Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta
Katedra aplikované matematiky a informatiky

Bakalářská práce

Data mining v ekonomickém výzkumu

Vypracoval: David Peroutka
Vedoucí práce: doc. RNDr. Václav Nýdl, CSc.

České Budějovice 2016

JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH
Fakulta ekonomická
Akademický rok: **2014/2015**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**David PEROUTKA**
Osobní číslo:	**E13538**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Ekonomická informatika**
Název tématu:	**Data mining v ekonomickém výzkumu**
Zadávající katedra:	**Katedra aplikované matematiky a informatiky**

Zásady pro vypracování:

Cíl práce: Cílem práce je zvládnutí principů data miningu se zaměřením na aplikace v ekonomickém výzkumu. Podrobně budou popsány způsob práce se zvoleným specializovaným software (např. "Rapid Miner") a ukázka konkrétní aplikace.

Metodický postup:

1. Zvolit dostupný software (např. "Rapid Miner"), provést jeho instalaci.
2. Prostudovat uživatelské návody, provést kontrolní výpočty.
3. Zpracovat vlastní podrobný popis používání software.
4. Pokusit se o zpracování souborů reálných dat, nejlépe z ekonomické oblasti.
5. Ukázat na konkrétním příkladu možnou interpretaci výsledků.
6. Závěr - zhodnoceni metody a možností aplikace v ekonomickém výzkumu.

Rozsah grafických prací: **dle potřeby**

Rozsah pracovní zprávy: **40 - 50 stran**

Forma zpracování bakalářské práce: **tištěná**

Seznam odborné literatury:

1. **Górecky, J. (2011).** *Data Miner* **(referát). Ostrava: FEI VŠ-TUO. Dostupné z: http://www.person.vsb.cz/archivcd/FEI/MAD/Priloha%201%20Referat%20Rapid%20Miner.pdf.**
2. **Hand, D. J. et al. (2001).** *Principles of Data Mining (Adaptive Computation and Machine Learning).*
3. **Procházka, M.** *Data mining - jiný pohled na problém.* **VTM. Dostupné z: http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem.**
4. **Williams, G. (2001).** *Data Mining with Rattle and R.* **Springer: New York.**
5. **DATAMIND. Dostupné z: http://www.datamind.cz/cz/blog/Data-mining-zdarma-rapid-miner-v-praxi.**
6. **eCOMMERCE SOFTWARE. http://sourceforge.net/projects/rapidminer/.**
7. **RAPIDMINER software. https://rapidminer.com/.**
8. **WIKIPEDIA: http://en.wikipedia.org/wiki/RapidMiner.**
9. **YOUTUBE video. https://www.youtube.com/user/RapidIVideos.**

Vedoucí bakalářské práce: **doc. RNDr. Václav Nýdl, CSc.**
Katedra aplikované matematiky a informatiky

Datum zadání bakalářské práce: **9. ledna 2015**

Termín odevzdání bakalářské práce: **15. dubna 2016**

doc. Ing. Ladislav Rolínek, Ph.D.
děkan

JIHOČESKÁ UNIVERZITA
V ČESKÝCH BUDĚJOVICÍCH
EKONOMICKÁ FAKULTA
Studentská 13 (26)
370 05 České Budějovice

L.S.

prof. RNDr. Pavel Tlustý, CSc.
vedoucí katedry

V Českých Budějovicích dne 27. března 2015

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47 zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to - v nezkrácené podobě vzniklé vypuštěním vyznačených částí archivovaných Ekonomickou fakultou - elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích, dne 11. 4. 2016 Podpis ……

Poděkování

Tímto bych rád poděkoval vedoucímu mé bakalářské práce, panu doc. RNDr. Václavu Nýdlovi, CSc., za jeho pomoc, ochotu, věnovaný čas a cenné rady při zpracování této bakalářské práce.

Obsah

1 Úvod

Ve stále se rozšiřujících oblastech informatiky, ekonomiky a práce s daty, vyvstává nutnost s těmito daty určitým způsobem nakládat a využívat je pro budoucí příležitosti. Data jsou různorodé informace, například o lidech, o zboží nebo o službách. Jejich shromažďováním získáváme informace o potřebách lidí, o zboží, které je pro lidi atraktivní nebo o službách, které vyhledávají. Získaná data musí být podle určitých pravidel a metod zpracována tak, aby jejich výsledky odpovídaly potřebám, které vedly k jejich třídění. Tomuto procesu se říká data mining, což lze volně přeložit jako dolování dat nebo těžbě dat.

V dnešní době je data mining používán jako nástroj pro nastavování prodejních strategií, různých uživatelských balíčků nabízených poskytovateli služeb a v mnoha dalších odvětví. Jedná se o přizpůsobování se zákazníkovým potřebám a touhám tak, aby pro něho byly lákavé a nakonec si tyto služby nebo zboží koupil. Každá dnešní domácnost většinou disponuje alespoň jedním osobním počítačem, tabletem nebo chytrým telefonem s přístupem na internet. To vše je spjato s usnadněním obchodování a umožňuje zvyšovat efektivitu procesů s ním spjatých. Například internetový obchod je založen na elektronickém zpracování a přenosu dat. Patří do něj elektronický prodej zboží a služeb, poprodejní služby, elektronická výměna nejrůznějších dat, vedení bankovních účtů, zpracování statistických informací, elektronické burzy cenných papírů, obchodní aukce a jiné. Firmy data mining používají jako marketingový nástroj pro zvětšení svých tržeb, zvýšení prodejů a uspokojení zákazníků.

V první části mé bakalářské práce popíši detailněji co to data mining je, kdy se poprvé objevil, jaká data se při tomto procesu používají a jako příklad uvedu několik metod a jejich specifikace. Objasním hlouběji jeho význam pro ekonomický sektor a jeho využití v tomto odvětví. Dále se seznámím se softwarem, který se používá pro data mining, popíši jeho uživatelské prostředí, formáty vstupních dat, které jsou podporovány tímto softwarem a přípravou dat pro práci s ním.

V praktické části bakalářské práce názorně ukáži na datech, které jsem si připravil, jak probíhá proces od začátku až k výsledkům. Data pochází z webových stránek Eurostat, kde jsou volně dostupná pro kohokoliv. Samozřejmě obsahují jen údaje, které odráží hospodaření a výsledky minulých let. Nejsou zde zveřejňovány žádné choulostivé

údaje, které by neměly být veřejnosti přístupné. Postupně názorně ukáži, jak vypadají jednotlivé metody znázorněné pomocí bloků, které představují určité funkce a jaké jsou výsledky jednotlivých metod.

V závěru poté zhodnotím výsledky, které jsem získal jednotlivými procesy v programu, a navrhnu jejich možné využití v budoucí tvorbě strategie. Vysvětlím, co výsledky znamenají a jak je nutné na ně nahlížet.

2 Data Mining

Se stále rychlejšími technologickými pokroky začíná data mining pronikat do podvědomí a nachází uplatnění v rychle se rozvíjejících oblastech na celém světě. Existence rozsáhlých databází dala vzniknout činnostem, které v těchto databázích, majících někdy i desetitisíce řádků, dokáží najít souvislosti, vazby a vztahy, které lze dále používat například v managementu, při tvorbě strategického plánu, ale i dalších neekonomických odvětvích.

Pro takové objemy dat nejsou standardní statistické metody příliš vhodné, bylo tedy potřeba nalézt metody, které dokáží odhalit i složitější nelineární vazby a to bez omezujících předpokladů. Prostředkem nalezení těchto struktur (pravidel, vzorů, asociací, atd.) bylo využití výpočetní techniky namísto statistických parametrů (středních hodnot, vah, atd.).

2.1 Pojem Data Mining

Pod pojmem data mining si každý může představit něco jiného. Na čem se většina autorů, kteří se nějakým způsobem zabývali nebo zabývají data miningem, shodne, je skutečnost, že se jedná o práci s velkým množstvím dat. V hlubších definicích se autoři rozcházejí a každý má svou vlastní představu, co si představit pod tímto pojmem.

Pojem data mining se začal objevovat v devadesátých letech; v roce 1991 napsal první definici data miningu Frawley: „Data mining je netriviální získávání předem neznámé a potencionálně užitečné informace, ukryté v datech.“ V některých případech je data mining též nazýván „dolováním informací“ nebo „vytěžováním dat“. Data mining se používá hlavně v oblastech, kde se sbírá velké množství dat. Typickým příkladem jsou:

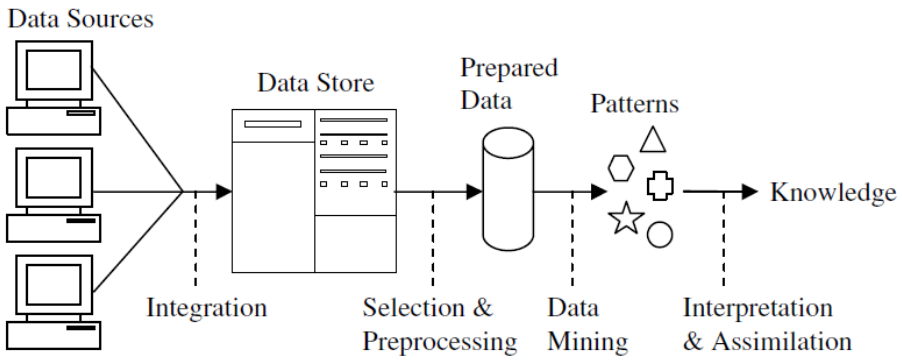
- Údaje o klientech
- Údaje o volání (operátoři)
- Údaje o tom, jak lidé nakupují (obchodní řetězce)

Podle Wittena a kol. (2011), data mining znamená získávání implicitních, předem neznámých a potencionálně užitečných informací z dat. Myšlenkou je vytvořit počítačové programy, které projdou data skrz celou databázi a hledají nejrůznější zákonitosti nebo

vzory, podle kterých by bylo možné dále postupovat. Samozřejmě se mohou vyskytovat problémy jako například to, že mnoho vzorů bude banálních nebo nezajímavých. Dalším nepřehlédnutelným faktorem je, že skutečná data nikdy nejsou dokonalá. To znamená, že mohou chybět některá data; v případě, že jsou data zadávána lidským faktorem, je velká pravděpodobnost, že se v datech budou vyskytovat překlepy nebo zkomolené výrazy, se kterými počítačové programy neumí pracovat (Witten a kol., 2011).

