Česká zemědělská univerzita v Praze
Provozně ekonomická fakulta

# Dataminingové techniky analýz vícerozměrných datových souborů

Disertační práce z oboru Systémové inženýrství

**Julie Poláčková**

Školitelka: prof. Ing. Libuše Svatošová, CSc.

Ráda bych na tomto místě poděkovala všem kolegům z katedry statistiky za cenné rady, kterými přispěli k vytvoření předkládané disertační práce. Především děkuji své školitelce prof. Ing. Libuši Svatošové, CSc. za odborné vedení v průběhu studia i při zpracování této práce.

Julie Poláčková

# Dataminingové techniky analýz vícerozměrných datových souborů

**Klíčová slova**

Data Mining, vícerozměrné datové soubory, shluková analýza, segmentace, SAS Enterprise Miner, IBM SPSS Modeler, transakční data, nákupní koš.

**Abstrakt**

Cílem předkládané disertační práce je zhodnocení různých metodických přístupů k analýze velkých datových souborů a vytvoření obecného metodického rámce pro segmentaci zákazníků pomocí zvolených dataminingových algoritmů. Postup modelování je demonstrován na příkladu segmentace nákupních košů zvoleného hypermarketu. Vlastní část práce zachycuje různé přístupy k přípravě dat, k modelování i ke zhodnocení výsledných shlukovacích modelů. Optimálním algoritmem pro segmentaci zákazníků byl zvolen postup využívající metodu *k*-průměrů aplikovanou na expertně vybrané a upravené podílové vstupní proměnné. Uvedená metoda *k*-průměrů prokázala ve všech realizovaných přístupech nejlepší segmentační vlastnosti, z čehož vyplývá, že ji lze obecně doporučit k realizaci segmentace zákazníků dle jejich nákupního chování. Výstupem práce je vytvoření doporučeného obecného postupu segmentace. Jde o určitý návod k realizaci segmentace transakčních údajů pro podporu marketingového rozhodování. Při tvorbě tohoto postupu byl kladen důraz především na jeho význam při praktickém využití.

# Data Mining Application in Multivariate Data Sets

**Keywords**

Data Mining, Multidimensional Data Sets, Cluster Analysis, Segmentation, SAS Enterprise Miner, IBM SPSS Modeler, Transactional Data, Shopping Basket.

**Abstract**

The aim of the dissertation is to evaluate different methodological approaches to the analysis of large data sets and to develop a general methodological framework for customer segmentation using selected data mining algorithms. The process of modelling is demonstrated by the example of segmentation of shopping baskets in a chosen hypermarket. The main part of the work describes different approaches to data preparation, modelling and evaluation of the resulting cluster models. The *k*-means method applied to expertly selected input variables has been selected as an optimal algorithm for customer segmentation. This method reached the best segmentation quality, which means that it can be generally recommended for customer segmentation modelling. The outcome of this dissertation is a recommended process for customer segmentation. It is a general instruction for the implementation of the transaction data segmentation to support marketing decision makers. In developing this procedure, the focus was mainly on practical applications.

# Obsah

# 1 ÚVOD

Data mining představuje proces extrakce inteligentní informace z velkého množství surových dat. Efektivním a maximálním využitím dostupných dat v organizaci se zabývá tzv. Knowledge Management neboli řízení znalostí. Tato vědní disciplína pokrývá zejména ty procesy v organizaci, které jsou synergií možností informačních technologií při zpracování dat a informací s tvůrčí a inovativní schopností lidských jedinců. V rámci procesu objevování znalostí v databázích jsou využívány postupy statistické analýzy dat, induktivního učení či dataminingu. Důležitost využívání informací zdůrazňuje i Kotler (2007) ve své nejznámější knize Marketing Management. Upozorňuje na skutečnost, že mezi daty, informacemi, znalostmi a moudrostí existují obrovské rozdíly. Pokud data nejsou zpracována v informace, které se transformují ve znalosti a později přemění v tržní moudrost, značná část jich přijde nazmar.

S nástupem počítačů získává data mining nad klasickým statistickým přístupem převahu. V nedávné době se nashromáždily ohromné datové sklady, které byly analyzovány pomocí klasických statistických postupů. Velké datové soubory však vyžadovaly nový a efektivnější přístup sloužící k podpoře rozhodování. Z tohoto důvodu se začaly rozvíjet algoritmy založené na umělé inteligenci spíše než na klasickém Fisherovském parametrickém modelu. V současné době se stále rozvíjejí nové a sofistikovanější techniky dataminingu, např. se jedná o metody napodobující fungování lidského mozku (NISBET et al. 2009).

Data mining tedy nevznikl jako nová akademická disciplína, ale jako reakce na technologický rozvoj v obchodní sféře. Primárním cílem bylo zefektivnit kombinaci využívání počítačů a datových zdrojů. Zaměřuje se na nalézání vhodných algoritmů sloužících k získání nových vzorů ze skutečných záznamů v databázi pomocí kombinace tradičních statistických analýz, umělé inteligence a technik strojového učení.

Rozdíly mezi dataminingem a klasickým statistickým přístupem popsal Nisbet et al. (2009) v knize *Handbook of Statistical Analysis and Data Mining Applications*. Tradiční statistické analýzy využívají minulé informace k určení budoucího stavu určitého systému, neboli k predikci, zatímco dataminingové analýzy využívají minulé informace k sestavení vzorů, které nevycházejí výhradně ze vstupních dat, ale také z logických souvislostí mezi těmito daty. Tento proces je také nazýván predikce, ale obsahuje nezbytné prvky, které chybí

statistické analýze. Jde o schopnost řádně vyjádřit, jaký by mohl být budoucí vývoj v porovnání se stavem minulým (v závislosti na předpokladech statistických metod). Souhrnně řečeno, data mining může poskytnout komplexnější porozumění datům pomocí nalezení vzorů, které dříve nebyly rozpoznány, a může vytvořit prediktivní modely, které se stanou užitečným nástrojem pro podporu rozhodování.

# 2 CÍL DISERTAČNÍ PRÁCE

Cílem disertační práce je zhodnocení různých metodických přístupů k analýze velkých datových souborů a vytvoření obecného metodického rámce pro segmentaci zákazníků pomocí zvolených dataminingových algoritmů. Postup modelování bude demonstrován na příkladu segmentace nákupních košů zvoleného hypermarketu. Nákupním košem se v této práci rozumí všechny položky z kategorie potravin (definované dle produktového listu v příloze č. 1) zaplacené zákazníkem v rámci jedné platby.

Hlavní cíl práce lze rozdělit do dílčích kroků dle realizovaného dataminingového procesu:

- Navrhnout vhodnou strukturu vstupní datové matice, která povede ke splnění definovaných cílů.
- Zvolit patřičné úpravy proměnných vstupujících do modelování, tak aby bylo dosaženo optimální segmentace zákazníků.
- Zhodnotit kvalitu vstupních dat dle předpokladů využitých modelů.
- Zvolit optimální algoritmus pro segmentaci zákazníků dle dostupných kritérií shlukování.
- Provést profilaci a logické zhodnocení získaných segmentů.
- Posoudit adekvátnost využitých softwarových nástrojů.

Adekvátnost využitých dataminingových systémů bude hodnocena na základě předem stanovených kritérií:

- náročnost instalace systému,
- orientace v uživatelském rozhraní,
- ovladatelnost,
- přehlednost generovaných výstupů,
- úplnost generovaných výstupů,
- přehlednost procesního diagramu (pracovní plochy),
- automatizace modelů,
- parametrizace modelů,
- přehlednost a úplnost nápovědy,
- odbornost nápovědy,

- řešené příklady v nápovědě,
- uživatelská podpora,
- náročnost na využívanou paměť a procesorový čas.

# 3 SOUČASNÝ STAV ZKOUMANÉ PROBLEMATIKY

Data mining, který se začal vyvíjet v 90. letech minulého století, představuje množinu metod a postupů převzatých z různých vědních oborů. Spojuje tradiční statistické analýzy, umělou inteligenci, strojové učení a vývoj rozsáhlých databází. Jednu z prvních definicí dataminingu vyslovil Frawley. Data mining definoval jako netriviální extrakci implicitních dříve neznámých a potenciálně užitečných informací z datového souboru.

> Data mining is "*...the non-trivial extraction of implicit, previously unknown, and potentially useful information from data.*" (FRAWLEY et al. 1991).

Pozdější definice částečně rozšířila definici původní: Data mining představuje aplikaci rozličných algoritmů za účelem nalezení vzorů či vztahů v datovém souboru.

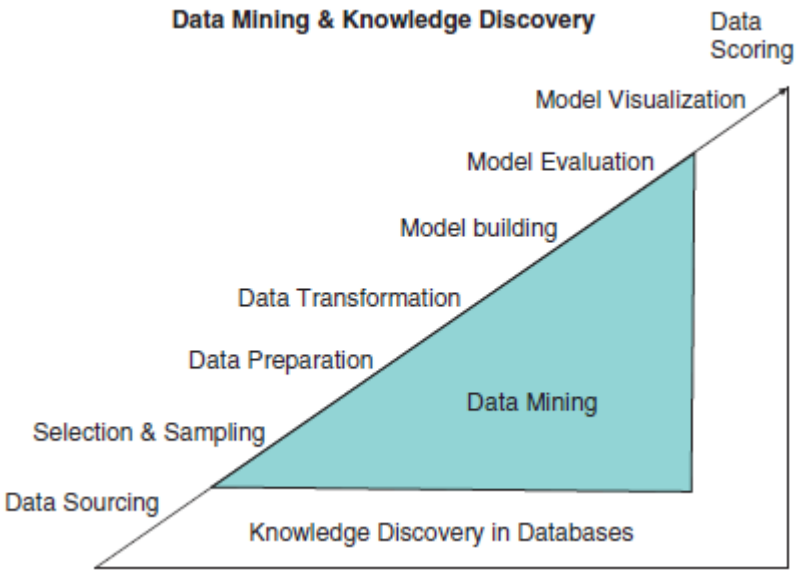
> Data mining „*…is an application of various algorithms for finding patterns or relationship in a data set.*" (FAYYAD et al., 1996).

Z novějších definic lze uvést např. Witten et al. (2011): Data mining je automatický či poloautomatický proces objevování užitečných vzorů ve velkých datových souborech.

> Data mining „*…is the process of discovering patterns, automatically or semiautomatically, in large quantities of data – and the patterns must be useful*." (WITTEN et al. 2011).

Rozdíly mezi statistickým modelováním, dataminingem a procesem objevování znalostí se zabýval Nisbet (2009). Pod pojmem statistické modelování rozumí užití parametrických statistických algoritmů k seskupení nebo předpovědi výstupu nebo události, které je založeno na vysvětlujících proměnných. Naopak data mining představuje užití algoritmů strojového učení k nalezení nejasných vzorů ve vztazích mezi datovými prvky v rozsáhlých, neuspořádaných datových souborech, které mohou nějakým způsobem přinést užitek. A proces objevování znalostí (Data Discovery) vysvětluje jako komplexní proces přistupování k datům zahrnující přípravu dat, modelování, nasazení modelu a monitorování. Tento rozsáhlý proces obsahuje také data mining (NISBET et al. 2009). Vztah mezi dataminingem a procesem objevování znalostí zachytil Fayyad již v roce 1996 (Obr. č. 1). Toto pojetí dataminingového procesu se blíží metodologii SEMMA, kterou navrhla společnost SAS (více v kapitole 4).

Obr. č. 1: Vztah mezi dataminingem a procesem objevování znalostí




Zdroj: Nisbet et al. (2009)

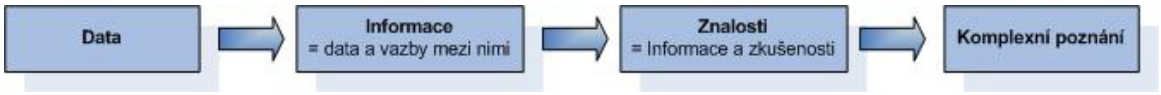
Z obrázku č. 1 je patrné, že koncepce data miningu neobsahuje pouze nalézání vztahů a vzorů, patří do ní i výběr a příprava datového souboru, hodnocení a interpretace výsledků a jejich zpracování do formy vhodné pro podporu rozhodování (NISBET et al. 2009).

## 3.1 Vztah mezi informacemi a znalostmi

Drucker (2000) upozorňuje na příchod nové informační revoluce, která se již primárně nesoustředí na shromažďování, ukládání, přenos a prezentaci dat, ale soustředí se na informace, na jejich význam a smysl. Dnešní společnost lze obecně charakterizovat jako společnost, kde znalosti jsou jak pro jednotlivce, tak i pro ekonomiku jako celek primárním zdrojem. Jsou to právě informace, které pracovníkům disponujícím znalostmi umožňují vykonávat svěřenou práci.

Proces řízení znalostí se stará o to, aby znalosti nebyly pouze hromaděny, ale aby byly racionálně využívány. K tomu patří především šíření znalostí a soustavná cílevědomá analýza účinnosti všech opatření souvisejících s řízením znalostí. Drucker (2000) dále uvádí, že nejsou-li informace organizované, zůstávají pouhými daty. Vztah mezi informacemi a znalostmi vyjadřuje Katolický (Obr. č. 2). Informace charakterizuje jako organizovaná data a vazby, které mezi nimi existují. Přidáme-li k informacím zkušenosti pracovníků, získáváme znalosti, ze kterých teprve vychází komplexní poznání. Komplexní poznání tedy představuje množinu znalostí, informací a dat vztahujících se k určité problematice.

Obr. č. 2: Vztah mezi daty, informacemi a znalostmi

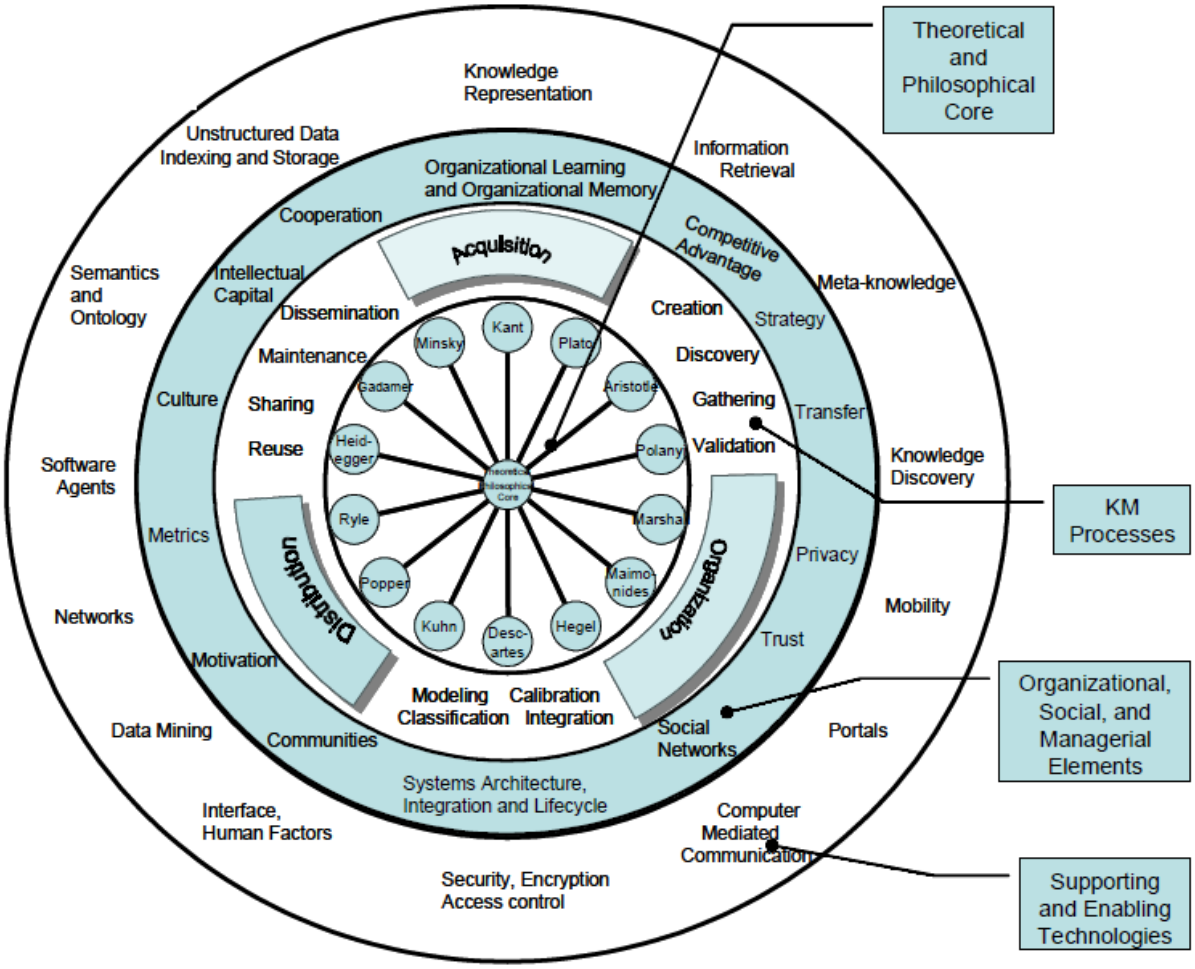



Zdroj: Katolický (2000)

Jedním z přístupů, jak charakterizovat obor řízení znalostí, který přináší ve své knize Schwartz (2006), je pomocí vrstvení. Obr. č. 3 poskytuje ucelený pohled na problematiku řízení znalostí. Jádro řízení znalostí je zastoupeno jednotlivými postupy a teoretickými východisky. Pochopení těchto postupů je základem úspěšného řízení znalostí. Na tato východiska poté navazuje praktická znalost procesů řízení znalostí, kterou zachycuje druhá vrstva. Procesní vrstva představuje pohled na různé etapy a aktivity, které jsou zahrnuty v řízení znalostí. Procesy by měly být pragmatické a komplexní, aby bylo možné dosáhnout konečného řešení, které by bylo zobecnitelné a opakovatelné.

Zmíněné procesy musí být implementovány a přizpůsobeny jednotlivým manažerským, sociálním a organizačním potřebám, které jsou uvedeny ve vrstvě č. 3. Funkcí této vrstvy je formování jednotlivých manažerských úvah a potřeb organizace tak, aby splnily vhodně definované, teoreticky zaměřené cíle. Aby jednotlivé procesy řízení znalostí vedly ke splnění potřeb organizace, musí být podpořeny relevantními informačními technologiemi, které jsou zachyceny ve vrstvě č. 4.

Obr. č. 3: Jednotlivé vrstvy procesu řízení znalostí podle Schwarze




Zdroj: Schwartz (2006)

## 3.2 Data mining jako metoda odhalování znalostí

Data mining je zpravidla zařazován mezi metody odhalování znalostí (Knowledge Discovery), Berry a Linoff (2004) však raději volí výstižnější označení, řadí data mining mezi metody pro vytváření znalostí. Data mining zkoumá a analyzuje velké množství dat za účelem objevení smysluplných vzorů a pravidel.

Rahman (2008) charakterizuje dolování dat jako proces extrakce inteligentní informace z velkého množství surových dat. Kolektivní úsilí při strojovém učení, umělá inteligence, statistické a databázové komunity využívají technologie získávání znalostí z databází k nalezení cenných informací v obrovském množství dat pro podporu inteligentního rozhodování. Data mining se tedy zaměřuje na nalezení algoritmů sloužících k získání nových vzorů ze skutečných záznamů v databázi.

Vlach (2006) vnímá data mining jako soustavu predikčních postupů k transformaci datových zdrojů na informaci podporující v organizaci řízení, rozhodování či plnění obchodních cílů. Využití těchto technik významným způsobem rozšiřuje možnosti práce s daty. Úspěch dataminingové úlohy je spíše než na množství dostupných dat závislý na správně definovaných cílech, kvalitě použitých dat, vhodném modelu apod. Smyslem a hlavním přínosem dataminingových úloh je tedy obohacení stávajících procesů o nové znalosti a jejich rychlé, akční využití. Vstupem do algoritmů přitom nemusí být jen databázová data, ale také data výzkumná, monitorovací, sbíraná ručním zápisem apod. Speciálními datovými vstupy jsou například záznamy o aktivitě na webových stránkách (weblogy) nebo nestrukturované textové dokumenty. Vzájemnou kombinací a analýzou výše uvedených dat jsou získány cenné informace pro kvalitnější řízení podnikových procesů. Bez ohledu na velikost organizace se při řešení úloh postupuje podle standardní metodologie pro řízení dataminingových projektů.

## 3.3 Oblasti využití datamingových technik

Swingler a Cairns (2008) udávají, že výpočetní inteligence nabízí nové příležitosti podnikům, které si přejí zlepšit efektivitu svého provozu. Dataminingové technologie poskytují pohled do budoucnosti, odpovídají na otázky: „*Co budou zákazníci kupovat?*“, „*U koho je nejvyšší pravděpodobnost realizace pojistné události?*“ nebo „*Jaký růst poptávky vyvolá následující reklamní kampaň?*“. Efektivní využití dataminingových technik pomáhá identifikovat potenciální příležitosti k realizaci zisků z prodeje, což může vést k vyšší návratnosti investic a tím k vytvoření konkurenční výhody. Pomocí prediktivního modelování je možné získat užitečné informace a lépe pochopit danou problematiku, což dopomáhá k tvorbě informovanějších obchodních rozhodnutí a doporučení (CHIU, TAVELLA 2008).

Dolování dat je proces výběru, prohledávání a modelování ve velkých objemech dat sloužící k odhalení dříve neznámých vztahů mezi daty za účelem získání obchodní výhody. Důležitým faktorem úspěšného nasazení metod dolování dat je stručná a srozumitelná prezentace výsledků ve formě přímo použitelné pro rozhodování. Jednou z častých oblastí aplikací je nalezení modelů, které jsou vhodné pro predikci budoucích hodnot atributů na základě nalezených vzorů v datech. Predikce tedy představuje odhadování současných či budoucích hodnot, které nejsou vzhledem ke složitosti a komplexnosti pozorovaného objektu běžnými způsoby odhalitelné či měřitelné (BERRY, LINOFF 2004).

Berry a Linoff (2004) dále uvádějí, že paměť organizace zajišťuje tzv. data warehouse, neboli datový sklad. Data mining poskytuje nástroje prohledávání dat, nalézání vzorů, navrhování pravidel, přicházení s novými nápady, myšlenkami, nalézání nových – vhodně položených – otázek a tvoření budoucích predikcí. K typickým dataminigovým úlohám patří klasifikace a přímá predikce budoucího chování individuálních subjektů (např. odhad pravděpodobnosti nákupu, odchodu zákazníků ke konkurenci nebo odpovědi na e-mail), segmentace (tvorba skupin s obdobnou charakteristikou, určení neobvyklých či podezřelých případů), detekce vztahů mezi položkami (analýza položek v nákupním košíku) nebo analýza sekvencí (typické průchody webovými stránkami). Široké uplatnění je rovněž v kontrole kvality (analýza příčin a detekce chyb kvality) nebo v predikci selhání systémů při monitoringu (BERRY, LINOFF 2004, VLACH 2006).

Prediktivní modelování představuje důležitou součást metod pro dolování dat. Pomáhá podnikům identifikovat zákazníky, kteří uvažují o odchodu ke konkurenci, lékařům pomáhá predikovat, kteří pacienti mají vysokou pravděpodobnost infarktu a pojišťovnám, kteří klienti jsou vysoce rizikoví z důvodu případného podvodu (CERRITO 2006). Berry a Linoff (2004) charakterizují prediktivní modelování jako analytické techniky, které mohou být použity k poznání potenciálních zákazníků pomocí dat o stávajících zákaznících. Podstatou prediktivního modelování je předpoklad, že data z minulosti obsahují informace, které je možné využít k poznání budoucího chování. Tento předpoklad je možné využít např. při predikci nákupního chování zákazníků, kteří se zpravidla nechovají nahodile, ale jejich chování odráží jejich potřeby, preference, tendence a obavy. Cílem prediktivního modelování v případě řízení vztahů se zákazníky je tedy nalezení vzoru v historických datech, které by osvětlilo tendence a potřeby zákazníků.

Dataminingové nástroje se také velmi často využívají v oblasti řízení vztahů se zákazníky (Customer Relationship Management, CRM). Šály (2003) ve svém článku zdůrazňuje, že naplnit CRM systém analytickými informacemi o zákazníkovi znamená dodat mu tak zvanou Customer Intelligence. Tedy hluboký vhled do struktury a chování zákazníka a integrace tohoto vhledu s kontaktním CRM tak, aby bylo možné ovlivnit chování zákazníka v okamžiku interakce s ním. Může se jednat o data na úrovni zákazníka, agregovaná data transakční úrovně, o všechny druhy demografických dat a další derivované datové elementy. V tomto kontextu vystupuje do popředí pojem segmentace zákazníků, což znamená rozčlenění zákazníků na podskupiny, které jsou s ohledem na kritéria segmentace vnitřně relativně homogenní a mezi sebou poměrně heterogenní. Výsledné segmenty jsou zpravidla dále profilovány, neboli označeny krátkým a výstižným názvem a poté dále analyzovány. Segmentace však zdaleka nepředstavuje pouze segmentaci zákazníků. Segmentovat lze například telefonní hovory podle jejich typů, stroje podle druhů údržby atd. (ŠÁLY 2003).

V kontextu s využíváním dataminingových nástrojů poukázal Hand (2005) na možné obtíže související s:

- kvalitou vstupních dat: špatná kvalita dat explicitně zapříčiní špatnou kvalitu výsledných modelů,
- příležitostmi: více příležitostí promění zdánlivě nemožné případy ve velmi pravděpodobné události,
- zásahy: vnější zásahy mohou znehodnotit model (např. vývoj modelů pro odhalování podvodů může vést k efektivnímu krátkodobému preventivnímu opatření, ale brzy poté mohou podvodníci změnit své chování tak, aby nedocházelo k těmto zásahům do jejich činnosti),
- oddělitelností informací: zpravidla je obtížné oddělit zajímavé informace od všedních, z čehož plyne důležitost stanovení vhodné predikované proměnné,
- samozřejmostí: některé objevené vzory v datech nejsou užitečné, jelikož jsou samozřejmé,
- nestacionaritou dat: nestacionarita nastane v případě, že se proces, který generuje datový soubor, je v čase nestálý.

## 3.4 Dataminingové softwarové nástroje

Mezi nejvyužívanější dataminingové softwarové nástroje patří SPSS Modeler (dříve Clementine), SAS Enterprise Miner a STATISTICA Data Miner. SPSS Clementine byl prvním dataminingovým nástrojem, který využíval grafické programovací prostředí. Tento software je využíván k modelování již od roku 1993. Intuitivní vizuální prostředí umožňuje relativně rychlý vývoj prediktivních modelů, které jsou poté implementovány do obchodního procesu. Výstupy těchto modelů jsou následně využívány k tvorbě informovanějších rozhodnutí. Stejně jako v případě Clementine, i SAS Enterprise Miner umožňuje v rámci dataminingového procesu vytvoření procesního diagramu. Důležitým prvkem je grafické uživatelské rozhraní, do kterého se dle potřeby přidávají uzly a spojnice těchto uzlů. Enterprise Miner obsahuje panel nástrojů, který organizuje jednotlivé nástroje dle metodologie SEMMA (NISBET et al. 2009).

Kromě těchto tří zmiňovaných dataminingových nástrojů eviduje server *KDnuggets.com*, dalších 97 společností, které se zabývají touto tematikou a poskytují dataminingové či analytické nástroje. Zmínit lze např. R-PLUS Enterprise Miner, TeraMiner™ či Oracle Data Mining (KD NUGGETS 2011).

K přenášení jednotlivých modelů mezi různými dataminingovými nástroji byl vyvinut jazyk PMML (Predictive Model Markup Language), který funguje na principu značkovacího XML jazyku. Většina dataminingových aplikací (včetně IBM SPSS Modeler a SAS Enterprise Miner) má pro tento jazyk implementovanou podporu (NISBET et al. 2009).

## 3.5 Ochrana osobních údajů v rámci dataminingu

Většina dataminingových analýz vychází z databází obsahujících reálná data o zákaznících. S těmito údaji by se mělo zacházet nanejvýš citlivě a tato otázka by neměla být opomíjena.

Nové technologické možnosti jsou velmi efektivní pro plánování marketingových aktivit, mohou však představovat hrozbu pro narušení soukromí zákazníků. Existuje tenká hranice mezi soukromím zákazníků a shromažďováním a následným statistickým vyhodnocováním

dostupných informací. Při analýze údajů o zákaznících by se měla striktně dodržovat pravidla na ochranu osobních údajů.

Nevládní nezisková organizace na ochranu lidských práv Iuridicum Remedium (2012) na svých stránkách uvádí, že „*…pokud chceme minimalizovat riziko zneužití osobních údajů, musíme minimalizovat jejich sdělované množství na skutečně nezbytnou míru. Jedině tak se lze vyvarovat škod plynoucích ze selhání lidského faktoru, aktuálního či budoucího zneužití.*“ Existuje i jiný úhel pohledu, a sice statistický. Z logiky věci vyplývá, že statistické vyhodnocení není možné bez poskytnutých údajů. Kompromisem by proto mělo být evidování pouze takových záznamů, které jsou skutečně potřebné pro následné vyhodnocení, a které zároveň nijak nenarušují soukromí zákazníků. Velké obchodní řetězce získávají detailní údaje o svých zákaznících snadno – nalákají je na věrnostní program a z něj plynoucí slevy z každého nákupu. Podstatné ale je, jaké informace si od důvěřivých zákazníků na oplátku za slevu vyžádají. K samotné analýze určené pro marketingové účely lze využívat agregované údaje, případně data očištěná od osobních údajů, která nijak nenaruší soukromí zákazníků.

Analýza dostupných informací je prospěšná především pro tvorbu marketingových kampaní, což může u zákazníků vzbuzovat nedůvěru. Společnost by proto měla shromažďovat pouze nezbytné údaje potřebné k vyhodnocení, např. pohlaví zákazníka, rok narození či místo bydliště určené regionem a velikostí obce. Společnosti by se měly vyvarovat požadavku na podrobnější údaje, jako je např. rodné číslo, přesné místo bydliště apod. Využitelnost věrnostních karet by měla být transparentní a výhodná pro obě strany. Zákazník získá za poskytnuté informace zpravidla slevu, společnost pak informace využije k vytvoření nákupního profilu zákazníků, čímž zvýší úspěšnost cílené marketingové kampaně. Z hlediska marketingu je využití databáze o zákaznících vhodnou alternativou k výsledkům písemných či telefonických dotazníků, průzkumů trhu a jiných zpětnovazebních informací, které mohou být do určité míry zkreslené a zpravidla neaktuální (zpožděné).

# 4 PROCES DATAMININGOVÉHO MODELOVÁNÍ

Dataminingový proces lze popsat pomocí několika oficiálně uznávaných přístupů. Jde např. o CRISP-DM, SEMMA či DMAIC (Six Sigma přístup). Metodologie SEMMA (Sample,

Explore, Modify, Model, Assess) je využívána především systémem SAS, naopak novější CRISP-DM (CRoss-Industry Standard Proces for Data Mining), která byla představena roku 1999, využívá konkurenční IBM SPSS (NISBET et al. 2009).

Autoři Azevedo a Santos (2008) provedli porovnání KDD (Knowledge Discovery in Databases) procesu neboli procesu objevování znalostí v databázích, který definoval Fayyad et al. (1996), s metodologiemi SEMMA a CRISP-DM. Ve svém článku uvádějí, že obě tyto metodologie mohou být považovány za implementace KDD procesu. Na první pohled se zdá, že CRISP-DM je komplexnější než SEMMA (viz tab. 1), v podstatě lze ale fázi porozumění problému (Business understanding) zařadit do fáze výběru (Sample), jelikož výběr nelze realizovat bez důkladné znalosti problematiky. Autoři dále konstatují, že pokud jde o celkový proces, stanovené požadavky (standardy) byly dodrženy, a že v obou případech metodologie usnadňují analytikům aplikaci dataminingového procesu v reálných systémech.

| KDD | SEMMA | CRISP-DM |
|---|---|---|
| Pre KDD | ---------- | Business understanding |
| Selection | Sample | Data Understanding |
| Pre processing | Explore | |
| Transformation | Modify | Data preparation |
| Data mining | Model | Modeling |
| Interpretation/Evaluation | Assessment | Evaluation |
| Post KDD | ---------- | Deployment |

**Tab. 1:** Porovnání metodologií KDD, SEMMA a CRISP-DM

Zdroj: Azevedo, Santos (2008)

## 4.1 SEMMA metodologie

Společnost SAS Institute vyvinula pro realizaci dataminingového procesu vlastní metodologii SEMMA, která se skládá z následujících kroků:

1) **Příprava vstupních dat** (SAMPLE)

   Příprava vstupních dat spočívá především ve výběru vzorků z rozsáhlých datových souborů, načtení a transformaci dat pocházejících z různých zdrojů či náhodného výběr dat. Kompletní vstupní datový soubor je v této fázi rozdělen na trénovací, validační a testovací množinu dat.

