitelný program vyvinutý na univerzitě Waikato na Novém Zélandě. Tento systém pracuje na principu knihoven programů v Javě. Weka nabízí mnoho algoritmů.



Obrázek 6 – aplikace explorer



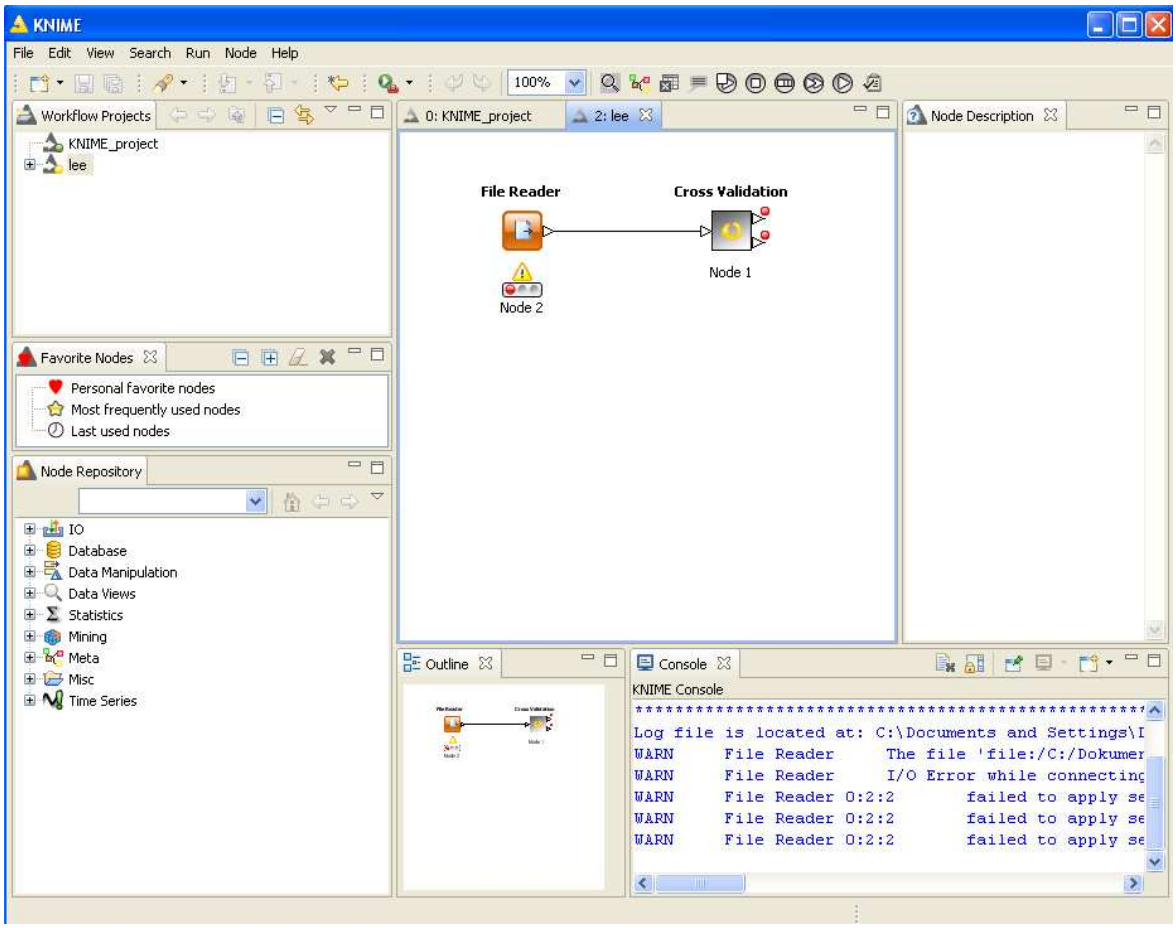
Obrázek *7 – textový výstup z metody k-Means*

Je zde sice nabídka na vizualizaci, ale většina výstupů zůstává v textové podobě.

4.3 KNIME

KNIMe byl vytvořen v Německu na Univerzitě Konstanz. Obsahuje množství metod pro analýzu dat s využití metod datamanin. Základní verze obsahuje mnoho variant pro zpracování dat I/O, předzpracování a čištění, modelování, analýzy a datamining. Integruje všechny analytické moduly prostředí WEKA a další pluginy.

KNIME je založena na platformě Eclipse. KNIME je distribuován pod dvěmi licenčními systémy. uvolněn pod duální licencí systému. The open source license (GPL) umožňuje KNIME ke stažení, distribuovány a používány volně.



Obrázek 8 – pracovní prostředí programu KNIME

II. PRAKTICKÁ ČÁST

5 POROZUMĚNÍ PROBLEMATICE, DEFINOVÁNÍ CÍLŮ

Téma nezaměstnanosti v dnešní společnosti je velmi aktuální. Jedním z mnoha důvodů je i fakt, že na osoby, které jsou v evidenci úřadů práce jsou vynakládány nemalé prostředky, které směřují na zvýšení schopnosti uplatnit se na trhu práce.

Vzhledem k charakteru trhu práce existují skupiny osob, které si umí nalézt práci bez pomoci ÚP a skupiny osob, které mají nějaký handicap, který jim ztěžuje možnost nalézt uplatnění na trhu práce a udržet si dlouhodobě zaměstnání. Faktory, které ovlivňují možnost uplatnění na trhu práce jsou například věk, zdravotní stav, zda-li se osoba stará o jinou závislou osobu (např. malé dítě). Pro tyto klienty úřadů práce se stát snaží jejich handicap snížit různými nástroji „Aktivní politiky zaměstnanosti“.

5.1.1 Cíle obecné

Cílem práce je nalezení charakteristických skupin osob evidence na Úřadu práce (ÚP) ve Zlíně, vzhledem k jejich schopnostem uplatnit se na trhu práce. Stanovit charakteristickou skupinu osob, které jsou schopni nalézt si práci samy, skupinu osob které je možné zaměstnat s pomocí ÚP a skupinu osob, které jsou velmi obtížně zaměstnatelné a jsou ohrožené dlouhodobou nezaměstnaností.