Bramer (2007), který ve své knize hovoří o „knowledge discovery process", což lze volně přeložit jako proces získávání znalostí, má za to, že data mining je pouze jednou částí z celého procesu získávání informací, ačkoliv se jedná o část centrální. Velkou váhu přikládá i činnostem předcházejícím samotnému data miningu a výsledné interpretaci získaných znalostí.

Obrázek 1: Proces získávání znalostí



Zdroj: Bramer, (2007)

Obrázek představuje mírně idealizovanou verzi kompletního procesu. Data vstupují do procesu z mnoha různých zdrojů a jsou integrována ve společném datovém skladišti. Část z nich se začne zpracovávat do standardizovaného formátu. Následně jsou tato data předána data miningovému algoritmu, který na základě svých vlastností vytvoří výstup. Tento výstup je považován za nové potencionálně užitečné znalosti (Bramer, 2007).

Cílem procesu dobývání znalostí je získat co nejvíce relevantních informací vhodných k řešení daného problému. Berka (2003) ve své knize uvádí jako příklad reálného problému otázku nalezení skupin zákazníků obchodního domu nebo skupin klientů banky, kterým by bylo možné nabídnout speciální služby. U zákazníků obchodního domu uvádí jako příklad zjištění, že zákazník kupuje potravinářské zboží odpovídající jisté

dietě. V případě klientů banky může jít o potencionální zájemce o hypoteční úvěr. Nalezené skupiny jsou interpretovány jako tzv. segmenty trhu v dané oblasti.

V konečném důsledku je tedy možné říci, že data mining slouží jako nástroj napomáhající pochopení významu dat a tvorbě předpovědí z nich.

2.2 Využití v ekonomickém výzkumu

Data mining zaujímá svou část i v ekonomickém odvětví. Hlavním cílem je zde neustále reagovat na změny trhu, být co nejvíce elastický k těmto změnám, reagovat na potřeby zákazníků, odhalovat souvislosti a umět na nich těžit, nacházet modely, podle kterých se lze do budoucna řídit a s tím spojené krátkodobé odhadování budoucích situací. Následující podkapitola pojednává o původu ekonomických dat, jejich zdrojích, uchovávání a možném využití v budoucnosti.

2.2.1 Ekonomická data

Vznik ekonomických dat je podmíněn ekonomickou činností, bez ní by žádná ekonomická data nebylo možné shromažďovat ani na základě jejich výsledků, získaných procesem data miningu, tvořit plány do budoucna. Existence těchto dat, zapříčinila proniknutí data miningu i do ekonomického sektoru, kde v dnešní době zaujímá svou nenahraditelnou pozici.

Většina dostupných ekonomických dat pochází z pozorování přírody. Podle Feelderse (2002), který je autorem článku o data miningu v ekonomické sféře, data nebyla získána prováděním kontrolovaných experimentů, ale pasivním pozorováním ekonomické reality. Podle něj mají omezující možnosti experimentů v ekonomice za následek mezeru mezi teorií a dostupnými daty.

Jako příklad uvádí odhad poptávkové křivky po pomerančích, kdy tvrdí, že není dostatečné pozorovat cenu pomerančů v různých časových intervalech a odpovídající zakoupené množství. Důvod je změna dalších faktorů, jako například cena substitutů, import atd. Aby byl odhad správný, musíme do něj zahrnout i tyto důležité vlivy.

Dalším užitečným rozdělením ekonomických dat je dělení na mikroekonomická a makroekonomická data. Mikroekonomická data se vztahují k jedné určité jednotce. Touto

jednotkou mohou být například firmy, domácnosti nebo jednotlivci. Makroekonomická data sdružují data na území daného státu a zabývají se konečným ekonomickým výsledkem, celkovou nezaměstnaností a průměrnými cenami veškerého zboží, vyprodukovaného v dané ekonomice. Mnoho dat z ekonomických aktivit je shromažďováno nejrůznějšímu subjekty. V případě makroekonomických dat je tak v největší míře činěno vládními orgány.

U nás je typickým příkladem Český statistický úřad, který obsahuje nejrůznější informace o státní ekonomice, srovnání se zahraničím, vědě a výzkumu. Tyto údaje jsou volně dostupné veřejnosti a lze s nimi dále pracovat, používat je v nejrůznějších pracích, tvořit příklady na jejich základě atd.

Ve Spojených státech je podobná možnost získávání informací o hospodaření země a to z webových stránek Data.gov. Jejich obsah je mnohem rozsáhlejší a poskytuje širší spektrum informací. Nabízí například data o zemědělství, ekosystémech, energetice, financích, oceánu, vědě a výzkumu a mnoho dalších.

2.2.2 Data miningové metody používané v makroekonomii

Používané metody mají řadu charakteristik, modelů a korelací mezi vstupními daty nebo pořadí poskytovaných dat, což zajišťuje informační nadbytečnosti a velmi jasně vymezený charakter existujících dat. Stancu a kol. (2012) ve své publikaci o data miningových metodách používaných na rumunských makroekonomických datech uvádí příklady využití těchto metod v následujících situacích:

- Pokud je velikost vstupních dat rozsáhlá
- Pokud jsou požadována stejnorodá data

Mezi hlavní tři metody používané pro práci s makroekonomickými daty patří:

- Analýza hlavních komponent
- Faktorová analýza
- Shluková analýza

Analýza hlavních komponent prochází podle Stancu a kol. (2012) několika kroky:

- Výpočet kovariační matice ze vstupních dat
- Maximalizování odchylek
- Výpočet proprietárních hodnot a jejich následné uspořádání od největší hodnoty po nejmenší
- Stanovení proprietárních vektorů spojených s proprietárními hodnotami
- Stanovení lineární kombinace v nové situaci

Faktorová analýza má velkou výhodu v nabízení možnosti stanovit množství nepozorovatelných charakteristik dat. Nejběžnějším faktorem je faktor vyjadřující změnu alespoň dvou proměnných. Na rozdíl od analýzy hlavních komponent, se faktorová analýza snaží modelovat korelaci, která již existuje mezi proměnnými (Stancu, a kol., 2012).

Kroky faktorové analýzy jsou následující:

- Určit minimální počet příčinných faktorů
- Otáčení faktorů v pořadí k nalezení faktoru řešení
- Vysvětlení společných faktorů
- Odhadnutí skóre matice faktorů

Shluková analýza je blíže popsána v kapitole 3.3.

2.2.3 Data mining v mikroekonomii

Rostoucí množství mikroekonomických dat, vztahujících se k individuálním spotřebitelům a jejich nákupnímu chování poskytuje skvělou příležitost pro data mining (Feelders, 2002). Hlavním cílem je pochopení vztahů například mezi spotřebitelem a dodavatelem nebo domácnostmi a dodavatelem apod. Využívá se zde zejména korelační matice a shluková analýza, ale i další metody podle potřeb informací, které jsou požadovány.

Snaží se odpovědět na otázky typu:

- Co určuje cenu konkrétního zboží
- Co určuje výstup konkrétní firmy nebo průmyslu
- Co určuje množství práce, kterou je konkrétní pracovník ochoten poskytovat

Každé rozhodnutí učiněné například bankou, státem nebo jednotlivcem musí být založeno na konkrétních datech s kompletní informační analýzou, která je z dat získána (Stancu, 2012).

2.3 Zdroje dat

Mezi hlavní zdroje dat patří databáze obecně. Ať už se jedná o databáze soukromé firmy, státu, fyzické osoby či jiného subjektu. Tato data musí být někde uložena. Jednou z možností jsou datové sklady a datová tržiště. Podle Inmona (1999), který v 80. letech zformuloval koncept datového skladu, je takový sklad:

- subjektově orientovaný,
- integrovaný,
- časově proměnný,
- leč stálý

soubor dat, který slouží pro podporu rozhodování.

Prvním charakteristickým rysem takového skladu je orientování na komodity, které jsou pro daný subjekt typické. Datový sklad neuchovává data, která nejsou vhodná pro podporu rozhodování na manažerské úrovni. Na základě vstupu dat do datového skladu je potřeba tato data integrovat a sjednocovat, což obnáší sjednocení názvů stejných ukazatelů, sjednocení měřítek, sjednocení kódování apod. Dalším takovým typem zdroje dat jsou produkční databáze, které se na rozdíl od datových skladů zabývají operacemi a transakcemi, jako například úvěry, fakturami, vklady, výběry apod. Uchovávají data potřebná pro operativní řízení bez ohledu na to, zda budou využitelná při budoucím strategickém rozhodování. Tyto dva typy jsou spolu provázány a fungují společně. Všechna data z datového skladu představují „časový snímek“ dat z produkčních databází

sejmutý v určitém okamžiku (Berka, 2003). Datový sklad je aktualizován offline v určitých časových intervalech (týdně, měsíčně, ročně) a je rovněž analyzován odděleně od produkčních databází. Vytvoření datového skladu zahrnuje úkony jako načtení dat, konverzi dat, čistění a transformaci. Data uložená v datovém skladu představují datový prostor, který je neutrální a není vytvářen za účelem konkrétních analýz. Proto se doporučuje vytvářet v závislosti na datovém skladu řadu specializovanějších datových tržišť, kam se z datového skladu přesunou data potřebná pro určitý typ analýzy. Tento proces je pak nazýván třívrstvou architekturou datového skladu (Symons, 2000).