2) **Průzkumová analýza dat** (EXPLORE)

   V tomto kroku probíhá průzkumová analýza datového souboru. Data podléhají statistickým šetřením, jejichž výsledky jsou vizualizovány. Dále jsou zkoumány vztahy mezi jednotlivými proměnnými a identifikovány důležité proměnné. V rámci výběru vhodných proměnných může být prováděna jejich segmentace pomocí shlukové analýzy.

3) **Příprava dat pro analýzu** (MODIFY)

   Tento krok spočívá ve výběru, vytváření a transformaci vstupních proměnných, identifikaci a ošetření odlehlých a vlivných pozorování a imputaci chybějících hodnot.

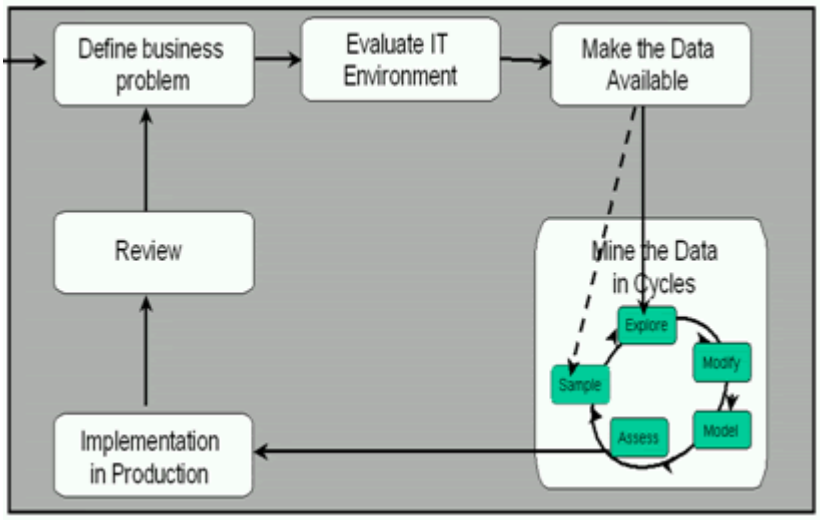
4) **Výběr a odhad modelu** (MODEL)

   Pomocí analytických nástrojů jsou vytvářeny modely pro predikci zvoleného výstupu. Tato fáze využívá lineární a logistickou regresi, rozhodovací stromy, neuronové sítě a další statistické techniky či genetické algoritmy.

5) **Interpretace a vyhodnocení výsledků** (ASSESS)

   Finální fáze dataminingového procesu porovnává a vyhodnocuje výsledky získané jednotlivými typy modelů a předkládá je uživateli zpravidla ve formě přehledných reportů a skórovacích kódů.

Obr. č. 4: Proces rozhodování pomocí metodologie SEMMA



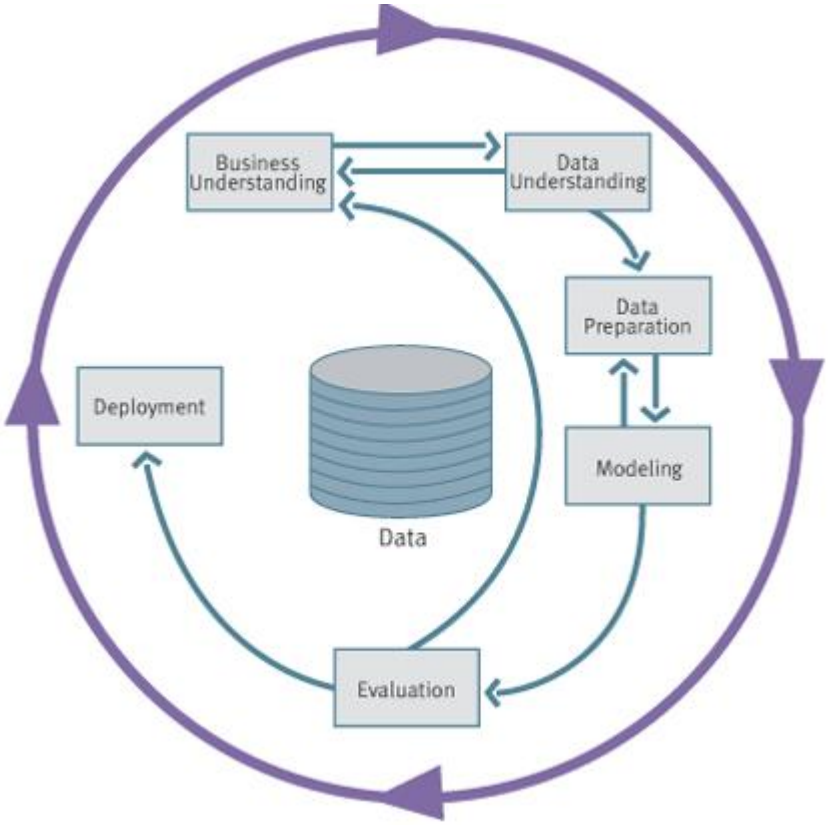

Zdroj: SAS Slovakia 2006

Komplexní proces rozhodování je uveden na obrázku výše (Obr. č. 4). Tento obrázek doplňuje SEMMA metodologii o další části sloužící k podpoře rozhodování. Samotnému procesu předchází definování problému, posouzení IT prostředí k modelování a příprava datových souborů. Finálním krokem je pak implementace výstupů modelu do praxe a následná revize výsledků. Více informací o SEMMA metodologii lze nalézt např. v (SAS DOCUMENTATION 2011).

## 4.2 CRISP-DM proces

Metodologie CRISP-DM vznikla v rámci výzkumného projektu Evropské komise. Kromě společnosti SPSS Inc. (USA) se na vzniku této metodologie podílely společnosti NCR Systems Engineering Copenhagen (USA a Dánsko), DaimlerChrysler AG (Německo) a OHRA Verzekeringen en Bank Groep B.V (Holandsko). Tato metodologie má, podobně jako dříve zmiňovaná SEMMA, poskytnout komplexní přístup k dataminingovému procesu (Obr. č. 5).

Obr. č. 5: Fáze CRISP-DM procesu




Zdroj: Chapman et al. 2000

Metodologie generuje hierarchickou posloupnost hlavních fází modelování. Každá fáze se dělí na další aktivity, operace a úkoly. Podrobnější informace lze nalézt přímo v metodice CRISP-DM (CHAPMAN et al. 2000) nebo v (NISBET et al. 2009).

Mezi jednotlivé kroky procesu patří:

1) Definování cílů (Business Understanding)

   Důležité je nalézt způsob, jak získat relevantní informace z nestrukturovaných dat a převést je do datového formátu, který bude sloužit k podpoře rozhodování. V této fázi se definují kritéria úspěchu. Podle těchto kritérií se hodnotí úspěšnost celého projektu. Cíle dataminingového modelu musí korespondovat s obchodními cíli. Patří mezi ně: vytvoření vhodné vstupní databáze, nasazení modelů, které budou generovat přidanou hodnotu apod. Tyto cíle se dělí na další konkrétnější cíle, např. sestavení vhodných datových souborů pro modelování, vytvoření seznamu predikovaných proměnných, aktualizace modelu s novými daty apod. Tyto konkrétnější cíle se dále

dělí na sadu dalších jednotlivých úkolů. Všechny tyto kroky by měly být zahrnuty v projektovém plánu. Tato fáze zahrnuje také posouzení integrace dat, kvality dat a analytických nástrojů, které jsou k dispozici. Také by měly být vytyčeny potenciální rizika a zváženo nasazení modelu do praktického využití.

2) Porozumění datům (Data Understanding)

   Tato fáze zahrnuje sběr dat a jejich integraci z různých datových zdrojů, popis a prozkoumání jednotlivých proměnných (popisná statistika, grafické znázornění, rozdělení klíčových proměnných, prozkoumání vztahů mezi proměnnými) a zhodnocení kvality dat (kontrola chybějících hodnot, odlehlých pozorování).

3) Příprava dat (Data Preparation)

   Hlavním úkolem této fáze je výběr datového souboru a jeho příprava pro modelování. Patří sem např. čištění dat, vytváření a odvozování nových proměnných, transformace proměnných, kontrola formátování, vážení, filtrování, redukce dimenzionality apod.

4) Modelování (Modeling)

   Modelovací fázi procesu lze dále rozdělit na jednotlivé úkoly:

   a. Výběr techniky modelování – výběr konkrétního algoritmu, architektury modelování (jednoduchý model nebo kombinace více modelů), specifikace předpokladů zvoleného algoritmu (předpoklady parametrických modelů, odlehlé hodnoty).

   b. Vytvoření testovacího návrhu – rozdělení vstupního souboru na trénovací a testovací část.

   c. Vytvoření modelu – nastavení parametrů zvoleného algoritmu, vytvoření rozličných typů modelů; jeden algoritmus poskytuje pouze jeden pohled na problém, více algoritmů přináší více různých pohledů (ensemble modeling).

   d. Posouzení modelů lze realizovat na reálných datech nebo pomocí hodnotících nástrojů (koincidenční tabulky, lift, ROI křivky apod.).

   e. Hodnocení modelů a stanovení dalších kroků.

5) Hodnocení výsledků (Evaluation)

   Vyhodnocení dosažených výsledků z hlediska obchodních cílů, případně otestování modelu v reálném prostředí a stanovení dalších kroků

6) Implementace (Deployment)

   Nasazení modelu do praxe zahrnuje tvorbu plánu nasazení, monitoring a tvorbu finálního reportu.

## 4.3 Postup dataminingového procesu

V postupu získávání znalostí z databází doporučuje Guidici (2009) nejprve definovat objekt analýzy, což ve většině případů nemusí být vůbec jednoduché. I když firemní cíle, jichž chce podnik dosáhnout, jsou obvykle jasné, mohou být obtížně vyčíslitelné. Jasné definování cíle a objasnění problému je mimořádně důležité pro vytvoření analýzy správným způsobem. Jakmile jsou stanoveny cíle analýzy, je podstatné získat potřebná data. Nejčastějšími a také nejlevnějšími zdroji dat je interní datový sklad. Další možností je nákup dat či jejich sběr pomocí dotazníkového šetření. Získaná data je nutné prozkoumat, odhalit existenci případných anomálií nebo dle potřeby data transformovat.

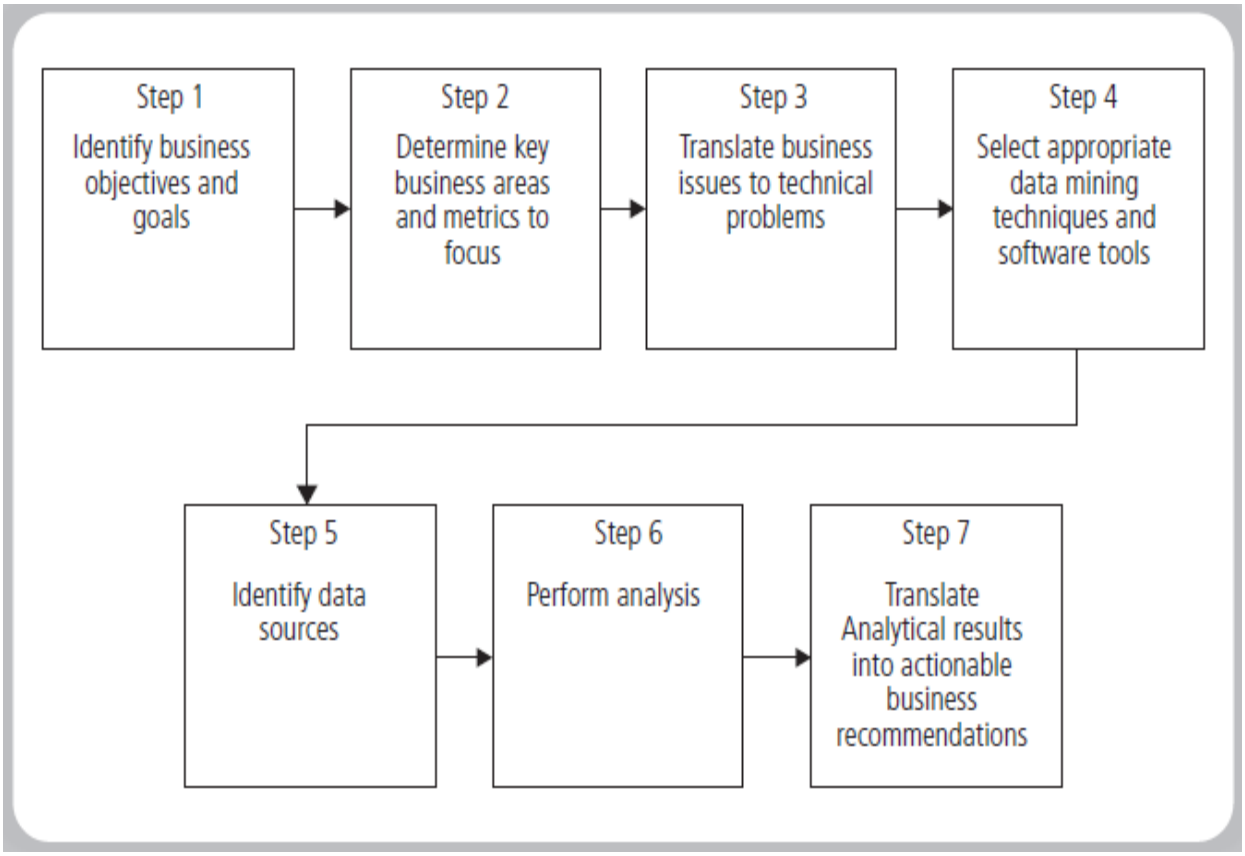
Velkou váhu lze přiřadit následné volbě vhodné statistické metody, jelikož existuje velké množství dostupných metod. Výběr vždy závisí na konkrétním cíli analýzy. K získání konečného řešení je nezbytné zvolit nejvhodnější model z různých dostupných analytických metod. Ke zvolení nejlepšího modelu se využívají odlišná hodnotící kritéria. Guidici (2009) doporučuje vždy aplikovat více metod a z nich poté zvolit tu nejvhodnější. Každá metoda má potenciál k vyzdvihnutí určitých aspektů, které mohou být jinou metodou ignorovány.

Důležitým krokem dataminingu je integrace výsledků analýzy do rozhodovacího procesu. Obchodní znalost, získávání pravidel a jejich následné využití v rozhodovacím procesu, nám dovoluje posunout se z analytické fáze k fázi rozhodovací. Pomocí získaných klasifikačních pravidel bude dále možné například rozlišit, který ze zákazníků bude více profitabilní.

Postup procesu objevování znalostí v databázi zachycuje následující schéma (Obr. č. 6). Dle Chiu a Tavelli (2008) by prvním krokem dataminingového procesu, stejně jako jakéhokoliv

jiného procesu, mělo být identifikování obchodních plánů a cílů. Druhý krok představuje stanovení klíčových obchodních oblastí a metrik, které budou využity pro měření výstupů. Třetím krokem je převedení obchodních problémů do technické terminologie. Nesprávný popis problematiky by mohl vést k plýtvání zdrojů a příležitostí. Konečně krok č. 4 představuje výběr vhodných dataminingových technik a softwarových nástrojů. Dále přichází na řadu identifikace datových zdrojů. Nejčastěji jsou využívány zdroje interní – databáze zákazníků, transakcí, marketingových aktivit apod. Krokem č. 6 je samotné provedení analýzy, pod čímž se skrývá vytvoření modelu, jeho ověření a následné testování. Posledním krokem je převedení výsledků analýzy do obchodních doporučení, což představuje vysvětlení hlavních závěrů analýzy v netechnické terminologii.

Obr. č. 6: Proces objevování znalostí v databázi podle Chiu a Tavella




Zdroj: Chiu, Tavella (2008)

Podrobněji definuje postup dataminingového procesu Hand et al. (2001), který uvádí, že komplexní proces zahrnuje následující všeobecné operace:

- Explorační analýza dat: zahrnuje interaktivní a vizuální techniky, které umožňují „pohled“ na data.
- Popisné modelování: jedná se o vyšší úroveň pohledu na datové soubory, zahrnuje odhad pravděpodobnostního rozdělení dat, popis vztahů mezi proměnnými, rozdělení dat do skupin (segmentace či shluková analýza).
- Prediktivní modelování (klasifikace a regrese): cílem je sestavit model, ve kterém jsou hodnoty vysvětlované proměnné vysvětlovány pomocí hodnot vysvětlujících proměnných.
- Objevování vzorů a pravidel: algoritmy spadající do této skupiny mají velmi široké uplatnění, jedná se např. o asociační pravidla, sekvenční analýzu, link analýzu (text mining).
- Vyhledávání dle obsahu (Retrieval by Content): tato aktivita začíná se známými vzory a snaží se objevit podobné vzory v novém datovém souboru. Jedná se o rozpoznávání vzorů, které je často využíváno v souvislosti s textovými či grafickými soubory.

## 4.4 Úprava datového souboru před modelováním

Přípravná fáze datové základny je jedním z nejdůležitějších kroků dataminingového procesu. Zde se připravuje struktura a formát vstupních dat k modelování. Provádí se explorační analýza a čistění datové matice, součástí je také analýza chybějících údajů. Některé modelovací algoritmy si s chybějícími údaji neumí poradit, proto je třeba tyto údaje identifikovat a případně provést nahrazení.

Přístup k datům lze řešit několika možnými způsoby, např. pomocí dotazování v jazyku SQL (Query-based data extraction), pomocí pokročilých dataminingových dotazovacích jazyků (Data Mining Query Language) nebo pomocí ODBC (uživatelské rozhraní poskytuje např. SAS Enterprise Miner, IBM SPSS Modeler či STATISTICA). Více informací lze nalézt např. (NISBET et al. 2009).

### 4.4.1 Příprava datové matice

Dataminingové nástroje zpravidla vyžadují, aby vstupní data byla prezentována ve formě záznamů. Je tedy nutné uspořádat datovou matici takovým způsobem, aby jednotlivé řádky tvořily jednotky (instance/záznamy) a sloupce představovaly jednotlivé proměnné (vlastnosti/atributy záznamů). Toho lze dosáhnout postupným propojováním datových zdrojů, restrukturalizací, případně transformováním vstupních proměnných (NISBET et al. 2009).

| | **Proměnná 1** | **Proměnná 2** | **…** | **Proměnná *p*** |
|---|---|---|---|---|
| **Jednotka 1** | $x_{11}$ | $x_{12}$ | | $x_{1p}$ |
| **Jednotka 2** | $x_{21}$ | $x_{22}$ | | $x_{2p}$ |
| **…** | | | $x_{ij}$ | |
| **Jednotka *n*** | $x_{n1}$ | $x_{n2}$ | | $x_{np}$ |

Tab. 2: Struktura datové matice

Zdroj: Field (2005)

Prvním krokem data miningu je tedy příprava datové matice neboli integrace dat. Podstatou je seskupit všechny dostupné datové zdroje do jednoho souboru instancí (záznamů). Zpravidla je tedy nutné denormalizovat data. Proces integrace datových zdrojů v podniku je nazýván data warehousing a velmi usnadňuje následnou přípravu dat pro data mining (WITTEN 2011).

### 4.4.2 Popisná analýza

Jednotlivé dataminingové metody se liší v závislosti na zkoumaných datech, pro kvantitativní data budou využity pravděpodobně odlišné statistické postupy než pro data kvalitativní. Meloun et al. (2005) však zdůrazňuje, že hledání struktury v datech v obou případech předchází explorační analýza, jež umožňuje ověřit předpoklady analyzovaných dat, např. normalitu, nekorelovanost, homogenitu a také nalezení vybočujících objektů, které mohou zkreslit výsledky provedené analýzy.

Přípravná fáze zahrnuje popis jednotlivých vstupních proměnných. Provádí se tedy jednorozměrná explorační analýza, která zahrnuje jednoduché popisné statistiky (průměr,

modus, medián, směrodatnou odchylku, variační rozpětí apod.), četnostní (frekvenční) tabulky, jednoduché grafy (histogram, boxplot, normální pravděpodobnostní graf) apod. Dataminingové nástroje většinou poskytují možnosti datového auditu, kde připravené algoritmy upozorní například na špatnou kvalitu dané proměnné, na existenci odlehlých pozorování či chybějících údajů. Nemělo by být opomenuto také ověření pravděpodobnostního rozdělení dané proměnné (NISBET et al. 2009).

Samostatnou část tvoří činnosti spojené s čištěním dat, které umožňují filtrování a opravu podezřelých či nesprávných údajů. Nejčastějším typem filtrování je odstraňování nepotřebných údajů – odlehlých pozorování. Tyto anomálie snižují predikční schopnosti modelu. Jejich odstranění by však mělo předcházet důkladné zvážení. Tyto hodnoty mohou být velmi užitečné například při predikci podvodného chování, kreditních rizik apod. (NISBET et al. 2009).

Mezi techniky odstraňování odlehlých pozorování patří např. „useknutí“ (*trimming*) datového souboru či nahrazení (*winsorisation*) hodnot přesahujících trojnásobek směrodatné odchylky (nebo mezikvartilového rozpětí) hodnotou předcházející. Pokročilejší techniky filtrování se využívají při analýzách časových řad. Efektivní způsob filtrování časových řad navrhl Masters (1995), který se inspiroval u technik ke zpracování signálů (high-frequency signal fluctuations). Pomocí filtrů odstraňuje údaje spadající nad definovanou nejvyšší úroveň přijatelnosti nebo naopak pod definovanou nejnižší úroveň přijatelnosti (více v MASTERS 1995).

### 4.4.3 Transformace proměnných

Transformace jednotlivých proměnných probíhá odlišně u kvantitativních a kvalitativních typů dat. U číselných proměnných se provádí především linearizace, která je vyžadována u některých typů modelovacích algoritmů (např. lineární regrese), či normalizace, tedy transformace všech číselných proměnných na stejnou škálu. Zpravidla se používá tzv. z-transformace, kdy je každá hodnota dané proměnné nahrazena z-hodnotou, která se spočítá jako odchylka od průměru vydělená směrodatnou odchylkou. Dále lze normalizovat data vydělením každé hodnoty průměrnou hodnotou dané proměnné, či vydělením každé hodnoty variačním rozpětím po odečtení minimální hodnoty (COLLICA 2011, NISBET et al. 2009).

Některé ordinální proměnné, které jsou ve škále např. 1 – 9, mohou být zpracovávány jako číselné proměnné, v takovém případě je ale nutná obezřetnost. U nominálních proměnných lze využít transformaci na tzv. dummy proměnné, které nabývají pouze alternativních hodnot (0, 1). Jedná se o jejich převedení na číselnou proměnnou. Tato operace však způsobuje snižování počtu stupňů volnosti, čímž dochází ke snížení obecnosti modelu. Větší množství dummy proměnných v modelu může u parametrických algoritmů a u algoritmů strojového učení způsobovat tzv. přeučení. K tomu dochází v případě, že model predikuje cílovou proměnnou s vysokou přesností na trénovacích datech (naučí se je „na zpaměť"), na nových (testovacích) datech však přesnost výrazně klesá (NISBET et al. 2009).

Obdobným případem transformace je také diskretizace (binning) číselných proměnných, neboli převedení číselné proměnné do kategorií. Tento postup, který může být výhodný pro některé typy modelovacích algoritmů, slouží k odstranění šumu z dat. Využívá se např. pro kategorizaci věku.

Kromě transformace je možné nové proměnné odvozovat z proměnných původních. Odvozovat lze také cílovou proměnnou (target). Často je zvolení vhodné cílové proměnné základem úspěšného modelu. Více transformačních technik lze nalézt např. v (NISBET et al. 2009, WITTEN 2011).

### 4.4.4 Problematika chybějících hodnot

S problematikou chybějící hodnoty (Missing values) se lze setkat v téměř každém datovém skladu. Prvním krokem by mělo být zjištění, zda jsou chybějící hodnoty rozmístěny v souboru náhodně. Nenáhodnost by například naznačovala skutečnost, že v případě dotazníkového šetření měli respondenti s některými otázkami problémy.

Existují různé druhy chybějících údajů, jako jsou neznámé hodnoty, irelevantní hodnoty, nezaznamenané hodnoty apod. Například systém SAS rozeznává přibližně 60 typů chybějících hodnot. Více informací o chybějících hodnotách lze nalézt v (FIELD 2005, NISBET 2009, WITTEN 2011).

Použitelné techniky pro vypořádání se s problematikou chybějících hodnot lze určit pomocí mechanismu výskytu chybějících hodnot. Tento pojem zavedl Rubin (1976), který rozlišoval tři možné případy. Prvním případem je situace, kdy chybějící hodnoty mají stejnou

pravděpodobnost výskytu pro všechny záznamy. Záznamy s chybějícími hodnotami nejsou nijak odlišitelné od těch bez chybějících hodnot. Za této situace hovoříme o tzv. MCAR (Missing Completely at Random) hodnotách. V případě tzv. MAR (Missing at Random) hodnot nezávisí příčina chybějící hodnoty na proměnné, v rámci níž se vyskytuje, nicméně může být závislá na jiných pozorovaných proměnných. Posledním případem, uvedeným v (RUBIN 1976), jsou MNAR (Missing Not at Random) hodnoty, kdy příčina výskytu závisí pouze na proměnné samotné. Konkrétní příčinou může být např. fakt, že pro daný záznam tato proměnná nebyla naměřena nebo byla data proměnné doplněna z externího zdroje pouze pro část záznamů. K těmto třem mechanismům byl později doplněn i čtvrtý případ, kdy příčinou chybějící hodnoty je nemožnost jejího fyzického měření. Tento případ se nazývá MBND (Missing By Natural Design) (PEJČOCH 2011, RUBIN 1976).

Jednou z možností, jak se vypořádat s chybějícími údaji, je vyřadit celý záznam z datové matice. Nevýhodou tohoto přístupu je redukce datové matice. V případě korelace mezi chybějícími údaji a cílovou proměnnou může dojít k tvorbě zkreslených (vychýlených) odhadů. Tento krok se provádí pouze v situacích, kdy chybějící údaje neovlivní výsledky zpracování (NISBET et al. 2009). Vyřazení záznamů může být aplikováno pouze na chybějící údaje typu MCAR a je vhodné jej provádět pouze při nízkém relativním počtu chybějících hodnot s maximální hranici 5% relativní četnosti u dané proměnné (PEJČOCH 2011).

V některých případech je možné neupravovat chybějící údaje, což však vyžaduje speciální postupy při matematických výpočtech a při použití statistických metod. Nisbet et al. (2009) uvádí, že nahrazení chybějících údajů je v odůvodněných případech přínosnější, než když se ponechají prázdná. Častěji se proto využívají možnosti imputace, neboli odhadnutí scházejících hodnot na základě platných hodnot jiných proměnných (HAIR, ANDERSON 2010). Imputace za použití podmíněného průměru, též nazývána Buckovou metodou, spočívá v doplnění více průměrných hodnot podmíněných hodnotami ostatních proměnných. Aplikace této metody vede ke konzistentním odhadům u MCAR a MAR. V případě MAR je ovšem nutné přijmout dodatečný předpoklad, že skutečnost výskytu chybějících hodnot nezávisí na ostatních proměnných (PEJČOCH 2011). Imputaci lze aplikovat také např. pomocí metody nejpravděpodobnější hodnoty (Maximum Likelihood Imputation), či jejím odhadem pomocí regresní analýzy nebo rozhodovacích stromů (Multiple Imputation).

### 4.4.5 Redukce datové základny

Redukce datové základny zahrnuje tvorbu výběrového vzorku (redukci záznamů) či redukci dimenzionality (proměnných). V případě velkých databází je efektivní testování jednotlivých algoritmů na menším počtu záznamů. K tvorbě výběrového souboru se využívá algoritmus jednoduchého náhodného výběru (Simple random sampling). Dalším typem redukce je partitioning, neboli rozdělení záznamů dle určité proměnné (např. segmentu). Pro každou kategorii (segment) je pak vybudován separátní model.

Vyvažování modelu je také jednou z možností redukce datové matice; nevyvážený datový soubor může způsobit komplikace u některých typů modelu. V případě, že je cílová proměnná nerovnoměrně zastoupena ve vstupním souboru, bývá žádoucí provést vyvážení modelu. K převažování dat dochází především při použití prediktivních algoritmů strojového učení, jako jsou rozhodovací stromy či neuronové sítě (NISBET et al. 2009). Principem je zvolení takového vzorku dat, aby cílová proměnná měla rovnoměrné zastoupení ve všech kategoriích. V případě málo zastoupené kategorie lze za účelem nalezení nejvhodnějšího modelovacího algoritmu provést duplikaci záznamů.

Datová matice zpravidla bývá také redukována z důvodu testování a ověřování kvality modelů. Vstupní soubor se proto dělí na dvě (v některých případech tři) části: trénovací, testovací (a ověřovací) část. Modelování probíhá na trénovací části datové matice, ověření přesnosti modelu je realizováno na testovací (případně ověřovací) části dat (NISBET et al. 2009).

Důležitá je také redukce dimenzionality neboli redukce postradatelných vstupních proměnných. Jednoduchou metodou pro výběr nepotřebných proměnných jsou korelační koeficienty. V případě, že se hodnota korelačního koeficientu mezi dvěma proměnnými blíží v absolutní hodnotě k 1, lze konstatovat, že v modelu existuje nežádoucí multikolinearita. Jedna ze dvou daných proměnných by proto měla být z modelu odstraněna. Dalším postupem vedoucím ke snížení dimenzionality je užití analýzy hlavních komponent (PCA). Tato technika je využívána pro identifikaci silných prediktorů v modelu. Odhaluje vztahy mezi jednotlivými proměnnými v datovém souboru. Výstupem jsou hlavní komponenty, které představují lineární kombinace vstupních proměnných. Následně jsou vstupní proměnné nahrazeny nižším počtem získaných komponent (NISBET et al. 2009). Tento přístup však může mít i svá úskalí. Jelikož se jedná o nesupervizovaný algoritmus, není tedy vztažený

k žádné cílové proměnné, je možné, že nalezené hlavní komponenty skryjí třídní odlišnosti ve vztahu původní proměnné k proměnné cílové (HAND et al. 2001).

Ke snížení dimenzionality lze využít i další sofistikované metody, kterými jsou např. algoritmus CHAID (Chi-Square Automatic Interaction Detection), singulární rozklad nebo Giniho index. Tento index se pohybuje v intervalu 0 – 1. Jde o míru nerovnosti v šíření hodnot v daném rozsahu proměnné. Podstatou této míry je, že proměnná s relativně vysokou nerovností v četnostním rozdělení hodnot má vyšší pravděpodobnost stát se dobrým prediktorem (NISBET et al. 2009).

Zmiňované techniky jsou přínosné zpravidla v případě číselných proměnných. Pro kategoriální data existují také vhodné nástroje ke snížení dimenzionality, především jde o nástroje grafické. Sílu asociace mezi dvěma proměnnými lze zachytit síťovým diagramem, v případě více proměnných lze využít např. vícerozměrné škálování (NISBET et al. 2009).

## 4.5 Modelování

Důležitým rozhodnutím v průběhu dataminingového procesu je výběr vhodného modelovacího algoritmu, případně zvolení skupiny algoritmů, která vytvoří lepší predikci než pouze jeden samotný model. Giudici (2009) dělí dataminingové modelovací techniky na dvě hlavní skupiny: deskriptivní a prediktivní metody. Cílem deskriptivních technik je výstižně popsat skupiny dat a jejich strukturu. Tyto metody nevyužívají žádné hypotézy o příčinné souvislosti mezi proměnnými. Naopak cílem prediktivních technik je popsat jednu nebo více proměnných ve vztahu k ostatním proměnným. Tyto metody pomáhají predikovat či klasifikovat budoucí výstupy cílové proměnné či proměnných (tzv. target) v souvislosti se změnami vysvětlujících neboli vstupujících proměnných. Hlavní zástupci těchto technik vychází nejen z oblasti klasických statistických metod, jako je lineární či logistická regrese, ale také ze strojového učení (např. neuronové sítě a rozhodovací stromy).