5.1.2 Cíle DM

Rozdělení datového souboru do jednotlivých skupin pomocí clusterizačního algoritmu (*k - means algoritmus).* Vstupními hodnotami bude vzdělání, věk, délka evidence, oblast, ke které přísluší, zda-li je osoba po mateřské dovolené, zda-li měla alespoň jednu rekvalifikaci, zdravotní omezení a výstupní hodnota bude způsob ukončení evidence.

Na základě těchto vstupních dat budou vytvořeny skupiny, podle kterých bude možné jednotlivé nové subjekty zařazovat.

Výsledky clusterizačního algoritmu budou porovnány se statistikami ÚP.

5.2 PRÍPRAVA DAT

Výsledkem této fáze je vytvoření finálního datového souboru s vybranými vstupy, vyčištěnými a naformátovanými daty. Tento proces zahrnuje:

5.3 Předběžné stanovení potřebných vstupů

Na základě zkušeností a výsledků statistik pravidelně vedených ÚP v ČR byly pro nadefinovány vstupy: vzdělání, věk, zdravotní omezení, způsob ukončení evidence, délka evidence na ÚP, pohlaví, oblast.

Dalším vstupem je vstup rekvalifikace. Rekvalifikace směřuje k lepšímu uplatnění na trhu práce na základě získání nových, případně doplnění stávajících znalostí. Do jaké míry má tento faktor vliv na schopnost zvýšit své uplatnění na trhu práce zatím není statisticky podloženo.

Další vstupe je označen jako „Dítě“. Jde o osoby, jejichž předchozí činnost byla péče o osobu do 4 let.Na základě statistických údajů, jde o skupinu znevýhodněnou na trhu práce.

5.4 Získání dat z databáze ÚP ve Zlíně

Tyto data byla získána z centrální databáze používané pro potřeby jednotlivých úřadů práce. Tato data budou muset být získána z různých výstupních sestav a poté pomocí jednoznačných identifikátorů spojena do finální tabulky. Dále je potřeba odstranit z tabulky všechny nepotřebné údaje, případně data spadající do ochrany osobních údajů.

5.4.1 Čištění dat

V jednotlivých datových souborech byly provedeny tyto změny:

Délka evidence – pokud obsahovala 1 den byl tento záznam smazán. Jde o případy buď osoby, která se zaevidovala na ÚP a nesplnila podmínky pro evidenci, nebo osoba, která se zaevidovala na ÚP a poté žádala převod evidence do jiného města. Tento počet osob byl v procentuálním vyjádření zanedbatelný.

Zdravotní stav – vzhledem ke změně legislativy v roce 2010, se v záznamech vyskytuje dvojí označení. Dle dřívějšího značení jsou Osoby zdravotně postižené (OZP), Osoby zdravotně postižené s částečně invalidním důchodem (OZP-ČID) a Osoby zdravotně postižené s plným invalidním důchodem (OZP-PID). Dle nového značení se jedná o osoby I., II. nebo III. stupně invalidity. U některých osob se objevila hodnota „nezadáno“. Tato hodnota se vyskytuje u osob, které nepředložily rozhodnutí o uznání OZP. Tato hodnota je považována za totožnou s hodnotou „bez zdravotního omezení“.

Pro použití shlukové analýzy byly jednotlivá data převedena do číselných hodnot, dle převodních tabulek. [6.1]

6 DATA

Data pro praktickou část byla získána z databáze Úřadu práce ve Zlíně. Každý řádek je záznam jednoho člověka který se zaevidoval na ÚP v letech 2005 - 2009. Jde o nově zaevidované osoby v daném roce. Z důvodu ochrany osobních údajů, nebyly použity žádná jména, rodná čísla, adresy ani žádné citlivé údaje. Všechny osobní údaje byly odstraněny z datového souboru pracovníkem ÚP. Z dat použitých v této práci není možné jednoznačně identifikovat konkrétní osobu.

6.1 Porozumění datům

Pro tuto práci byla použita data která ovlivňují možnost člověka uspět na trhu práce - věk, vzdělání, zdravotní omezení, město, ve kterém žije, zda-li prošel rekvalifikačním kurzem, zda-li si nalezl práci sám, nebo s pomocí ÚP nebo zatím neuspěl na trhu práce. Posledním podstatným důležitým údajem, jak se ukázalo během práce byl i rok evidence na ÚP.

Pro rok 2005 - 2009 jsem zjišťovala četnosti jednotlivých vstupů. U jednotlivých vstupů jsem označila atributy s největší četností.

Tab. 2 - zjišťování procentuální zastoupení jednotlivých vstupů pro rok 2005- 2009

Vstup	Hodnoty	2005	2006	2007	2008	2009
Vzdělání	Základní	13,92%	13,56%	14,63%	16,64%	16,62%
	Střední odborné	41,57%	41,74%	38,32%	36,41%	43,84%
	ÚSO s maturitou	33,43%	33,00%	35,42%	32,64%	29,27%
	VOŠ	1,84%	1,32%	1,17%	1,54%	1,09%
	Vysokoškolské	9,24%	10,18%	10,25%	12,68%	9,18%
Způsob ukončení evidence	Zůstává v evidenci	2,00%	3,18%	5,61%	15,95%	49,95%
	Našel si práci sám	69,17%	82,31%	55,70%	50,14%	29,56%
	Umístěn s pomocí ÚP	15,56%	0,42%	24,91%	21,11%	13,96%
	Ostatní	13,26%	14,10%	13,78%	10,99%	6,53%
Délka evidence	Do 6 měsíců	56,60%	61,48%	65,08%	59,11%	65,37%
	Do 12 měsíců	21,91%	20,01%	18,04%	18,32%	30,27%
	Nad 12 měsícků	21,17%	18,52%	16,88%	22,58%	4,36%
Oblast	Otrokovicko	17,68%	17,18%	17,30%	19,21%	18,76%
	Zlínsko	63,74%	63,41%	62,57%	61,88%	59,59%
	Slavičínsko	6,72%	6,68%	8,40%	7,33%	8,03%
	Valašské Klobouky	10,90%	11,97%	10,93%	10,70%	12,22%
	Ostatní	0,95%	0,77%	0,78%	0,87%	1,37%
Zdravotní stav	Bez omezení	92,05%	91,16%	90,15%	90,89%	93,28%
	OZP	7,95%	8,84%	9,85%	9,11%	6,72%
Pohlaví	Muž	46,67%	43,99%	43,12%	45,82%	54,43%
	Žena	53,33%	56,01%	56,88%	54,18%	45,57%
Absolvent	Ne	88,01%	88,33%	88,21%	88,43%	91,52%
	Ano	11,99%	11,67%	11,79%	11,57%	8,48%

Rekvalifikace	Ne	93,08%	90,00%	92,32%	93,80%	96,88%
	Ano	6,92%	10,00%	7,68%	6,20%	3,12%
Dítě	Ne	99,25%	99,12%	99,32%	99,62%	99,85%
	Ano	0,75%	0,88%	0,68%	0,38%	0,15%

6.1.1 Způsob ukončení evidence

Tento atribut nabývá hodnot 0 – 3.