3 Metodika

Pro svou bakalářskou práci jsem měl možnost vybrat si z více dostupných programů zaměřených na práci s velkým množstvím dat. Jedním z nejznámějších je RapidMiner. Existuje mnoho dalších programů, například Statistica, ten však nabízí mnohem větší možnosti a stává se tím náročnější na práci. Pro mou bakalářskou práci jsem si vybral RapidMiner, protože je mi svým prostředím blíže a práce s ním mě baví.

3.1 RapidMiner Studio

RapidMiner je software, který byl vytvořen za účelem usnadnění práce s velkými databázemi obsahujícími data nejrůznějšího charakteru. Mezi jeho nesporné výhody patří bezplatná dostupnost na internetových stránkách výrobce, intuitivní ovládání, grafické prostředí, ve kterém je práce opravdu zábava a řada dalších. Nynější verze dostupná pro veřejnost je verze 7.0. Dostupná je pro operační systémy Windows, Mac OS a Linux. Instalace je typická jako u většiny software, vyberete adresář, ve kterém chcete program mít a vyčkáte na dokončení instalace. Po prvotním spuštění je možnost projít si tutoriál, který seznamuje uživatele s programem a jeho funkcemi.

3.1.1 Uživatelské prostředí

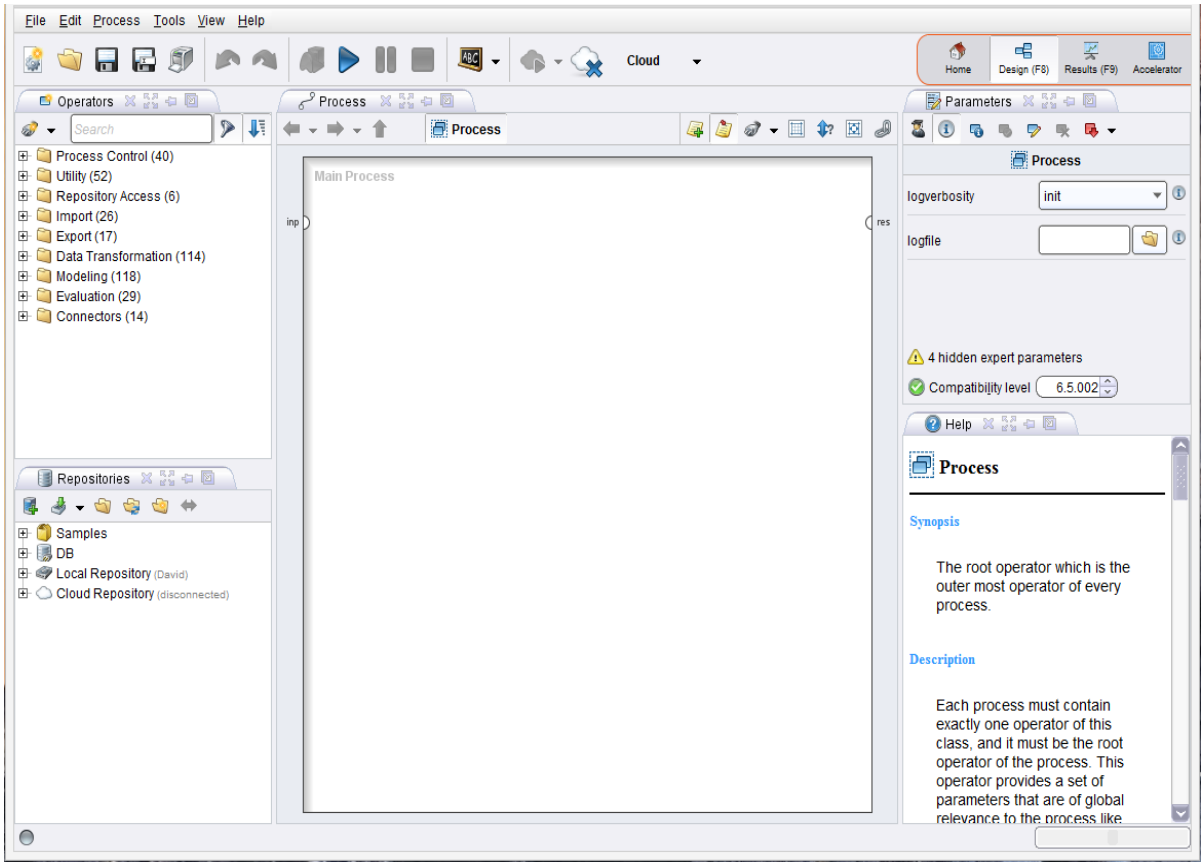
Jak bylo krátce zmíněno výše, uživatelské prostřední je vytvořeno tak, aby v programu mohl pracovat i člověk, který s ním nemá větší zkušenosti. Všemu napomáhá grafické prostředí, ve kterém dochází k vybrání správných bloků, jenž reprezentují funkci ve zvolené metodě a jejich následné spojení do jednoho funkčního celku. Bloky lze vybírat z nabídky Operators v levém horním rohu okna programu, která nabízí velké množství funkcí. V levém dolním rohu se nachází okno Repositories, ve kterém se ukrývají veškerá data potřebná při práci. Mohou a většinou i budou se tam nacházet data, která se musí do úložiště nahrát z vnějšku. Hlavní a největší okno je okno Process, umístěné ve středu, ve kterém dochází ke spojování daných bloků, jenž byly vybrány. Další nespornou výhodou je nápověda v pravém dolním rohu sloužící pro informativní účely týkající se vybraného bloku. Stačí vybrat si blok a přesunout ho do

okna Process, tím aktivujete nápovědu, ve které se dozvíte, co vše vybraný blok zvládne, stručný popis jeho funkce, s jakými vstupy pracuje a jaké jsou jeho výstupy.

Každý blok má vstup a výstup, kromě prvního bloku. Ten představuje vstupní data, se kterými se bude pracovat a která budou použita pro danou metodu. Z toho vyplývá, že z bloků vzniká jeden řetězec po sobě jdoucích procesů tvořících finální metodu. Poslední blok vždy musí být připojen do bodu „res“, nacházejícího se na pravém okraji okna Process.

Po spojení všech bloků, by měl každý blok mít v levém dolním rohu zelené kolečko. Pokud se někde vyskytla chyba nebo byly zvolené špatné atributy, může se stát, že blok bude mít červená kolečko, což značí nějakou chybu, kterou je potřeba odstranit, jinak nedojde ke správnému průběhu procesu a výsledky nebudou mít očekávanou podobu. Pokud mají všechny bloky zelené kolečko, je možno spustit proces. Proces spustí tlačítko v levé horní části okna programu, které má tvar modré šipky. Po spuštění procesu se změní prostředí z návrhové části do části výsledků. Výsledky jsou pro každou metodu jiné, čili je těžké o nich hovořit obecně.

Obrázek 2: Uživatelské prostředí RapidMineru



Zdroj: vlastní zpracování

3.1.2 Import vstupních dat

Potřebná data pro práci mohou pocházet z různých míst a v řadě různých podob. Proto je důležité, importovat je do prostředí RapidMineru tak, aby splňovala předpoklady pro úspěšnou práci s nimi. Hlavní důraz se klade na formát vstupních dat a poté také na jejich samotnou strukturu, zda obsahují prázdné buňky nebo ne, zda veškerá data uvnitř souboru mají stejnou formu nebo zda se mísí různé datové typy.

Abychom byli schopni pracovat s daty efektivně s vidinou úspěšného konce, je nutné data do RapidMineru importovat. Slouží k tomu průvodce, který se nachází v okně Repositories a jmenuje se „Import data“. Na výběr je z několika formátů, nejpoužívanější jsou formáty CSV a XLS. Průvodce zajistí správnou podobu dat pro vybraný formát a tím se eliminují případné problémy s daty během průběhu procesu. V průvodci je možnost data různě formovat, vynechávat řádky nebo sloupce, přeskakovat komentáře v souboru, vynechávat záhlaví a zápatí souboru nebo například měnit datové typy sloupců, které průvodce automaticky rozpoznává a defaultně volí podle jejich vlastností.

3.1.3 Formáty vstupních dat

Data používaná pro data mining bývají často rozsáhlá, obsahují tisíce řádků a desítky sloupců. Je proto žádoucí, aby měla určitý formát. Do prostředí RapidMineru lze importovat 6 různých typů formátů dat.

1. **XLS**

   Představuje nejpoužívanější formát uchovávání dat. Jedná se o dokument vytvořený v programu Microsoft Excel. Každý si jistě dokáže představit tabulku z Excelu a v ní data například o prodeji zboží.

2. **CSV**

   Jsou data, která mají mezi každou hodnotou čárku. Zkratka vychází z anglického názvu „Comma-separated values“, což v předkladu do češtiny znamená „Hodnoty oddělené čárkou“. Na operačním systému Windows je lze otevřít například v poznámkovém bloku nebo v programu PS Pad.

3. **XML**

   Z anglického „Extensible Markup Language“, což lze do češtiny přeložit jako „Rozšiřitelný značkovací jazyk“. Tento jazyk je určen pro výměnu dat mezi aplikacemi. Nezabývá se vzhledem, ale pouze věcným obsahem.

4. **Binární soubory**

   Soubory obsahující kombinace binárních čísel.

5. **Databázové tabulky**

   Databázové tabulky z různých programů, kterými disponují firmy zabývající se činností, která je spjata s uchováváním informací.