Kromě zmiňovaného členění na dvě hlavní kategorie, existují i další členění. Berry a Linoff (2004) rozdělují modelovací algoritmy do šesti následujících skupin:

- Klasifikace

- Odhady
- Predikce
- Porovnávání
- Shlukování
- Profilování

První tři skupiny se označují jako přímý data mining, jehož cílem je nalézt hodnotu konkrétní cílové proměnné. Tato skupina metod je někdy označována také jako učení s učitelem či prediktivní modelování. Rozdíl mezi predikcí a klasifikací popisují Han a Kamber (2000) takto: „Pokud predikujeme hodnotu klasifikační proměnné, potom se jedná o klasifikaci, pokud ale odhadujeme hodnotu spojité proměnné, již mluvíme o predikci.“ Takto položená definice by však vedla k rozporu mezi zařazením lineární regresní analýzy, která by patřila mezi predikční metody, a logistické regresní analýzy, která by se řadila k metodám klasifikačním.

Odhalení struktury v datech je cílem dalších dvou metod – porovnávání a shlukování. Tyto dvě metody jsou označovány jako nepřímý data mining, tedy učení bez učitele; neobsahují cílovou proměnnou (target), jejíž výstupní hodnotu bychom se snažili odhadnout pomocí predikčních technik. Profilování, které tvoří poslední skupinu, je popisná úloha, která může zahrnovat jak deskriptivní tak také predikční algoritmy.

## 4.6 Vybrané dataminingové techniky modelování

Mezi využívané techniky patří kromě klasických metod, jako je shluková analýza, analýza hlavních komponent či regresní analýza, také metody nestatistické. Například se jedná o neuronové sítě, podpůrné vektory či rozhodovací stromy.

### 4.6.1 Shluková analýza

Rencher (2002) uvádí, že s čím dál větší dostupností počítačů a jednotlivých statistických paketů se přechází od jednoduchých „filtračních“ segmentačních nástrojů, které využívaly základní demografické údaje, na sofistikovanější dataminingové metody. V případě malého počtu (2-3) dimenzí lze shluky jednoduše vizualizovat a rozpoznat vlastním okem, s růstem

dimenzí však roste náročnost vizuálně shluky rozeznat. Čím více je dimenzí, tím více roste důležitost geometrických analýz a vhodných algoritmů (RENCHER 2012, COLLICA 2007).

Dle Řezankové (2007) patří shluková analýza (Cluster analysis, CLU) mezi nejvýznamnější segmentační techniky. Jedná se o nástroj nepřímého objevování znalostí. Automatický algoritmus detekuje existující strukturu dat bez ohledu na konkrétní cílovou proměnnou. Härdle a Simar (2007) uvádějí, že shluková analýza poskytuje nástroje a metody pro seskupování jednotlivých pozorování různorodého souboru do homogennějších podsouborů dle příslušných kritérií.

Cílem shlukové analýzy je nalézt optimální seskupení, kdy jednotlivá pozorování nebo objekty každého shluku jsou vzájemně podobné, avšak jednotlivé shluky navzájem rozdílné. Shluková analýza poskytuje cestu k získávání znalostí o struktuře dat. Shlukovací techniky jsou využívány k nalezení vzorů v datech; tyto vzory však nejsou na první pohled patrné. V některých případech se nepodaří homogenní shluky v datovém souboru nalézt, jindy je počet nalezených shluků příliš vysoký. Důvodem k realizaci shlukové analýzy je tedy předpoklad, že ve zkoumaných datových souborech nalezneme smysluplná přirozená seskupení - podtřídy (RENCHER 2002, BERRY, LINOFF 2004).

Jain (2010) udává, že shluková analýza se nejčastěji využívá v následujících třech hlavních případech:

- Popis struktury dat: cílem je nahlédnou do struktury dat, porozumět datům, vytvořit hypotézy, odhalit případné anomálie v datech, či identifikovat charakteristické rysy souboru.
- Klasifikace: identifikace míry podobnosti mezi objekty.
- Komprese: využívá se jako metoda pro organizaci dat a ke shrnutí problematiky prostřednictvím vzorů v jednotlivých shlucích.

Před vlastní shlukovou analýzou je třeba řešit otázku, zda je žádoucí data standardizovat. V případě, že jsou proměnné uváděny ve stejných jednotkách a vykazují shodnou variabilitu, není standardizace nutností. Je však potřebné respektovat skutečnost, že většina měr vzdáleností je velmi citlivá na měřítka (stupnice), vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s vyšší směrodatnou odchylkou mají větší vliv na míru podobnosti či nepodobnosti (MELOUN, MILITKÝ 2004). V takovém případě je

standardizace vstupních proměnných žádoucí. Shlukovat je možné i hlavní komponenty nebo faktory získané pomocí faktorové analýzy (ŘEZANKOVÁ 2007).

Tato dataminingová technika není zpravidla využívána samostatně, nýbrž ve spojení s dalšími metodami. Jakmile jsou identifikovány výsledné shluky (segmenty), je aplikována další metoda, pomocí které se zjistí význam jednotlivých shluků (profilace). K lepšímu porozumění existujících rozdílů mezi segmenty je vhodné navíc popsat každou skupinu popisnými charakteristikami.

Analýza shluků objektů není charakteru statistického testování. Požadavky normality, linearity, homoskedasticity, které jsou tolik důležité v ostatních vícerozměrných technikách, zde nemají význam. Důležitá je však reprezentativnost vzorku a vliv multikolinearity. Multikolinearita se chová jako neviditelný proces vážení, který silně ovlivňuje analýzu. Uživatel musí proto ověřit přítomnost multikolinearity, a když je její přítomnost prokázána, je třeba zredukovat počet znaků nebo použít vhodnou míru, jakou je např. Mahalanobisova vzdálenost (MELOUN, MILITKÝ 2004).

Mezi běžné metody shlukování vektorů pozorování patří hierarchické shlukování a metoda rozkladu (partitioning). Hierarchické shlukování začíná s $n$ shluky, kdy každé pozorování tvoří samostatný shluk, a končí jedním shlukem, který zahrnuje všechna pozorování. V každém kroku jsou dvě nejbližší pozorování nebo shluky pozorování sloučeny do jednoho nového shluku. Tato metoda je označována aglomerativní. Opačný, méně používaný, přístup se nazývá divizivní. Tento proces je nevratný, žádný ze dvou shluků, které byly spojeny (rozděleny), již nemůže být později oddělen (sloučen); žádné dřívější chyby tedy nemohou být opraveny. Možným přístupem je provést po hierarchickém shlukování proceduru partitioning, kde mohou být jednotlivé jednotky přeřazovány z jednoho shluku do shluku jiného (RENCHER 2002, ŘEZANKOVÁ et al. 2007).

**Míry vzdáleností**

S měřením vzdáleností mezi shluky souvisí využití různých vzdálenostních měr aplikovaných na různé typy proměnných. Existují míry jen pro textová data, pro binární proměnné,

kategorické proměnné atd. Základní typy proměnných lze rozdělit na čtyři skupiny (COLLICA 2007):

- Kategorické (nominální) proměnné – u kategorických proměnných platí, že existuje rozdíl mezi jedním či druhým objektem, ale tento rozdíl nelze kvantifikovat. V matematické terminologii lze stanovit, že $X \neq Y$, ale už nelze určit, že $X < Y$ nebo $X > Y$.
- Ordinální (pořadové) proměnné – ordinální proměnné jsou seřazené dle určitého specifického pořadí, které ale nic neříká o vzdálenosti mezi jednotlivými kategoriemi. Tento vztah se nazývá tranzitivita. V případě, že $A > B$ a $B > C$, pak musí platit, že $A > C$.
- Intervalové (rozdílové) proměnné – v případě intervalových proměnných lze měřit vzdálenost mezi jednotlivými pozorováními.
- Poměrové (podílové).

Každá míra vzdálenosti musí dodržet následující pravidla:

1. $D(X,Y) = 0$ právě tehdy, když $X=Y$.
2. $D(X,Y) \geq 0$ pro všechna X a všechna Y.
3. $D(X,Y) = D(Y,X)$.
4. $D(X,Y) \leq D(X,Z) + D(Z,Y)$.

První pravidlo značí, že v případě nulové vzdálenosti musí být oba body identické. Druhé pravidlo udává, že všechny vzdálenosti musí být nezáporné. Vektory mají rozsah a směr, vzdálenost mezi vektory však nesmí nabývat záporné hodnoty. Třetí pravidlo vyjadřuje symetrii, tedy skutečnost, že vzdálenost X od Y je stejná jako vzdálenost od Y k X. V případě ne-euklidovské geometrie toto pravidlo nemusí vždy platit. Čtvrté pravidlo je známo jako trojúhelníková nerovnost. Udává, že délka jedné strany trojúhelníku nemůže být delší než součet délek zbývajících dvou stran (ANDENBERG 1973).

Běžnou vzdálenostní metrikou mezi dvěma vektory X a Y je euklidovská vzdálenost, která představuje délku přepony pravoúhlého trojúhelníku a její výpočet je založen na Pythagorově větě (MELOUN, MILITKÝ 2004, RENCHER 2002).

Platí, že vzdálenost:

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} \left(x_{kj} - x_{lj}\right)^2} . \tag{4.1}$$

V případě měření vzdáleností u kategorických proměnných se využívá transformace na tzv. *dummy* proměnné, neboli proměnné nabývající pouze hodnot 0 a 1. Pro intervalová data se zpravidla využívá míra vzdálenosti nazvaná Minkowskeho metrika. Tato metrika je také nazývána jako $L_p$ norma a je definovaná vzorcem č. 4.2:

$$d_p(x_k, x_l) = \left( \sum_{i=1}^{i=n} \left| x_{ik} - x_{il} \right|^p \right)^{1/p} . \tag{4.2}$$

V případě, že $p = 1$, jedná se o Manhattanskou vzdálenost, nazývanou také míra městských bloků. V případě, že $p = 2$, pak jde o euklidovskou vzdálenost a když $p = 3$, mluvíme o Čebyševově míře (ANDERBERG 1973, COLLICA 2007, DUDA et al. 2001).

Mírou, která není závislá na měřících jednotkách a navíc také neobsahuje nadměrný vliv korelovaných proměnných, je známá Mahalanobisova vzdálenost:

$$d_{Ma}(x_k, x_l) = \sqrt{\left(x_k - x_l\right)^T C^{-1} \left(x_k - x_l\right)}. \tag{4.3}$$

Více informací o mírách pro měření vzdálenosti objektů lze nalézt např. v (HEBÁK et al. 2007, MELOUN, MILITKÝ 2004).

**Hierarchické shlukování**

K nejužívanějším postupům uplatňovaným ve shlukové analýze patří vytváření hierarchické posloupnosti rozkladů dat. Ne/podobnost shluků je možné stanovit dle různých aglomerativních algoritmů (vzdáleností), např.:

- Metoda nejbližšího souseda (Nearest Neighbor / Single Linkage)

  V této metodě je vzdálenost mezi dvěma shluky *A* a *B* definována jako minimální vzdálenost mezi bodem *A* a bodem *B* (minimální euklidovská vzdálenost):

$$D(A,B) = \min d(y_A, y_B). \tag{4.4}$$

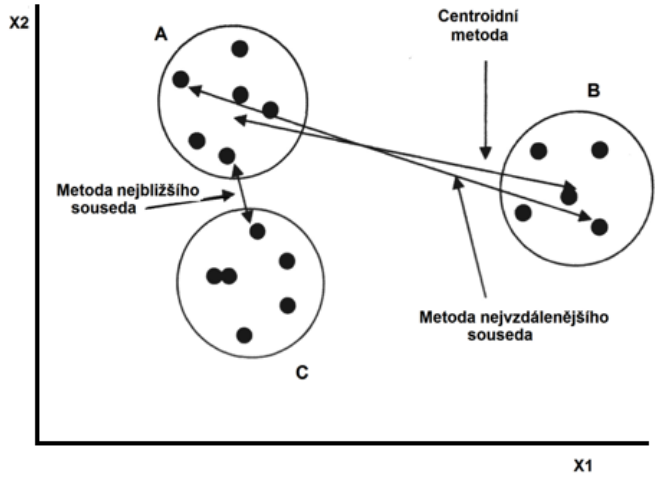
V každém kroku je zjištěna vzdálenost mezi každým párem shluků; dva shluky s nejmenší vzdáleností se spojí. Tato metoda je robustní k odlehlým pozorováním v datech, její tendence k tvorbě řetězců však může vést až ke zcela mylným závěrům. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.

- Metoda nejvzdálenějšího souseda (Farthest Neighbor / Complete Linkage)
  Vzdálenost mezi dvěma shluky *A* a *B* je definována jako maximální vzdálenost mezi body *A* a *B*:

$$D(A,B) = \max d(y_A, y_B). \tag{4.5}$$

V každém kroku je nalezena maximální vzdálenost mezi páry shluků a pár s nejmenší hodnotou nalezené vzdálenosti se sloučí. Tato metoda je citlivá na odlehlé hodnoty a má tendenci vytvářet kompaktní shluky se stejnou formou a počtem objektů.

Obr. č. 7: Tři nejběžněji využívané metody hierarchického shlukování




Zdroj: Collica 2007, vlastní zpracování

- Metoda průměrné vazby (Average Linkage)

  V této metodě je vzdálenost mezi shluky definována jako vzdálenost průměrů jednotlivých shluků:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j). \tag{4.6}$$

  Opět v každém kroku spojíme dva páry s nejmenší naměřenou průměrnou vzdáleností. Tato metoda má tendenci spojovat shluky s malým rozptylem.

- Centroidní metoda (Centroid Linkage)

  V této metodě je vzdálenost mezi shluky definována jako euklidovská vzdálenost mezi vektory průměrů (centroidy) dvou shluků:

$$D(A,B) = d(\bar{y}_A, \bar{y}_B). \tag{4.7}$$

  V každém kroku poté opět spojíme dva shluky s nejmenší vzdáleností mezi vektory. Tato metoda je více robustní k odlehlým hodnotám než většina ostatních hierarchických metod.

- Mediánová metoda (Median Linkage)

  V případě, že se chceme vyhnout větší váze průměrných vektorů v závislosti na velikosti shluků (jak tomu je u metody centroidní), je možné využít medián k vyjádření nové vzdálenosti mezi ostatními shluky:

$$M(A,B) = \frac{1}{2}(\bar{y}_A + \bar{y}_B). \tag{4.8}$$

  V každém kroku shlukové analýzy jsou pak sloučeny shluky s nejmenší vzdáleností mezi mediány.

- Wardova metoda (Incremental Sum of Squares Method)

  Wardova metoda využívá reziduální a meziskupinový rozptyl. Shluky spojuje takovým způsobem, aby součet přírůstků reziduálního rozptylu byl minimální – minimalizuje tedy ztrátu informace při spojení dvou tříd. V porovnání s centroidní metodou je Wardova metoda vhodnější pro spojování malých shluků nebo shluků podobné

velikosti; je však velmi citlivá na odlehlé hodnoty (MELOUN et al. 2005, RENCHER 2002, SAS DOCUMENTATION 2000).

Určité rozdíly mezi výsledky různých shlukovacích technik je možné vysvětlit tím, že některé metody prostor mezi objekty "zužují" tvorbou řetězících se shluků na nízké shlukovací úrovni, jiné prostor "rozšiřují" tvorbou kompaktních shluků na vysoké shlukovací hladině, další metody prostor zachovávají (MARHOLD, SUDA 2002).

**Nehierarchické shlukování**

Metody shlukové analýzy lze dělit do dvou skupin, první skupina metod vychází z matice vzdáleností mezi objekty (metody založené na vzdálenostech). Do této skupiny patří hierarchické shlukování. Druhá skupina metod vychází přímo z původní zdrojové matice, kde každý řádek představuje vektor charakterizující určitý objekt (metody vektorového prostoru). Na tomto principu funguje také nejčastěji využívaná nehierarchická metoda pro detekci shluků, metoda *k*-průměrů. Jedná se o iterační algoritmus, který minimalizuje součet vzdáleností každého objektu od těžiště shluku. Cílem je získat množinu shluků, které jsou kompaktní a navzájem dobře separované. První, kdo použil tento termín, byl J. B. MacQueen. MacQueen (1967) popsal proces rozdělení (partitioning) n-dimenzionální populace do *k* souborů na základě výběru. Jako výhody této metody uvedl jednoduchou programovatelnost, nenáročný výpočet a využitelnost pro velké datové soubory za použití výpočetní techniky. Princip této metody popisuje následovně: Metoda *k*-průměrů začíná s *k* skupinami, kdy každá skupina obsahuje náhodně vygenerovaný bod, k tomuto bodu jsou dále přiřazovány nové body, jejichž průměrná vzdálenost je nejmenší. Po přiřazení nového bodu je průměr skupiny odhadnut tak, aby započítal nově přiřazený bod. V každé fázi tedy jednotlivé průměry reprezentují skupinu, ze které byly vypočítány.

Nezávisle na sobě tento algoritmus využili již dříve Steinhaus (1956), Lloyd (navržen v roce 1957, publikován 1982), Ball a Hall (1965). Klasický algoritmus *k*-průměrů zavedl Hartigan v roce 1975 (více v HARTIGAN 1975, HARTIGAN, WONG 1979). Přestože byl tento algoritmus využit poprvé před několika desítkami let, stále patří k nejčastěji využívaným

technikám shlukování. Snadnost implementace, jednoduchost, účinnost a empirické úspěchy představují hlavní důvody pro jeho popularitu (JAIN 2010).

Základní smysl této metody je jednoduchý. Algoritmus poskytuje fixní počet ($k$) shluků, jednotlivá pozorování rozřazuje do shluků tak, aby shluky byly co nejvíce rozdílné a zároveň minimalizovaly vnitroskupinovou variabilitu (NISBET et al. 2009).

Uvedený parametr $k$ slouží ke zvolení $k$ počtu bodů (těžišť), které jsou náhodně vybrány jako centra shluků. Jednotlivé objekty jsou následně přiřazeny k nejbližšímu centru v závislosti na Euklidovské vzdálenosti. Následně se pro každý shluk spočítá těžiště jednotlivých pozorování, které se stane novým centrem shluku. Poté se celý proces opakuje s novými centry. Iterační proces pokračuje do té doby, než se jednotlivá centra shluků stabilizují a zůstávají neměnná. Nakonec jsou tedy jednotlivá pozorování přiřazena k nejbližšímu centru shluku, čímž se minimalizuje celková čtvercová vzdálenost (WITTEN 2011). Cílem techniky $k$-průměrů je optimalizace následující podmínky:

$$KM(X,C) = \sum_{i=1}^{n} \min_{j \in (1,\ldots k)} \left\| v_i - c_j \right\|^2 , \tag{4.9}$$

kde:

$v_i$ = $i$-tý datový vektor,

$c_i$ = $j$-tý shlukový centroid,

$X$ = datová matice,

$C$ = matice centroidů (TABOADA JIMENEZ 2007).

Algoritmus fungování metody k-průměrů jednoduše popsal Duda et al. (2001):

Start: inicializuj $n$, $k$ a $u_1$, $u_2$, $u_3$, … $u_k$

klasifikuj $n$ výběrů v závislosti na nejbližším $\mu_i$

přepočítávej $\mu_i$

dokud se $\mu_i$ mění

vrať hodnoty proměnných $u_1$, $u_2$, $u_3$, … $u_k$

Konec:

Kde $\mu_i$ značí průměr, $n$ je počet výběrů a $k$ představuje počet shluků.

Algoritmus k-průměrů představuje efektivní metodu, která minimalizuje vnitroskupinovou variabilitu. Je však citlivá na inicializované centrum shluku, které se může změnit v závislosti na změnách vstupních dat. I malá změna může způsobit naprostou změnu center. Nevýhodou této metody je skutečnost, že je nutné před začátkem analýzy zvolit vhodnou hodnotu parametru $k$. Witten (2011) doporučuje jako jeden z možných přístupů zvolit různé hodnoty $k$ a následně vybrat tu nejlepší možnost.

Klíčovou vlastností tohoto algoritmu je skutečnost, že centroid je počítán na základě stávající příslušnosti ke shluku nikoliv na základě jeho příslušnosti na konci iteračního procesu.

Nedostatky metody $k$-průměrů popsal Collica (2007):

- Problémy v případě překrývajících se shluků.
- Citlivost na odlehlá pozorování – centra mohou být kvůli odlehlým pozorováním zkreslená.
- Nelze vyčíslit pravděpodobnost příslušnosti objektu k danému shluku.

V případě algoritmu $k$-průměrů existuje mnoho variací a postupů realizace této metody. Existují metody zárodečných bodů (Seeds), metody výpočtu následujícího centroidu (methods of computing the next centroid), případně lze využít hustotu pravděpodobnosti místo míry vzdálenosti asociovaných objektů v rámci shluku. U metody zárodečných bodů uživatel na základě svých věcných znalostí určí, které objekty mají tvořit zárodky nově vytvořených

shluků a systém rozdělí objekty do shluků podle jejich Eukleidovské vzdálenosti od těchto typických objektů. Existuje několik postupů zadávání zárodků shluku a zařazování objektů do shluku (MELOUN, MILITKÝ 2004).

Metoda *k*-průměrů pracuje s kvantitativními proměnnými, nicméně byla navržena i její modifikace, metoda *k*-modů, která se využívá pro shlukování nominálních dat. Kombinace metody k-průměrů a *k*-modu pro smíšená data se nazývá metoda *k*-prototypů (HEBÁK et al. 2007). V metodě k-prototypů je použita speciální míra nepodobnosti, která kombinuje kvadratickou Euklidovskou vzdálenost, použitou pro kvantitativní data, s mírou užívanou pro pouze kategoriální data v metodě k-modů, založené na koeficientu prostého nesouhlasu. Tento koeficient je definován jako poměr počtu proměnných, u nichž jsou u obou objektů rozdílné hodnoty, a celkového počtu proměnných.

Nejedná se o jediné rozšíření metody *k*-průměrů. V literatuře jsou uváděny dvě nejznámější varianty techniky *k*-průměrů: ISODATA (BALL, HALL 1965) a FORGY (FORGY 1965). Obě tyto techniky přiřazují každý sledovaný objekt právě do jednoho z vytvořených shluků. Metoda fuzzy c-průměrů byla navržena Dunnem (1973) a dále rozpracována Bezdekem (1981). Jedná se o rozšíření metody *k*-průměrů, kde každý bod může příslušet více shlukům zároveň (JAIN 2010).

Na myšlence metody *k*-průměrů je založen také algoritmus EM (Expectation Maximization). Místo jednoznačného přiřazení objektu k určitému shluku jsou objektům přiřazeny váhy, které reprezentují pravděpodobnosti příslušnosti k jednotlivým shlukům. V případě existence odlehlých pozorování není vhodné využívat metody založené na průměru, z tohoto důvodu byly vyvinuty další metody, v nichž je shluk reprezentován jeho konkrétním objektem, *medoidem*, jenž je umístěn nejblíže středu shluku. Tento princip využívá metoda *k*-medoidů (HEBÁK et al., 2007).

Mezi moderní techniky shlukování se řadí i nesupervizovaný algoritmus založený na principu neuronových sítí zvaný Kohonenovy mapy. Tento algoritmus, hledající středy shluků, je velmi podobný principu algoritmu *k*-průměrů. Více o této technice v kapitole věnované neuronovým sítím.

### Fuzzy shluková analýza

Fuzzy příslušnost ke shluku vyjadřuje pravděpodobnost, že záznam patří do daného shluku, případně do více shluků. K vyjádření pravděpodobnostního rozdělení se zpravidla využívá normální (Gaussovo) rozdělení. Tyto varianty metody *k*-průměrů jsou nazývány jako směs gaussovských rozložení (Gaussian mixture models) (COLLICA 2007).

Metoda fuzzy shlukování vychází z matice nepodobností. Pro každý *i*-tý objekt a *h*-tý shluk je počítána míra příslušnosti $u_{ih}$. Míry příslušnosti jsou definovány pomocí minimalizace účelové funkce (ŘEZANKOVÁ 2007).

### Požadavky na metody shlukování

Zvolená metoda shlukování by měla splňovat určité požadavky, kterými jsou především *přiměřená časová náročnost*, *nezávislost na pořadí vstupů do analýzy*, *schopnost zhodnocení platnosti nalezených shluků* a *interpretovatelnost*. Dále by vybraná metoda měla být robustní v následujících oblastech: *dimenzionalita*, *šum a odlehlá pozorování*, *statistické rozdělení*, *tvar, velikost a hustotu shluků*, *oddělení shluků* a *typy proměnných*. V současnosti využívané techniky splňují zpravidla vždy pouze některé z uvedených předpokladů (HEBÁK et al. 2007).

### Modifikace hierarchických metod

Mezi metody vycházející z modifikace hierarchických shlukovacích postupů patří techniky frakcionizace a refrakcionizace. Frakcionizace spočívá v rozdělení datového souboru do podsouborů (frakcí) a aplikování hierarchické metody na každou frakci. Shluky vzniklé ve frakcích jsou dále shlukovány do předem stanovených *k* skupin stejnou metodou shlukové analýzy. Zbylé objekty jsou přiřazeny do vytvořených *k* shluků na základě centroidů. Problémy spojené s metodou frakcionizace, jako je např. potřeba předem specifikovat počet shluků, či využití metaobjeků, které nemusí být vhodným reprezentantem skupiny, vedly k vytvoření algoritmu refrakcionizace. Tento algoritmus se liší od výše zmíněného tím, že

shluky vzniklé frakcionizací vytvářejí frakce pro následující iteraci. Součástí algoritmu je odhad počtu shluků *k* (HEBÁK et al. 2007).

### 4.6.2 Asociační pravidla

Cílem asociačních pravidel je detekovat skryté vztahy neboli asociace mezi specifickými hodnotami kategorických proměnných ve velkých datových souborech. Tato pravidla umožňují například identifikovat položky, které zákazníci kupují společně; mezi těmito položkami existuje asociace. Asociační algoritmy mohou být využity k analýze kategoriálních proměnných, dichotomických proměnných a/nebo vícenásobných cílových proměnných. Algoritmus nevyžaduje přesnou definici kategorií v datech, kontingenční tabulky mohou být sestrojeny bez nutnosti specifikovat počet proměnných nebo kategorií. Tato technika je velmi přínosná především při aplikaci na velké datové sklady (NISBET et al. 2009).

Asociační pravidla jsou často používána při analýze nákupního koše, kdy dochází k modelování asociační struktury nákupního chování. Tato analýza využívá asociační techniky k nalezení skupin položek, které mají tendenci vyskytovat se společně v jednotlivých nákupních koších - v souboru položek. Tato technika má explorační charakter a je založena na zkoumání párových marginálních poměrů šancí. Cílem je identifikace asociací mezi nakupujícími zákazníky vybírajícími z různých produktů, obvykle v rámci určité jednotky (supermarketu). Analýza dat sestává ze všech provedených transakcí uskutečněných klienty v dané jednotce za určitý čas. Nalezené asociace je možné využít při plánování marketingové strategie. Do analýzy asociací může vstupovat i časové hledisko, například asociace dvou různých produktů může existovat pouze v určitý den v týdnu (GIUDICI 2002, WITTEN 2011).

Výhodou asociačních pravidel je srozumitelnost a názornost dosažených výsledků. Nevýhodou naopak je skutečnost, že tyto výsledky nemusí být vždy užitečné, v některých případech jsou triviální (nepřekvapivé) či logicky nevysvětlitelné (BERRY, LINOFF 2004).

Asociačním pravidlem je nazýván výraz ve formě (Položka A) => (Položka B), neboli položka A implikuje položku B. Cílem analýzy je určit sílu asociačních pravidel v rámci souboru položek. Tato síla asociace je měřena pomocí ukazatelů *support* a *confidence*.

Ukazatel podpora (*Support*) určuje pravděpodobnost, že se dané položky objeví společně (GEORGES 2009). Pro pravidlo *A* => *B* se určí pomocí následujícího vzorce

$$Support = \frac{Transakce, které\ obsahují\ obě\ položky\ A\ a\ B}{Všechny\ transakce}. \qquad (4.10)$$

Spolehlivost (*Confidence*) představuje podmíněnou pravděpodobnost: když transakce obsahuje položku *B*, tak také obsahuje položku *A* (GEORGES 2009). Berry a Linoff (2004) definují spolehlivost takto:

> „*...confidence is the ratio of the number of the transactions supporting the rule to the number of transactions where the conditional part of the rule holds*".

Spočítá se podle vzorce:

$$Confidence = \frac{Transakce, které\ obsahují\ obě\ položky\ A\ a\ B}{Transakce, které\ obsahují\ položku\ A}. \qquad (4.11)$$

Další charakteristikou je *lift* neboli zlepšení. Tento ukazatel udává, o kolik lépe pravidlo predikuje výsledek oproti náhodnému výběru (tzn. v případě, že by neexistoval žádný vztah mezi danými položkami). Berry a Linoff (2004) definují zlepšení takto:

> „*Lift is the ratio of the density of the target after application of the left-hand side to the density of the target in the population.*"

Spočítá se podle vzorce:

$$Lift(A \rightarrow B) = \frac{Confidence\,(A \rightarrow B)}{Support\,(B)} = \frac{Support\,(A \rightarrow B)}{Support\,(A)\,Support\,(B)}. \qquad (4.12)$$

Ukazatel *lift* může být interpretován jako obecná míra asociace mezi dvěma položkami. Hodnota vyšší než 1 indikuje pozitivní závislost, hodnota rovna 1 indikuje nulovou korelaci (položky jsou nekorelované) a hodnota nižší než 1 indikuje negativní korelaci. *Lift* je symetrická míra, tato míra je tedy stejná u pravidla *A* => *B* i u *B* => *A* (GEORGES 2009). V případě, že je hodnota *liftu* větší než 1, pak nalezené pravidlo předpovídá lépe než náhodný výběr.

Podobnou mírou je *Excess*, která představuje rozdíl mezi počtem jednotek podporujících pravidlo a očekávanou hodnotou. Výhodou této míry je, že je uvedena v původních jednotkách (BERRY, LINOFF 2004).

### 4.6.3 Logistická regresní analýza

Další důležitou skupinou metod, kterou je možné využít pro predikci, je logistická regresní analýza. Logistická regrese představuje prediktivní model pro kvalitativní proměnnou. Dle Giudiciho (2009) se jedná o jeden z nejvýznamnějších prediktivních dataminingových modelů. Představuje alternativní metodu klasifikace v případě, že nejsou splněny předpoklady vícerozměrného normálního modelu. Klasická logistická regrese, také nazývaná logitové modely, využívá binární odezvu, tudíž je předpokládáno binomické rozdělení. Obecnou logistickou regresi je však možné aplikovat i na vícekategoriální proměnnou. Loglineární modely jsou aplikovány na data s Poissonovým rozdělením (MELOUN et al. 2005).

Dle typu vysvětlující proměnné je možné rozlišit logistickou regresi:

1) binární – binární vysvětlovaná proměnná, nabývající dvou možných obměn;
2) ordinální – ordinální vysvětlovaná proměnná, nabývající tří a více hodnot;
3) nominání – nominální závisle proměnná s více než třemi stavy.

Logistická regrese se liší od lineární v tom, že predikuje pravděpodobnost, zda daná událost nastala či nikoliv. Vypočtená pravděpodobnost je potom rovna buď 0 (událost nenastala), nebo 1 (událost nastala). K vytvoření této vazebné podmínky je využita logitová transformace, která vede na sigmoidální vztah mezi závisle proměnnou $y$ a vektorem nezávisle proměnných $x$. Logitová transformace vychází z tzv. *poměru šancí*.

Pomocí logistické regrese lze predikovat hodnoty výstupní kategoriální proměnné. Vysvětlujícími proměnnými jsou v tomto případě kategoriální či spojité proměnné. Kategoriální vysvětlující proměnné bývají označovány jako prediktory, kvalitativní jako faktory. Pomocí logitové transformace je z binární proměnné (0, 1) přes odpovídající pravděpodobnosti (P1, D1) ležící v intervalu (0, 1) vytvořena spojitá veličina „šance“ $C$, definovaná v intervalu (0, nekonečno), jejíž logaritmus může nabývat libovolných hodnot na celé reálné ose. To znamená, že ln $C$ je spojitá veličina, neomezená konečným intervalem

a lze ji přímo použít jako vysvětlující proměnnou v klasických lineárních regresních modelech.“ (MELOUN et al. 2005).