Tab. 3 - tabulka pro vstup „Způsob ukončení evidence“

Hodnota	Způsob ukončení evidence na ÚP
0	Zůstává v evidenci
1	Umístěný jinak (našel si práci sám)
	Neposkytování součinnosti při zprostředkování zaměstnání.
	Na vlastní žádost (§ 29/b)
	Zahájil SVČ bez příspěvku ÚP
2	Umístěn na dotované místo v rámci projektu ESF
	Umístěn v rámci projektu ESF
	Umístěný ÚP - nástup do zaměstnání
	Umístěný ESF - VPP (OP LZZ)
	Umístěný ÚP na chráněné pracovní místo
	Umístěný ÚP na chráněné pracovní místo – SVČ
	Umístěný ÚP na chráněnou pracovní dílnu
	Umístěný ÚP na SÚPM – SVČ
	Umístěný ÚP na SÚPM vyhrazené - příspěvek na mzdu
	Umístěný ÚP na SÚPM zřízené - příspěvek na zřízení
	Umístěný ÚP na VPP
3	Nástup na soustavnou přípravu na povolání
	Nástup výkonu trestu odnětí svobody (§ 29/c)
	Úmrtí (§ 29/d)
	Sankční vyřazení

6.1.2 Věk

Věk je jedním z důležitých atributů ovlivňující možnost získání zaměstnání na trhu práce. Na základě statistik, které ÚP pravidelně vede jsou skupiny ohrožené dlouhodobou nezaměstnaností osoby do 20 let věku (absolventi škol), osoby nad 55 let věku u mužů a 50 let u žen. Atribut „věk“ byl rozdělen do pěti skupin tak, aby věkové skupiny byly rovnoměrné rozděleny a zároveň aby obsahovaly skupiny ohrožené zvýšenou nezaměstnaností. Poslední pátá skupina je ovlivněna rozdílným odchodem mužů a žen do starobního důchodu.

Tento atribut nebývá hodnot 1 – 5.

Tab. 4 - tabulka pro vstup „Věk“

Hodnota	Věk
1	18-25 let
2	26-35 let
3	36-45 let
4	46-55 let
5	56-64 let

6.1.3 Zdravotní stav

Zdravotní stav je jeden z faktorů, který ovlivňuje možnost uplatnění na trhu práce. Ve zdroji dat je uvedeno, zda-li má osoba rozhodnutí o přiznání částečného nebo plného invalidního důchodu, nebo zda-li toto rozhodnutí nemá (osoba je bez zdravotního omezení nebo má jiné zdravotní omezení).

Tento atribut nebývá hodnot 0 – 1.

Tab. 5 - tabulka pro vstup „Zdravotní stav“

Hodnota	Zdravotní stav
0	Bez zdravotního omezení
	Jiné zdravotní omezení
1	OZP
	Invalidita

6.1.4 Vzdělání

Vzdělání je do jisté míry také omezující faktor při hledání zaměstnání. Největší množství osob evidovaných ÚP jsou osoby se středním vzděláním.

Tento atribut nebývá hodnot 0 – 4.

Tab. 6 - tabulka pro vstup „Vzdělání“

Hodnota	Kód	Vzdělání
0	A	Bez vzdělání
	B	Neúplné základní
	C	Základní + praktická škola
	D	Nižší střední
	E	Nižší střední odborné
1	H	Střední odborné (vyučen)
	J	Střední nebo střední odborné bez maturity a bez vyučení
2	K	ÚSV
	L	ÚSO (vyučení s maturitou)
	M	ÚSO s maturitou (bez vyučení)
3	N	Vyšší odborné
4	R	Bakalářské

	T	Vysokoškolské
	V	Doktorské (vědecká výchova)

Pro další dělení dat např. podle vystudovaného oboru není možné získat data v potřebné kvalitě.

6.1.5 Délka evidence

Tyto data jsou důležitá ve vztahu, k atributu způsob ukončení evidence. Tento atribut nabývá hodnot 1-3.

Tab. 7 - tabulka pro vstup „Délka evidence"

Hodnota	Délka evidence	Evidence dle počtu dnů
1	do 6 měsíců	0 - 183 dnů
2	do 12 měsíců	184 - 365 dnů
3	nad 12 měsíců	více jak 365 dnů

Například pro rok 2005 byla u osob, které si našly práci sami struktura délky evidence následující:

Tab. 8 - tabulka pro vstup „Délka evidence" v roce 2005, u osob, které si našly práci samy

Hodnoty	Délka evidence	Procentuální vyjádření
1	do 6 měsíců	60,42%
2	do 12 měsíců	21,58%
3	nad 12 měsíců	18,00%

Z daných údajů je možné usuzovat, že osoba s evidencí do 6 měsíců si pravděpodobně najde práci sama.

U osob, které se nově zaevidovaly v roce 2009 je hodnota „Délka evidence nad 6 měsíců“ a „Délka evidence nad 12 měsíců“ zkreslená. Nelze vyhodnotit evidenci nad 12 měsíců u osob, které se zaevidovaly na ÚP od dubna 2009 a evidenci nad 6 měsíců u osob, které se zaevidovaly na ÚP v prosinci 2009.

6.1.6 Pohlaví

Tento atribut nabývá těchto hodnot 0-1.