6. **Accessové databázové tabulky**

   Tabulky z prostředí Microsoft Access, které vznikaly a stále vznikají například v knihovnách, kde se v nich uchovávají informace o vypůjčených knihách a vypůjčitelích nebo ve společnostech, které disponují sklady, ve kterých sdružují různé zboží a výrobky.

3.1.4 Typy proměnných

Typů proměnných je celá řady. Nedostatečné pochopení rozdílů mezi různými typy proměnných může vést k problémům při jakékoliv práci s nimi. Můžeme rozlišit šest hlavních typů (Bramer, 2007).

1. **Nominální proměnné**

   Hodnoty použité k přiřazení objektů do kategorií. Například můžeme přiřadit 10 lidem čísla od 1-10. Hodnotami mohou být texty nebo číselné kódy. Nelze u nich provádět aritmetické operace.

2. **Binární proměnné**

   Jsou specifickým případem nominálních dat, které nabývají pouze dvou možných hodnot: pravda nebo nepravda, 1 nebo 0.

3. **Pořadové proměnné**

   Pořadová čísla jsou podobná číslům nominálním s výjimkou toho, že u pořadových čísel jsou proměnné uspořádány do smysluplného pořadí. Například: malá, střední, velká.

4. **Celočíselné proměnné**

   Jsou to proměnné, které představují skutečné celočíselné hodnoty. Například: Počet plátců DPH. Na rozdíl od nominálních proměnných s celočíselnými lze provádět aritmetické operace. Lze je sčítat, odečítat a podobně.

5. **Intervalové proměnné**

   Jsou to proměnné nabývající numerických hodnot a lze vyjádřit o kolik je jedna hodnota větší (resp. menší) než hodnota druhá. Mezi dobře známá intervalové proměnné patří Fahrenheitova teplotní stupnice a Celsiova teplotní stupnice.

6. **Poměrové proměnné**

   Poměrové proměnné jsou podobné proměnným intervalovým s tím rozdílem, že u poměrových má smysl se ptát nejen na rozdíl, ale i podíl hodnot. Typickým příkladem je výška – více než to, že jsou třeba Evropané v průměru o několik centimetrů vyšší než Asiaté, nám řekne sdělení, o kolik procent jsou vyšší.

3.2 Jednotlivé fáze procesu

Následující podkapitola popisuje postup procesu získávání znalostí z databází. Vysvětluje jednotlivé kroky, kterými je třeba projít, aby byl proces úspěšný.

3.2.1 Porozumění problematice

V této fázi je cílem pochopení cílů úlohy a požadavků na řešení formulovaných z manažerského hlediska. Manažerská formulace musí být následně převedena do zadání úlohy. V této fázi se rovněž provádí inventura zdrojů, hodnotí se možná rizika, náklady a přínos použitých metod, stanovuje se i předběžný plán prací (Berka, 2003).

3.2.2 Porozumění datům

Fáze porozumění datům je zahájena prvotním sběrem dat. Dalším krokem jsou činnosti, které umožní získat určitou základní představu o datech, jež jsou k dispozici (posuzuje se kvalita dat, vytipování zajímavých podmnožin apod.). Nejčastěji se zjišťují různé četnosti atributů, průměrné hodnoty, minima a maxima apod.

3.2.3 Příprava dat

Pro mnoho případů stačí pouze data extrahovat z databáze nebo jiného místa, kde se nacházejí a následně je použít. Naopak v dalších případech může být právě příprava dat klíčovým a nejtěžším úkolem. Typickým příkladem, který ve své knize Bramer (2007) uvádí, může být extrahování z přepisů výslechů svědků, podezřelých z nezákonných činností. Jejich správnost ovlivní celý soudní proces a proto je tato činnost nesmírně důležitá (Bramer, 2007).

Pro přípravu dat je tedy typické čištění dat, selekce dat, jejich transformace, vytváření, integrování a formátování dat. Většina autorů se shoduje na tom, že tato fáze je nejpracnější částí celého procesu získávání znalostí z databází.

3.2.4 Modelování

V tomto kroku se aplikují na data nejrůznější metody. Existuje řada metod pro řešení úloh, proto je klíčové vybrat tu nejhodnější a vhodně nastavit parametry. Doporučuje se používat více různých metod a kombinovat jejich výsledky (Berka, 2003). Může se stát, že použití metod povede k nutnosti data nějakým způsobem modifikovat, a tedy k návratu k přípravě dat z předcházející fáze.

K této fázi dále patří ověřování nalezených znalostí z pohledu metod. To může představovat například testování klasifikačních znalostí na nezávislých datech.

3.2.5 Vyhodnocení výsledků

V této fázi jsme se dostali do bodu, kdy jsme nalezli znalosti a ověřili je testováním na nezávislých datech. Tyto výsledky je ale ještě potřeba vyhodnotit z pohledu manažera, zda byly splněny cíle formulované při zadávání úlohy. Závěrem této fáze je přijmutí rozhodnutí o způsobu využití výsledků.

3.2.6 Využití výsledků

Nyní je nejdůležitějším krokem upravit získané znalosti do podoby použitelné pro zákazníka. Podle typu úlohy může využití výsledků na jedné straně znamenat prosté sepsání závěrečné zprávy, na straně druhé pak zavedení systému (hardwarové, softwarové, organizační) pro automatickou klasifikaci nových případů (Berka, 2003).

3.3 Metody

V této kapitole se zaměřím na čtyři metody data miningu. Je to korelační matice, shluková analýza, asociační pravidla a rozhodovací stromy. Metod je samozřejmě velké množství a pro každý případ se volí metoda nejlépe vyhovující dané situaci, ale pro mou práci jsem si zvolil metody víceméně nejznámější a ve stručnosti popíši jejich vlastnosti. Nebudu se zabývat matematickým vyjádřením algoritmů pro všechny metody, neboť jsou velice obsáhlé a poměrně složité. Mým cílem je zde nastínit vlastnosti vybraných metod a jejich principy. V prostředí RapidMineru tyto metody vykonává jediný blok a je proto nadbytečné zaobírat se zde algoritmy jako takovými.

3.3.1 Korelační matice

Korelační matice patří mezi nejzákladnější metody data miningu. Její hlavní využití je v případě, pokud hledáme vazby mezi některými z proměnných, za účelem jejich ovlivnění v budoucím rozhodování. Pokud vypočteme koeficienty korelace pro všechny dvojice atributů a sestavíme je do symetrické čtvercové matice R typu n x n, dostaneme korelační matici. Tyto koeficienty, vyjadřují sílu vazby mezi každými dvěma proměnnými. Rozlišuje se kladná a záporná korelace na základně znaménka u korelačního koeficientu.

Korelační koeficient nabývá hodnot z intervalu <-1; 1>, přičemž tyto krajní meze značí nejsilnější možnou lineární vazbu mezi dvěma proměnnými. Znamená to, že dvě proměnné jsou na sobě maximálně závislé a ovlivňují se jakoukoli změnou. Pokud je hodnota korelačního koeficientu velmi blízká nule, vyplývá z toho, že dané dvě proměnné na sebe působí zanedbatelnou silou, čili změna jedné proměnné se neprojeví na proměnnou druhou.

V případě kladné korelace je znaménko korelačního koeficientu kladné a značí, že pokud se jedna z proměnných zvětší, nastává zvětšení i u druhé proměnné. Pokud se první proměnná zmenší, nastává změna v podobě zmenšení i u druhé proměnné. Záporná korelace se vyznačuje záporným znaménkem u korelačního koeficientu a vyjadřuje, že pokud se první proměnná zvětší, druhá se bude zmenšovat a naopak.

3.3.2 Shluková analýza

Shluková analýza se vyznačuje seskupováním objektů, které jsou podobné ostatním objektům, ale zároveň odlišné od objektů patřících do jiného shluku (Bramer, 2007). Na první pohled jednoduchá úloha skrývá řadu problémů. Neexistuje jednoznačná definice podobnosti objektů a neexistuje ani jednoznačná definice shluku. Výsledky jsou většinou formulovány jen jako hypotézy o klasifikaci zkoumaných objektů. (Šarmanová, 2012) Shluková analýza nachází uplatnění i v ekonomické sféře (Bramer, 2007).

Například:

- Seskupovat země s podobnou výší důchodů.
- V marketingu například vytvářet shluky zákazníků se stejnou potřebou.
- Seskupovat korporace podle výše jejich obratů

Podle Stancu (2012) má tato technika následující výhody:

- Studování významných spojení mezi daty
- Uchovávání nejdůležitějších charakteristik z databází
- Shrnutí informací získaných z dat

Metody shlukové analýzy:

- Hierarchická metoda
- Metoda K-středů
- Hustotní metoda
- Mřížková metoda
- Modelová metoda

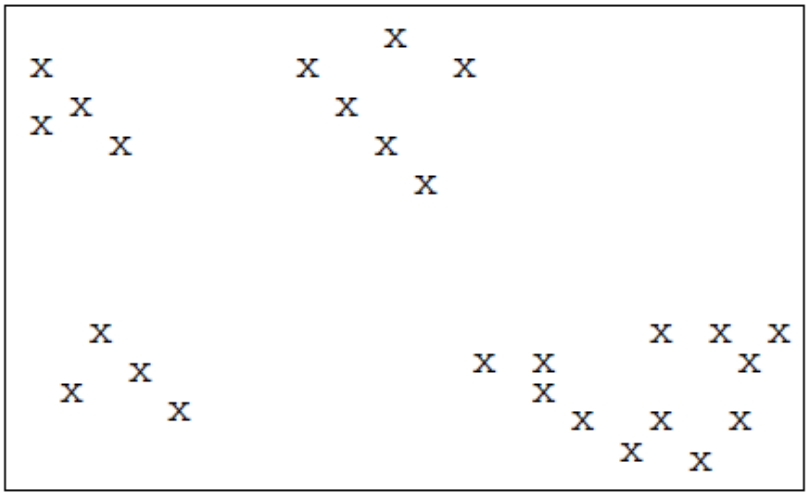
Metoda hierarchického shlukování je založena na principu „zdola nahoru“. Začíná se tedy v situaci, kdy každý objekt tvoří jeden samostatný shluk. Postupně se pak jednotlivé shluky spojují, až skončíme s jedním shlukem obsahujícím všechny objekty. Proces hierarchického shlukování většinou bývá zachycen v podobě tzv. dendrogramu. Ten ukazuje postupné spojování shluků počínaje očíslovanými objekty.