Rovnice jednoduché logistické regrese lze zapsat:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \varepsilon_i)}}. \tag{4.13}$$

Tuto rovnici lze ekvivalentně rozšířit na rovnici vícenásobné lineární regrese.

Obdobou Pearsonova korelačního koeficientu je v logistické regresi ukazatel *log-likelihood*, neboli logaritmus věrohodnostní funkce, který vychází ze součtu pravděpodobností spojených s predikovanými a skutečnými výstupy. Tato charakteristika je analogická k reziduálnímu součtu čtverců odchylek ve vícenásobné regresi. Vysoká hodnota věrohodnostní funkce (*log-likelihood*) indikuje, že model slabě popisuje skutečnost – tzn. existuje velký počet nevysvětlených pozorování. Tato charakteristika slouží především k porovnání více modelů (FIELD 2005).

Postup, vyčíslující logistické koeficienty, porovnává pravděpodobnost události odehrané $L_{(1)}$ vůči pravděpodobnosti události neodehrané $L_{(0)} = 1 - L_{(1)}$ využitím *pravděpodobnostního poměru* $L_{(1)}/L_{(0)}$, ve kterém je pravděpodobnost $L_{(1)}$ vyjádřena logistickou funkcí:

$$L_{(1)} = \frac{1}{1 + e^{\ln\left(\frac{L_{(1)}}{L_{(0)}}\right)}}. \tag{4.14}$$

Vícenásobný logistický regresní model (logit), lze tedy zapsat jako logaritmus poměru šancí:

$$\ln \frac{L_{(1)}}{L_{(0)}} = b_0 + b_1 x_1 + \ldots + b_p x_P. \tag{4.15}$$

Uvedený model lze upravit do ekvivalentního tvaru:

$$L_{(0)} = \frac{1}{1 + Exp[-(b_0 + b_1 x_1 + \ldots + b_p x_p)]}. \tag{4.16}$$

Parametr $b_i$ určuje míru a typ závislosti mezi šancí přijetí a vysvětlující proměnnou $x_i$. Kladné znaménko koeficientu $b_i$ zvyšuje pravděpodobnost $L_{(1)}$ a záporné ji naopak snižuje. Je-li $b_i$ kladné, funkce *Exp* je větší než 1 a pravděpodobnostní poměr se tudíž bude zvyšovat.

Podobně je-li $b_i$ záporné, je funkce *Exp* menší než 1 a pravděpodobnost se sníží. V případě, že je koeficient roven nule, pravděpodobnost se nezmění (MELOUN et al. 2005).

Změnu šancí *(Exp(B))* je možné vyjádřit podílem šance po jednotkové změně prediktoru k původní šanci:

$$\Delta odds = \frac{odds\ after\ a\ unit\ change\ in\ the\ predictor}{original\ odds} . \qquad (4.17)$$

Pro otestování významnosti parciálního regresního koeficientu $b_i$ ($H_0$: $\beta_i = 0$), se využívá Waldovo testové kritérium:

$$z = \frac{b_i}{SE_b} , \qquad (4.18)$$

které testuje statistickou významnost odhadů regresních koeficientů. $SE_b$ v tomto případě představuje směrodatnou odchylku regresního koeficientu $b_i$.

Waldovu statistiku je však nutné využívat obezřetně, v případě vysoké absolutní hodnoty regresního koeficientu a vysoké hodnoty směrodatné odchylky má směrodatná chyba tendenci růst, což způsobuje, že se Waldova statistika stává podhodnocená. Testové kritérium představuje příliš malou hodnotu, která vede k nekorektnímu zamítnutí nulové hypotézy o významnosti regresního koeficientu. Rozhodnutí učiněná na základě Waldovy charakteristiky je proto vhodné podložit vyhodnocením změny logistického modelu v případě zařazení či nezařazení proměnné do modelu (FIELD 2005).

### 4.6.4 Neuronové sítě

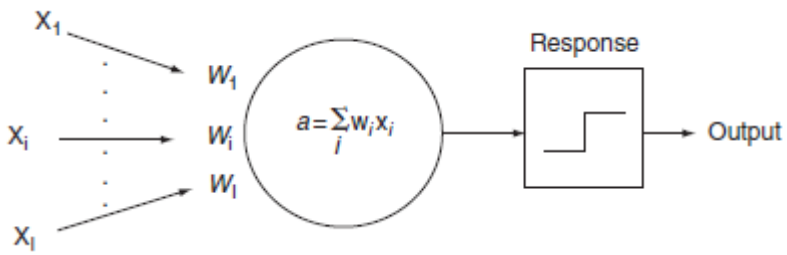
Neuronové sítě patří společně s např. genetickými algoritmy, strojovým učením, fuzzy logikou a rozpoznáváním vzorů mezi disciplíny umělé inteligence. Neuronové sítě představují typický příklad využití prediktivního modelování pro vytvoření předpovědi, kdo by mohl zakoupit produkt na základě známých dat předchozího prodeje výrobku. Realizace modelu neuronových sítí spočívá ve zpracování historických dat obsahujících nákupní chování zákazníků v minulosti a pomocí toho je možné určit řadu vážených hodnot, které korelují s napozorovanými vstupními skupinami (RAHMAN 2008).

Dle Awada (2009) se jedná o velmi účinnou robustní klasifikační techniku, kterou je možné využít v různých oblastech. Uměle vytvořené neuronové sítě simulují nervový systém člověka, skládající se z velkého množství vysoce propojených spolupracujících jednotek (neuronů), které vytvářejí naše reakce a emoce. Stejně tak jako lidé, i neuronové sítě se zkušenostmi učí. Proces učení neuronové sítě představuje upravování jednotlivých vah v modelu.

Existují různé typy neuronových sítí, nejčastěji se využívají: Mnohovrstvový perceptron (Multilayer Perceptron), Lineární sítě (Linear Network), Bayesovské sítě (Bayesian Network), Pravděpodobnostní sítě (Probabilistic Network), Zobecněné regresní modely (Generalized Regression) a Kohonenovy samoorganizující se mapy (Kohonen self organizing maps).

Podstata umělých neuronových sítí je založena na racionálním napodobení struktury principů činnosti biologických neuronových sítí pomocí technických nebo programových prostředků. Simulují schopnost lidského myšlení učit se, kdy proces učení lze definovat jako proces automatického navazování nových spojů. Umělé neuronové sítě se skládají z umělých neuronů, jejichž předobrazem je biologický neuron. Neurony jsou vzájemně propojeny a navzájem si předávají signály a transformují je pomocí určitých přenosových funkcí, zpravidla lineárních nebo logistických. Lineární přenosové funkce se využívají pro numerické odhady (např. regresi), logistické naopak pro klasifikační problémy. Neuron má libovolný počet vstupů, ale pouze jeden výstup. Neuronová síť je popsána pomocí dynamicky se měnícího orientovaného grafu s ohodnocenými hranami a uzly (NISBET et al. 2009, ŘEZANKOVÁ 2007).

Obr. č. 8: Architektura neuronu s počtem vstupů $X_i$



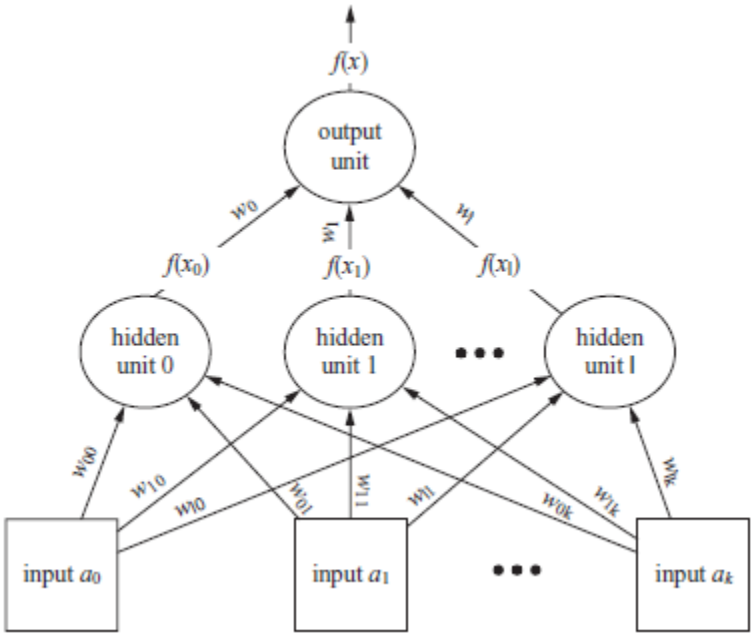

Zdroj: NISBET et al. (2009)

Obr. č. 8 zachycuje architekturu neuronu s počtem vstupů ($X_i$). $X_i$ značí počet vstupních proměnných a reprezentuje počet ostatních neuronů připojených do lidského neuronu. $W_i$ představuje číselnou váhu propojenou s každou vazbou a určuje sílu propojení. Tato síla spojení představuje blízkost spojení mezi dvěma neurony v dané oblasti a nazývá se synapse (BISHOP 1995).

Jednotlivé umělé neurony jsou propojeny a tvoří procesní strukturu. Tato forma sítě, ve které je každá vstupní proměnná propojena s jednou nebo více dalšími, se nazývá umělá neuronová síť (Artificial Neural Network). V případě, že jsou jednotlivé vstupní proměnné propojeny pomocí agregační funkce a logistické aktivační funkce přímo k výstupnímu uzlu, je matematické zpracování analogické logistické regresní analýze s binárním výstupem. Neuronové sítě bez skrytých vrstev tedy v podstatě poskytují obdobné výsledky jako regresní analýza. Neuronová síť může být konstruována např. pouze s jedním výstupním uzlem a být nastavena jako regrese nebo binární klasifikátor. Alternativně může být sestavena s více výstupními modely a fungovat jako seskupovací algoritmus (CERRITO 2006, NISBET et al. 2009).

Cerrito (2006) uvádí, že komplexnost modelu roste s přibývajícími skrytými vrstvami a dodatečnými vstupními proměnnými. Obr. č. 9 zachycuje váhy ($W_{ij}$), které jsou umístěny ke každému spojení mezi jednotlivými vrstvami. Tyto váhy společně s prostřední (skrytou) vrstvou umožňují modelovat nelineárních vztahy mezi vstupními a výstupními uzly. Tato vlastnost přináší velký význam pro řešení dataminingových problémů. Vyšší počet uzlů ve střední (skryté) vrstvě poskytuje vyšší kapacitu k rozpoznání nelineárních vztahů v datovém souboru. S růstem počtu těchto uzlů však roste exponenciálně i čas učení, s čímž souvisí vyšší pravděpodobnost přeučení modelu (NISBET et al. 2009).

Obr. č. 9: Architektura neuronové sítě s jednou skrytou vrstvou




Zdroj: Witten et al. (2011)

**Kohonenovy samoorganizující se mapy**

V roce 1988 Kohonen přišel s úplně novou myšlenkou neuronové sítě, která nepotřebuje ani učitele, ani nemá zbloudilé asociace, ale sama o sobě mění své vnitřní chování, aniž by se od člověka dozvěděla, zda se chová dobře nebo špatně. V takovém případě se hovoří o samoorganizujících se mapách (Self-Organizing Map, SOM) založených na soutěživosti mezi paralelními neuronovými buňkami. Sám Kohonen tyto sítě realizoval elektronicky a použil je k analýze lidské řeči (KUKAL 2005).

Principem tohoto algoritmu je opakované třídění vstupů do té doby, než jsou rozdíly mezi třídami maximální. Tato technika je používána jako alternativa ke klasickému seskupování dat (clustering) za předpokladu, že počet případů nebo kategorií není příliš velký. V případě velkého datového souboru zabere učení této sítě velmi dlouhou dobu (NISBET et al. 2009). Tato neuronová síť je schopná naučit se rozpoznávat shluky dat a přiřadit k sobě podobné třídy. Slouží např. k vyhledávání shluků na webových stránkách nebo v textových

dokumentech. Po rozpoznání shluků v datech může být použita také pro klasifikaci (ŘEZANKOVÁ 2007).

Kohonenovy mapy se skládají z jednotek uspořádaných do dvourozměrného vizualizačního prostoru. Cílem je zobrazit vícerozměrná data do dvourozměrného prostoru tak, aby byly podobné objekty umístěny v mapě co nejblíže. Zpravidla se jednotky uspořádávají do pravoúhlé mřížky. Každé jednotce je přiřazen modelový vektor z vícerozměrného vstupního prostoru. Vstupní položka (charakterizovaná svým vektorem) je zařazena do jednotky s nejvíce podobným modelovým vektorem. Kohonenova mapa může být náhodně inicializována – náhodné vektory ve vstupním prostoru jsou přiřazeny každému modelovému vektoru. Alternativu představuje např. inicializace na základě dvou hlavních představitelů dat (více v KOHONEN et al. 2001, WEINLICHOVÁ, FEJFAR 2010).

Algoritmus samoorganizujících se map je založen na kompetiční strategii učení. Vstupní vrstva slouží k distribuci vstupních vzorů $p_i$, $i = 1, 2, \ldots, n$. Neurony v kompetiční vrstvě jsou reprezentanty vstupních vzorů a jsou organizovány do topologické struktury. Ta určuje, které neurony spolu sousedí. Nejprve jsou vypočteny Euklidovské vzdálenosti $d_j$ mezi vzorem $p_i$ a vahami synapsí $w_{i,j}$ všech neuronů v kompetiční vrstvě. Je vybrán ten vítězný neuron, pro který je vzdálenost $d_j$ od vzoru $p_i$ minimální. Výstup tohoto neuronu je aktivní, zatímco výstupy ostatních neuronů jsou neaktivní. Cílem učení je aproximovat hustotu pravděpodobnosti vstupních vektorů pomocí konečného počtu reprezentantů. Po nalezení reprezentantů je každému vzoru přiřazen reprezentant vítězného neuronu (OLEJ, HÁJEK 2007).

**Výhody a nevýhody neuronových sítí**

Cerrito (2006) označuje neuronové sítě jako černou skříňku. Model neuronové sítě není možné popsat rovnicí či souborem rovnic, ani není prezentován ve výstižném formátu, jaký je dostupný například v regresní analýze. Neuronové sítě také nepoužívají žádné testování hypotéz, žádné p-hodnoty nejsou k dispozici pro porovnání jednotlivých vstupních proměnných. Jednotlivé dataminingové softwary se však již pokouší tzv. „*black box*“, jak je někdy neuronová síť nazývána, otevřít a poskytují uživateli alespoň souhrnnou informaci o tom, jaký přínos mají jednotlivé proměnné pro konstrukci neuronové sítě. Další nevýhodou

neuronových sítí je relativně velká časová a hardwarová náročnost při učení a aplikaci těchto sítí (NISBET et al. 2009).

Naopak výhodou neuronových sítí je skutečnost, že jsou obecnými klasifikátory, poradí si s problémy, které mají mnoho parametrů a jsou schopny klasifikovat objekty i v případě, že rozdělení těchto objektů v *n*-dimenzích je velmi komplexní. Také si poradí s velkým množstvím nelinearity u vysvětlujících proměnných. Mohou být využity i pro číselné vysvětlované proměnné. Nemají žádné požadavky na pravděpodobnostní rozdělení vstupních dat. Jsou vhodné ke zkoumání nelineárních vztahů, jelikož skryté vrstvy poskytují schopnost efektivně modelovat vysoce nelineární funkce (NISBET et al. 2009).

### 4.6.5 Analýza hlavních komponent

Analýza hlavních komponent je obecně diagnostický nástroj pro identifikaci a hodnocení zvláštností posuzovaných a analyzovaných datových souborů. Je doporučována především pro průzkumovou analýzu dat. Dále se využívá společně s některými dalšími technikami analýzy vícerozměrných dat, ale i jako samostatná metoda rozboru vztahů v množině vzájemně závislých proměnných. Tato metoda byla původně zavedena K. Pearsonem už v roce 1901 jako popisná statistická metoda, sloužící především k redukci vícerozměrných dat. Hotelling (1933) zobecnil postup aplikací komponentní analýzy na náhodné vektory a navrhl její použití pro rozbor kovarianční struktury proměnných (HEBÁK et al. 2007).

Podstatou analýzy hlavních komponent je snížit dimenzionalitu souboru dat, a zároveň uchovat informace přítomné v původním datovém souboru (LAVINE 2000). Hlavními cíli této analýzy jsou na jedné straně nalezení správného rozměru souboru dat a tím bez výrazné ztráty informace zlepšení kvality analýzy a na straně druhé nalezení nových proměnných. Jedná se v podstatě o transformaci původních proměnných $x_i$, $i = 1, ..., m$, do menšího počtu latentních proměnných $y_j$. Tyto proměnné mají vhodnější vlastnosti, jejich počet je výrazně nižší, vystihují téměř celou proměnlivost původních proměnných a jsou vzájemně nekorelované. Latentní proměnné jsou u této metody nazvány hlavními komponentami a jde o lineární kombinace původních proměnných: první hlavní komponenta $y_1$ popisuje největší část proměnlivosti čili rozptylu původních dat, druhá hlavní komponenta $y_2$ zase největší část rozptylu neobsaženého v $y_1$ atd.

Při výpočtu hlavních komponent se vychází z populace, ve které náhodné veličiny $X_1$, $X_2$, …, $X_p$ mají vícerozměrné normální rozdělní s $p$-členným vektorem středních hodnot $\boldsymbol{\mu}$ a s kovarianční maticí $\boldsymbol{\Sigma}$ (hodnosti $p$). První hlavní komponenta je definována:

$$Y_1 = \varpi_1^T (x - \mu), \tag{4.19}$$

kde vektor $\omega_1$ je určen maximalizací rozptylu komponenty $Y_1$ přes všechny vektory $\omega_1$ tak, aby byla splněna normalizační podmínka

$$\varpi_1^T \varpi_1 = 1. \tag{4.20}$$

Analogicky jsou definovány i všechny další komponenty.

Pro použití hlavních komponent ve statistické analýze je třeba vypočítat komponentní skóre, tedy hodnoty hlavních komponent pro každou výběrovou jednotku. Označíme-li jako $x_i$ vektor hodnot i-té jednotky souboru, $i = 1, 2, \ldots, n$, komponentní skóre r-té hlavní komponenty u i-té jednotky je

$$y_{ir} = \omega_r^T (x_i - \mu), \; r = 1, 2, \ldots, R, \; i = 1, 2, \ldots, n. \tag{4.21}$$

Podrobnější informace lze nalézt např. v (HEBÁK et al. 2007, MELOUN et al. 2005, RENCHER 2002).

### 4.6.6 Rozhodovací stromy

Berry a Linoff (2004) definují rozhodovací stromy jako nástroj klasifikace a prediktivního modelování. Tyto postupy tedy nejsou určeny pouze k analýze kvantitativních dat, ale také pro data kvalitativní. Užívají se jako alternativa k diskriminační či regresní analýze nebo neuronovým sítím. Model rozhodovacího stromu je složen z množiny pravidel sloužících k rozdělení heterogenní populace do menších homogenních skupin podle specifické cílové proměnné. Principem je vytvoření stromové struktury na základě rozhodovacích pravidel. Dle Giudiciho (2009) je možné dodat, že cílem dělení je maximalizace homogenity v jednotlivých skupinách.

Výstupem rozhodovacího stromu je konečné rozštěpení jednotlivých pozorování. K dosažení tohoto výsledku je nutné stanovit kritérium pro zastavení procesu dělení. Konečný výstup z analýzy je tvořen stromem podobným dendrogramu, jenž je výstupem hierarchického

shlukování. Každá větev stromu je reprezentována jako klasifikační pravidlo (GIUDICI 2009).

Výhodou rozhodovacích stromů je skutečnost, že jsou reprezentovány právě zmiňovanými pravidly. Jelikož rozhodovací stromy kombinují možnosti explorační analýzy a modelování, jsou vhodným prvním krokem v modelovacím procesu a to i v případě, že finální model bude získán jinou technikou. Další výhoda použití rozhodovacích stromů spočívá v jednoduché interpretaci výsledků, možnostech vytvoření komplexních vstupně-výstupních asociací a také ve schopnosti automaticky se vypořádat s problematikou chybějících hodnot bez nutnosti imputace. Chybějící údaj je v případě rozhodovacího stromu považován za samostatnou kategorii. Stromy mohou být generovány/trénovány automaticky za pomoci různých algoritmů nebo vytvořeny ručně, tzv. systematickým štěpením (BERRY, LINOFF 2004).

Giudici (2009) zařazuje rozhodovací stromy mezi neparametrické prediktivní modely, jelikož nevyžadují žádné předpoklady pravděpodobnostního rozdělení vysvětlující proměnné. To v podstatě znamená, že tyto modely lze aplikovat na jakékoliv vysvětlované a vysvětlující proměnné. Ale tato flexibilita může mít i nevýhody. Sekvenční charakter rozhodovacích stromů a jejich algoritmická složitost je může učinit závislými na získaných údajích, a proto i malá změna v datech může významně ovlivnit strukturu stromu. Je proto náročné vytvořenou stromovou strukturu navrženou pro konkrétní případ zobecnit.

Rozhodovací stromy je možné dělit do dvou skupin, regresní stromy, kde vysvětlovaná proměnná je spojitá, a klasifikační stromy, kde cílovou proměnnou může být spojitá, diskrétní či nominální veličina. Běžně využívaným zástupcem rozhodovacích stromů je algoritmus CHAID (Chi-squared Automatic Interaction Detection), který je využíván k analýze kategoriálních dat. Pro štěpení jednotlivých uzlů využívá tento algoritmus test věrohodnostního poměru, tzn. v každém kroku je vybrána taková proměnná, která má největší vliv na hodnoty vysvětlované proměnné (ŘEZANKOVÁ et al. 2007). Druhým hojně využívaným algoritmem je CART (Classification and Regression Trees). Dva hlavní aspekty CART algoritmu jsou dělící kritéria a prořezávání (GIUDICI 2009).

### 4.6.7 Další využívané dataminingové metody

**Diskriminační analýza**

Narozdíl od shlukové analýzy umožňuje další z využívaných technik – diskriminační analýza (Discriminant analysis, DA) – zařazení objektů do již existujících, pevně stanovených, skupin (segmentů). Techniky diskriminační analýzy se dělí na dva základní typy, jedny umožňují hodnocení rozdílů mezi předem stanovenými skupinami objektů, cílem druhé skupiny technik je pak rozřazení objektů do skupin na základě charakteristických znaků. Objekty jsou zařazovány do již existujících tříd podle míry podobnosti, např. dle nejmenší Mahalanobisovy vzdálenosti. K rozřazování do skupin se využívá tzv. diskriminační funkce; každá primární třída je charakterizovaná svou funkcí hustoty pravděpodobnosti $f_j(x)$. V případě, že data vykazují silnou nenormalitu, např. přítomnost binárních proměnných, je možné použít k výpočtu pravděpodobnosti, že objekt je členem dané třídy, logistický model - neboli logistickou diskriminaci (MELOUN et al. 2005).

**Korespondenční analýza**

Korespondenční analýza představuje populární grafickou techniku využívanou k analýze vztahů mezi kategoriemi jedné či více proměnných v kontingenčních tabulkách. Pomocí nástrojů korespondenční analýzy je možné popsat asociace nominálních či ordinálních proměnných a získat grafické znázornění souvislostí ve vícerozměrném prostoru (RAMOS–CARVALHO 2010). Beh (2010) spatřuje největší výhodu této metody právě v její schopnosti graficky znázornit propojenost jednotlivých kategorií. Základem vytváření subjektivní (korespondenční) mapy jsou tzv. latentní veličiny. Polohy bodů v subjektivní mapě přímo vyjadřují asociaci; vzdálenosti mezi body (neboli vzdálenost řádkových a sloupcových profilů) je možné přenést do dvojrozměrné euklidovské roviny, ve které body odpovídají jednotlivým kategoriím (RENCHER 2002).

Hebák (2007) dodává, že korespondenční analýza zobrazuje korespondence kategorií jednotlivých proměnných a poskytuje společný obraz řádkových i sloupcových kategorií ve stejných dimenzích. Na rozdíl od většiny ostatních vícerozměrných metod umožňuje korespondenční analýza zpracování kategorizovaných nemetrických dat i nelineárních vztahů. Představuje obdobu faktorové analýzy, místo faktorů je však sledován vliv jednotlivých

kategorií, jejich vzájemná podobnost či asociace s kategoriemi ostatních proměnných (RENCHER 2002).

Cílem korespondenční analýzy je dle Hebáka et al. (2007) „*...redukce mnohorozměrného prostoru vektorů řádkových a sloupcových profilů při maximálním zachování informace obsažené v původních datech.*“ V subjektivním mapování bývá nejčastěji využíváno dvojrozměrného (roviny) či maximálně trojrozměrného zobrazení vzdáleností v euklidovském prostoru. Častěji než euklidovská vzdálenost se využívá Pearsonova statistika chí-kvadrát. Blízké řádkové body indikují řádky, které mají podobné profily v celém řádku, blízké sloupcové body indikují sloupce s podobnými profily směrem dolů přes všechny řádky. A řádkové body, které jsou v těsné blízkosti sloupcových bodů, představují kombinace, které se objeví častěji, než by se očekávalo u nezávislého modelu, ve kterém řádkové kategorie nejsou vztaženy ke sloupcovým (MELOUN et al. 2005).

Rozptýlenost bodů je možné posuzovat dle ukazatele inercie, který odpovídá váženému průměru chí-kvadrát vzdáleností řádkových (respektive sloupcových) profilů od svého průměru (MELOUN et al. 2005). Singulární hodnota a inercie odpovídá vlastnímu číslu v analýze hlavních komponent = míra „variability“ mezi profily vysvětlená danou dimenzí řešení nebo danou kategorií. Podle toho určíme potřebný počet dimenzí. Odlišnost profilů, měřená pomocí míry založené na chí-kvadrát statistice, je to, co se projeví v grafu jako vzdálenost mezi položkami stejné proměnné. Vzdálenost mezi položkami různých proměnných jsou obrazy standardizovaných reziduí na průsečíku položek.

Jako hlavní předpoklad pro použití korespondenční analýzy uvádějí autoři Meloun a Militký (2005) kromě porovnatelnosti objektů také úplnost datové matice. Řešení vychází z matice standardizovaných reziduí, kterou je možné vytvořit na základě některé z normalizačních metod. Výběr normalizační metody závisí na preferencích výzkumníka. Při preferencích vztahů mezi řádkovými kategoriemi je využívána analýza řádkových profilů, při upřednostnění sloupcových kategorií vycházíme z analýzy sloupcových profilů. Kombinací těchto dvou analýz je metoda symetrická, která umožňuje vzájemné srovnání řádkové a sloupcové kategorie. Tato metoda je preferována pokud je cílem vytvořit bodový graf sloupcových a řádkových profilů, tzv. symetrické mapy (ŘEZANKOVÁ 2007).

**Podpůrné vektory**

Relativně novým způsobem analýzy vícerozměrného datového souboru je seskupení jednotlivých bodů do vektorů (neboli řádků v záznamech). Tyto vektory mohou být zobrazeny jako *n*-dimenzionální prostor, kde *n* je počet atributů (vysvětlujících proměnných). Tento přístup, nazvaný statistická teorie učení, využívá výhod lineární algebry (NISBET et al. 2009). Nejznámější implementaci představuje technika podpůrných vektorů (Support Vectore Machine). Algoritmus navržený Aizermanem, Bravermanem a Rozonoerem rozvinul Vapnik (1992), který nazval výpočetní model provádějící klasifikaci pomocí jader *support vector machine* (SVM) neboli algoritmus podpůrných vektorů. Původně se jednalo o binární klasifikátor pro separabilní problém, v roce 1995 byla tato metoda rozšířena i pro neseperabilní problémy. Principem je nalezení takové přímky, roviny či nadroviny, která oddělí instance obou tříd. Kromě této techniky již byla představena i SVM regrese (NISBET et al. 2009, VAPNIK 1998).

V terminologii neuronových sítí je SVM síť s jednou skrytou vrstvou tvořenou jádrovými jednotkami a s jednou prahovou výstupní jednotkou. V současné době jsou jádrové metody předmětem intenzivního výzkumu jak v oblasti teorie, tak v oblasti aplikací. Jádra jsou vhodnými prvky výpočetních modelů ze dvou důvodů: umožňují měnit skalární součin a tím geometrii prostorů dat a vyjádřit míru nežádoucích oscilací funkcí vstup-výstup (KŮRKOVÁ 2008, VAPNIK 1998).

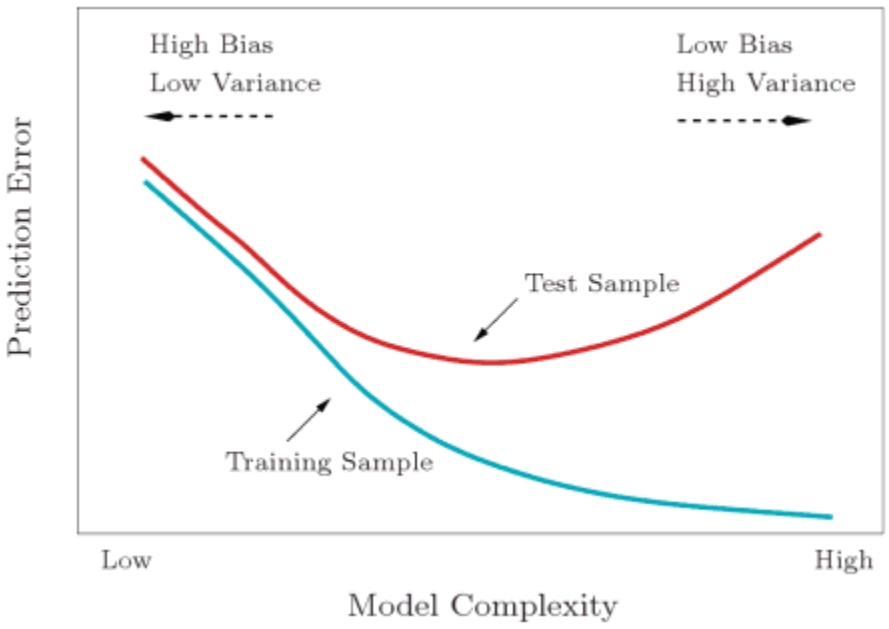
## 4.7 Validace získaných predikčních modelů

Cílem validace dataminingových modelů je zvolit takový model, který je nejjednodušší a zároveň vysvětluje největší množství informace. Mezi validační kritéria založená na statistickém testování patří různé typy vzdáleností mezi jednotlivými modely. Jde např. o Chí-kvadrát či Euklidovskou vzdálenost. Tyto vzdálenosti slouží k sestavení kritérií založených na nesouladu (discrepancy) statistického modelu. Mezi další kritéria patří kritéria založená na hodnotících (scoring) funkcích. Sem patří například střední kvadratická chyba (Mean Square Error, MSE) (GIUDICI 2009).

Přesnost a predikční schopnost získaných modelů je možné zkoumat například pomocí míry nesprávné klasifikace (Misclassification Rate), Giniho indexu, Akaikeho informačního kritéria či střední kvadratické chyby (CERRITO 2006). Jako nástroj pro ověření kvality prediktivního modelu je také možné zvolit Gain diagram. Tento graf porovnává model získaný dataminingovými postupy s modelem náhodného výběru. Zachycuje procentuální podíl pravděpodobnosti výskytu cílové proměnné.

S množstvím proměnných vstupujících do modelu se zpravidla zvyšuje jeho přesnost, roste ale i pravděpodobnost nadhodnocení modelu. Z tohoto důvodu informační kritéria penalizují počet proměnných v modelu. Výsledkem by měl být kompromis mezi složitostí modelu a jeho přesností (viz Obr. č. 10).

Obr. č. 10: Vztah mezi chybou predikce a složitostí modelu




Zdroj: Hastie et al. (2009)

Informační kritérium AIC (Akaike Information Criterion) navrhl Akaike v roce 1974. Toto kritérium je založeno na penalizaci složitosti modelu. Výsledkem by měl být kompromis mezi složitostí modelu a jeho přesností. Obdobným kritériem je bayesovské informační kritérium

(Bayesian Information Criterion, BIC), které zkonstruoval Schwarz v roce 1978. U tohoto kritéria je penalizace přidaných proměnných větší, než v případě AIC (GIUDICI 2009).