Tab. 9 – tabulka pro vstup „Pohlaví“

Hodnota	Pohlaví
0	Muž
1	Žena

6.1.7 Oblast

Tento atribut se dělí do 3 oblastí v rámci Zlínského kraje, vzhledem k místu bydliště a oblasti, kde si hledá zaměstnání. Tento atribut nabývá těchto hodnot 0 - 4:

Tab. 10 - tabulka pro vstup „Oblast“

Hodnota	Oblast
0	Otrokovice
1	Zlín
2	Slavičín
3	Valašské Klobouky
4	Jiné

Jednotlivé oblasti se shodují s oblastí působností ÚP ve Zlíně a jeho dislokovaných pracovišť [PŘÍLOHA P1] Skupina „4“ zahrnuje osoby, které mají např. bydliště na území jiného okresu, nebo jiné republiky.

6.1.8 Absolvent

Tento atribut udává, zdali osoba před nástupem na ÚP udává jako svou předcházející činnost „Příprava na povolání“. V případě že se jedná o absolventa má index „1“ v opačném případě má index „0“.

6.1.9 Rekvalifikace

Tento atribut má hodnotu „0“ v případě, že neabsolvoval žádnou rekvalifikaci, a hodnotu „1“ v případě, že absolvoval jednu nebo i více rekvalifikací. Zda-li tento atribut má vliv na získání zaměstnání není zatím statisticky podložené.

6.1.10 Dítě

Tento atribut udává, zdali se jedná o osobu, která jako poslední činnost vykonávala péči o dítě do 4 let věku s indexem „1“, v ostatních případech je uveden index „0“. To znamená, že tato osoba se zaevidovala na ÚP přímo po mateřské dovolené. Vzhledem k tomu, že tyto osoby pečují o osobu do 15 let věku mají omezené možnosti při hledání zaměstnání převážně vzhledem k možnosti pracovat na směny nebo k možnosti vzdálenějšího dojíždění do zaměstnání. Vzhledem však k malému zastoupení v použitém datovém souboru (téměř 1%) se domnívám že, význam tohoto atributu bude pro daný výsledek zanedbatelný.

7 MODELOVÁNÍ

Prvním bodem modelování je stanovení algoritmu. Pro splnění cílů byl vybrán clusterizační algoritmus a regresní algoritmus. Data byla testována pomocí programu Rapid Miner.

Nejdříve byla sestavena korelační matice, za účelem zjištění závislosti jednotlivých atributů.

Attributes	VzdelKod	ZdravStav	DelkaEv	absolvent2...	rekval2007	Dite	Gender	Age
VzdelKod	1	-0.133	-0.104	0.175	0.049	0.022	0.063	-0.182
ZdravStav	-0.133	1	0.310	-0.103	-0.016	-0.023	-0.034	0.297
DelkaEv	-0.104	0.310	1	-0.160	0.212	0.073	0.130	0.310
absolvent20	0.175	-0.103	-0.160	1	-0.056	-0.029	-0.032	-0.396
rekval2007	0.049	-0.016	0.212	-0.056	1	0.275	0.019	0.016
Dite	0.022	-0.023	0.073	-0.029	0.275	1	0.057	0.004
Gender	0.063	-0.034	0.130	-0.032	0.019	0.057	1	0.012
Age	-0.182	0.297	0.310	-0.396	0.016	0.004	0.012	1

Attributes	VzdelKod	ZdravStav	absolvent2...	Gender	Age
VzdelKod	1	-0.133	0.175	0.063	-0.182
ZdravStav	-0.133	1	-0.103	-0.034	0.297
absolvent20	0.175	-0.103	1	-0.032	-0.396
Gender	0.063	-0.034	-0.032	1	0.012
Age	-0.182	0.297	-0.396	0.012	1

Obrázek 9 – *korelační matice na upravených datech*

Dle obr. 4 se hodnoty korelačního koeficientu oscilují kolem 0, s maximální hodnotou 0,3 což značí, že zde není mezi jednotlivými atributy téměř žádná lineární závislost (korelační koeficient se nerovná 1), která by se dala změřit pomocí této metody.

Pro korelační koeficienty platí, že jsou platné pouze v rozmezí daném použitými daty proto byla struktura dat změněna. Byly použity data bez úprav (kap. Data), aby bylo ověřeno, že nedošlo k chybě při úpravách dat. Výsledná korelační matice, však ukázala podobné výsledky jako přecházející matice.

Attributes	VzdělKód	ZdravStav	ZpusUkonč	colDélkaEv	absolvent	rekval	dítě	POHLAVÍ	VĚK	ID_oblasti
VzdělKód	1	-0.136	-0.101	-0.081	0.187	0.082	0.022	0.036	-0.152	-0.023
ZdravStav	-0.136	1	-0.008	0.330	-0.103	-0.023	-0.022	-0.009	0.268	-0.008
ZpusUkonč	-0.101	-0.008	1	-0.245	0.107	-0.053	-0.027	-0.154	-0.102	-0.070
colDélkaEv	-0.081	0.330	-0.245	1	-0.139	0.156	0.074	0.144	0.256	0.059
absolvent	0.187	-0.103	0.107	-0.139	1	-0.006	-0.034	-0.031	-0.413	-0.005
rekval	0.082	-0.023	-0.053	0.156	-0.006	1	0.283	0.068	-0.003	0.009
dítě	0.022	-0.022	-0.027	0.074	-0.034	0.283	1	0.083	-0.014	-0.012
POHLAVÍ	0.036	-0.009	-0.154	0.144	-0.031	0.068	0.083	1	-0.014	-0.016
VĚK	-0.152	0.268	-0.102	0.256	-0.413	-0.003	-0.014	-0.014	1	-0.023
ID_oblasti	-0.023	-0.008	-0.070	0.059	-0.005	0.009	-0.012	-0.016	-0.023	1

Obrázek 10 – *korelační matice bez upravených dat*

Name	Type	Statistics	Range	Missings
ZpusUkonc	integer	avg = 5.98294 +/- 5.99376	[1.00000 ; 19.00000]	0
VzdelKod	integer	avg = 6.66278 +/- 2.90136	[0.00000 ; 13.00000]	0
ZdravStav	integer	avg = 0.08269 +/- 0.27544	[0.00000 ; 1.00000]	0
DelkaEv	integer	avg = 163.23424 +/- 165.24851	[0.00000 ; 1092.00000]	0
absolvent2007	integer	avg = 0.12241 +/- 0.32779	[0.00000 ; 1.00000]	0
rekval2007	integer	avg = 0.07427 +/- 0.26223	[0.00000 ; 1.00000]	0
Dite	integer	avg = 0.00604 +/- 0.07752	[0.00000 ; 1.00000]	0
Gender	integer	avg = 0.56498 +/- 0.49581	[0.00000 ; 1.00000]	0
Age	integer	avg = 35.03886 +/- 12.56107	[16.00000 ; 67.00000]	0

Obrázek 11 - Statistika získaná preprocesingem

Z výsledků je patrno že příliš velký rozsah u střední hodnoty linearity ukazuje, že použití lineární regrese na těchto datech není možná.