Metoda k-středů nebo k-means, je další metodou shlukové analýzy, kterou v roce 1967 uvedl John MacQueen. Jejím principem je přiřazení každému zkoumanému objektu právě jeden ze sady shluků. Prvním krokem je rozhodnutí kolik shluků chceme z dat udělat. Tato hodnota se označuje písmenem k a bývá řádově jednomístné celé číslo, ale existují i výjimky. Dalším krokem je zvolení k bodů, které jsou považovány za těžiště k shluků, přesněji řečeno za těžiště potencionálních shluků, které v tomto okamžiku nemají žádné objekty. Výběr těchto bodů je naprosto libovolný, ale pro správnou funkci metody se doporučuje vybírat body, které jsou od sebe poměrně daleko. Nyní můžeme přiřazovat objekt po objektu do shluků, které mají těžiště nejblíže danému objektu. Ve chvíli, kdy jsou všechny objekty přiřazeny do patřičných shluků, dostáváme právě k shluků, založených na původních k těžištích, které v tuto chvíli již nejsou pravými těžišti shluků (Bramer, 2007).

Souhrnně lze tedy napsat následující postup:

- Zvolíme hodnotu k (počet shluků)
- Zvolíme hodnotu k (počet bodů, představující těžiště k shluků)
- Přiřazujeme objekty do shluků, podle nejbližších bodů, dokud nejsou přiřazeny všechny objekty

Obrázek 3: Objekty shlukové analýzy



Zdroj: Bramer, (2007)

3.3.3 Asociační pravidla

Asociační pravidla jistým způsobem zkoumal již počátkem 90. let Agrawal (1993). Zabýval se analýzou nákupního košíku, při které zjišťoval, jaké druhy zboží si v supermarketech zákazníci kupují současně. Hlavní myšlenkou je tedy hledání vazeb (asociací) mezi různými položkami sortimentu prodejny. Přitom není upřednostňován žádný speciální druh zboží jako závěr pravidla (Berka, 2003).

V případě asociačních pravidel není žádný atribut (sloupec tabulky) vyčleněn jako cíl klasifikace. Asociační pravidla hledají „všechny zajímavé“ asociace (implikace, ekvivalence) mezi hodnotami různých atributů. (Šarmanová, 2012)

Využívají se algoritmy postavené na konstrukci IF – THEN, příkladem může být případ IF nezaměstnaný = ano THEN příjem = nízký apod. Kodratoff (1999) uvádí ve své knize další metody získávání znalostí, které jsou zaměřeny spíše na charakteristiky různých pravidel, než na algoritmy samotné.

Nejznámějším algoritmem hledání asociačních pravidel je algoritmus „apriori“. Tento algoritmus navrhl Agrawal v návaznosti na analýzu nákupního košíku (Agrawal, 1996). Podstatou je hledání často se opakujících množin položek. Jde o kombinaci kategorií, které dosahují předem zadané četnosti v datech (Berka, 2003).

Algoritmus „apriori“ můžeme popsat následovně:

První průchod algoritmem spočítá výskyt jednoprvkových množin a určí jednoprvkové frekventované množiny. Druhý průchod k se pak skládá ze dvou fází. Nejprve se frekventovaná množina L_{k-1} nalezená k-1 průchodem použije pro vygenerování kandidátských frekventovaných množin C_k. Pak se prochází databáze a pro jednotlivé kandidátské množiny je spočítána podpora. Aby se urychlil tento výpočet, potřebují se rychle a efektivně identifikovat kandidáti z množin C_k v dané transakci. (Šarmanová, 2012)

Další metodou je metoda GUHA (General Unary Hypothesis Automaton), která pochází z Čech, kde na ní v 80. letech 20. století pracovali Hájek, Havránek a Chytil. Tato metoda dokáže pracovat s neúplnou informací a vytváří hypotézy na základě empirických dat (Burda, 2004).

Její princip byl formulován již v roce 1966 jako „pomocí počítače generovat všechny hypotézy zajímavé na základě daných empirických dat“. Zajímavost je dána logickým tvarem hypotézy a způsobem, jakým ji podporují data. Data mají tvar obdélníkové matice, její řádky odpovídají objektům a sloupce atributům. Ve většině implementací mají hypotézy tvar „ϕ souvisí s ψ“, kde ϕ, ψ jsou logické kombinace atributů. Souvisení je dáno pomocí nějakého zobecněného kvantifikátoru, často daného nějakou statistikou užívanou při testování statistických hypotéz (Hájek, Havel, Chytil, 1966).

GUHA je svým způsobem jedna z prvních metod data miningu, ale zůstala celkem neznámá, proto se o ní v nynější literatuře většinou ani nedočteme. Svým způsobem je to škoda, protože teorie samotné metody obsahuje mnoho podnětů, přístupů a skutečností, na které data mining postupně přicházel sám.

3.3.4 Rozhodovací stromy

Způsob prezentování znalostí v podobě rozhodovacích stromů je dobře znám z řady oblastí. Indukce rozhodovacích stromů patří k nejzákladnějším algoritmům z oblasti symbolických metod strojového učení (Berka, 2003).

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 dokáží záznamy rozdělit a snižují tak nejistotu.

Podstatou této metody je rozdělování zkoumaných dat na stále menší a menší podmnožiny (uzly stromu) tak, aby v těchto podmnožinách převládaly příklady jedné třídy (Berka, 2003). Uzel na nejvyšší úrovni je označován pojmem kořenový. Větví nazýváme možný výsledek testu, externí uzly označované jako listy reprezentují jednotlivé třídy. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu, který odpovídá zařazení příkladu do třídy.

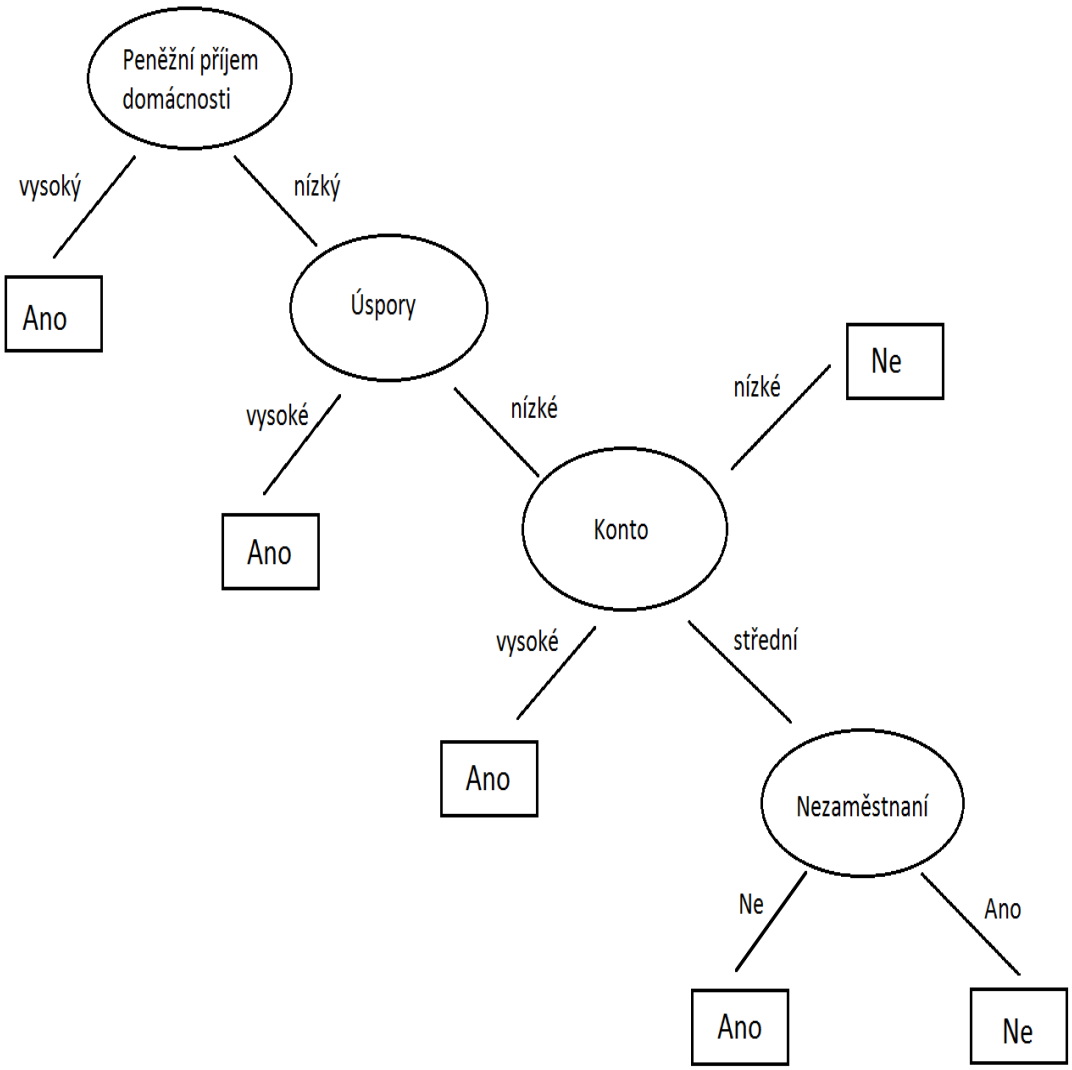
Konečným výsledkem je poté strom s rozhodovacím uzlem a s koncovým uzlem. Rozhodovací uzel má dvě a více větví a koncový uzel reprezentuje rozhodnutí nebo klasifikaci.