Další technikou validace je rozdělení souboru na trénovací a testovací soubor. Nevýhodou tohoto přístupu je nutnost většího počtu pozorování, aby při rozdělení nedocházelo k výrazné ztrátě informace. Nejčastěji využívanými přístupy k hodnocení modelů jsou techniky založené na křížovém ověřování (Cross-validation). Výhodou těchto kritérií je jejich aplikovatelnost na různé modely. Křížovou validací se rozumí rozdělení pozorování do *k* nezávislých podsouborů. Principem je vytvoření *k* modelů, kdy při každém modelování je jeden z podsouborů využit pro validaci. Z výstupů se zhodnotí stabilita a predikční schopnost zvolené modelovací techniky.

Meloun et al. (2005) dále doporučuje využít pro hodnocení kvality získaného modelu analýzu ROC (Receiver Operating Characteristics), tedy matici záměn (viz Tab. 3) a křivku ROC, neboli křivku prahové operační charakteristiky. Plocha pod křivkou představuje kritérium kvality zvoleného predikčního modelu. V grafu ROC je na ose *y* zachyceno procento správně zařazených objektů nazvané *pozitivní podíl* (senzitivita). Na ose *x* je pak zachyceno procento nesprávně zařazených objektů nazvané *falešný podíl* (1 - specificita).

| | | Predikované případy | |
|---|---|---|---|
| | | Ano | Ne |
| Skutečnost | Ano | A | B |
| | Ne | C | D |

Tab. 3: Jednoduchá matice záměn

Zdroj: Meloun et al. (2005)

Sensitivita (Ano x Ano) udává podmíněnou pravděpodobnost predikce, že událost nastane, v případě, že opravdu nastala. Vypočítáme ji podle vzorce:

$$\frac{A}{A+B}. \tag{4.22}$$

Specificita (Ne x Ne) naopak udává podmíněnou pravděpodobnost predikce, že událost nenastane, v případě, že opravdu nenastala. Vypočítáme ji podle vzorce:

$$\frac{D}{C+D}. \tag{4.23}$$

Více informací k validaci predikčních modelů lze získat např. v (MELOUN et al. 2005, OSLON a DELEN 2008).

## 4.8 Hodnocení technik shlukování

Pomocí různých shlukovacích technik s různorodými možnostmi jejich nastavení lze získat velké množství výsledných shlukovacích modelů. Je proto důležité zvolit vhodné kritérium k porovnání získaných řešení (BAE 2010).

K hodnocení výsledných shluků není žádoucí využívat obyčejné testy významnosti, jako je například analýza rozptylu. Tyto testy nejsou platné (validní) pro testování rozdílů mezi jednotlivými shluky, jelikož jejich předpoklady jsou výrazně porušovány (předpoklad jednorozměrného normálního rozdělení vede u rozsáhlejších souborů k výraznému zvýšení chyby prvního druhu).

Vhodnějším přístupem je předpoklad vícenásobného normálního rozdělení. Ani tento přístup však není uspokojivý, jelikož existuje typicky vyšší pravděpodobnost zamítnutí nulové hypotézy v případě, že data pochází z rozdělení s nižší špičatostí než má normální rozdělení. Otázkou předpokladů jednorozměrných normálních rozdělení a vícenásobného normálního rozdělení se zabývali např. Hartigan (1978) a Arnold (1979).

Typickou mírou meziskupinové variability je rozptyl. Toto pravidlo platí pro hierarchické i nehierarchické shlukování (COLLICA 2007). Často využívaný kritériem je proto ukazatel $R^2$ neboli index determinace. Index determinace značí podíl rozptylu, který je vysvětlen pomocí vytvořených shluků. Jedná se o nejjednodušší kritérium počítané dle vzorce:

$$R^2 = 1 - \frac{W}{T}, \tag{4.24}$$

kde $W$ představuje vnitroskupinovou variabilitu a $T$ značí meziskupinovou variabilitu.

Využívaným shlukovacím kritériem, založeném na součtu čtverců vzdáleností, je i kubické shlukovací kritérium (The Cubic Clustering Criterion - CCC). Toto kritérium, využívané softwarovými nástroji společnosti SAS, navrhl v roce 1983 Sarle. Kritérium je využíváno k odhadu výsledného počtu shluků u Wardovy metody minimalizace rozptylu, u metody *k*-průměrů, či jiných metod využívajících minimalizaci meziskupinového součtu čtverců. Toto kritérium bylo empiricky odvozeno pomocí metod Monte Carlo (SAS, 2008).

Kubické kritérium vychází z nulové hypotézy, že data mají rovnoměrné (uniform) rozdělení vycházející z hyperboxu, tento přístup navrhl Sarle (1983). CCC vychází z předpokladu, že rovnoměrné rozdělení v hyper-obdélníku bude rozděleno do shluků ve tvaru přibližně hyper-kostky. Kritérium lze využit k testování hypotéz a k odhadnutí počtu shluků v populaci.

> $H_0$: data pochází z rovnoměrného rozdělení na hyperboxu.
>
> $H_A$: data pochází ze směsi sférických vícerozměrných normálních rozdělení se shodnými rozptyly a shodnými pravděpodobnostmi výběru (SAS Technical Report, 1983).

V případě platnosti alternativní hypotézy, index determinace ($R^2$) odpovídá maximálně-pravděpodobnostnímu kritériu. Kubické shlukovací kritérium se spočítá dle následujícího vzorce

$$CCC = \ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] \times K, \tag{4.25}$$

kde $E(R^2)$ představuje očekávané $R^2$, $R^2$ je naměřené $R^2$ a $K$ je transformace stabilizující rozptyl (SARLE, 1983).

Další z využívaných ukazatelů, pseudo F statistiku, navrhli v roce 1974 Cainski a Harabasz. Tato statistika je zkonstruována tak, aby zachycovala „těsnost“ vytvořených shluků. Představuje poměr součtu čtverců mezi skupinami shluků k vnitroskupinovém součtu čtverců, vypočítá se tedy dle následujícího vzorce:

$$F = \frac{(T - P_G)/(G - 1)}{P_G/(n - G)}, \tag{4.26}$$

kde $G$ představuje počet shluků, $T$ je celkový součet čtverců odchylek a $P_G$ je vnitroskupinový součet čtverců. Vyšší hodnota tohoto ukazatele obvykle značí vhodnější řešení shlukování.

Obecně lze konstatovat, že lokální extrémy kubického shlukovacího kritéria a pseudo F statistiky značí vhodné rozdělení shluků. Tato kritéria jsou však využitelná pouze v případě kompaktních shluků, ideálně shluků s vícenásobným normálním rozdělením.

K hodnocení kvality nalezených řešení je možné také využít např. Silhouettovu míru (siluetu), která kombinuje principy koheze a separace shluků (LLETÍ et al. 2004). Grafickou podobu siluety navrhl Rousseeuw (1987), který definuje tuto míru takto:

> "*The Silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters*".

Tato statistika poskytuje klíčovou informaci o dobrém a špatném shluku. Dle Melouna a Militkého (2004) se hodnota siluety $s$ vypočte tímto postupem:

1. Objekt $i$ je ve shluku $A$ a má průměrnou vzdálenost $a$ ke všem objektům ve svém shluku. Je-li ve shluku $A$ jediný objekt, je $a = 0$.
2. Sousední shluk $B$ obsahuje objekty, které jsou nejbližší k objektu $i$ ve shluku $A$ a $b$ je průměrná vzdálenost mezi objektem $i$ a všemi objekty ve shluku $B$.
3. Silueta $s$ objektu $i$ se vyčíslí: když shluk $A$ obsahuje pouze jeden objekt, je $s = 0$. Když $a < b$, je $s = 1 - a/b$. Když $a > b$, je $s = b/a - 1$. Když $a = b$, je $s = 0$.

Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku $A$ se všemi vzdálenostmi objektů nejbližšího souseda $B$ dle pravidla:

1. Je-li $s$ blízko hodnotě +1, objekt $i$ je dobře klasifikován do shluku $A$, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda $B$.
2. Je-li $s$ blízko hodnoty nula, objekt $i$ se nachází kdesi uprostřed mezi shluky $A$ a B, a čistě náhodou byl přiřazen do shluku $A$.
3. Je-li $s$ blízko hodnotě -1, objekt $i$ je špatně klasifikován. Vzdálenosti k ostatním objektům ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda $B$.

Přehlednou statistikou pro určení počtu shluků je průměrná silueta *s*, počítaná přes všechny objekty. Tato hodnota sumarizuje, jak těsně prokládá shlukové uspořádání analyzovaná data. Ideální počet shluků pak maximalizuje průměrnou hodnotu siluety.

Meloun a Militký (2004) rozlišují následující typy shlukových uspořádání:

- *s* od 0.71 do 1.00 ⇒ silná a dobrá struktura,
- *s* od 0.51 do 0.70 ⇒ ještě přijatelná struktura,
- *s* od 0.26 do 0.50 ⇒ slabá struktura, asi umělá; je třeba najít novou, lepší,
- *s* od -1.00 do 0.25 ⇒ naprosto nevhodná struktura.

Silhouettova míra je aplikována v data miningovém systému IBM SPSS Modeler.

# 5 ZVOLENÉ METODY DISERTAČNÍ PRÁCE

V rámci disertační práce budou zhodnoceny různé metodické přístupy k segmentaci zákazníků dle jejich nákupního chování. Konkrétně se tato práce zaměří na porovnání jednotlivých shlukovacích algoritmů a zhodnocení jejich využitelnosti při řešení segmentace nákupních košů. K analýze byl využit datový soubor obsahující 61.904 transakčních záznamů nakoupených potravin zákazníky vybraného hypermarketu. Využitý datový soubor má následující podobu:

Obr. č. 11: Struktura vstupní datová matice

| EAN | Name | Price | PCount | Total | WGR | Payment | BasketN | Club | VANR | BASVANR | AKTIONSNR | WGR01 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8595121800117 | AL DR.H.BÍLÝ JOG150G | 4,50 | 1,00 | 4,50 | 635 | 0 | 832506321 | 0 | 243828003 | 243828003 | neni | 6351611 |
| 5901677000186 | KA MILLANO-KOK. 100G | 5,90 | 1,00 | 5,90 | 615 | 0 | 832506321 | 0 | 264419006 | 264419006 | neni | 6157111 |
| 287287 | PAPRIKA BÍLÁ 1KG | 19,00 | 0,19 | 3,61 | 736 | 0 | 832506321 | 0 | 87700008 | 87700008 | neni | 7363522 |
| 287473 | CIBULE ŘEZANÁ 1KG | 8,90 | 0,61 | 5,43 | 736 | 0 | 832506321 | 0 | 424567004 | 424567004 | 328001 | 7363617 |
| 100458 | VITAL HOUSKA 45G | 3,90 | 2,00 | 7,80 | 754 | 0 | 832506321 | 0 | 427378003 | 427378003 | neni | 7541111 |
| 600 | MASO/UZENINY PRODEJ | 19,00 | 1,00 | 19,00 | 600 | 0 | 832506321 | 0 | neni | neni | neni | neni |
| 2001800007372 | KOBLIHA DŽEM 45G | 5,50 | 2,00 | 11,00 | 754 | 3 | 832506322 | 0 | 337708006 | 337708006 | 328001 | 7541111 |
| 287111 | ST.HROZNY RÉVY B.1KG | 39,90 | 1,33 | 53,07 | 736 | 3 | 832506322 | 0 | 106363009 | 106363009 | neni | 7362312 |
| 8594006880060 | LK NÁLEV NA OKUR. 1L | 21,00 | 1,00 | 21,00 | 610 | 3 | 832506322 | 0 | 482657006 | 482657006 | neni | 6106213 |
| 8594006880060 | LK NÁLEV NA OKUR. 1L | 21,00 | 1,00 | 21,00 | 610 | 3 | 832506322 | 0 | 482657006 | 482657006 | neni | 6106213 |

Zdroj: Vlastní zpracování

Datový soubor (Obr. č. 11) zahrnuje následující vstupní proměnné:

- identifikační číslo transakce (EAN),
- název zboží (Name),
- jednotkovou cenu uvedeného zboží (Price),
- množství zboží v nákupního koši (PCount),
- celkovou cenu (Total),
- produktovou kategorii zboží (WGR),
- typ platby (Payment),
- identifikační číslo nákupního koše (BasketN),
- typ probíhající akce (AKTIONSNR)
- podkategorii zboží (WGR01).

Uvedená vstupní datová matice bude v rámci přípravné fáze dataminingového procesu dále restrukturalizována a následně agregována na úroveň jednotlivých nákupních košů

a produktových kategorií. Zvolená úroveň produktových kategorií bude vycházet z produktové hierarchie daného hypermarketu, která je uvedena v příloze č. 1.

K samotné realizaci dataminingových algoritmů byly využity možnosti dvou nejrozšířenějších softwarových nástrojů. Jde o dataminingové nástroje IBM SPSS Modeler 14.2 (dále jen Modeler) a SAS Enterprise Miner 6.2 (dále jen Enterprise Miner). Proces analýzy probíhající v systému Modeler bude vycházet z metodiky CRISP-DM. Naopak systém SAS vychází z navržené metodologie SEMMA.

Jednotlivé fáze dataminingového procesu jsou doplněny o názornou ukázku postupu analýzy ve zvolených softwarových nástrojích. Průběh analýzy je rozdělen do následujících kroků:

**1) Definování cílů**

Cílem realizovaného dataminingového procesu je segmentovat zákazníky dle jejich nákupního chování tak, aby vytvořené shluky byly homogenních, logické a jasně popsatelné.

**2) Porozumění datům**

V druhé fázi procesu je posouzena integrita vstupní datové matice. Jednotlivé vstupní proměnné jsou popsány pomocí technik explorační analýzy dat a grafického znázornění.

**3) Příprava dat**

V rámci této fáze je provedena příprava vstupního datového souboru pro modelování. Využita byla restrukturalizace datové základy a následná agregace souboru dle nákupního koše a zvolených produktových kategorií. Došlo k odvození nových podílových proměnných a k filtrování nepotřebných proměnných. Zvažována byla možnost transformace dat. U číselných proměnných jde především o možnost jejich linearizace či diskretizace; u nominálních proměnných pak transformace na tzv. dummy (alternativní) proměnné.

Předpokladem shlukovacích technik je nezávislost vstupních proměnných. Z tohoto důvodu byla sledována přítomnost multikolinearity v modelu. Jednou z možností, jak nežádoucí vliv multikolinearity odstranit je využití analýzy hlavních komponent. Výstupem analýzy jsou ortogonální lineární kombinace původních vstupních proměnných. Další možností je redukce počtu vstupní proměnných, případně jejich transformace. V rámci modelovací fáze byly porovnány výsledky shlukování hlavních komponent a redukované datové matice, kdy byly využity nově vypočítané podílové proměnné, které nevykazovaly vzájemnou závislost.

Existující odlehlá pozorování byla upravena pomocí techniky winsorizace. Na závěr přípravné fáze byla provedena standardizace vstupních proměnných, aby nedocházelo ke zkreslení výsledků shlukové analýzy.

**4) Modelování**

Modelovací fáze dataminingového procesu se dělí do následujících pěti kroků:

a. Výběr techniky modelování

   V rámci modelování byly porovnávány tři dostupné shlukovací algoritmy. Jedná se o metodu *k*-průměru, metodu dvoustupňového shlukování a Kohonenovy mapy.

b. Vytvoření testovacího návrhu

   V případě volby supervizovaného modelu je vstupní soubor rozdělen na trénovací a testovací (validační) část, pro seskupovací algoritmus nemá toto dělení opodstatnění.

c. Vytvoření modelu

   V rámci tohoto kroku jsou pomocí vybraných segmentačních algoritmů zkonstruovány segmentační modely.

d. Posouzení modelů

   Posouzení modelů je provedeno pomocí statistických charakteristik: statistiky silueta, pseudo F statistiky a kubického shlukovacího kritéria. Dále je provedeno grafické posouzení segmentace pomocí komponentních skóre, které jsou vyneseny do 3D grafu.

e. Hodnocení modelů a stanovení dalších kroků.

   Hodnocení modelů probíhá na základě komplexní znalosti dané problematiky a logického zhodnocení výsledných segmentů.

**5) Hodnocení výsledků**

Na závěr jsou dosažené výsledky segmentace vyhodnoceny z hlediska předem stanovených cílů.

## 5.1 Modelování pomocí nástroje IBM SPSS Modeler

V modelovacím nástroji od společnosti IBM SPSS jsou implementovány tři algoritmy umoňující segmentační seskupování: metoda *k*-průměrů, dvoustupňová seskupovací metoda a Kohonenovy mapy.

### 5.1.1 Dvoustupňová seskupovací metoda

Základem dvoustupňové shlukové analýzy (TwoStep Cluster analysis) je metoda BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*). Tento algoritmus, který je založen na podobném principu jako frakcionizace, uspořádá objekty do podshluků, které jsou charakterizovány pomocí shlukovacích vlastností označovaných zkratkou CF (*Cluster Features*). Tyto podshluky jsou dále shlukovány do *k* skupin pomocí tradiční hierarchické shlukové analýzy. Nevýhodou tohoto postupu je citlivost na pořadí objektů vstupujících do analýzy (Hebák et al., 2007).

V předshlukové fázi dvoustupňový algoritmus automaticky standardizuje všechny číselné vstupní proměnné na stejnou škálu s nulovým průměrem a jednotkovým rozptylem. Algoritmus také automaticky detekuje a vyřadí odlehlá pozorování z analýzy, čímž předchází zkreslení získaných výsledků. Za potenciální odlehlá pozorování označí algoritmus málo četné shluky, hranice četnosti shluků je stanovena zvoleným procentem. V případě, že tyto málo četné shluky nejsou v následujícím iteračním postupu spojeny s jinými shluky, jsou uznány za extrémy a vyloučeny ze shlukové analýzy. Tyto objekty jsou zařazeny do samostatného shluku „-1“, který je označen za šum (Modeler Help, 2011).

Dvoustupňový algoritmus tedy probíhá ve dvou fázích. V prvním kroku jsou objekty rozděleny do malých shluků (podshluků), jejichž počet je podstatně menší než počet objektů původního souboru. Algoritmus vytváří podsluky na základě modifikovaného CF-stromu, který se skládá z několika úrovní uzlů, kdy každý uzel obsahuje určitý počet vstupů. Ve druhém kroku jsou vzniklé podshluky seskupeny do stanoveného počtu shluků. V systému SPSS se tento krok realizuje pomocí hierarchické shlukové analýzy. V obou krocích mohou být využity dvě různé míry podobnosti. Euklidovská vzdálenost se využívá v případě kvantitativních spojitých proměnných. Analýzy zahrnující kvantitativní i kategoriální proměnné využívají míru nepodobnosti typu věrohodnostní poměr (*log-likelihood*). Předpokladem využití věrohodnostního poměru je normalita rozdělení spojitých náhodných veličin a multinomické pravděpodobnostní rozdělení u kategoriálních proměnných. Zároveň se předpokládá nezávislost vstupních veličin.

Výhodou tohoto algoritmu je skutečnost, že optimálně určí výsledný počet shluků a zároveň není časově náročný. K nalezení optimálního počtu shluků využívá algoritmus shlukové kritérium (Clustering Criterion), kterým může být Bayesovo informační kritérium (BIC) nebo Akaikeho informační kritérium (AIC). Více informací o dvoustuňové metodě lze nalézt např. v Hebák et al., 2007, nebo Modeler Help, 2011.

### 5.1.2 Metoda *k*-průměrů

Metodu *k*-průměrů lze v Modeleru realizovat pomocí uzlu *K-Means*. Jedná se o algoritmus nesupervizovaného modelování, který dělí datový soubor do rozdílných podskupin. Tento algoritmus nevyužívá cílovou proměnnou, ale rozkrývá strukturu vzorů obsaženou ve vstupních datech.

V prvním kroku algoritmus definuje počáteční centra shluků. Jednotlivé hodnoty jsou přiřazeny k nejvíce podobnému shluku. Po rozdělení celého souboru do shluků jsou jednotlivé centra shluků přepočítány na základě přiřazených jednotek. Jednotlivá pozorování jsou zkontrolována dle příslušnosti k nejbližšímu shluku, případně jsou přeřazena do jiného, bližšího, shluku. Tento postup je opakován, dokud nedojde k dosažení maximálního počtu iterací, případně dokud změna po každé iteraci nedosáhne zadané hraniční hodnoty. Uvedený postup rozdělení jednotek do shluků je závislý na pořadí vstupních dat, v případě změny struktury jednotek může dojít k rozdílům ve finálních modelech.

Algoritmus *k*-průměru se řadí mezi metody s nejnižší časovou náročností u velkých datových souborů.

V úvodu modelování je možné nastavit výsledný počet shluků, přednastaven je celkový počet 5 shluků. Dále lze označit možnost generování vzdálenosti jednotlivých záznamů od center shluků, k nimž přísluší. Následuje nastavení technických specifikací pro vytvoření modelu. Lze zvolit optimalizaci pro zvýšení výkonu, případně zvolit možnost rychlého výpočtu, který nezatěžuje disk za účelem zvýšení výkonu. Přednastavenou možností je volba optimalizace paměti, která šetří operační paměť na úkor rychlosti.

Expertní nastavení umožňuje ladit proces učení. Manuálně lze změnit kritérium pro zastavení procesu učení. Přednastavenou volbou je maximální počet 20 iterací či změna < 0,000001. Proces je zastaven, jakmile nastane alespoň jedna z uvedených událostí.

Dále lze specifikovat hodnotu přepočtu nominálních proměnných na dichotomické. Přednastavenou hodnotou je druhá odmocnina z hodnoty 0,5 (přibližně 0,707107), tato hodnota poskytuje vhodnou váhu pro přepočtené dichotomické proměnné.

### 5.1.3 Kohonenovy mapy

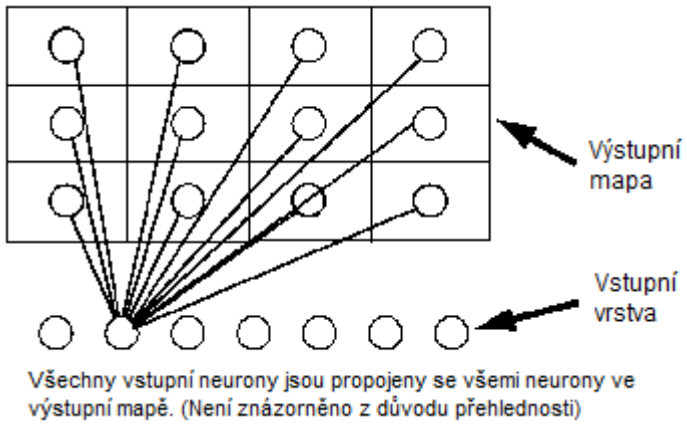
Uzel *Kohonen* představuje typ neuronové sítě využívaný ke shlukování. Tato technika je zpravidla označována jako samoorganizující mapy. Používá se pro seskupení jednotlivých pozorování do vzájemně rozdílných skupin, v případě, že tyto skupiny nejsou před začátkem modelování známy. Vstupní jednotky (označované jako neurony) jsou organizovány do vstupní a výstupní vrstvy. Všechny vstupní neurony jsou propojeny se všemi výstupními neurony. Tato propojení jsou definována pomocí váhy a stupně asociace. V průběhu procesu

učení mezi sebou jednotlivé výstupní neurony soupeří. Vstupující záznam je vždy přiřazen k vítěznému výstupnímu neuronu. Výstupem analýzy je mapa uspořádaná ve formě dvojdimenzionální mřížky neuronů. Detailnější popis metodiky neuronových sítí je uveden v literární rešerši.

Výstupem Kononenových map je několik málo jednotek zastupujících mnoho pozorování, tyto jednotky jsou označovány jako silné, a několik jednotek označovaných za slabé, které nekorespondují s žádným pozorováním. Silné jednotky reprezentují pravděpodobná centra vytvořených shluků.

Struktura Kohonenovy neuronové sítě je znázorněna na níže uvedeném obrázku (Obr. č. 12).

Obr. č. 12: Architektura Kohonenovy mapy




Zdroj: Modeler Help, 2011, vlastní zpracování

Nastavení uzlu umožňuje určit, zda při každém průchodu uzlem vytvořit kompletně nový model, nebo zda pokračovat se sítí, která byla úspěšně vypočítána dříve. Uzel také poskytuje možnost zobrazit graf zpětné vazby (*feedback graph*), který vizualizuje zastoupení dvojdimenzionálního pole. Síla každého uzlu je reprezentována barvou. Silné jednotky jsou zobrazeny červenou barvou, slabé naopak bílou. Tento graf je časově náročný, z tohoto důvodu může způsobit prodloužení procesu učení. Obdobně jako u uzlu *k*-průměrů lze i u Kohonenových map definovat kritérium, které ukončí proces učení. Lze také nastavit maximální dobu učení v minutách či optimalizaci rychlosti a paměti.

Uzel umožňuje nastavit náhodná centra shluků. Jednotlivé sekvence náhodných hodnot užité k inicializaci vah sítě jsou v takovém případě aktualizovány při každém spuštění daného uzlu. Tato volba může zapříčinit různé výsledky analýzy při každém průchodu uzlem i při neměnném nastavení.

Expertní nastavení umožňuje zvolit velikost dvojdimenzionální mapy. Dále lze nastavit míru útlumu učení (Learning rate decay), která může být lineární či exponenciální. Tato míra představuje vážící faktor, jenž vykazuje po celou dobu učení klesající průběh, postupuje od hrubých vzorů až k jemnější úrovni detailu. Učení Kohonenovy sítě se dělí na dvě fáze. V první fázi jsou zachyceny pouze hrubé vzory v datech. Ve druhé fázi probíhá ladění, které se využívá pro upravení detailů v mapě. Pro každou fázi lze definovat tři parametry:

- Sousedství (Neighborhood) – zadaná hodnota definuje počáteční radius sousedství. Určuje počet „blízkých" jednotek, které jsou dále aktualizovány v rámci procesu učení dle příslušnosti k vítěznému neuronu. Daný rádius se s každou iterací monotónně snižuje. Konečný radius sousedství udává hladkost zobrazení.
- Eta – inicializuje počáteční hodnotu učení. V průběhu první fáze se hodnota Eta 1 snižuje na hodnotu inicializovaného parametru Eta 2. Hodnota parametru Eta 2 klesá až na nulu. V průběhu druhé fáze hodnota parametru Eta 2 opět klesá do nuly. Tyto dva poslední kroky se opakují, dokud není učení sítě ukončeno. Hodnota Eta 1 v první fázi by měla být vyšší než hodnota Eta 2 ve druhé fázi.
- Cykly – definují počet cyklů každé fáze učení (Modeler Help, 2011).

## 5.2 Modelování pomocí nástroje SAS Enterprise Miner

Jedním z hlavních úkolů segmentace je profilování zákazníků v rámci segmentu. Jedná se o základní popis společných charakteristik jednotlivých zákazníků v rámci segmentu. Porovnání segmentových profilů umožní porozumět skupině zákazníků v každém segmentu. K podrobnějšímu popisu segmentů lze využít prostředků klasické explorační analýzy. SAS EM poskytuje základní nástroje pro explorační analýzu, pro úvodní představu jsou využívány jednoduché grafy, např. histogram či sloupcový graf, poté základní statistické charakteristiky jako počet pozorování, průměr, medián, směrodatná odchylka, rozdělení vstupních

proměnných, či počet chybějících údajů. Tento krok poskytuje prvotní pohled na data, zda jsou vhodná pro dataminingové modelování, informace o rozložení vstupních proměnných apod. K explorační analýze v SAS EM slouží uzly *StatExplore* a *MultiPlot*.

### 5.2.1 Transformace vstupních proměnných

SAS EM poskytuje široké množství transformačních technik, od jednoduché standardizace, přes logaritmizaci či kategorizaci až po transformaci založenou na maximalizaci normality. Tato technika je vhodná především v případě využití modelovacích algoritmů vyžadujících normální rozdělení. Transformací se upraví proměnné s asymetrickým rozdělením s těžkými konci či extrémními hodnotami, případně s rozdělením, které je špičatější nebo naopak plošší než je rozdělení normální. Před samotnou realizací shlukové analýzy je doporučováno standardizovat vstupní proměnné např. pomocí uzlu *Filter*. Standardizace je doporučována v případě využití proměnných s velkými rozptyly, které mají větší vliv na výsledné seskupení než proměnné s rozptyly nízkými.

### 5.2.2 Míry vzdálenosti

Nejčastěji využívanou metodou měření vzdálenosti je převedení jednotlivých proměnných na numerické hodnoty, aby je bylo možné zachytit ve vícerozměrném poli (body v prostoru). Tento postup je žádoucí, jelikož vzdálenost mezi body v prostoru lze měřit pomocí Euklidovské geometrie a jednoduché vektorové algebry. Jednoduše řečeno, objekty, které jsou blíže k sobě, jsou si podobnější než objekty vzájemně vzdálené. K výpočtu matice vzdáleností lze v SAS EM využít proceduru DISTANCE, která počítá nejen euklidovskou vzdálenosti, ale i jiné běžně využívané míry. Pomocí procedury MDS (Multi-dimensional scaling routine) a makra %PLOTIT lze získané vzdálenosti zobrazit graficky. Pomocí škálování je možné vizualizovat vzdálenosti ve dvourozměrném prostoru (COLLICA 2011).

### 5.2.3 Přístupy ke shlukování

Mezi jednoduché techniky segmentace patří přístup založený na segmentaci buněk (*Segmentation using a cell based approach*), přístup založený na pravidlech a jednoduché shlukování. Segmentace na základě buněk vyžaduje kategorizaci všech vstupních proměnných a je využitelná pouze pro nízký počet proměnných. Každá jednotlivá kombinace

vstupů pak představuje samostatný shluk. Přístup založený na pravidlech vychází z rozhodovacího stromu.

Do této skupiny zařazuje Collica (2011) také algoritmus *k*-průměrů založený na mírách vzdálenosti (MDS). Mezi pokročilé přístupy řadí využití Kohonenových samoorganizujících map, kombinované segmentace (Ensemble methods) a segmentace transakčních záznamů nebo časových řad (COLLICA 2011).

Uzlu *Cluster*, který realizuje shlukovou analýzu, by měl předcházet výběrový (Sampling) uzel, který náhodně vybere jednotky k vytvoření modelu, a transformační (Transform) uzel, který provede standardizaci či jinou transformaci vstupních proměnných,

### 5.2.4 Shluková analýza pomocí uzlu *Cluster*

Pomocí uzlu *Cluster* lze realizovat disjunktní shlukovou analýzu pomocí euklidovské vzdálenosti. Tato vzdálenost je počítána na základě jedné či více kvantitativních proměnných a vygenerovaných zárodečných bodů, které jsou aktualizovány na základě použitého algoritmu. Daný uzel akceptuje binární, nominální, ordinální a intervalové typy proměnných. Vstupující kategorizované proměnné jsou před vlastní analýzou překódovány na číselné dummy proměnné. Pro kódování ordinálních proměnných lze využít např. metodu pořadí, či indexování. Pro nominální znaky EM nabízí metodu GLM a referenční metodu. Pomocí shlukovacího algoritmu jsou objekty rozděleny do skupin tak, že každé pozorování patří do nejvýše jednoho shluku.

Uzel Cluster umožňuje realizovat přípravnou fázi modelování, např. procházení vstupní datové matice, tvorbu jednoduché explorační analýzy a definování vstupních proměnných, imputaci chybějících pozorování nebo standardizaci vstupních proměnných.

Dále lze nastavit jednotlivé parametry modelování. Nastavení počtu výsledných shluků lze definovat manuálně či automaticky. V případě nastavení automatického zjišťování počtu shluků, algoritmus nejprve provede předshlukovou analýzu s maximálním počtem definovaných shluků, poté jsou využity vícenásobné průměry shluků, které vstupují do následné analýzy využívající aglomerativní hierarchické algoritmy. Nejmenší počet shluků, který splňuje uvedená kritéria, je zvolen jako finální. Prvním předpokladem je, že počet shluků bude vyšší než přednastavená hodnota minimálního počtu shluků a že hodnota

kubického shlukovacího kritéria (CCC) je vyšší než předem specifikovaná mez. V případě, že zmíněná kritéria nejsou splněna, je stanoven takový počet shluků, který představuje první lokální maximum.

Uživatelské nastavení umožňuje nastavit minimální a maximální hodnoty uvedených kritérií. Při automatickém zjišťování výsledného počtu shluků lze zvolit jednu ze tří metod počítajících vzdálenosti mezi shluky: metodu průměrné vazby, centroidní metodu a Wardovu metodu, která je nastavena jako výchozí.