Výsledné hodnoty lineární regrese aplikované na daná data, ukazují, že nebyla nalezena závislost mezi výstupem (způsob ukončení evidence) a jednotlivými vstupy.

Role	Name	Type	Statistics	Range	Missings
label	ZpusUkonc	integer	avg = 5.983 +/- 5.994	[1.000 ; 19.000]	0
regular	VzdelKod	integer	avg = 6.663 +/- 2.901	[0.000 ; 13.000]	0
regular	ZdravStav	integer	avg = 0.083 +/- 0.275	[0.000 ; 1.000]	0
regular	DelkaEv	integer	avg = 163.234 +/- 165.249	[0.000 ; 1092.000]	0
regular	absolvent2007	integer	avg = 0.122 +/- 0.328	[0.000 ; 1.000]	0
regular	rekval2007	integer	avg = 0.074 +/- 0.262	[0.000 ; 1.000]	0
regular	Dite	integer	avg = 0.006 +/- 0.078	[0.000 ; 1.000]	0
regular	Gender	integer	avg = 0.565 +/- 0.496	[0.000 ; 1.000]	0
regular	Age	integer	avg = 35.039 +/- 12.561	[16.000 ; 67.000]	0

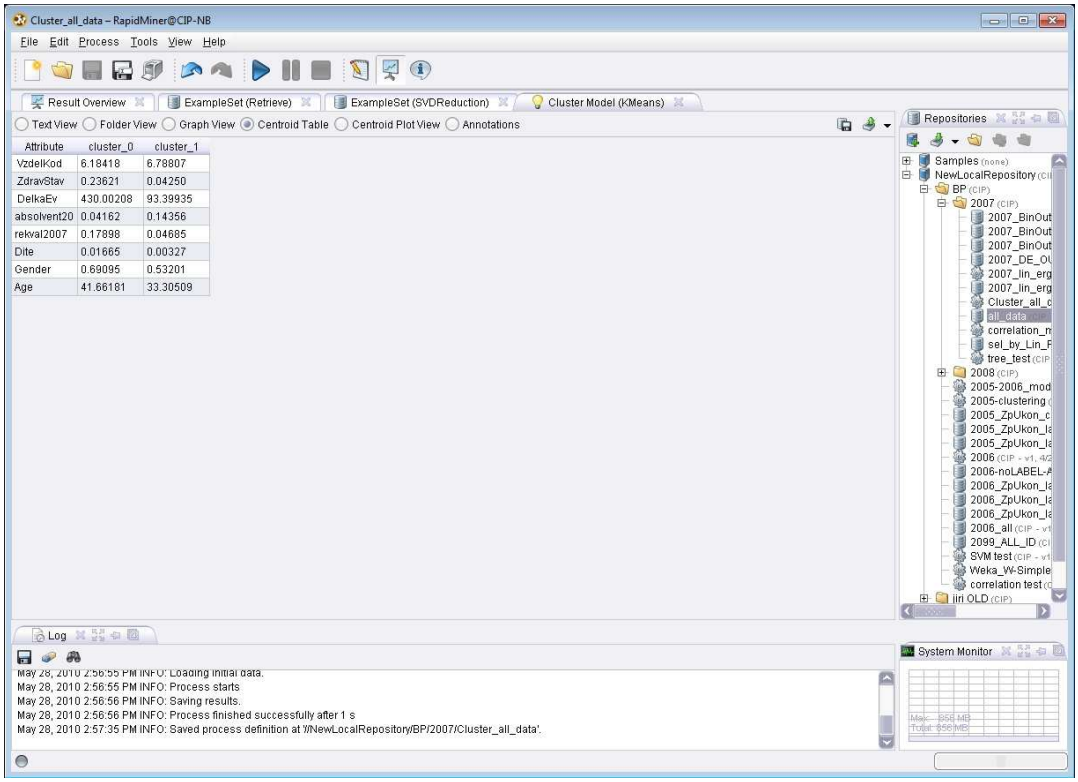
Attribute	Coefficient	Std. Error	Std. Coeffici...	t-Stat	Significance
VzdelKod	-0.373	0.029	-0.181	-12.662	0
ZdravStav	1.365	0.310	0.063	4.396	0.000
absolvent20	1.394	0.261	0.076	5.344	0.000
Gender	-1.586	0.172	-0.131	-9.199	0
Age	-0.035	0.007	-0.072	-5.076	0.000
(Intercept)	10.292				

Role	Name	Type	Statistics	Range	Missings
label	ZpusUkonc	integer	avg = 5.983 +/- 5.994	[1.000 ; 19.000]	0
regular	VzdelKod	integer	avg = 6.663 +/- 2.901	[0.000 ; 13.000]	0
regular	ZdravStav	integer	avg = 0.083 +/- 0.275	[0.000 ; 1.000]	0
regular	absolvent2007	integer	avg = 0.122 +/- 0.328	[0.000 ; 1.000]	0
regular	Gender	integer	avg = 0.565 +/- 0.496	[0.000 ; 1.000]	0
regular	Age	integer	avg = 35.039 +/- 12.561	[16.000 ; 67.000]	0

Attribute	Coefficient	Std. Error	Std. Coeffici...	t-Stat	Significance
VzdelKod	-0.373	0.029	-0.181	-12.662	0
ZdravStav	1.365	0.310	0.063	4.396	0.000
absolvent20	1.394	0.261	0.076	5.344	0.000
Gender	-1.586	0.172	-0.131	-9.199	0
Age	-0.035	0.007	-0.072	-5.076	0.000
(Intercept)	10.292				