Berka (2003) se ve své práci zmiňuje o obecném schématu algoritmu pro tvorbu rozhodovacích stromů. O tomto algoritmu hovoří jako o postupu, který je často nazýván „top down induction of decision trees (TDIDT)“ a specifikuje ho následovně:

- Zvol jeden atribut jako kořen dílčího stromu.
- Rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu.
- Existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči.

Klíčovou otázkou, kterou si stále odborníci kladou, zůstává, jak vybrat vhodný atribut pro větvení stromu.

Obrázek 4: Úplný rozhodovací strom



Zdroj: vlastní zpracování, na základě podkladů (Berka, 2003)

3.3.5 Další metody

Výše uvedenými metodami škála rozhodně nekončí. V dnešní době se nabízí velké množství dalších metod, které mají svá uplatnění v různých odvětví, nemusí se jednat vůbec o ekonomický sektor. Data mining se v určité podobě používá například v informatice, kde může napomáhat při různých testech nebo dokonce i v kriminalistice, kde pomáhá predikovat nezákonné delikty. Jak již bylo řečeno v úvodu, oblast použití se

s vývojem technologií a postupem času neustále rozšiřuje a zaujímá svou nepopiratelnou část jako nástroj, používaný stále větším počtem společností.

4 Praktická část

V praktické části se zaměřím na využití některých metod, které byly popsány výše, abych demonstroval jejich použití v praxi a přiblížil tím tak jejich podstatu. Mým cílem není objevení nových vztahů ani jiných vazeb, ale spíše srozumitelné seznámení s problematikou data miningu a činností, které jsou s tímto tématem spjaty. Dojde na využití softwaru RapidMiner, o kterém jsem hovořil na začátku své práce a také k interpretování výsledků, které získám při využití metod.

4.1 Data

Data použitá pro tuto práci jsem získal na webových stránkách Eurostatu, což jsou internetové stránky, obsahující statistiky nejrůznějších charakterů o všech zemích Evropské Unie (http://ec.europa.eu/eurostat/data/database). Nachází se tam velké množství informací o všech zemích EU, ale také o dění ve světě a vlivech těchto skutečností na evropské státy.

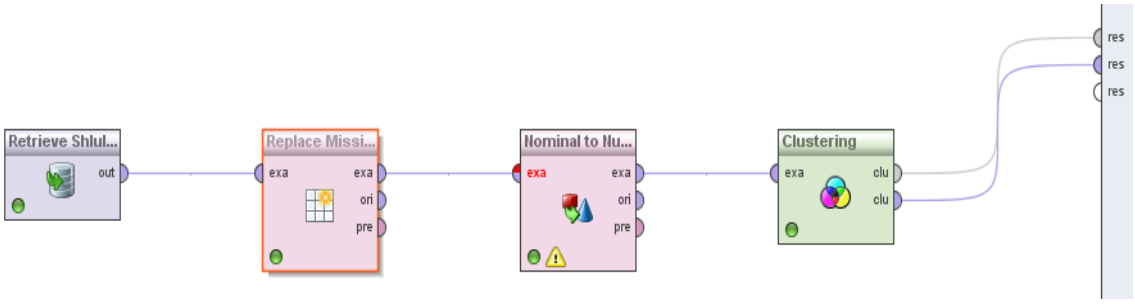
Zvolil jsem data, která odráží makroekonomické ukazatele České republiky a Rakouska za období od roku 2005 do roku 2014, která použiji pro korelační matice a souhrnná ekonomická data všech členských zemí Evropské Unie za rok 2013, na kterých ukáži princip shlukové analýzy. V případě shlukové analýzy se stalo, že nebyla dostupná všechna data pro všechny země, proto tam, kde byla data nedostupná, musí být pomocí bloku „Replace missing values“ nahrazena. Tento blok nabízí nahrazení chybějících údajů nulou, minimem nebo maximem z údajů pro daný atribut nebo jejich průměrem. Tato skutečnost ovlivňuje i výsledky, protože ve své podstatě neodráží skutečný stav.

Nejdůležitější je dobře si data připravit, to znamená, projít si je, pokud možno smazat přebytečné informace, které nejsou žádoucí, vybrat pouze ta data, která potřebujeme pro práci a zajistit jejich správný formát. Tyto činnosti lze provést i v samotném RapidMineru, ale pokud je to možné doporučuji data si takto připravit, před samotným zahájením procesu. Předchází se tak zbytečným komplikacím, které vedou k neefektivní práci a protahovaní celého procesu. Na druhou stranu jsme jenom lidé, a pokud přehlédneme prázdné buňky nebo chybné znaky, bloky v RapidMineru zajišťující jejich nalezení a eliminování nám budou velice užitečné. S takto připravenými a přetříděnými daty je možné přistoupit s samotným metodám a ukázat je na názorných příkladech.

4.2 Shluková analýza

Jednou z metod, které ve své práci uvedu, bude shluková analýza. Její formulace pomocí RapidMineru je znázorněna na obrázku níže. Jako vždy jsou na místě prvního bloku vstupní data, která jsou použita pro daný proces. V mém případě za daty následuje blok s názvem „Replace missing values“, který zajišťuje zmíněné nahrazení chybějících dat, kvůli bloku „Clustering“, který s nimi neumí pracovat a vrací chybové hlášení. Následuje blok „Nominal to Numerical“, který zajistí, že veškerá data, která přichází na vstup, budou zpracována a pokud některá z nich nejsou číselná, budou převedena tak, aby z výstupu vycházely pouze číselné hodnoty. Poslední blok, který nese název „Clustering“, což lze přeložit jako shlukování, zajišťuje právě tu nejdůležitější část celého procesu a to přiřazení veškerých vstupních dat do patřičných shluků (clusterů). Ve vlastnostech tohoto bloku je možnost nastavit počet těchto shluků a tím zajišťovat podrobnější rozdělení dat.

Obrázek 5: Sestavení bloků shlukové analýzy v prostředí RapidMineru



Zdroj: vlastní zpracování

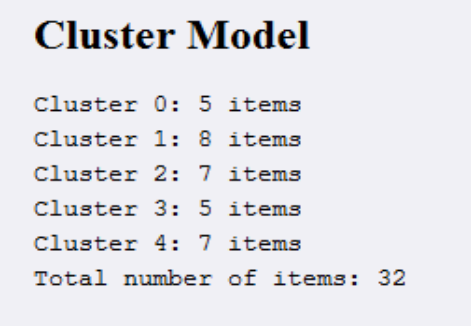
Data použitá pro tuto metodu pochází z databáze Economy and finance, dostupná z této stránky: http://ec.europa.eu/eurostat/data/database. Tato databáze obsahuje nejrůznější ekonomické a finanční údaje všech zemí Evropské Unie za různá časová období.

Pro svůj příklad jsem si sestavil datový soubor obsahující jednotlivé země EU a k nim čtrnáct atributů odrážející jejich ekonomiku a finance za rok 2013. Atributy jsou následující: celkový státní příjem (% z HDP), směnný kurz, index cen nemovitostí, počet letišť, tvorba hrubého kapitálu podle odvětví, výdaje domácností, zaměstnanost podle odvětví, splavné vodní cesty, vládní příjmy, vládní výdaje, index cen spotřebitelů, platební bilance, investice a obchod se službami. Údaje je možné stáhnout v různých datových formátech. Eurostat nabízí klasický excelový dokument, csv nebo tsv dokument nebo také pdf a html dokumenty, které však nejsou vhodné pro další práci s údaji, ale spíše pro jejich prezentování.

V bloku „Replace missing values“ jsem zvolil nahrazování chybějících údajů průměrem z dostupných údajů pro daný atribut, aby bylo zajištěno provedení rozřazení. Celkově chybělo v datech přibližně 10% údajů. Některé země poskytovaly údaje u všech atributů, u jiných chyběly např. dva údaje. Dále jsem v bloku „Clustering“ zvolil třídění do pěti shluků. Defaultně je v tomto bloku nastaveno třídění na dva shluky.

Po spuštění procesu se dostáváme k výsledkům a vlastnímu rozřazení jednotlivých zemí do patřičných shluků.

Obrázek 6: Počet zemí v jednotlivých shlucích



Zdroj: vlastní zpracování.

Celkově jsem pracoval s 32 zeměmi. Jak byly země rozřazeny je vidět na obrázku 6. Při tomto dělení je patrné, že nejvíce zemí, které mají podobná data, patří do shluku 1 resp. (clusteru 1). Naopak nejméně zemí je v shlucích 0 a 3. Toto rozdělení je však tvořeno na základě dat, která jsem vybral. Země v těchto shlucích jsou si podobné svými celkovými státními příjmy (% z HDP), směnovými kurzy vůči euru, počtem letišť, indexem cen nemovitostí atd. Určitý vliv zde má i skutečnost, že data nebyla úplná, to znamená, že byla vyhodnocena na základě chybějících údajů, které by v případě jejich doplnění mohly změnit konečnou strukturu jednotlivých shluků.