Dále lze definovat: maximální počet shluků pro předshlukovou analýzu (přednastavená hodnota je 50), minimální a maximální počet výsledných shluků, hraniční hodnotu kubického shlukovacího kritéria (CCC), přednastavenou hodnotou je 3.

V rámci inicializace center shluků lze nastavit inicializační metodu a minimální hodnotu vzdálenosti (radius) mezi centry shluků. Mezi inicializační metody patří: MacQueenova metoda, která je nastavena jako výchozí, metoda prvních kompletních případů, metoda plného nahrazení, metoda hlavních komponent a metoda částečného nahrazení. V pokročilém nastavení umožňuje EM dále definovat jednotlivé úrovně učení, maximální počet iterací, maximální počet kroků či konvergenční kritérium.

Výsledné shluky lze zkoumat graficky pomocí grafických výstupů uzlu *Cluster*. Vygenerovaný report obsahuje graf rozdělení jednotlivých shluků, stromový diagram a graf vzdáleností shluků (SAS Enterprise Miner 12.1 Reference Help, 2011).

### 5.2.5 Shluková analýza pomocí uzlu SOM/*Kohonen*

Obdobně jako metoda *k*-průměrů i Kohonenovy mapy se řadí mezi techniky nesupervizovaného modelování. Uzel SOM/*Kohonen* patří k explorační fázi dataminingového procesu v metodologii SEMMA. Tento uzel poskytuje tři různé techniky. Jedná se o shlukovací techniku založenou na vektorové kvantizaci (Kohonen vector quantization), Kohonenovy samoorganizující mapy, které jsou určeny primárně pro redukci dimenzionality v datech, a dávkové samoorganizující mapy.

Sítě založené na vektorové kvantizaci jsou konkurenceschopné sítě, které mohou být vnímány jako nesupervizované modely určené k odhadům hustoty pravděpodobnosti nebo jako autoklasifikační techniky (Kohonen 2001, Hecht-Nielsen 1990). Tato technika je velmi úzce

spjata s technikou *k*-průměrů. Kohonenovo učení neuronové sítě spočívá v algoritmu, který nalezne nejbližší střed shluků každé jedné učící se jednotky. Tato jednotka je následně přiřazena k danému shluku a „vítězné“ centrum shluku se posune blíže k dané jednotce. Posun závisí na podílu vzdálenosti mezi centrem shluku a dané jednotky; zmíněný podíl je specifikován parametrem učení (*learning rate*). Index *n* vítězného shluku je určen na základě následujícího vzorce:

$$n = \arg\min \left\| C_j^s - X_i \right\| \tag{4.27}$$

kde $C_j^s$ je centrum shluku *j* v kroku *s*, $X_i$ je vektor vstupních proměnných pro učící se jednotku *i*, a $L^s$ je parametr učení pro krok *s*.

Kohonenův aktualizovaný vzorec má tvar:

$$C_n^{s+1} = C_n^s (1 - L^S) + X_i L^s . \tag{4.28}$$

Pro všechny „nevítězné“ shluky platí, že $C_j^s + 1 = C_j^s$.

Tento postup je velmi obdobný jako MacQueenův algoritmus *k*-průměrů. Rozdíl je pouze ve skutečnosti, že parametr učení v MacQueenově algoritmu je tvořen převrácenou hodnotou počtu případů, které byly přiřazeny k vítěznému shluku. Tato redukce parametru učení způsobuje, že každý střed shluku je tvořen průměrnou hodnotou všech jednotek, které k danému shluku přísluší, což zaručuje konvergenci algoritmu na optimální hodnotu chybové funkce (součtu čtverců euklidovské vzdálenosti mezi centry shluku a jednotlivými případy) s rostoucím počtem učících se jednotek, jejichž počet se blíží nekonečnu. Kohonenova učící se síť má fixní parametr učení, který nekonverguje (SAS Enterprise Miner 12.1 Reference Help, 2011).

### 5.2.6 Profilace vytvořených shluků

Profilování a příprava dat slouží k lepšímu pochopení skutečností, které se v analyzovaných datech skrývají. K profilaci již vytvořených segmentů slouží uzel *Segment Profile*. Pro využití tohoto uzlu musí být definována segmentační proměnná s nastavenou rolí cluster nebo segment. Tento uzel poskytuje rozdělení vstupních proměnných v rámci jednotlivých segmentů a několik grafických výstupů. Další možností profilování segmentů je využití popisného uzlu *StatExplore* (COLLICA 2011).

### 5.2.7 Hodnotící kritéria shlukování

SAS využívá k hodnocení výsledného počtu shluků kubické shlukovací kritérium (Cubic Clustering Criterion – CCC). Kubické kritérium vychází z nulové hypotézy, že data mají rovnoměrné rozdělení vycházející z hyperboxu. Vyšší hodnota tohoto kritéria představuje lepší rozdělení jednotek do jednotlivých shluků.

Dalším v SASu využívaný ukazatelem je pseudo F statistika, která určuje těsnost vytvořených shluků. Lokální extrémy těchto ukazatelů značí vhodné rozdělení shluků.

# 6 VÝSLEDKY DISERTAČNÍ PRÁCE

V praktické části disertační práce je demonstrováno využití adekvátních shlukovacích technik na konkrétním příkladu seskupování zákazníků zvoleného hypermarketu. Pomocí dostupných metod shlukové analýzy je realizována segmentace zákazníků dle jejich nákupního koše.

Uvedené výsledky práce jsou strukturovány do podkapitol dle dataminingové procesní metodiky. První podkapitola se zaměřila na přípravu datového souboru pro modelování. Jedná se o jednu z nejdůležitějších a časově nejnáročnějších fází dataminingového procesu.

Druhá podkapitola byla věnována modelování. Aplikace zvolených technik shlukování byla demonstrována ve dvou dataminingových nástrojích – Modeler a Enterprise Miner. V rámci této podkapitoly byla hodnocena kvalita nalezených řešení pomocí Silhouetovy míry, kubického shlukovacího kritéria a grafického znázornění získaných shluků. Pomocí nástroje na redukci dimenzionality – analýzy hlavních komponent – byly vstupní proměnné transformovány do menšího počtu vzájemně nekorelovaných hlavních komponent. Tyto komponenty byly následně využity pro grafické znázornění homogenity nalezených segmentů. Získané segmenty byly následně profilovány a byla zhodnocena jejich logická správnost.

Poslední část práce se zaměřila na porovnání využitých dataminingových nástrojů IBM SPSS Modeler a SAS Enterprise Miner.

## 6.1 Příprava datové matice

V přípravné fázi dataminingového procesu byla sestavena vstupní datová matice a provedeno zhodnocení jednotlivých vstupních proměnných pomocí explorační analýzy a příslušných grafických nástrojů. Byla zhodnocena kvalita vstupních dat, existence odlehlých pozorování a nežádoucí multikolinearity. Vstupní datový soubor, čítající 61 904 záznamů, obsahoval následující vstupní proměnné:

- identifikační číslo transakce (EAN),
- název zboží (Name),

- jednotkovou cenu uvedeného zboží (Price),
- množství zboží v nákupního koši (PCount),
- celkovou cenu (Total),
- kategorii zboží (WGR),
- typ platby (Payment),
- identifikační číslo nákupního koše (BasketN),
- typ probíhající akce (AKTIONSNR)
- podkategorii zboží (WGR01).

Obr. č. 13 zachycuje strukturu vstupní datové matice.

Obr. č. 13: Struktura vstupního datového souboru

| EAN | Name | Price | PCount | Total | WGR | Payment | BasketN | AKTIONSNR | WGR01 |
|---|---|---|---|---|---|---|---|---|---|
| 8595121800117 | AL DR.H.BÍLÝ JOG150G | 4.500 | 1.000 | 4.500 | 635 | 0 | 832506321 | neni | 6351611 |
| 5901677000186 | KA MILLANO-KOK. 100G | 5.900 | 1.000 | 5.900 | 615 | 0 | 832506321 | neni | 6157111 |
| 287287 | PAPRIKA BÍLÁ 1KG | 19.900 | 0.190 | 3.900 | 736 | 0 | 832506321 | neni | 7363522 |
| 287473 | CIBULE ŘEZANÁ 1KG | 8.900 | 0.610 | 5.400 | 736 | 0 | 832506321 | 328001 | 7363617 |
| 100458 | VITAL HOUSKA 45G | 3.900 | 2.000 | 7.800 | 754 | 0 | 832506321 | neni | 7541111 |
| 600 | MASO/UZENINY PRODEJ | 19.000 | 1.000 | 19.000 | 600 | 0 | 832506321 | neni | neni |
| 2001800007372 | KOBLIHA DŽEM 45G | 5.500 | 2.000 | 11.000 | 754 | 3 | 832506322 | 328001 | 7541111 |
| 287111 | ST.HROZNY RÉVY B.1KG | 39.900 | 1.330 | 53.200 | 736 | 3 | 832506322 | neni | 7362312 |
| 8594006880060 | LK NÁLEV NA OKUR. 1L | 21.000 | 1.000 | 21.000 | 610 | 3 | 832506322 | neni | 6106213 |
| 8594006880060 | LK NÁLEV NA OKUR. 1L | 21.000 | 1.000 | 21.000 | 610 | 3 | 832506322 | neni | 6106213 |

Zdroj: Vlastní zpracování

Do dataminingové analýzy vstupují následující číselné proměnné: množství zboží v nákupním koši (PCount; dle měrné jednotky, např. ks, kg, l), celková cena za položku v Kč (Price) a následně je pomocí restrukturalizace a agregace vypočítána celková cena nákupního koše v Kč (realizovaná platba). Z nominálních proměnných byl zvolen typ marketingové akce (AktionsNR) a produktová kategorie zboží (WGR), která slouží k restrukturalizaci datové matice. Hierarchie produktových kategorií zboží je uvedena v příloze č. 1. Nepotřebné proměnné byly zredukovány v uzlu *Filter*.

V úvodní fázi přípravy dat byla proměnná *typ akce* transformována na dichotomickou proměnnou, kdy 1 značí, že zboží bylo v akci, 0 naopak představuje skutečnost, že na zboží nebyl deklarován žádný typ akce. Vytvořená proměnná *akce* tedy vyjadřuje, zda určitý výrobek byl v akci či nikoliv. V rámci dataminingového nástroje Modeler byla tato změna provedena pomocí uzlu *Derive* a funkce *Derive as flag*. Četnostní rozdělení je zachyceno na Obr. č. 14.

Obr. č. 14: Typy probíhajících akcí u jednotlivých výrobků

| Value | Proportion | % |
|---|---|---|
| 328001 | | 13.04 |
| 328010 | | 0.22 |
| 328030 | | 4.43 |
| neni | | 82.3 |

Zdroj: Vlastní zpracování

Následně byly doplněny názvy jednotlivých kategorií produktů dle produktového listu (viz příloha č. 1). Jako hlavní kategorie bylo zvoleno sedm typů produktů: alkoholické a nealkoholické nápoje, balené potraviny, drogerie, masné výrobky, ovoce a zelenina, čerstvé pečivo a ostatní čerstvé potraviny. Rozdělení kategorií v datovém souboru zachycuje následující obrázek (Obr. č. 15).

Obr. č. 15: Rozdělení potravin do sedmi produktových kategorií

| Value | Proportion | % |
|---|---|---|
| Alkoholické a nealkoholické nápoje | | 12.26 |
| Balené potraviny | | 26.88 |
| Drogerie | | 8.41 |
| Masné výrobky | | 5.34 |
| Ovoce a zelenina | | 8.93 |
| Čerstvé pečivo | | 7.97 |
| Čerstvé potraviny | | 30.21 |

Zdroj: Vlastní zpracování

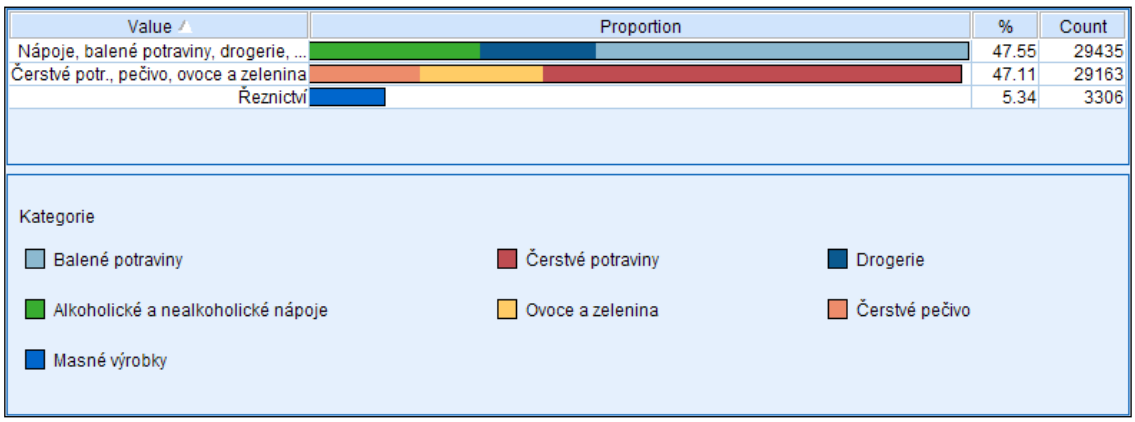
Důležitým krokem úvodní přípravy dat je výběr produktové úrovně, na které bude segmentace probíhat. Pro účely této analýzy byly zvoleny tři produktové kategorie dle produktového listu (viz příloha č. 1), které vykazovaly výrazně lepší segmentační vlastnosti než volba sedmi produktových kategorií. Dalším důvodem volby tří kategorií je praktické hledisko. Pro účely marketingového plánování je vhodnější zvolit nižší počet dobře vyprofilovaných segmentů, na které mohou cílit marketingové kampaně. Nižší počet vstupních proměnných je vhodný také pro většinu shlukovacích technik, které vykazují lepší vlastnosti při využití nižšího počtu dimenzí.

Za účelem segmentace byly zvoleny následující produktové kategorie, které vychází z produktové hierarchie (viz příloha č. 1):

- první kategorie (FOOD1) obsahuje alkoholické a nealkoholické nápoje, balené potraviny (konzervy, slané a sladké pečivo) a drogérii,
- druhá kategorie (FOOD2) zahrnuje čerstvé potraviny (např. balené uzeniny, masné výrobky, mléčné výrobky apod.), pečivo a ovoce a zeleninu,
- poslední kategorie (FOOD3) představuje produkty řeznictví.

Rozdělení tří hlavních kategorií je patrné z Obr. č. 16.

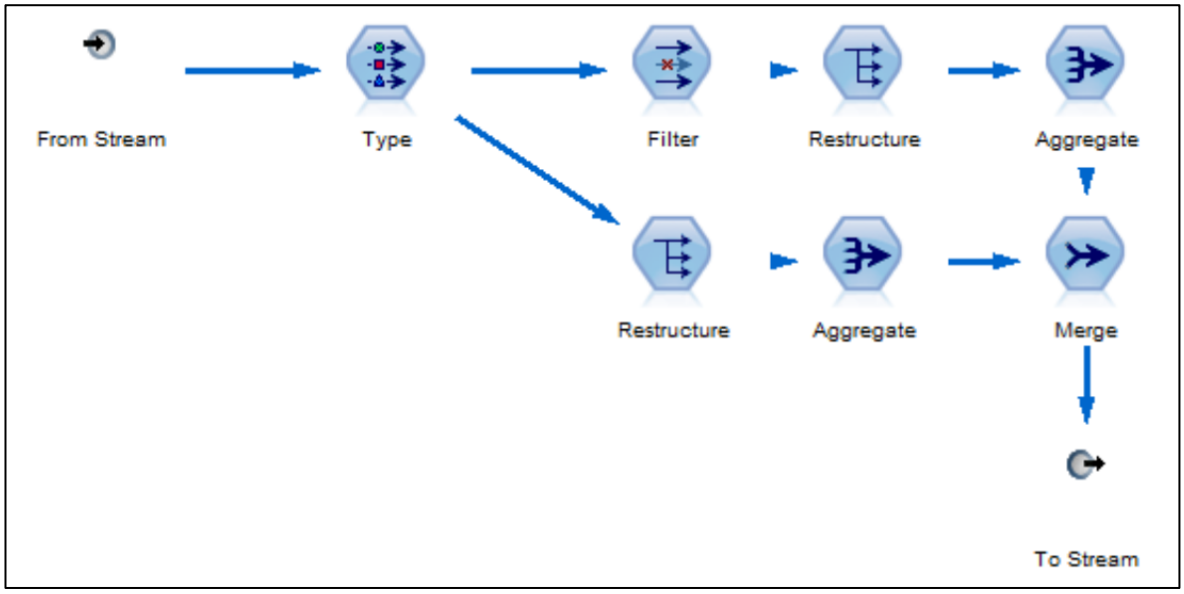
Obr. č. 16: Složení tří hlavních produktových kategorií




Zdroj: Vlastní zpracování

V následujícím kroku byla restrukturalizována datová matice. Konkrétně byla provedena restrukturalizace jednotlivých transakcí tak, aby každý řádek odpovídal jednomu nákupnímu koši. Restrukturalizace byla realizována v Modeleru pomocí uzlu *Restructure* a probíhala ve dvou krocích. V první procesní větvi byly restrukturalizovány produktové kategorie (FOOD1 – FOOD3). Z jedné kategorické proměnné tak vznikly tři proměnné, které obsahují cenu daného výrobku za příslušnou kategorii. V druhém kroku byla restrukturalizována proměnná akční zboží. Nově vytvořená proměnná zachycuje hodnotu akčního zboží v nákupním koši. Proces restrukturalizace je zachycen na následujícím schématu (Obr. č. 17).

Obr. č. 17: Proces restrukturalizace záznamů v nástroji Modeler 14.2




Zdroj: Vlastní zpracování

Po restrukturalizaci následovala agregace proměnných. Jednotlivé záznamy byly agregovány pomocí uzlu *Aggregate*, kde klíčovou proměnnou tvořilo identifikační číslo nákupního koše. V tomto kroku byla vytvořena nová proměnná *počet položek* v nákupním koši. Vzniklé datové soubory byly spojeny pomocí uzlu *Merge*. Restrukturalizovanou datovou matici zachycuje Obr. č. 18.

Obr. č. 18: Restrukturalizovaná datová matice

| BasketN | Aktion | Total | FOOD1_Total_Sum | FOOD2_Total_Sum | FOOD3_Total_Sum | N_polozek |
|---|---|---|---|---|---|---|
| 832504727 | $null$ | 39.900 | 39.900 | $null$ | $null$ | 1 |
| 832504729 | 31.700 | 79.600 | 27.900 | 51.700 | $null$ | 4 |
| 832504730 | $null$ | 232.800 | 232.800 | $null$ | $null$ | 2 |
| 832504731 | $null$ | 82.600 | 12.900 | 42.500 | 27.200 | 5 |
| 832504733 | $null$ | 10.800 | 2.900 | 7.900 | $null$ | 2 |
| 832504734 | $null$ | 72.700 | 59.300 | 13.400 | $null$ | 4 |
| 832504735 | 170.500 | 352.600 | 187.000 | 124.800 | 40.800 | 15 |
| 832504736 | 136.100 | 335.800 | 198.500 | 90.500 | 46.800 | 12 |
| 832504737 | 40.400 | 176.600 | $null$ | 94.200 | 82.400 | 9 |
| 832504739 | $null$ | 199.700 | 20.900 | 115.700 | 63.100 | 8 |
| 832504741 | $null$ | 12.900 | 12.900 | $null$ | $null$ | 1 |
| 832504743 | 18.900 | 97.900 | 26.800 | 71.100 | $null$ | 6 |
| 832504744 | $null$ | 66.800 | 66.800 | $null$ | $null$ | 2 |
| 832504745 | $null$ | 65.000 | 65.000 | $null$ | $null$ | 1 |
| 832504747 | 190.500 | 1290.500 | 468.200 | 466.700 | 355.600 | 34 |

Zdroj: Vlastní zpracování

Uvedené chybějící hodnoty „*$null$*“, které vznikly v průběhu restrukturalizace dat, byly nahrazeny nulou, která vyjadřuje nulovou hodnotu zboží v dané kategorii. Pomocí uzlu *Derive* byla vytvořena nová proměnná *podíl akčního zboží* (*Akcni_zbozi*), jež byla spočítána pomocí následujícího vzorce:

$$Podíl\ akčního\ zboží = (Aktion/Total)*100. \quad (6.1)$$

Tato proměnná je uváděna v procentech, jedná se tedy o procentuální podíl akčního zboží v nákupním koši. Výsledná datová matice po odstranění nulových hodnot je uvedena v následujícím obrázku (Obr. č. 19).

Obr. č. 19: Upravená datová matice pro modelování

| BasketN | Total | FOOD1_Total_Sum | FOOD2_Total_Sum | FOOD3_Total_Sum | N_polozek | Akcni_zbozi |
|---|---|---|---|---|---|---|
| 832504727 | 39.90 | 39.900 | 0.000 | 0.000 | 1 | 0.000 |
| 832504729 | 79.60 | 27.900 | 51.700 | 0.000 | 4 | 39.824 |
| 832504730 | 232.80 | 232.800 | 0.000 | 0.000 | 2 | 0.000 |
| 832504731 | 82.60 | 12.900 | 42.500 | 27.200 | 5 | 0.000 |
| 832504733 | 10.80 | 2.900 | 7.900 | 0.000 | 2 | 0.000 |
| 832504734 | 72.70 | 59.300 | 13.400 | 0.000 | 4 | 0.000 |
| 832504735 | 352.60 | 187.000 | 124.800 | 40.800 | 15 | 48.355 |
| 832504736 | 335.80 | 198.500 | 90.500 | 46.800 | 12 | 40.530 |
| 832504737 | 176.60 | 0.000 | 94.200 | 82.400 | 9 | 22.877 |
| 832504739 | 199.70 | 20.900 | 115.700 | 63.100 | 8 | 0.000 |
| 832504741 | 12.90 | 12.900 | 0.000 | 0.000 | 1 | 0.000 |
| 832504743 | 97.90 | 26.800 | 71.100 | 0.000 | 6 | 19.305 |

Zdroj: Vlastní zpracování

### 6.1.1 Popisná analýza

Explorační analýza předchází procesu hledání struktury ve vstupní datové matici. Umožňuje ověřit předpoklady analyzovaných dat, např. homogenitu, nekorelovanost a nalezení vybočujících objektů, které mohou zkreslit výsledky provedené analýzy, apod.

V rámci jednorozměrné analýzy jsou využívány jednoduché popisné charakteristiky, četnostní tabulky a jednoduché grafy. Dataminingové nástroje zpravidla poskytují možnosti datového auditu, kde připravené algoritmy upozorní například na špatnou kvalitu dané proměnné, na existenci odlehlých pozorování či chybějících údajů. Nemělo by být opomenuto také ověření pravděpodobnostního rozdělení vstupních proměnných.

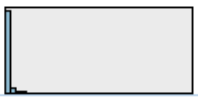
Do analýzy vstupuje šest připravených vstupních proměnných, které budou v jednotlivých analýzách využity pod následujícími názvy:

- celková cena nákupního koše (Total),
- hodnota produktů v kategorii 1 (FOOD1),
- hodnota produktů v kategorii 2 (FOOD2),
- hodnota produktů v kategorii 3 (FOOD3),
- počet položek v nákupním koši (N_polozek),
- podíl akčního zboží v nákupním koši (Akcni_zbozi).

### 6.1.2 Datový audit vstupních proměnných

V rámci využitého nástroje Modeler 14.2 lze k hodnocení kvality vstupní datové matice využít prostředků datového auditu, který lze realizovat prostřednictvím uzlu *Data audit*. Z výstupů datového auditu (Obr. č. 20) je patrné, že vstupní proměnné jsou spojité náhodné veličiny, které vykazují výrazně asymetrické rozložení. Z tohoto důvodu je vhodné zvážit možnost transformace či diskretizace proměnných.

Obr. č. 20: Výstup datového auditu vstupních proměnných

| Field | Sample Graph | Measurement | Min | Max | Mean | Std. Dev | Skewness |
|---|---|---|---|---|---|---|---|
| Total |  | Continuous | 2.300 | 11481.100 | 393.660 | 553.351 | 5.103 |
| FOOD1_Total_Sum |  | Continuous | 0.000 | 11441.100 | 203.395 | 352.544 | 8.953 |
| FOOD2_Total_Sum |  | Continuous | 0.000 | 2386.700 | 137.378 | 194.123 | 2.979 |
| FOOD3_Total_Sum |  | Continuous | 0.000 | 10148.400 | 52.887 | 187.676 | 32.798 |
| N_polozek |  | Continuous | 1 | 151 | 12.557 | 15.084 | 2.478 |
| Akcni_zbozi |  | Continuous | 0.000 | 100.000 | 18.503 | 25.377 | 1.681 |

Zdroj: Vlastní zpracování

### 6.1.3 Transformace proměnných

V přípravné fázi je třeba řešit otázku, zda je žádoucí vstupní proměnné standardizovat. U shlukové analýzy je většina využívaných měr vzdáleností citlivá na měřítka, vedoucí k různé numerické velikosti znaků. Obecně platí pravidlo, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají větší vliv na míru podobnosti. Z výstupů datového auditu je patrná různá variabilita v rámci datové matice. Z tohoto důvodu byly

vstupní proměnné standardizovány na stejnou škálu s nulovým průměrem a jednotkovým rozptylem. Tento krok byl proveden v rámci uzlu *Auto Data Prep*.

Vhodným prostředkem k normalizaci vstupních proměnných je využití logaritmické transformace. V tomto případě však není vhodné tuto transformaci vstupních proměnných aplikovat, jelikož by došlo k zásadní ztrátě informace a snížení interpretovatelnosti výsledků. Proměnné vyjadřující kategorie výrobků obsahují velké množství nulových hodnot, kdy zákazník nekoupil žádnou položky spadající do dané kategorie. Při použití logaritmické transformace jsou nulové hodnoty přetransformovány na neznámé hodnoty, jelikož logaritmus nuly nelze definovat. Z tohoto důvodu nebyla možnost využití logaritmické transformace zahrnuta do modelu. Tento problém lze řešit využitím dataminingového nástroje Enterprise Miner, který umožňuje logaritmickou transformaci proměnných obsahujících nulové či záporné hodnoty bez generování chybějících hodnot. Ani v tomto případě však logaritmická transformace nebyla využita z důvodu složitosti vytvořeného modelu a obtížnému uplatnění v praxi.

### 6.1.4 Odlehlá pozorování

Samostatnou část explorační analýzy tvoří činnosti spojené s čištěním dat, které umožňují filtrování a opravu podezřelých či nesprávných údajů. Nejčastějším typem filtrování je odstraňování nepotřebných údajů – odlehlých pozorování. Tyto anomálie snižují kvalitu výstupního modelu. Jejich odstranění by však mělo předcházet důkladné zvážení, jelikož mohou být nositeli jedinečných informací. Při segmentaci velkých datových souborů je však odstranění extrémů žádoucí. V případě ponechání odlehlých pozorování v souboru, tvoří tyto hodnoty zpravidla samostatné shluky. V rámci analýzy byla využita technika nahrazení, hodnoty přesahujících trojnásobek směrodatné odchylky byly nahrazeny hodnotou předcházející. K tomuto účelu byl využit uzel *Automated Data Preparation*.

### 6.1.5 Problematika chybějících hodnot

Vstupní datová matice neobsahovala chybějící pozorování, z tohoto důvodu nebylo nutné využít dostupných technik imputace dat.

### 6.1.6 Multikolinearita

Multikolinearita se chová jako neviditelný proces multiplicity, který silně ovlivňuje analýzu. Z tohoto důvodu byla ověřována přítomnost vzájemné závislosti vstupních proměnných. Detekována byla pomocí korelačních polí a Spearmanova korelačního koeficientu.

Z uvedené tabulky Spearmanových korelačních koeficientů (Tab. 4) je patrné, že multikolinearitu lze identifikovat u proměnné počet položek a celkové ceny nákupního koše i jednotlivých kategorií zboží. Proměnná počet položek v košíku bude z tohoto důvodu vyřazena z následné shlukové analýzy, jelikož nese obdobnou informaci jako proměnná hodnota nákupního koše.

| Spearmanovy korelační koeficienty | | | | | | |
|---|---|---|---|---|---|---|
| | TOTAL | FOOD1 | FOOD2 | FOOD3 | N_POLOZEK | AKCNI ZBOZI |
| TOTAL | 1,000 | 0,860 | 0,760 | 0,530 | 0,890 | 0,300 |
| FOOD1 | 0,860 | 1,000 | 0,450 | 0,300 | 0,740 | 0,250 |
| FOOD2 | 0,760 | 0,450 | 1,000 | 0,440 | 0,830 | 0,320 |
| FOOD3 | 0,530 | 0,300 | 0,440 | 1,000 | 0,520 | 0,120 |
| N_POLOZEK | 0,890 | 0,740 | 0,830 | 0,520 | 1,000 | 0,350 |
| AKCNI_ZBOZI | 0,300 | 0,250 | 0,320 | 0,120 | 0,350 | 1,000 |

Tab. 4: Korelační matice vstupních proměnných

Zdroj: Vlastní výpočty

Nezanedbatelná je také existující vzájemná závislost mezi hodnotou nákupního koše a produktovými kategoriemi. Kategorie potravin byly proto převedeny na podíl jednotlivých potravin na hodnotě celého nákupního koše, čímž došlo k redukci vzájemné závislosti mezi

kategoriemi a celkovou cenou nákupního koše. Pomocí uzlu *Derive* byly vytvořeny podílové proměnné:

- podíl první kategorie na celkové hodnotě košíku,
- podíl druhé kategorie na celkové hodnotě košíku,
- podíl třetí kategorie na celkové hodnotě košíku.

Tyto proměnné uváděné v procentech byly spočítány pomocí následujícího vzorce:

$$Podíl\ i\text{-}té\ kategorie = (FOOD_i/Total)*100, \tag{6.2}$$

kde $i = 1, 2, 3$.

Korelační matice vytvořených podílových proměnných je uvedena v Tab. 5.

| Spearmanovy korelační koeficienty | | | | | |
|---|---|---|---|---|---|
| | TOTAL | AKČNÍ_ZBOŽÍ | FOOD1 | FOOD2 | FOOD3 |
| TOTAL | 1,000 | 0,304 | 0,206 | -0,224 | 0,429 |
| AKČNÍ_ZBOŽÍ | 0,304 | 1,000 | 0,050 | 0,013 | 0,096 |
| FOOD1 | 0,206 | 0,050 | 1,000 | -0,866 | -0,227 |
| FOOD2 | -0,224 | 0,013 | -0,866 | 1,000 | -0,147 |
| FOOD3 | 0,429 | 0,096 | -0,227 | -0,147 | 1,000 |

Tab. 5: Korelační matice podílových vstupních proměnných

Zdroj: Vlastní výpočty

### 6.1.7 Redukce datové základny

Redukce datové základny zahrnuje tvorbu výběrového vzorku či redukci dimenzionality. V úvodní fázi byl dataminingový proces realizován na náhodně generovaném výběrovém vzorku. Výstupní modely však byly vypočítány na základě celého datového souboru.