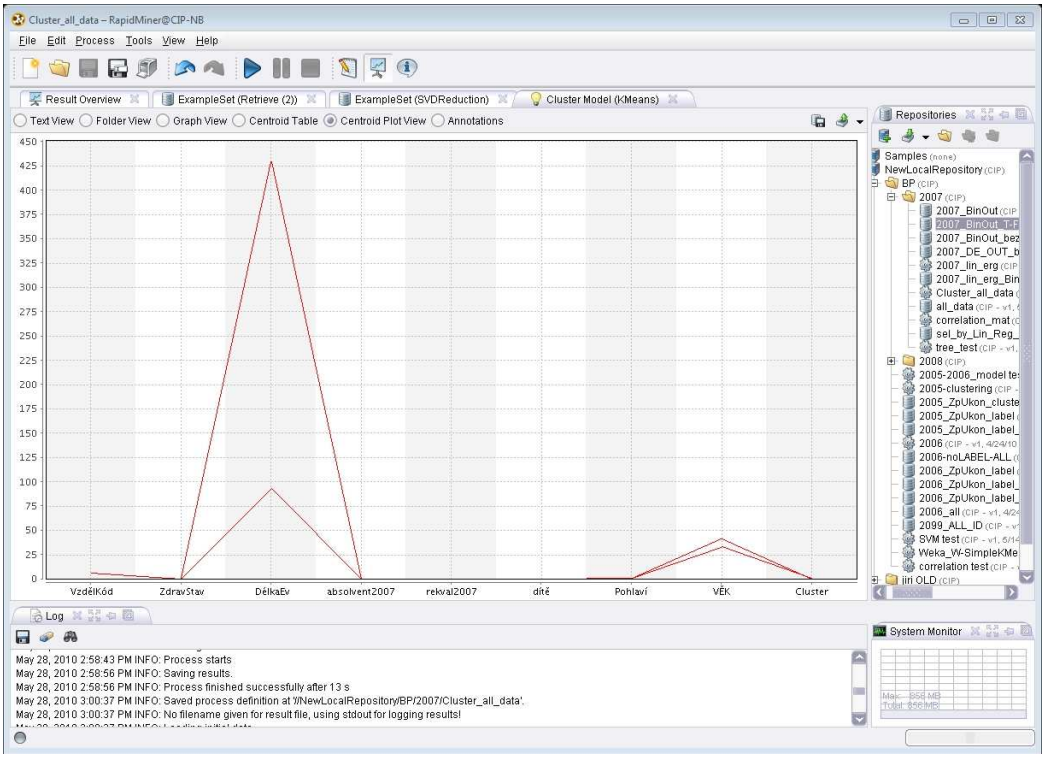
Obrázek 12 –výsledky lineární regrese

V další fázi, byly ověřeny výsledky metody lineární regrese aplikací clusterizačního algoritmu na daná data.



Obrázek 13 – tabulka clasterizace na dva klastry

Výsledkem clusterizačního algoritmu k-menans, kdy k=2 bylo nalezení dvou clastrů. Při k>2 následně docházelo k řazení jedné hodnoty atributu do více clastrů.



Obrázek 14 - Centroid plot Clusterizace na 2 clustry

Z grafické zobrazení centroidů pro předcházející rozdělení na dva clastry je patrná dominance jednoho parametru (délka evidence) a zcela minimální závislost ostatních proměnných.

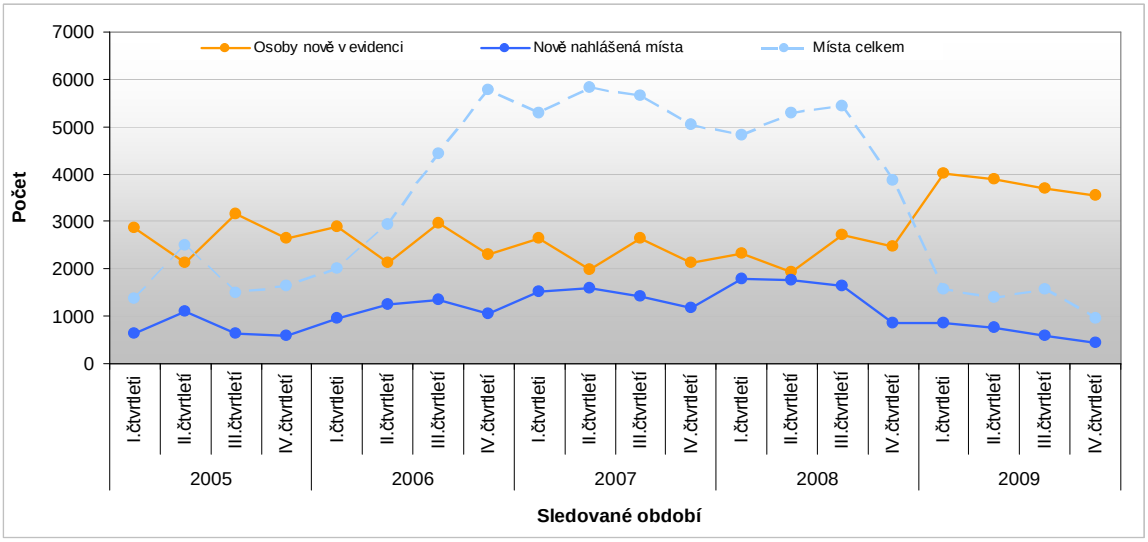
I když byly použity metody ověřené a platné, nepodařilo se na daných datech aplikací Lineární regresní analýzy ani k-means clusterizačního algoritmu, nenalezl závislosti u daných dat.

Během modelingu nebyly nalezeny žádná závislost mezi danými atributy, přesto nelze jednoznačně konstatovat, že mezi těmito daty žádná závislost neexistuje. Lze pouze konstatovat, že tato závislost nebyla danými metodami nalezena.

8 VYHODNOCENÍ

Součástí této fáze je i revize použitých dat. Tyto data pravděpodobně nebyla vhodná případně dostačující pro datamining. Data by bylo vhodné doplnit dalšími o další atributy např.: oborem, ve kterém osoba vykonávala praxi, případně obor, který vystudovala, nebo požadované zaměstnání. Praxe, kterou vykonával není v databázi jednoznačně zadávána a tudíž tento vstup není možné použít. Obor, který dotyčný člověk vystudoval, také není možné získat z databáze bez chybějících dat.