Obrázek 7: 5 shluků obsahující roztříděné země

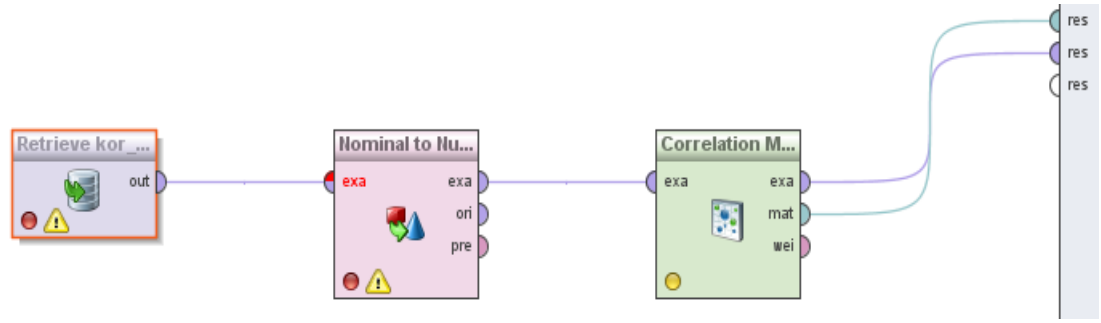
Zdroj: vlastní zpracování

Na obrázku výše jsou vidět jednotlivé země roztříděné do pěti shluků. Pokud by došlo k nějakým změnám v celém procesu, změní se i výsledná struktura. Například pokud by se změnil počet shluků nebo zredukoval počet proměnných nebo vypočítaly odvozené proměnné např. na jednoho obyvatele apod.

4.3 Korelační matice

Druhou metodou bude korelační matice. Její sestavení v prostředí RapidMineru není nijak složité. Opět se začíná vstupními daty, za kterými následuje blok „Nominal to Numerical“ a za ním blok, zajišťující vytvoření samotné matice „Correlation Matrix“. Mezi první dva bloky by se mohl případně přidat i blok „Select Attributes“, ale v mém případě to nebylo nutné, protože veškerá data, která jsem měl připravená, jsem chtěl použít. V mém případě byla data úplná a nechybí v nich žádný údaj. Samotné sestavení pro můj příklad vypadá následovně.

Obrázek 8: Sestavení bloků korelační matice



Zdroj: vlastní zpracování

Pro tento příklad jsem použil stejné makroekonomické údaje o České republice a Rakousku. Vytvořím korelační matici pro každou zem, aby byly patrné rozdíly v závislostech mezi stejnými atributy. Údaje jsem opět získal na stránkách Eurostatu, přesněji v sekci „Tables on EU policy“. Jedná se o časové řady údajů za roky 2005-2014. Během těchto let se pozorovalo devět makroekonomických ukazatelů, kterými jsou: míra nezaměstnanosti, nominální index cen nemovitostí, míra aktivity (15-64 let), míra nezaměstnanosti mladých lidí, podíly na vývozních trzích, aktuální zůstatek na účtu, míra zaměstnanosti, index cen nemovitostí a míra populace ohrožena chudobou nebo sociálním vyloučením. Stejně jako v případě dat pro shlukovou analýzu, i zde jsem zvolil formát .xls, který jsem následně importoval do prostředí RapidMineru za pomoci průvodce.

Po spuštění procesu se nám zobrazí výsledné korelační matice, které jsou na obrázcích níže. Z tabulek je patrné, které atributy na sebe působí nejvíce, čili hodnota jejich korelačního koeficientu je blízká hodnotě 1 nebo -1 a naopak, které na sebe nepůsobí prakticky vůbec a hodnota jejich korelačního koeficientu je blízká hodnotě 0. Například jednu z nejsilnějších korelačních vazeb, v tabulce pro Českou republiku, mezi sebou mají

atributy celková míra nezaměstnanosti a míra nezaměstnanosti mladých lidí, což je logické, protože tyto dva atributy se mění souběžně a změna jednoho atributu se zákonitě projeví i na atributu druhém. Další silnou vazbou je vztah mezi aktuálním zůstatkem na účtu a mírou aktivity lidí ve věku 15-64 let. V těchto případech se jedná o kladnou korelaci, která vyvolá stejnou změnu u obou atributů, například poroste-li nezaměstnanost mladých lidí, poroste i celková nezaměstnanost. V druhém případě, kdy hovoříme o záporné korelaci, se jako nejsilnější ukázaly atributy roční procentní změna nominálního indexu cen nemovitostí a míry nezaměstnanosti mladých lidí. Naopak korelaci, jejíž korelační koeficient je blízký hodnotě 0 jsem objevil mezi atributy roční procentní změnou nominálního indexu cen nemovitostí a mírou populace ohroženou chudobou nebo sociálním vyloučením.

Obrázek 9: Výsledná korelační matice pro Českou republiku

Current acc...	Export mark...	House pric...	Unemploy...	Nominal ho...	Employmen...	Activity rate (...	Youth unem...	People at ri...
1	-0.354	-0.384	0.191	-0.408	0.055	0.779	0.360	0.102
-0.354	1	0.583	-0.145	0.645	0.614	-0.708	-0.415	0.728
-0.384	0.583	1	-0.653	0.548	0.770	-0.298	-0.799	0.251
0.191	-0.145	-0.653	1	-0.678	-0.410	0.178	0.942	0.355
-0.408	0.645	0.548	-0.678	1	0.335	-0.626	-0.821	-0.007
0.055	0.614	0.770	-0.410	0.335	1	-0.115	-0.499	0.620
0.779	-0.708	-0.298	0.178	-0.626	-0.115	1	0.369	-0.255
0.360	-0.415	-0.799	0.942	-0.821	-0.499	0.369	1	0.168
0.102	0.728	0.251	0.355	-0.007	0.620	-0.255	0.168	1

Zdroj: vlastní zpracování

Obrázek 10: Výsledná korelační matice pro Rakousko

Current acc...	Export mark...	House pric...	Unemploy...	Nominal ho...	Employmen...	Activity rate (...	Youth unem...	People at ri...
1	0.437	-0.675	-0.479	-0.476	0.413	-0.371	-0.350	-0.159
0.437	1	-0.386	0.175	-0.850	0.359	-0.954	0.357	-0.719
-0.675	-0.386	1	0.291	0.321	-0.509	0.230	0.270	0.031
-0.479	0.175	0.291	1	-0.132	-0.482	-0.145	0.890	-0.449
-0.476	-0.850	0.321	-0.132	1	-0.088	0.878	-0.416	0.383
0.413	0.359	-0.509	-0.482	-0.088	1	-0.372	-0.591	-0.354
-0.371	-0.954	0.230	-0.145	0.878	-0.372	1	-0.332	0.679
-0.350	0.357	0.270	0.890	-0.416	-0.591	-0.332	1	-0.383
-0.159	-0.719	0.031	-0.449	0.383	-0.354	0.679	-0.383	1

Zdroj: vlastní zpracování

Pokud porovnáme obě tabulky vzájemně, jsou patrné změny ve velikostech koeficientů u všech atributů. Za zmínku stojí, že údaje, u kterých se v případě České republiky našla nejsilnější kladná korelace, jsou totožné s údaji pro Rakousko, ačkoliv se jejich korelační koeficient mírně liší. Naopak nejsilnější záporná korelace se v případě Rakouska objevila mezi atributy podíly na vývozních trzích a mírou aktivity (15-64 let). V některých případech se změnila i polarita korelačního koeficientu. Tímto jsem poukázal na skutečnost, že každá země má odlišné závislosti, které jsou zapříčiněny například politikou státu, mentalitou obyvatel a podobnými vlivy.

Stejně tak jako hodnota korelačního koeficientu, i barevný odstín daného políčka napomáhá rozeznat sílu vazby mezi atributy, přičemž nejtmavší políčka odpovídají nejsilnějším korelacím (kladným i záporným) a naopak nejsvětlejší odpovídají nejslabším. To umožňuje dobrou orientaci pro uživatele.

5 Výsledky práce

V této kapitole předvedu výsledky, ke kterým jsem došel ve vlastním zkoumání, ale i výsledky jiných autorů, kteří se nějakým způsobem zabývali data miningem v ekonomice. Dále vysvětlím význam svých výsledků a navrhnu jejich využití v budoucnosti.

5.1 Výsledky jiných autorů

Nejdříve bych rád uvedl výsledky, ke kterým došli jiní autoři při svém zkoumání dat. Například Stancu a kol. (2012), kteří se zabývali rumunskými makroekonomickými daty. Aplikovali na tato data tři metody, analýzu hlavních komponent, faktorovou analýzu a shlukovou analýzu. Makroekonomická data obsahovala údaje o hrubém domácím produktu Rumunska a Evropské unie, o exportu, importu, populaci a o směnném kurzu.

Výsledkem analýzy hlavních komponent je podle autorů skutečnost, že informace může být v hlavní komponentě zachována na 97,39%, což znamená, že vztahy mezi proměnnými jsou velice silné a mohou být prezentovány v novém neformálním prostoru skrze komponentu. Dále autoři tvrdí, že všechny původní proměnné, které použitý data set obsahoval, lze vyjádřit prostřednictvím první komponenty s 2,61% rizikem ztráty informací. Další poznatky této metody jsou:

- Export a import rostou současně
- Export a HDP na evropské úrovni mají tendenci růstu ve stejném směru
- Pokud úroková míra bude nadále klesat, export bude růst

Výsledkem faktorové analýzy je možnost zapsat faktor jako lineární kombinaci makroekonomických indikátorů bereme-li v úvahu jejich přidružené koeficienty. Autoři přišli na to, že rumunský hrubý domácí produkt bude nejvíce ovlivněn změnou úrokové míry. Pokud se tato míra jednoprocentně zvětší, hrubý domácí produkt klesne řádově o stovky milionů euro.

Shluková analýza je užitečná při třídění dat, která mají podobné vlastnosti. Autorům se podařilo na základě korelační matice objevit devět komponent, které syntetizují původní proměnné. Objevili tři shluky (clustery), přičemž druhý shluk byl největší.