Dále byla využita analýza hlavních komponent, která představuje nástroj redukce dimenzionality. V tomto případě však byla využita z důvodu redukce multikolinearity v modelu. Výstupem analýzy jsou nekorelované hlavní komponenty, které představují lineární kombinace vstupních proměnných. Tyto hlavní komponenty pak slouží jako vstupní proměnné pro následující shlukovou analýzu. Výsledky analýzy hlavních komponent zachycují následující tabulky (Tab. 6, Tab. 7).

| Komponenta | Vlastní čísla | % vysvětlené variability | Kumulativní % |
|---|---|---|---|
| 1 | 3,69 | 61,5 | 61,5 |
| 2 | 1,03 | 17,19 | 78,68 |
| 3 | 0,71 | 11,79 | 90,47 |
| 4 | 0,44 | 7,32 | 97,8 |
| 5 | 0,11 | 1,89 | 99,69 |
| 6 | 0,02 | 0,31 | 100 |

Tab. 6: Vektor vlastních čísel

Zdroj: Vlastní výpočty

První hlavní komponenta, vysvětlující 61,5 % variability původních vstupních proměnných, značí velikost nákupního koše. Druhá komponenta, vysvětlující 17 % variability, vystihuje podíl akčního zboží v koši. Třetí komponenta určuje hodnotu nakoupeného zboží v řeznictví a čtvrtá pak nejvíce souvisí s prvními dvěma produktovými kategoriemi. Vyšší hodnotu čtvrté komponenty dosáhnou koše s vyšší hodnotou zboží v první produktové kategorii (nápoje, balené potraviny, drogérie) a naopak nízkou hodnotu lze pozorovat u košů se zbožím v druhé kategorii (čerstvé potraviny, pečivo, ovoce a zelenina).

| | Komponenta | | | | |
|---|---|---|---|---|---|
| Proměnná | 1 | 2 | 3 | 4 | 5 |
| Total | 0,98 | 0,01 | -0,08 | 0,10 | 0,10 |
| FOOD1 | 0,84 | 0,08 | -0,37 | 0,39 | 0,05 |
| FOOD2 | 0,87 | 0,05 | 0,01 | -0,47 | 0,14 |
| FOOD3 | 0,61 | -0,30 | 0,71 | 0,19 | 0,01 |
| N_polozek | 0,95 | 0,01 | -0,07 | -0,14 | -0,28 |
| Akcni_zbozi | 0,05 | 0,97 | 0,25 | 0,05 | -0,01 |

Tab. 7: Matice komponentních zátěží

Zdroj: Vlastní výpočty

## 6.2 Modelování

Pro nalezení segmentů zákazníků s obdobným nákupním chováním, lze využít shlukovou analýzu**,** jejímž cílem je nalezení optimálního seskupení dat, kdy jednotlivá pozorování nebo objekty každého shluku jsou vzájemně podobné, ale jednotlivé shluky co nejvíce rozdílné. V tomto případě jde tedy o nalezení optimálního seskupení zákazníků s podobnými nákupními zvyklostmi.

V první části této kapitoly budou uvedeny výstupy shlukové analýzy realizované v dataminingovém nástroji IBM SPSS Modeler 14.2, druhá část pak zachytí proces modelování v nástroji SAS Enterprise Miner 12.1.

Každá fáze procesu modelování poskytuje nespočet různých přístupů a možností. V rámci modelování v dataminingovém nástroji Modeler byly porovnávány tři dostupné shlukovací algoritmy. Jedná se o metodu *k*-průměru, metodu dvoustupňového shlukování a Kohonenovy mapy. Nejprve jsou zachyceny různé možnosti tvorby dataminingového procesu, v závěru kapitoly je pak zvolen model vykazující nejvhodnější segmentační vlastnosti.

V průběhu modelování v nástroji Modeler byly porovnány následující možné postupy segmentace:

- Aplikace shlukové analýzy na standardizované vstupní proměnné.
- Aplikace shlukové analýzy na nově vytvořené podílové vstupní proměnné.
- Aplikace shlukové analýzy na vytvořené hlavní komponenty.
- Aplikace shlukové analýzy na diskretizované vstupní proměnné.
- Aplikace shlukové analýzy pomocí Kohonenových map.
- Aplikace shlukové analýzy na nižší úroveň produktové kategorie.

Následně byla provedena fáze modelování v nástroji Enterprise Miner, který pro účely segmentace poskytuje dvě modelovací techniky: metodu *k*-průměru a Kohonenovy mapy.

### 6.2.1 Modelování v prostředí dataminingového nástroje Modeler

K samotné realizaci segmentace nákupních košů v prostředí dataminingového nástroje Modeler byly využity následující techniky shlukování:

- Metoda *k*-průměrů (*K*-Means Clustering),
- Metoda dvoustupňového shlukování (TwoStep Clustering) – modifikace hierarchické metody,
- Kohonenovy mapy (Self-Organizing Map).

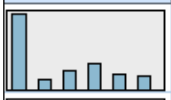
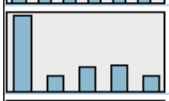
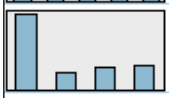
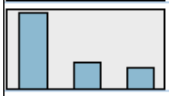
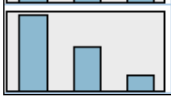
Nástroj Modeler umožňuje porovnávat kvalitu vytvořených modelů pomocí hodnotící Silhouettovy míry (silueta). Čím vyšší je hodnota tohoto ukazatele, tím lépe jsou shluky separovány.

**Aplikace shlukové analýzy na standardizované vstupní proměnné**

Nejprve byly shlukovány pouze standardizované proměnné, které byly upraveny o odlehlá pozorování. Shluková analýza byla realizovaná pomocí automatického uzlu *Auto Cluster*, který umožňuje porovnávat modely vytvořené různými algoritmy s různými parametry. V rámci uzlu je automaticky přednastavena hodnota 20 iterací. V rámci trénovací fáze byly testovány i modely s vyšším počtem iterací (50, 100, 200), výsledky se však nelišily. Pro shlukovací proces je tedy 20 iterací dostačujících.

Z výstupů automatického uzlu (viz Obr. č. 21) je patrné, že nejlepší míru siluety dosáhly modely *k*-průměrů. Pouze o tři desetiny zaostával nejlepší model dvoustupňového shlukování. Nejvyšší hodnotu siluety (s) dosáhl model *k*-průměrů se šesti vytvořenými shluky (s = 0,54). Velmi obdobné výsledky vykázal stejný model i s rozdělením do čtyř či pěti výsledných shluků.

Obr. č. 21: Výsledné seskupovací modely standardizovaných vstupních proměnných

| Graph | Model | Silhouette | Number of Clusters | Smallest Cluster (%) | Largest Cluster (%) |
|---|---|---|---|---|---|
|  | K-means 3 | 0,54 | 6 | 5 | 48 |
|  | K-means 1 | 0,53 | 5 | 9 | 49 |
|  | K-means 2 | 0,53 | 4 | 11 | 55 |
|  | TwoStep 2 | 0,51 | 3 | 15 | 62 |
|  | K-means 4 | 0,44 | 3 | 12 | 61 |
|  | TwoStep 3 | 0,39 | 3 | 10 | 57 |

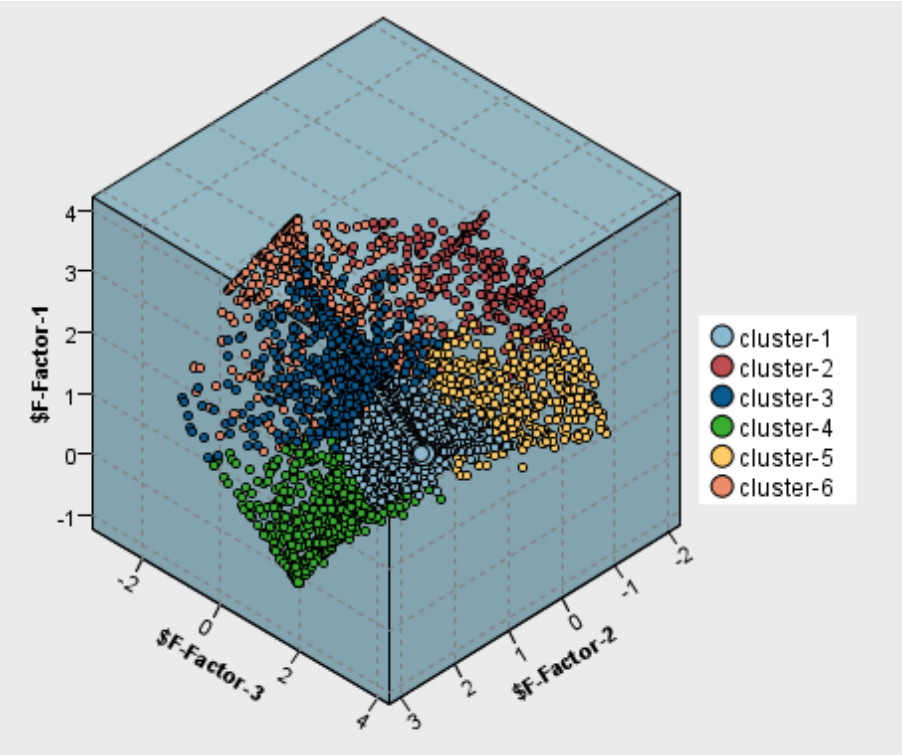
Zdroj: Vlastní výpočty

Z výše uvedených histogramů (viz Obr. č. 21) je patrné, že první shluk vždy obsahuje přibližně polovinu zákazníků, což poukazuje na nerovnoměrné rozložení jednotlivých shluků.

### Grafické hodnocení výsledného modelu

Kvalitu nalezeného řešení lze posoudit také grafickým vyjádřením. Ke snížení dimenzionality při tvorbě grafického znázornění byla využita analýza hlavních komponent. Výsledná skóre prvních tří komponent byla vynesena do grafu. Jednotlivé hodnoty byly obarveny dle příslušnosti ke shluku. V níže uvedeném grafu (Obr. č. 22) jsou zachyceny výsledné čtyři shluky vytvořené metodou *k*-průměrů. Z grafického vyjádření je patrné, že jednotlivé shluky jsou zřetelně separované, avšak velikost shluků je rozdílná. První nejčetnější shluk dosahuje nejnižších hodnot všech tří komponent.

Obr. č. 22: Výsledné shluky metody *k*-průměrů vynesené dle komponentního skóre




Zdroj: Vlastní výpočty

V následující tabulce je zachycena hodnota testového kritéria pro F-test a výsledná důležitost proměnných při tvorbě shluků. Důležitost vychází z vypočtené *p*-hodnoty. V případě, že je vypočítaná *p*-hodnota menší než 0,05, pak je proměnná označena jako *důležitá*. Tento případ indikuje méně než 5% pravděpodobnost, že výsledky mohou být vysvětleny pouhou náhodou.

Přestože nástroj Modeler poskytuje možnost otestovat významnost rozdílů mezi jednotlivými shluky pomocí klasického F-testu (viz Tab. 8), není tento přístup v případě větších datových souborů doporučovaný, jelikož jsou porušeny stanovené předpoklady testování, viz např. Hartigan (1978).

| Proměnná | F-Test | DF | Důležitost | |
|---|---|---|---|---|
| Total | 4 444,45 | 3,49 | 1,00 | Důležitá |
| FOOD1 | 2 097,72 | 3,49 | 1,00 | Důležitá |
| FOOD2 | 1 714,31 | 3,49 | 1,00 | Důležitá |
| FOOD3 | 4 849,09 | 3,49 | 1,00 | Důležitá |
| Akcni_zbozi | 3 647,68 | 3,49 | 1,00 | Důležitá |

Tab. 8: Testování důležitosti proměnných při tvorbě shluků získaných metodou *k*-průměru

Zdroj: Vlastní výpočty

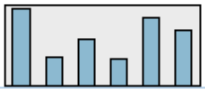
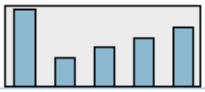
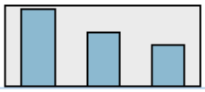
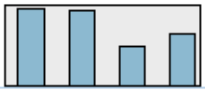
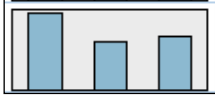
**Aplikace shlukové analýzy na nově vytvořené podílové vstupní proměnné**

Dalším možným přístupem je aplikace shlukové analýzy na podílové kategorie. Ceny za produktové kategorie byly převedeny na podíl jednotlivých kategorií na celkové hodnotě nákupního koše. Vstupní proměnné byly standardizovány a odlehlé hodnoty upraveny.

Tento postup nepřinesl výrazné zhoršení výsledného modelu (viz Obr. č. 23). Nejvyšší hodnoty siluety (0,51) dosáhl model *k*-průměrů se šesti výslednými shluky, což je o tři desetiny méně, než tomu bylo při využití hodnotových proměnných. Z uvedených histogramů

je však patrné rovnoměrnější rozdělení zákazníků v rámci jednotlivých shluků, než tomu bylo při pouhé standardizaci proměnných.

Obr. č. 23: Výsledné modely shlukové analýzy podílových vstupních proměnných

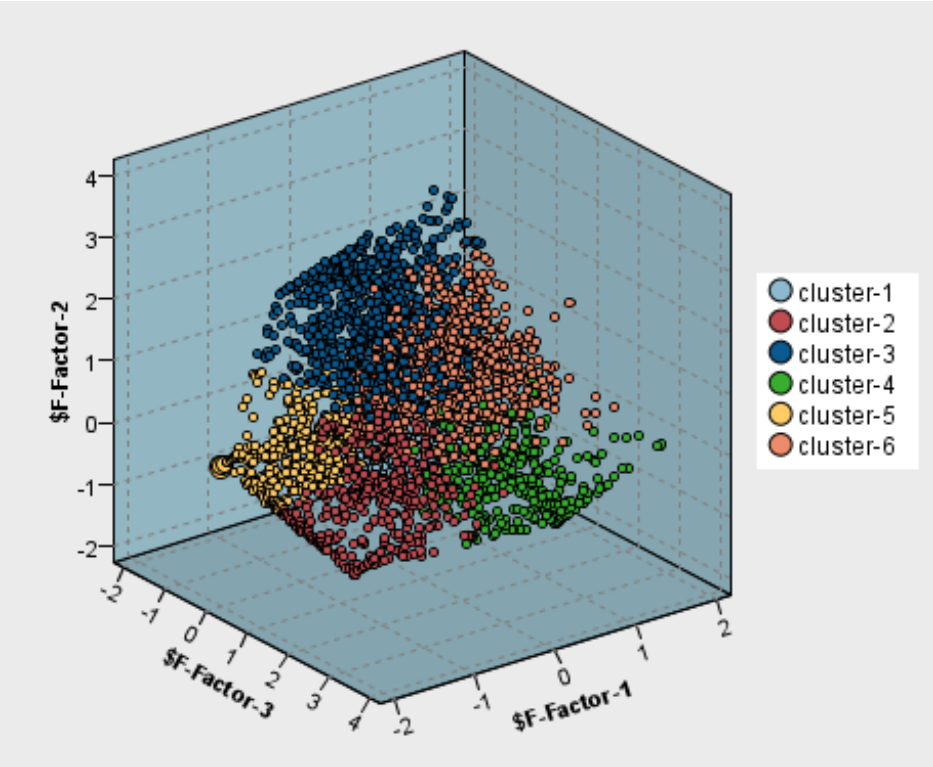
| Graph | Model | Silhouette | Number of Clusters | Smallest Cluster (%) | Largest Cluster (%) |
|---|---|---|---|---|---|
|  | K-means 3 | 0,51 | 6 | 8 | 25 |
|  | K-means 1 | 0,49 | 5 | 10 | 30 |
|  | K-means 4 | 0,48 | 3 | 23 | 45 |
|  | K-means 2 | 0,48 | 4 | 15 | 32 |
|  | TwoStep 3 | 0,4 | 3 | 10 | 46 |
|  | TwoStep 2 | 0,39 | 3 | 26 | 43 |

Zdroj: Vlastní výpočty

**Grafické zhodnocení výsledného modelu**

Grafické posouzení nalezeného řešení bylo opět realizováno pomocí získaného skóre tří hlavních komponent. Jednotky, představující nákupní koše, byly obarveny dle příslušnosti ke shluku. Na Obr. č. 23 je zachyceno šest výsledných shluků vytvořených metodou *k*-průměrů. Uvedený graf poukazuje na vytvoření homogenních shluků obdobné velikosti.

Obr. č. 24: Výsledné shluky metody *k*-průměrů vynesené dle komponentních skóre

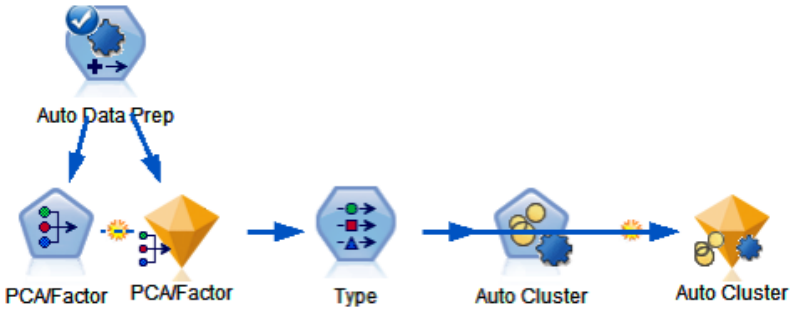



Zdroj: Vlastní výpočty

**Aplikace shlukové analýzy na vytvořené hlavní komponenty**

Odstranění multikolinearity ze vstupní datové matice lze docílit také využitím hlavních komponent jako vstupních proměnných ke shlukování. Následující obrázek zachycuje proces shlukování hlavních komponent. Nejprve proběhla standardizace a odstranění odlehlých pozorování. Na upravená data byla aplikována analýza hlavních komponent. Následně byly shlukovány nekorelované hlavní komponenty. Proces shlukování zachycuje následující schéma (Obr. č. 25).

Obr. č. 25: Postup shlukové analýzy aplikované na hlavní komponenty




Zdroj: Vlastní výpočty

Modely vycházející z hlavních komponent přinesly velmi obdobné výsledky jako první realizovaná shluková analýza. Nejvyšší hodnotu siluety (0,54) lze pozorovat u modelu *k*-průměrů se čtyřmi výslednými shluky (Obr. č. 26).

Obr. č. 26: Výsledné modely shlukové analýzy aplikované na hlavní komponenty

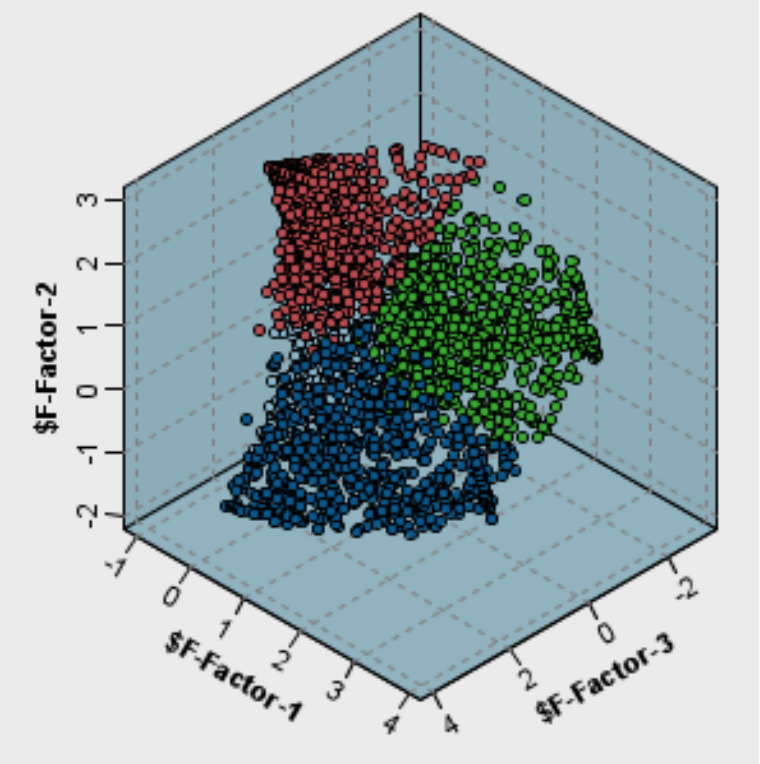
| Graph | Model | Silhouette | Number of Clusters | Smallest Cluster (%) | Largest Cluster (%) |
|---|---|---|---|---|---|
| | K-means 2 | 0,54 | 4 | 11 | 54 |
| | K-means 1 | 0,53 | 5 | 5 | 52 |
| | K-means 4 | 0,51 | 6 | 5 | 48 |
| | TwoStep 1 | 0,49 | 3 | 12 | 57 |
| | K-means 3 | 0,49 | 3 | 12 | 69 |

Zdroj: Vlastní výpočty

**Grafické zhodnocení výsledného modelu**

Posouzení nalezeného řešení byla opět realizováno pomocí grafického vyjádření tří hlavních komponent. Jednotlivé hodnoty byly obarveny dle příslušnosti ke shluku. V níže uvedeném grafu jsou zachyceny výsledné čtyři shluky vytvořené metodou *k*-průměrů.

Obr. č. 27: Výsledné shluky metody *k*-průměrů vynesené dle komponentních skóre




Zdroj: Vlastní výpočty

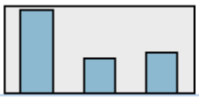
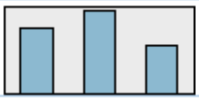
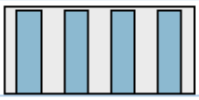
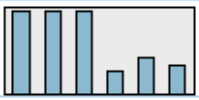
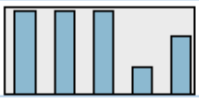
Metoda *k*-průměrů rozdělila jednotlivé hodnoty do čtyř zřetelných shluků. Z grafu jsou patrné tři samostatné shluky, čtvrtý, umístěný v pozadí, dosahuje nízkých hodnot všech tří komponentních skóre.

**Aplikace shlukové analýzy na diskretizované vstupní proměnné**

Dalším možným přístupem realizace shlukové analýzy je její aplikace na diskretizované vstupní proměnné. Výhodou tohoto přístupu je odstranění nežádoucího zešikmení jednotlivých proměnných, nevýhodou využití diskretizace je naopak možná ztráta užitečných informací. Ke shlukování kategorizovaných dat je vhodná metoda dvoustupňového shlukování, jež kromě euklidovské míry využívá i věrohodnostní funkci (*log-likelihood*).

Jednotlivé vstupní proměnné byly převedeny na ordinální znaky. Diskretizaci proměnných jednoduše umožňuje uzel *Binning*. V rámci uzlu proběhla transformace spojitých vstupních proměnných do čtyř kategorií na základě kvantilového rozložení.

Obr. č. 28: Výsledné seskupovací modely aplikované na diskretizované proměnné

| Graph | Model | Silhouette | Number of Clusters | Smallest Cluster (%) | Largest Cluster (%) |
|---|---|---|---|---|---|
|  | TwoStep 1 | 0,49 | 3 | 21 | 53 |
|  | TwoStep 3 | 0,39 | 3 | 24 | 42 |
|  | K-means 2 | 0,34 | 4 | 24 | 25 |
|  | K-means 3 | 0,33 | 6 | 6 | 25 |
|  | K-means 1 | 0,33 | 5 | 7 | 25 |
|  | TwoStep 2 | 0,33 | 3 | 26 | 46 |

Zdroj: Vlastní výpočty

Z výsledků modelování (obrázek č. 28) je patrné, že metoda dvoustupňového shlukování je vhodnějším nástrojem pro shlukování ordinálních znaků. Naopak metodu *k*-průměrů není vhodné v tomto případě využívat, jelikož vstupující kategoriální znaky jsou dále

transformovány na tzv. dummy proměnné, čímž se výrazně zvyšuje rozměr dané úlohy. Segmentační kvalita modelů vytvořených metodou *k*-průměrů je z tohoto důvodu oslabena.

Přestože dvoustupňová metoda vykazuje relativně vysokou hodnotu siluety (s = 0,49), zahrnuje skupinu nezatříděných údajů. V prvním případě se jedná o 60 % nezařazených zákazníků. Z praktického hlediska není tedy využití tohoto přístupu přínosné.

**Aplikace shlukové analýzy na nižší úroveň produktové kategorie**

V tomto kroku byla provedena segmentace na sedmi produktových kategoriích, celkové hodnotě nákupního koše a podílu akčního zboží. Pro účely této analýzy bylo zvoleno sedm produktových kategorií, jejich rozdělení zachycuje Obr. č. 29. Z jedné kategorické proměnné tedy vzniklo sedm proměnných, které obsahují cenu výrobků v příslušné produktové kategorii. Modelování předcházela explorační analýza, která zahrnovala detekci nežádoucí multikolinearity. Pomocí korelačních koeficientů bylo zjištěno, že mezi uvedenými vstupními proměnnými neexistuje multikolinearita. V přípravné fázi byla také zkoumána existence odlehlých pozorování, které byly, stejně jako v předchozích případech, nahrazeny hodnotou předcházející pomocí metody winsorizace.

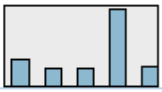
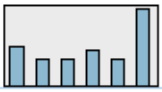
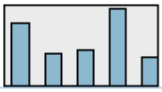
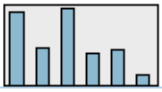
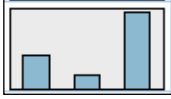
Obr. č. 29: Rozdělení potravin do sedmi produktových kategorií

| Value | Proportion | % | Count |
|---|---|---|---|
| Alkoholické a nealkoholické nápoje | | 12.26 | 7591 |
| Balené potraviny | | 26.88 | 16637 |
| Drogerie | | 8.41 | 5207 |
| Masné výrobky | | 5.34 | 3306 |
| Ovoce a zelenina | | 8.93 | 5527 |
| Čerstvé pečivo | | 7.97 | 4932 |
| Čerstvé potraviny | | 30.21 | 18704 |

Zdroj: Vlastní výpočty

Tento přístup dle očekávání nepřinesl zlepšení modelu. Podle hodnoty ukazatele silueta byl zvolen jako nejvhodnější model *k*-průměrů s pěti výslednými shluky (s = 0,45). Výsledky modelování jsou uvedeny na Obr. č. 30.

Obr. č. 30: Výsledné seskupovací modely aplikované na sedm produktových kategorií

| Graph | Model | Build Time (mins) | Silhouette | Number of Clusters | Smallest Cluster (%) | Largest Cluster (%) | Smallest/ Largest |
|---|---|---|---|---|---|---|---|
|  | K-means 1 | < 1 | 0,45 | 5 | 10 | 50 | 0,22 |
|  | K-means 3 | < 1 | 0,44 | 6 | 11 | 33 | 0,33 |
|  | TwoStep 1 | < 1 | 0,39 | 5 | 11 | 33 | 0,35 |
|  | TwoStep 3 | < 1 | 0,37 | 6 | 3 | 29 | 0,11 |
|  | K-means 4 | < 1 | 0,34 | 3 | 10 | 62 | 0,18 |

Zdroj: Vlastní výpočty

**Realizace shlukové analýzy pomocí algoritmu Kohonenovy mapy**

Využitý algoritmus samoorganizujících map funguje na podobném principu jako algoritmus *k*-průměrů. Cílem je zobrazit vícerozměrná data do dvourozměrného prostoru tak, aby byly podobné objekty umístěny v mapě co nejblíže. Dle předpokladů nelze očekávat výrazně lepší segmentační kvality, než tomu bylo v předchozích přístupech. Metoda samoorganizujících se map je vhodná pro segmentační modely s vyšším počtem vytvořených shluků.

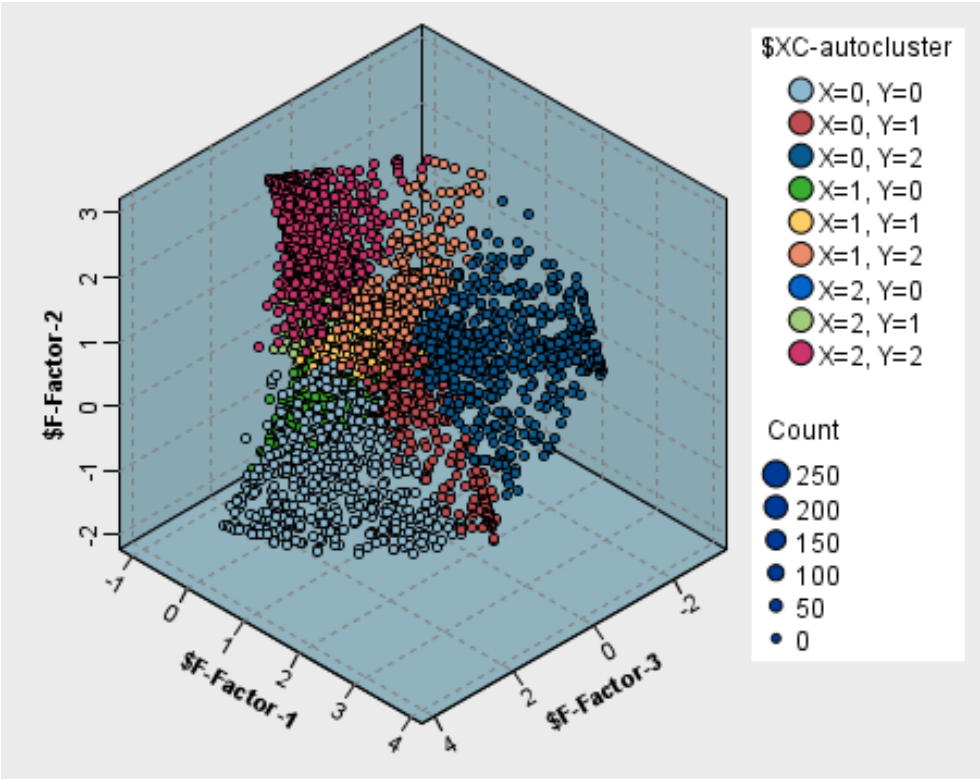
Nejvyšší hodnotu siluety získal model o velikosti mřížky 3x3, tedy celkem s devíti shluky. Tento model dosáhl průměrné hodnoty siluety s = 0,4.

Výsledný model samoorganizujících se map byl získán s následujícími parametry:

- Mřížka 3x3.
- Fáze 1: Sousedství 3, Eta 0,4, Cykly 20.
- Fáze 2: Sousedství 1, Eta 0,1, Cykly 150.

Princip tohoto algoritmu je zachycen na následujícím obrázku (Obr. č. 31). Jednotky jsou uspořádávány do pravoúhlé mřížky, což se projevilo na výsledném dělení jednotek do shluků.

Obr. č. 31: Výsledné shluky metody Kohonenových map vynesené dle komponentních skóre




Zdroj: Vlastní výpočty

Rozdělení zákazníků do vyššího počtu různých segmentů je z marketingového cílení nepraktické. Také segmentační vlastnosti a profilace shluků poukazují na skutečnost, že v první fázi segmentace není tento model vhodný pro nasazení do praxe. Jeho využití by však bylo možné zvažovat v případě následné hlubší analýzy zákazníků.

### 6.2.2 Zhodnocení modelů vytvořených pomocí nástroje Modeler

Z výše uvedených postupů byly zvoleny jako nejvhodnější modely vycházející z upravených vstupních proměnných. Prvním zvoleným modelem je model vycházející z hlavních komponent. Tento postup je z metodického hlediska korektní, jelikož byla v rámci procesu odstraněna nežádoucí multikolinearita mezi vstupními proměnnými. Také hodnoty průměrné siluety u metody *k*-průměrů poukazují na kvalitní model. Nevýhodou tohoto přístupu jsou snížené možnosti interpretace výsledných shluků.

Druhým vhodným přístupem je aplikace shlukové analýzy na vytvořené podílové kategorie. Nevýhodou tohoto kroku je existující multikolinearita mezi vstupními proměnnými, jeho interpretační možnosti však tuto nevýhodu předčí. Výsledný model *k*-průměrů se šesti shluky dosáhl hodnoty siluety (0,51), což je o tři desetiny méně, než tomu bylo při využití hodnotových proměnných, model ale vykazuje rovnoměrnější rozdělení zákazníků do jednotlivých shluků.

| Shluky vycházející z hlavních komponent | Shluky vycházející z podílových kategorií | | | | | |
|---|---|---|---|---|---|---|
| | cluster-1 | cluster-2 | cluster-3 | cluster-4 | cluster-5 | cluster-6 |
| cluster-1 | 24% | 0% | 8% | 0% | 22% | 1% |
| cluster-2 | 0% | 9% | 1% | 8% | 0% | 0% |
| cluster-3 | 0% | 0% | 6% | 0% | 0% | 5% |
| cluster-4 | 2% | 0% | 0% | 1% | 1% | 12% |

Tab. 9: Četnostní tabulka výsledného shlukování metody *k*-průměrů aplikované na hlavní komponenty a podílové proměnné

Zdroj: Vlastní výpočty

Tab. 9 uvádí porovnání četností výsledných shluků hlavních komponent a podílových proměnných. Z tabulky četností je patrná tendence algoritmu jednotky seskupovat do obdobných shluků nehledě na transformaci vstupních proměnných. První shluk hlavních

komponent je složen z objektů spadajících do prvního, pátého a částečně i třetího shluku podílových proměnných. Druhý shluk hlavních komponent je tvořen nákupními koši z druhého a čtvrtého shluku podílových proměnných. Složení třetích shluků se částečně překrývá a částečně zahrnuje i koše z šestého shluku podílových proměnných. Čtvrtý a šestý shluk se také z velké části překrývají.

### Zhodnocení modelu *k*-průměrů vycházejícího z hlavních komponent

Počet iterací potřebných pro tvorbu výsledných modelů hlavních komponent se čtyřmi a pěti shluky zachycuje následující (Tab. 10).