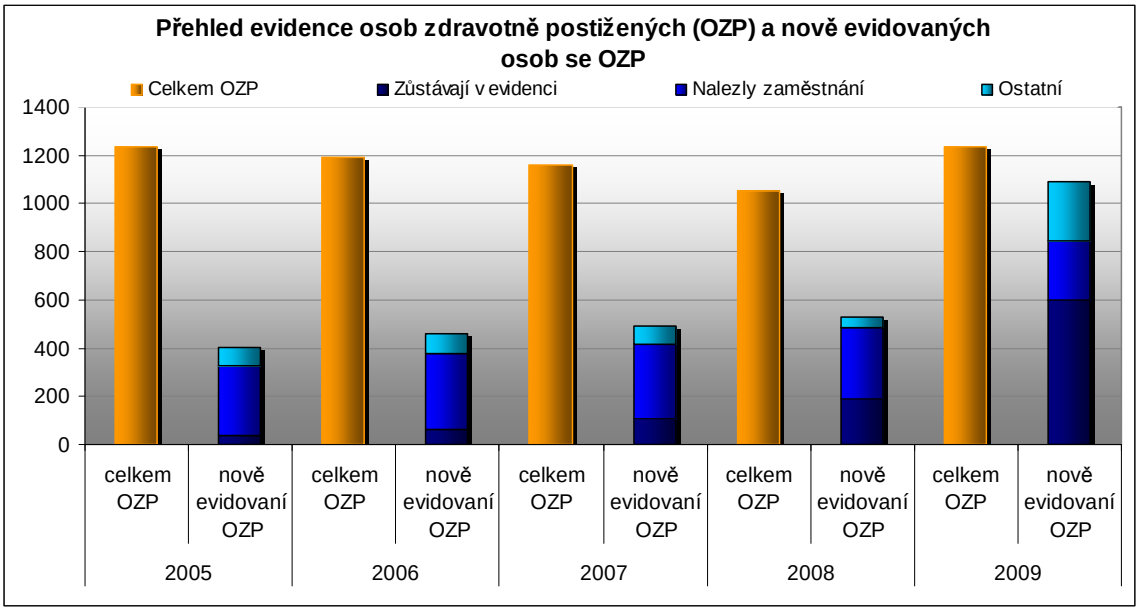
Dalším faktorem, který mohl ovlivnit získaný výsledek dataminingu jsou i vnější vlivy působící na trh práce. Například závislost nabídky volných míst a počet osob, hledající zaměstnání (poptávky po volných místech).



Graf 1 – graf vývoje počtu osob nově v evidenci, nově nahlášených míst a volných míst celkem

Dalším faktorem, který mohl mít vliv na výsledek dataminingu je i regulace některých oblastí formou změn zákonů, nařízení, případně financování některých nástrojů určených pro podporu zaměstnanosti.

V grafu č. 2 je zobrazeno množství lidí OZP v evidenci celkem a množství lidí s OZP, kteří nově nastoupily do evidence. Zatímco trend u nově evidovaných osob s OZP je rostoucí, tento trend není kopírován celkovým množství osob OZP v evidenci, i za předpokladu, že trend osob, kterým se nedaří nalézt místo na trhu práce roste.



Graf 2 – graf vývoje množství osob OZP a osob OZP, které nově přišli do evidence v roce 2005 - 2009

Dalším důležitým faktorem, který mohl ovlivnit výsledek dataminingu jsou i vnitřní faktory osob, jejich životní situace, životní názory a postoje.

V této fázi je třeba přistoupit k opětovné analýze vstupů, nalezení nových vhodných vstupů, vyhodnotit míru vnějších a vnitřních vlivů na hledané závislosti ve spolupráci s odborníky na trh práce a sociology. To však z časových důvodů není možné zpracovat v rámci této práce.

ZÁVĚR

Dolování dat z databází je používáno hlavně z důvodů existence velkého množství dat, jejichž analýza přesahuje schopnosti člověka. Z těchto důvodů jsou používány sofistikované počítačové techniky pro jejich analýzu.

V teoretické části jsou popsány jednotlivé kroky dolování dat z databází a vztah mezi dolováním dat z databází a dataminingem. Tento vztah pomohla upřesnit až mezinárodní konference Knowledge discovery from database v Montrealu (1995), kde definovala datamining jako jeden z kroků v procesu dolování dat z databází zahrnující výběr a aplikaci metod pro vyhledávání zajímavých vztahů v datech.

V teoretické části je popsána metoda CRoss-Industry Standard Process for Data Mining, jednotlivé fáze, její úkoly a vztahy. Na závěr teoretické části jsou popsány úlohu, které lze pomocí DM řešit a k nim vhodné algoritmy.

Cílem práce bylo aplikovat DM na sociologická data z ÚP ve Zlíně a nalézt mezi nimi vazby a popsat je. Data obsahovala nově zaevidované osoby za rok 2005 – 2009. Cílem DM bylo nalezení charakteristických skupin osob, které se dostanou do evidence na Úřadu práce (ÚP) ve Zlíně, vzhledem k jejich schopnosti uplatnit se na trhu práce, specifikovat atributy osob, které jsou schopni si práci nalézt sami, které je možné zaměstnat s pomocí nástrojů ÚP, případně skupiny, které jsou velmi obtížně zaměstnatelné a jsou ohrožené dlouhodobou nezaměstnaností.

Velký důraz byl kladen na vyčištění a přípravu dat. Následná aplikace modelu však nebyla úspěšná. Data nevykazovala žádné jednoznačné závislosti mezi sebou. Pro další pokus byla data přetransformována a použita původní data bez rozdělení do skupin hodnot. Testovány byly také různé varianty výstupních hodnot (např. ukončení evidence, délka evidence) a různé algoritmy - lineární regrese a metody clusteringu. Přesto se nepodařilo nalézt dostatečně významné závislosti mezi jednotlivými atributy.

Datamining je velmi silný nástroj, ale má svá omezení, obzvláště v oblasti výzkumu sociálních dat, kde jsou schopnosti a jednání lidí závislá na více faktorech. Analyzovaná data bohužel ani zdaleka neposkytují dostatek informací pro podrobnou analýzu a sofistikované vyhodnocení.