Další takovou prací, zabývající se ekonomickým pohledem na data mining je práce Kleinberga a kol. (1998), kteří se zaměřili na mikroekonomický pohled na data mining. Zajímalo je vytvoření přesného rámce pro automatické hodnocení data miningových operací. Za tímto účelem, navrhli sadu zásad a idejí a vytvořili rámce čtyř odlišných stylů. Jako příklad teorií, metodik a nástrojů, které mohou být vyvinuty, uvádí jejich rámec optimalizující problémy s koeficienty nelineárně závislými na datech a jejich definici zajímavostí uvnitř tohoto rámce.

Záměrně se však vyhýbají uvádění a dokazování přesných výsledků, kvůli příliš širokému potencionálnímu rozsahu zajímavých teorémů, které jsou přímočarými aplikacemi těchto myšlenek.

5.2 Vlastní výsledky

Vlastní výsledky, ke kterým jsem došel pomocí použití metod shlukové analýzy a korelační matice, neobjevují žádné nové nebo dosud neobjevené vztahy a skutečnosti, ale pouze reflektují vlastnosti použitých dat. Za stěžejní část celé práce považuji právě data samotná a jejich přípravu pro použití v metodách. Málokdy naleznete přesně taková data, jako potřebujete pro svou práci, a tak nezbývá než si je upravit podle vlastních potřeb. Tato činnost je časově náročná a ne příliš záživná, jedná se hlavně o nalezení těch správných dat, jejich vytřídění, formátování a v neposlední řadě načtení do prostředí RapidMineru. Celkově považuji právě práci s daty za nejnáročnější část z celé práce, na které jsem strávil nejvíce času.

Výsledky získané pomocí shlukové analýzy představují rozřazení zemí Evropské Unie do pěti shluků na základě vstupních dat. Pokud bychom hledali nějaké řešení pro vylepšení nebo růst například celkových příjmů, které jsou jedním z atributů dat, je vždy nutné nahlížet na každý shluk zvlášť a pro každý takový shluk tvořit samostatné strategické plány, protože co platí pro jeden shluk, nemusí platit pro druhý. Ve výsledcích je patrné, že svou roli hrají i chybějící údaje o některých členských zemích.

V případě korelačních matic jsou výsledky celkem jasně viditelné. Odlišnosti obou zemí jsou patrné z tabulek. Nejsilnější korelaci u údajů pro Českou republiku jsem objevil u celkové míry nezaměstnanosti a míry nezaměstnanosti mladých lidí. Z výsledků je

patrné, že tato korelace je téměř rovna jedné, z čehož vyplývá, že jakákoliv změna, na jednom z atributů se zásadně projeví i na atributu druhém.

Tyto matice jsou pouze jednou z částí při tvorbě budoucích strategií, není proto vhodné tvořit závěry pouze na jejich výsledcích. Mohou mít pouze informativní účely nebo být použity pro další hlubší šetření.

6 Závěr

V budoucím světě bude data mining nacházet stále širší uplatnění v mnoha oblastech. Bude se dále rozvíjet a pronikat do stále většího spektra podniků, kde bude mít nejrůznější uplatnění. Postupem času se stane nezbytným nástrojem většiny podniků díky rozmanitým lidským potřebám a stále se rozrůstajícímu sortimentu zboží a služeb. Zařadí se do řad marketingových strategií na zákazníky, kde v dnešní době působí například slevy, množstevní akce, dárky k nákupům nad určitou hodnotu a podobné zdánlivě lákavé dárky.

Cílem této bakalářské práce bylo seznámení se s problematikou data miningu jako takového, s některými metodami, vybrat si libovolný software a zpracovat příklad s využitím některé z výše uvedených metod. V teoretické části jsem poukázal na autory různých publikací, kteří se zajímali o data mining v rámci ekonomiky a na jejich výsledky, ke kterým se dopracovali. Popsal jsem částečně i RapidMiner, software, který jsem si nakonec zvolil pro svou práci, jeho prostředí a nástroje, ale pouze do té míry, která byla nezbytná pro názornou ilustraci principu metod. Tento software skrývá mnoho dalších užitečných funkcí a možností, které nemusí být využity pouze pro potřeby data miningu. Proto jsem uvedl jen základní popis uživatelského prostředí a v praktické části pak popsal jednotlivé kroky, které předcházely vytvoření modelové metody. Celkově mě práce v jeho uživatelském prostředí bavila, protože není nijak složité na ovládání, pokud má uživatel alespoň určitou rámcovou představu o tom, co od tohoto programu očekává.

V praktické části jsem předvedl použití metod na reálných datech, která se týkala české ekonomiky za minulé roky a souhrnných údajů o členských zemích Evropské Unie, která jsem získal na webových stránkách Eurostat. Využil jsem metod, které jsem popisoval výše, abych ještě více objasnil jejich smysl a tím nabídl hlubší pochopení této problematiky. Výsledky, ke kterým jsem se dopracoval v praktické části, mohou být použity pro další šetření podniků nebo jen pro informativní účely.

Jak jsem již poukazoval dříve, cílem této práce není objevení nových vztahů nebo dosud nezjištěných skutečností, ale pouze hlubší porozumění procesu získávání znalostí z databází, vymezení pojmů a názorný příklad použití některých metod.

I. Summary and keywords

This bachelor thesis is specialized in sorting enormous amount of data using several methods. Data mining is a process consisting in collecting knowledge from databases or data warehouses. Data can come from service providers or it can be a local database of certain company or national data which are focused on economic research. Data used in this bachelor thesis come from freely available website Eurostat.

The theoretic part is dedicated to data itself, their format and methods and RapidMiner, which is the tool for sorting data. The practical part is focused on real methods which are built in this software. In the end results are interpreted. Every single method shows different results and this results can be used for management or future economic decision to help business economics.

Keyword: Data mining, databases, sorting data, RapidMiner, Data mining methods

II. Seznam použitých zdrojů

Agrawal, R., Arun S., Tomasz I. (1993). *Mining association rules between sets of items in large databases.* S. 207-216. ISBN 0-89791-592-

Berka, P. (2003). *Dobývání znalostí z databází*. Praha: Academia, 2003. 366 s. ISBN 80-200-1062-9.

Burda, M. (2004). *Asociační pravidla (metoda GUHA)*[online]. [cit. 2016-03-14] Dostupné z: http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/GUHA

Bramer, M. (2013). *Principles of Data Mining.* London: Springer. ISBN 978-1-4471-4883-8.

Datamind [online].[cit. 2016-2-27] Dostupné z: http://www.datamind.cz/cz/blog/Data-mining-zdarma-rapid-miner-v-praxi

eCommerce software. http://sourceforge.net/projects/rapidminer

Eurostat database. Dostupné z: http://ec.europa.eu/eurostat/data/database

Feelders, A., (2002). *Data Mining in Economic Science* [online]. [cit. 2016-2-15] Dostupné z: http://www.staff.science.uu.nl/~feeld101/dmecon.pdf

Frank, E., Hall, M. A., Witten, I. H. (2011). *Data Mining, Practical Machine Learning Tools and Techniques,* 3rd edition. Dostupné z: http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/Texts/[7]%20Data%20Mining%20-%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20(3rd%20Ed).pdf

Frawley, W. J., Christopher J. M., Gregory P. (1991). *Knowledge Discovery in Databases: An Overview* [online]. [cit. 2016-3-15] Dostupné z: http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929

Górecky, J. (2011). *Data Miner (referát)*. Ostrava: FEI VŠ-TUO.

Hand, D. J. et al. (2001). *Principles of Data Mining.* Boston: The MIT Press. ISBN 026208290x.

Hájek, P., M., K. Chytil, T. Havránek (1983). *Metoda GUHA: automatická tvorba hypotéz.* Praha: Academia.

Inmon, W. H., Claudia I., R. H. Terdeman (1999). *Exploration Warehousing*. ISBN 0471374733

Kleinberg, J., C. Papadimitriou, and Raghavan, P (1998). *A Microeconomic View of Data Mining. Data Mining and Knowledge Discovery*. Dostupné z: http://link.springer.com/article/10.1023/A:1009726428407#page-1

Kodratoff, Y. (1999). *Comparing Machine Learning and Knowledge Discovery in Databases*: An Application to Knowledge Discovery in Texts. *Lecture Notes on AI (LNAI) - Tutorial series*.

Kotlárová, J. (2010). *Doporučovací systémy pro internetové obchody* (diplomová práce). Brno: MUNI-FI.

Metody dobývání znalostí [online]**.** [cit. 2016-2-25] Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody#a-pravidla

Procházka, M. *Data mining – jiný pohled na problém.* **VTM** [online]**.** [cit. 2016-2-28] dostupné z: http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem

Rapid Miner software. http://rapidminer.com

Stancu, **S., a kol. (2012).** *Using Data Mining Techniques in Macroeconomic Analysis on Romania's Case.* In: Latest Advances in Information Science, Circuits and Systems, Series 4, s. 116-121

Symons, V. (2000). *Three tiered Data Warehouse structure.* White Paper. DM Review.

Šarmanová, J. (2012). *Metody analýzy dat.* Studijní text, VŠB-TU, Ostrava.

Theoretical Foundations of Data Mining [online]**.** [cit. 2016-2-27] Dostupné z: http://www.tutorialspoint.com/data_mining/dm_themes.htm

Úvod do data miningu [online]. [cit. 2016-3-18] Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf

Wikipedia: http://en.wikipedia.org/wiki/RapidMiner

Williams, G. (2001). *Data mining with Rattle and R.* **Springer: New York.**

YouTube video: http://www.youtube.com/user/RapidIVideo

III. Seznam obrázků a tabulek