ZÁVĚR V ANGLIČTINĚ

Knowladge Discovery in Databases is mainly used for reason of the existence of large amounts of data, whose analysis goes beyond human´s capacity. For this reason, are used sophisticated information technology, for their analysis.

In the theoretical parts of this work are describes the steps of the discovery knowladge form database and relation between discovery knowladge form database and datamining. This relacion has been clarify by The International Conference of Knowledge Discovery from Database in Montreál (1995), where datamining has been defined as the stamp in the process Knowladge Discovery in Databases, including the selection and applicatin of methods for search interesting relations of data.

The theoretical parts is describes the method, phases, tasks and relation of CRoss-Industry Standard Process for Data Minig. In the end of theoretical part are descripted the problems, can be solved by the Datamining and their appropirate algorithms.

The goal of work was convert Datamining into sociological data from the Labour Office in Zlin, search relation and describe them. The data include newly registred persones in the Labour Office in Zlin, from 2005 till 2009. Datamining´s goal was searched characteristic groups of registr´s people at the Loubour Office (LO) in Zlín, due to their potenciality find a job. Specify the attributes of persons, who had potencial to search job themselves, which can be employed by tool of the LO in Zlín, or groups are very difficult employable and in dangere of long-term unemployment.

Great emphasis was placed on cleaning and preparing the data. The application of the model was not succeful. The data wasn´t showed the significant relation between themselves. For the another test was transformed data and used the original data without groupings of values. The tested was also different variations of the output values (eg end of registration, length of registration). Or as a result of these changes the data wasn´t showed the significant relation.

I belive that the dataminig is a very powerful tool, but it has its limitations, especially in the field of social research data, where the reaction of people are dependent on many external factor.

SEZNAM POUŽITÉ LITERATURY

[1] HAN, J., KAMBER, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006. 550s. ISBN 978-1-55860-489-6

[2] FAYYAD, M., PIATETSKY-SHAPIRO, G., SMYTH P., UTHURUSAMY R. Advances in knowladge discovery and data mining, California 1996. 611s. ISBN 0-262-5697-6

[3] KLÍMEK, P. Aplikovaná statistika, Studijní pomůcka pro distanční studium, Zlín, Univerzita Tomáše Bati ve Zlíně, 2005. ISBN 80-7318-304-8

[4] KLÍMEK, P. Získávání znalostí z podnikových dat (data mining) = Knowledge discovery in company data (data mining): teze disertační práce. Zlín: Univerzita Tomáše Bati ve Zlíně, 2005. 35 s. ISBN 8073182416.

[5] HORNICK, M. F., MARCADÉ. E., VENKAYALA, S. Java Data Mining: Strategy Standard, and Practice, A Practi Guide for Architecture, Design and Implementation, Morgan Kaufmann, 2007. 520s. ISBN 978-0-12-370452-8

[6] *Cross Industry Standard Process for Data Mining*, Dotupnný z WWW: <http://www.crisp-dm.org/Process/index.htm>

[7]*Data Minig SOLUTIONS.* Dostupný z WWW: <http://datamining.xf.cz/view.php?cisloclanku=2002102801>

[8] *Data Mining SOLUTINOS.* Dostupný z WWW: http://datamining.xf.cz/view.php?cisloclanku=2002102807

[9] *Dobývání znalostí z databází.* Dostupný z WWW: <http://sorry.vse.cz/~berka/docs/izi456/kap_1.pdf>

[10] EuroMISE centrum – Kardio (Evropské centrum pro medicínskou informatiku statistiku a epidemiologii – Kardio). Dostupný z WWW: http://euromise.vse.cz/kdd/index.php?page=metody

[11] *Rozhodovací stromy.* Dostupný z WWW http://sorry.vse.cz/~berka/docs/izi456/kap_5.1.pdf

[12] *Asociační pravidla*. Dostupný z WWW: http://sorry.vse.cz/~berka/docs/izi456/kap_5.2.pdf

[13] Statistická data trhu práce. Dostupná z WWW: <http://portal.mpsv.cz/sz/stat/nz>

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

KDD	Knowledge discovery from databases
DM	Datamining
ÚP	Úřad práce
CRIPS-DM	CRoss-Industry Standard Process for Data Minig
MPSV	Ministerstvo práce a sociálních věcí
SUPM	Společensky účelná pracovní místa
SVČ	Samostatně výdělečná činnost
VPP	Veřejně prospěšné práce
ESF	Evropský sociální fond
OZP	Osoby zdravotně postižené
PID	Plný invalidní důchod
CID	Částečný invalidní důchod
OP LZZ	Operační program lidské zdroje a zaměstnanost

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

PŘÍLOHA P1 PRACOVIŠTĚ ÚP VE ZLÍNĚ DLE MÍSTA BYDLIŠTĚ

0 Otrokovice	1 Zlín			2 Slavičín	3 Valašské Klobouky
Bělov	Biskupice	Jasenná	Podkopná Lhota	Bohuslavice nad Vláří	Brumov-Bylnice
Halenkovice	Bohuslavice u Zlína	Kaňovice	Pozlovice	Haluzice	Drnovice
Komárov	Bratřejov	Karlovice	Provodov	Jestřabí	Křekov
Napajedla	Březnice	Kašava	Racková	Lipová	Návojná
Oldřichovice	Březová	Kelníky	Sazovice	Loučka	Nedašova
Otrokovice	Březůvky	Lhota u Malenovic	Sehradice	Petrůvka	Lhota
Pohořelice	Dešná	Lhotsko	Slušovice	Rokytnice	Nedašov
Spytihněv	Dobrkovice	Lípa	Šarovy	Rudimov	Poteč
Tlumačov	Dolní Lhota	Ludkovice	Tečovice	Slavičín	Tichov
Žlutava	Doubravy	Luhačovice	Trnava	Slopné	Újezd
	Držková	Lukov	Ublo	Šanov	Valašské Klobouky
	Fryšták	Lukoveček	Velký Ořechov	Štítná nad Vláří	Vlachova Lhota
	Horní Lhota	Lutonina	Veselá	Vlachovice	Vysoké Pole
	Hostišová	Machová	Vizovice		
	Hrobice	Mysločovice	Vlčková		
	Hřivínův	Neubuz	Všemina		
	Újezd	Ostrata	Zádveřice-Raková		