**Počet shluků: 4**

| Iterace | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Chyba | 0,404 | 0,147 | 0,216 | 0,091 | 0,030 | 0,014 | 0,006 | 0,003 | 0,001 | 0,001 |
| Iterace | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Chyba | 0,001 | 0,001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Počet shluků: 5**

| Iterace | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Chyba | 0,398 | 0,163 | 0,217 | 0,082 | 0,028 | 0,018 | 0,019 | 0,019 | 0,016 | 0,010 |
| Iterace | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Chyba | 0,008 | 0,003 | 0,003 | 0,001 | 0,001 | 0,002 | 0,001 | 0,001 | 0 | 0 |

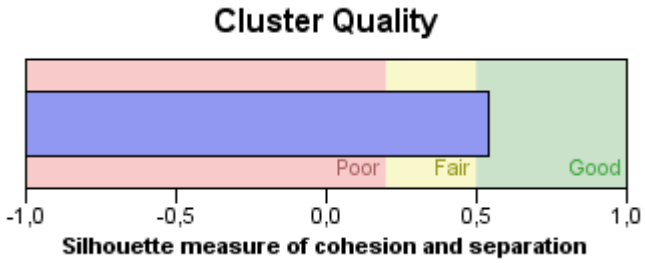
Tab. 10: Počet iterací při aplikaci metody *k*-průměrů na hlavní komponenty

Zdroj: Vlastní výpočty

V každém kroku iteračního procesu se vytvářejí dočasné shluky, jednotlivá pozorování jsou pak přiřazena do nejbližšího shluku. Iterační proces se zastaví, když je maximální relativní změna ve středu shluku menší nebo rovna konvergenčnímu kritériu a vyhovuje případným dalším podmínkám. Z výstupů vyplývá, že v případě čtyř shluků bylo dosaženo efektivního řešení s nižším počtem iterací než u pěti výsledných shluků.

Také v případě průměrné siluety vykázal model *k*-průměrů se čtyřmi výslednými shluky vyšší hodnotu než model s pěti shluky. Kvalitu nalezeného řešení vyjádřenou pomocí Silhouettovy míry koheze a separace shluků zachycuje následující graf (Obr. č. 32).

Obr. č. 32: Zhodnocení kvality modelu *k*-průměrů se čtyřmi shluky




Zdroj: Vlastní výpočty

**Zhodnocení modelu *k*-průměrů vycházejícího z podílových kategorií**

Počet iterací potřebných pro tvorbu výsledného modelu aplikovaného na podílové kategorie se šesti shluky zachycuje následující tabulka (Tab. 11). Z tabulky vyplývá, že v tomto případě bylo dosaženo efektivního řešení se stejným počtem iterací jako v případě aplikace algoritmu *k*-průměrů na hlavní komponenty s výslednými čtyřmi shluky.

**Počet shluků: 6**

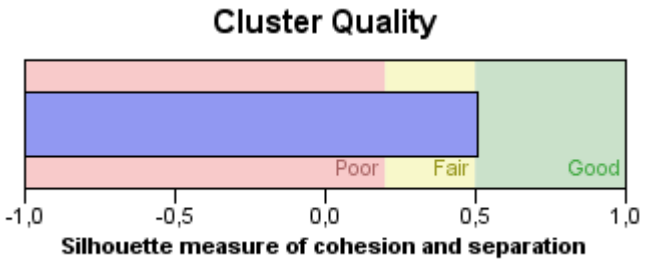
| Iterace | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Chyba | 0,679 | 0,296 | 0,106 | 0,032 | 0,009 | 0,003 | 0,001 | 0,001 | 0,001 | 0,001 |
| Iterace | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Chyba | 0,001 | 0,002 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Tab. 11: Počet iterací při aplikaci metody *k*-průměrů na podílové kategorie

Zdroj: Vlastní výpočty

Z níže uvedeného schématu (Obr. č. 33) vyplývá, že daný postup vykazuje nižší míru průměrné siluety než je tomu u modelu hlavních komponent, tento rozdíl je však zanedbatelný.

Obr. č. 33: Zhodnocení kvality modelu *k*-průměrů se šesti shluky




Zdroj: Vlastní výpočty

### 6.2.3 Profilace vytvořených shluků

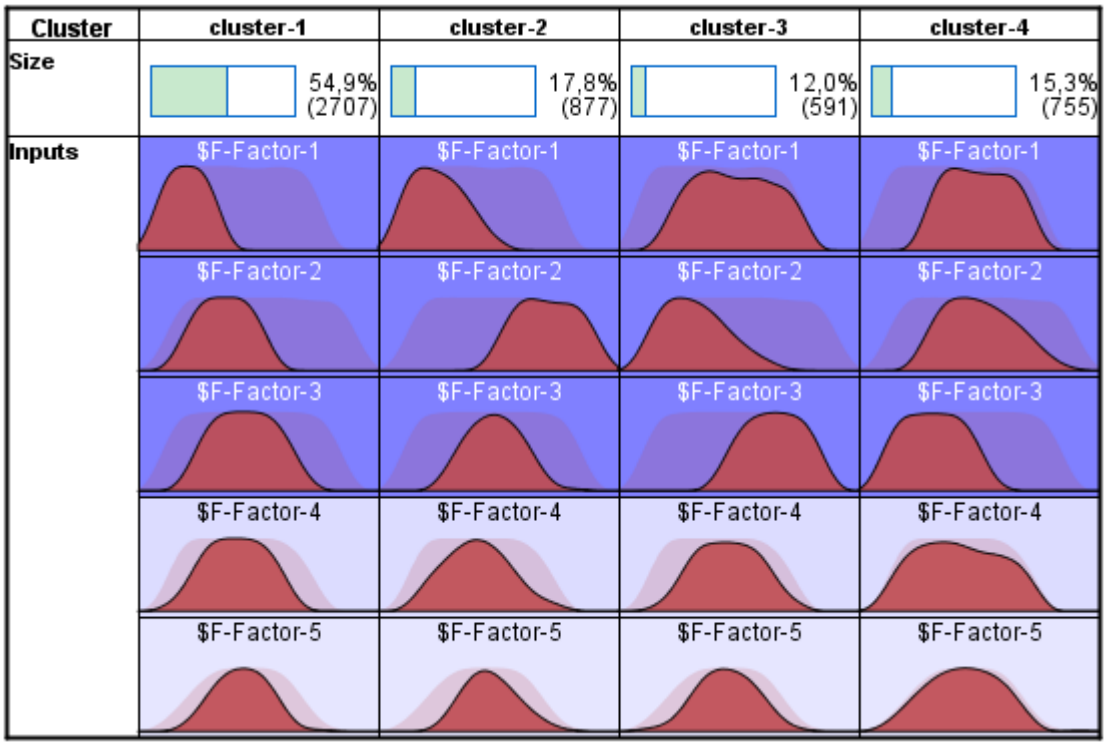
Výše uvedený proces segmentace rozdělil skupinu všech nákupních košů do menších homogenních celků, které se vzájemně liší potřebami zákazníků, jejich charakteristikami a nákupním chováním. Následná definice segmentů by měla usnadnit řídícím

a marketingovým pracovníků oslovit každý segment odlišnými nabídkami, které budou směřovány právě na potřeby zákazníků v daném segmentu. Finálním krokem segmentace je tedy identifikace a podrobnější analýza nalezených segmentů.

**Profilace segmentů vytvořených na základě hlavních komponent**

Nejdůležitější proměnnou z hlediska segmentace je první hlavní komponenta, která značí velikost nákupního koše. Koreluje se všemi vstupními proměnnými s výjimkou podílu akčního zboží, které je nejvíce zastoupeno ve druhé komponentě. Třetí komponenta představuje zastoupení třetí produktové kategorie FOOD3, tedy produkty řeznictví. Z níže uvedeného profilačního grafu (Obr. č. 34) je patrné, že interpretace tohoto přístupu je z důvodu transformace vstupních proměnných nepraktická. Pro názornost byla uvedena i tabulka průměrů původních netransformovaných proměnných.

Obr. č. 34: Profilace vytvořených segmentů




Zdroj: Vlastní výpočty

První nejčetnější shluk reprezentuje zákazníky s malým nákupním košem, tedy běžný denní nákup. Jednotlivé produktové kategorie jsou v koši zastoupeny rovnoměrně. Tito zákazníci se také nesoustředí na probíhající akce. Druhý, výrazně menší segment, zahrnující 18 % zákazníků, lze charakterizovat výrazným zastoupením akčního zboží v první produktové kategorii. Jedná se tedy o zákazníky, kteří se zaměřují převážně na akční zboží.

Segment č. 3, zastupující 12 % zákazníků, se vyznačuje velký mi a drahými nákupními koši. Průměrná hodnota koše je 1.032,- Kč. Z produktových kategorií je výrazněji zastoupena kategorie FOOD3, tedy produkty řeznictví. Velmi obdobný je i čtvrtý segment, který lze taktéž charakterizovat drahými nákupními koši. Oproti předcházejícímu segmentu však obsahuje výrazně vyšší zastoupení produktové kategorie FOOD1, tedy alkoholické i nealkoholické nápoje, drogérii, či balené zboží.

| Proměnná | cluster-1 | cluster-2 | cluster-3 | cluster-4 |
|---|---|---|---|---|
| Total | 145,837 | 220,774 | 1031,725 | 983,568 |
| FOOD1 | 75,934 | 123,065 | 385,813 | 610,917 |
| FOOD2 | 56,305 | 86,840 | 324,680 | 340,146 |
| FOOD3 | 13,599 | 10,869 | 321,232 | 32,505 |
| N_polozek | 5,646 | 6,927 | 28,672 | 31,258 |
| Akcni_zbozi | 5,294 | 63,289 | 13,202 | 17,989 |

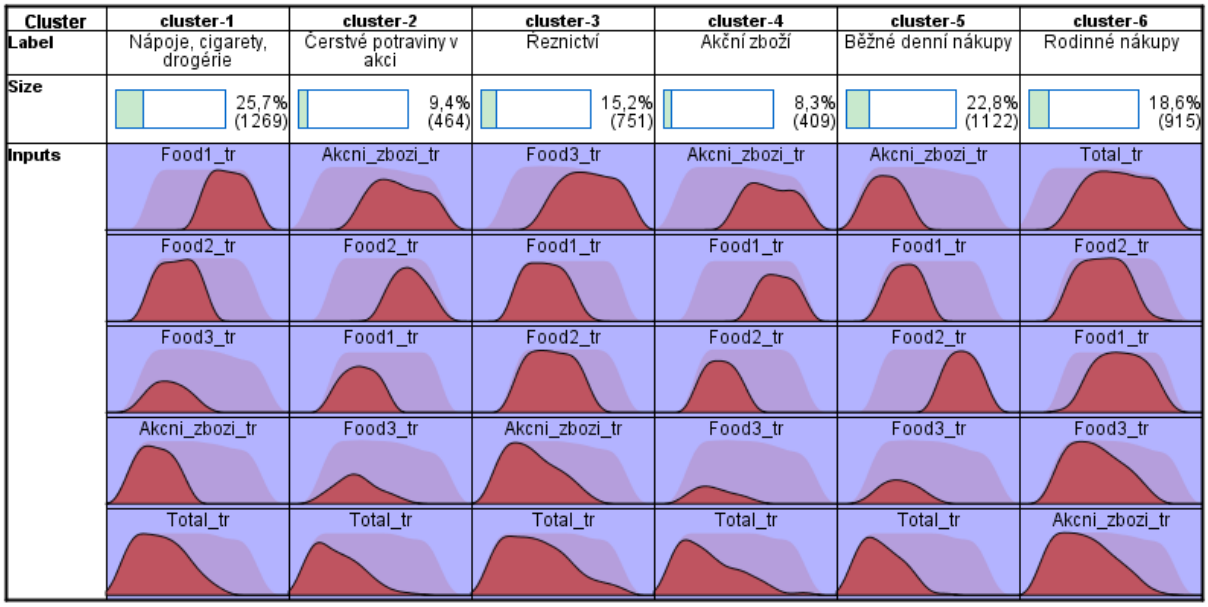
Tab. 12: Průměrné hodnoty vstupních proměnných v rámci vytvořených shluků

Zdroj: Vlastní výpočty

**Profilace shluků vytvořených z podílových proměnných**

V následujícím přehledu (Obr. č. 35) je uvedena profilace vytvořených segmentů na základě podílových kategorií.

Obr. č. 35: Profilace vytvořených segmentů




Zdroj: Vlastní výpočty

**První** a zároveň největší vytvořený **segment** zahrnuje jednu čtvrtinu všech zákazníků (25,7 %). Tento segment lze charakterizovat spíše menšími nákupními koši o průměrné hodnotě 200,- Kč a nízkým podílem akčního zboží. Jde o koše obsahující převážně výrobky z první produktové kategorie FOOD1 (nápoje, drogérii, balené zboží). Zbývající dvě produktové kategorie tvoří pouze 16 % celkové hodnoty koše.

**Druhý shluk** se od prvního segmentu neliší velikostí nákupního koše, ale jeho složením a velikostí. Obsahuje vysoký podíl akčního zboží a především produkty kategorie FOOD2, tedy čerstvé potraviny, pečivo, ovoce a zeleninu. Celkově zahrnuje v průměru každého desátého zákazníka.

**Třetí shluk** zahrnující 15 % zákazníků je charakterizován vysokým podílem řeznického zboží (kategorie FOOD3). Zákazníci patřící do toho shluku zpravidla příliš nevyužívají akčního zboží. Průměrná hodnota jejich nákupního koše je 404,- Kč, jedná se tedy o druhý nejhodnotnější segment.

**Čtvrtý**, nejméně zastoupený, **segment** se vyznačuje nejvyšším podílem akčního zboží. Jedná se o zákazníky reagující na probíhající akce. Nakupují produkty především z kategorie FOOD1, tedy alkoholické i nealkoholické nápoje, drogérii, konzervy.

V pořadí již **pátý segment** lze označit za segment běžných denních nákupů. Tento segment zahrnující 23 % zákazníků se řadí na druhé místo dle četnosti. Jedná se o nejmenší nákupní koše o průměrné hodnotě 106,- Kč, které z 90 % zahrnují položky druhé produktové kategorie FOOD2, tedy čerstvé potraviny, pečivo, ovoce a zeleninu. Obdobně jako první segment se vyznačuje nízkým podílem akčního zboží.

Poslední **šestý segment** je charakterizován výrazně vyšší hodnotou nákupního koše, než tomu bylo u segmentů předcházejících. Tento segment zahrnuje téměř pětinu zákazníků, kteří v průměru utratí 1.146,- Kč. Rozložení produktových kategorií je rovnoměrné, lze tedy předpokládat, že jde o velké rodinné nákupy.

Zatím co schéma výše (viz Obr. č. 35) zachycuje vizuální profilaci segmentů, v níže uvedené tabulce (Tab. 13) jsou vyjádřeny průměrné hodnoty jednotlivých proměnných v rámci daných segmentů.

| Proměnná | cluster-1 | cluster-2 | cluster-3 | cluster-4 | cluster-5 | cluster-6 |
|---|---|---|---|---|---|---|
| Total | 201,318 | 190,390 | 404,326 | 308,151 | 105,748 | 1146,010 |
| Akcni_zbozi | 5,133 | 59,649 | 12,243 | 67,457 | 4,278 | 16,879 |
| Food1 | 83,749 | 16,391 | 22,944 | 84,282 | 9,411 | 54,005 |
| Food2 | 15,181 | 80,718 | 31,313 | 14,278 | 89,683 | 36,244 |
| Food3 | 1,070 | 2,891 | 45,742 | 1,440 | 0,906 | 9,751 |

Tab. 13: Průměrné hodnoty podílových proměnných v rámci vytvořených shluků

Zdroj: Vlastní výpočty

### 6.2.4 Zhodnocení uvedených technik shlukování

V kapitole věnované shlukové analýze byly porovnávány různé přístupy využívané k segmentaci zákazníků. Z uvedených výstupů je patrné, že segmentační analýza poskytuje různé možnosti realizace a záleží především na zkušenostech analytika, jaký postup zvolí k získání finálního řešení.

V první fázi byly vyloučeny některé běžně využívané postupy shlukování jako nevhodné. V případě shlukování původních standardizovaných proměnných nastal problém s nežádoucí multikolinearitou obsaženou ve vstupním datovém souboru. S tím také souviselo nerovnoměrné rozložení výsledných shluků, kdy první nalezený shluk zahrnoval přibližně 50 % všech zákazníků. Tento problém byl vyřešen využitím podílových kategorií, tedy převodem jednotlivých produktových kategorií z korunového vyjádření na procentuální. Druhou možností, jak odstranit závislost vstupních proměnných, je využití nekorelovaných hlavních komponent jako vstupů do shlukové analýzy. Tento přístup společně s využitím podílových kategorií se následně ukázal jako nejvhodnější.

Naopak dále využitý model aplikovaný na diskretizované vstupní proměnné nepřinesl kvalitní výsledky. Jednotlivé proměnné byly převedeny na ordinální znaky na základě kvantilového rozložení. Výsledný model dvoustupňového shlukování sice přinesl srovnatelnou hodnotu siluety (0,49), obsahoval však pouze přibližně 40 % zákazníků. Zbývající jednotky zůstaly nezařazené. Kvůli výrazné ztrátě informace byl tento model označen jako nevhodný pro analýzu nákupních košů.

Zlepšení modelu nepřineslo ani využití nižší úrovně produktové kategorizace či seskupování pomocí algoritmu kohonenových map.

Finální model byl tedy vybírán ze dvou výsledných modelů *k*-průměrů. Tato segmentační technika vykazovala nejlepší segmentační vlastnosti z hlediska siluety, tedy hodnotícího kritéria, které kombinuje principy koheze a separace shluků. První ze zvažovaných modelů vycházel z podílových vstupních proměnných, druhý z hlavních komponent. Nespornou výhodou prvního přístupu je jednoduchost modelu a možnost snazší interpretace, dále rovnoměrnost vytvořených shluků. Využití hlavních komponent je sice metodicky korektnější, avšak složitější a obtížně aplikovatelné v praxi. Z tohoto důvodu byl jako vhodnější označen model *k*-průměrů vycházející z podílových proměnných.

Kompletní proces modelování v nástroji IBM SPSS Modeler je zachycen na procesním digramu v příloze č. 4.

### 6.2.5 Modelování v nástroji Enterprise Miner

Pro účely modelování pomocí modulu Enterprise Miner byla využita již upravená datová matice v nástroji Modeler. Do procesního toku tedy vstupovaly již restrukturalizované a agregované proměnné. Analýza byla provedena pouze na podílových produktových kategoriích.

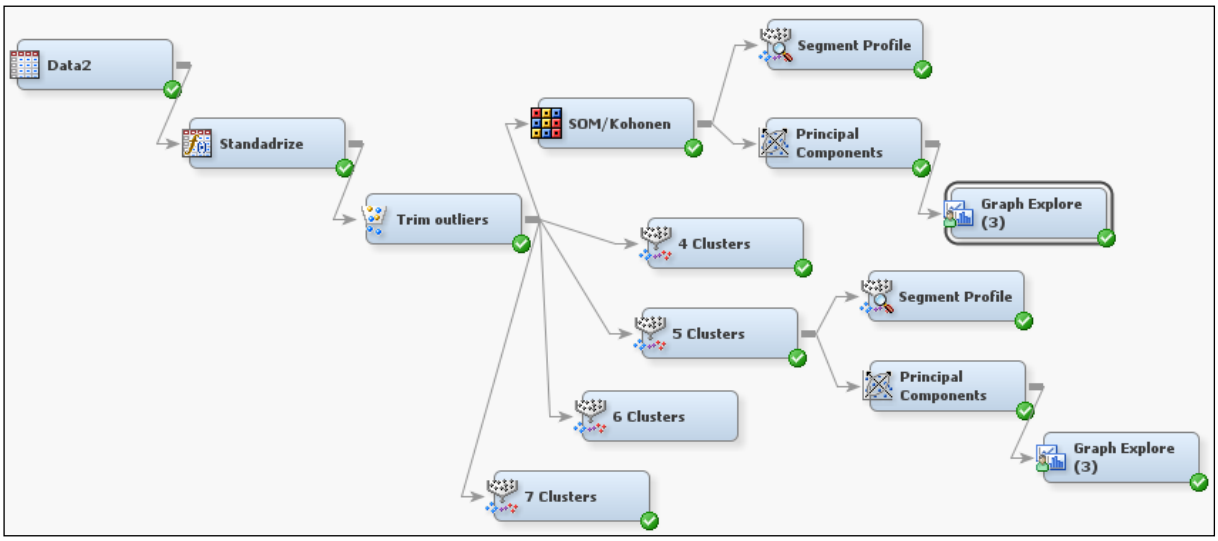
Prvním krokem modelování byla úprava proměnných, která spočívala ve standardizaci proměnných a odstranění odlehlých pozorování. Transformace číselných proměnných byla provedena pomocí uzlu *Transform Variables*. Tento uzel zahrnuje 18 možných transformací spojitých proměnných. Mezi nejběžněji využívané transformace patří: standardizace, logaritmování, exponenciální transformace, kategorizace apod. EM nabízí i možnost transformace metodou maximalizace normality. Z důvodu jednoduchosti a interpretační snadnosti byla zvolena, stejně jako v případě nástroje Modeler, pouze standardizace vstupních proměnných.

Odstranění odlehlých pozorování bylo provedeno pomocí uzlu *Filter*. Tento uzel poskytuje nástroje pro vypořádání se s extrémními hodnotami. Lze využít metodu průměrné absolutní odchylky, useknutí extrémních percentilů, či metodu modálních center. Přednastavenou metodou je nejčastěji využívaná metoda vzdálenosti *n* (tří) směrodatných odchylek od průměru. Tento uzel dále poskytuje možnosti vypořádání se s chybějícími údaji či nominálními proměnnými. Zvolena byla možnost useknutí (trimming) jednotek přesahujících vzdálenost tří směrodatných odchylek od průměru.

V prostředí Enterprise Mineru lze využít dva algoritmy shlukování: metodu *k*-průměrů a Kohonenovy mapy. Výsledný počet shluků byl volen na základě dostupných kritérií shlukování. Samotná shluková analýza představuje pouze první krok procesu, v druhém kroku byla provedena profilace vytvořených segmentů pomocí uzlu *Segment Profile*. Zachycení výsledných segmentů v trojrozměrném grafu umožňuje uzel *Graph Explore*. K tomuto účelu byla, stejně jako v prostředí Modeler, využita analýza hlavních komponent. Do grafu pak byla

vynesena komponentní skóre prvních tří hlavních komponent. Realizace analýzy hlavních komponent proběhla pomocí uzlu *Principal Component*. Celý proces shlukování je uveden v následujícím přehledu (Obr. č. 36).

Obr. č. 36: Proces shlukové analýzy v Enterprise Mineru




Zdroj: Vlastní výpočty

**Shlukování pomocí *k*-průměrů**

Realizace shlukové analýzy metodou *k*-průměrů probíhá pomocí modelovacího uzlu Cluster. K inicializaci zárodečných středů metody *k*-průměrů byla využita metoda plného nahrazení (Full Replacement), která vychází z procedury FASTCLUS v základním nástroji SAS/STAT. Tato technika poskytla nepatrně lepší výsledky jednotlivých shlukovacích statistik než druhá využívaná MacQueenova metoda.

Výběr počtu shluků byl založen na dostupných statistikách systému Enterprise Miner. Výsledné hodnoty jednotlivých kritérií jsou uvedeny v následující tabulce (Tab. 13).

| Kritérium | 4 shluky | 5 shluků | 6 shluků | 7 shluků |
|---|---|---|---|---|
| CRITERION | 0,740 | 0,560 | 0,540 | 0,500 |
| PSEUDO_F | 1224,970 | **2470,660** | 2272,090 | 2341,220 |
| ERSQ | 0,360 | 0,440 | 0,510 | 0,540 |
| CCC | 39,180 | **143,630** | 126,380 | 139,130 |
| TOTAL_STD | 1,000 | 1,000 | 1,000 | 1,000 |
| WITHIN_STD | 0,740 | 0,560 | 0,540 | 0,500 |
| RSQ | 0,450 | 0,680 | 0,710 | 0,760 |
| RSQ_RATIO | 0,810 | 2,170 | 2,490 | 3,080 |

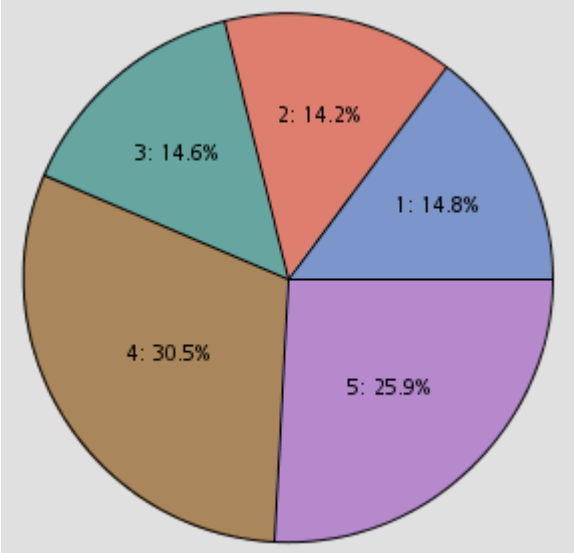
Tab. 14: Statistiky vytvořených shluků pomocí metody *k*-průměrů

Zdroj: Vlastní výpočty

Počet shluků lze určit na základě hodnot kubického shlukovacího kritéria (CCC) či pseudo F statistiky. Ideální počet shluků odpovídá lokálnímu maximu daných kritérií. V tomto případě tvoří lokální maximum 5 shluků. Uvedených pět shluků vykazuje taktéž největší nárůst indexu determinace (RSQ), který v případě pěti shluků dosahuje 68 %.

Na základě uvedených hodnot statistiky pseudo F a kubického shlukového kritéria (CCC) bylo tedy zvoleno pět výsledných shluků. Četnostní rozdělení shluků zachycuje obrázek níže (Obr. č. 37).

Obr. č. 37: Rozložení výsledných shluků metody *k*-průměrů




Zdroj: Vlastní výpočty

K tvorbě těchto shluků nejvíce přispěly proměnné: *hodnota nákupního koše v Kč*, *podíl položek v kategorii FOOD1* a *podíl položek v kategorii FOOD2*. Důležitost jednotlivých vstupních proměnných je uvedena v následující tabulce (Tab. 15).

| Proměnná | Důležitost |
|---|---|
| TOTAL | 1,000 |
| FOOD1 | 0,962 |
| FOOD2 | 0,946 |
| FOOD3 | 0,892 |
| AKCNI_ZBOZI | 0,862 |

Tab. 15: Důležitost proměnných

Zdroj: Vlastní výpočty

### 6.2.6 Profilace shluků vytvořených metodou *k*-průměrů

Profilování vytvořených segmentů slouží k lepšímu pochopení toho, co se v analyzovaných datech skrývá. Slouží marketingovým pracovníkům k zacílení na specifické segmenty. K profilaci již vytvořených segmentů byl implementován uzel *Segment Profile node*. Pro využití tohoto uzlu musí být definována segmentační proměnná s nastavenou rolí cluster nebo segment. Tento uzel poskytuje rozdělení vstupních proměnných v rámci jednotlivých segmentů a přehledný výstup ve formě histogramů, který porovnává rozdělení původních (standardizovaných) vstupních proměnných s rozdělením těchto proměnných v rámci vytvořených shluků. Grafická profilace segmentů je uvedena na Obr. č. 38. Další možnosti profilace segmentů nabízí uzel *StatExolore*.

Na vytvoření **1. segmentu** se nejvíce podílela proměnná podíl produktů v kategorii FOOD3, tedy produktů řeznictví. Jde o zákazníky se středně velkými nákupními koši, které obsahují především produkty řeznictví, doplněny jsou čerstvými potravinami (FOOD2) a v menším množství i nápoji či drogérií (FOOD1). Tento segment, zahrnující téměř 15 % zákazníků, lze označit za segment menších rodinných nákupů.

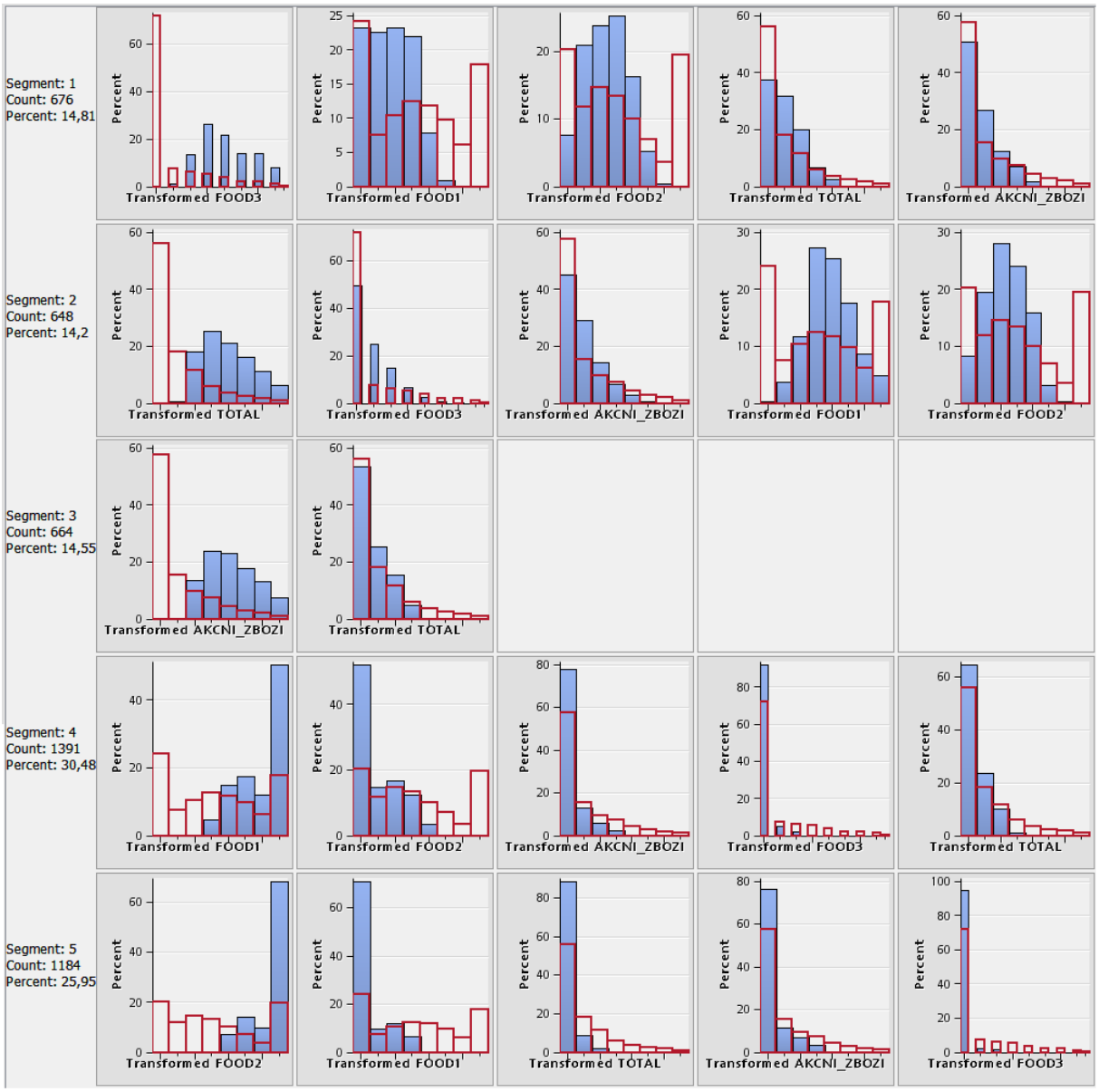
**2. segment** se vyznačuje drahými nákupními koši, ve kterých jsou obsaženy všechny produktové kategorie. Tento segment, zahrnující 14 % košů, lze tedy označit jako segment velkých rodinných nákupů.

**3. segmen**t byl vytvořen pouze na základě dvou vstupních proměnných, a sice podílu akčního zboží a celkové hodnotě koše. Je charakteristický malými nákupními koši s nejvyšším podílem akčního zboží.

Z hlediska zastoupení je nejčetnější **4. segment**. Představuje malé nákupní koše obsahující především produkty kategorie FOOD1, tedy nápoje, drogerii či balené zboží. Lze předpokládat, že jde o náhodný nákup, tedy o zákazníky, kteří svůj nákup neplánovali, ale pouze se zastavili pro chybějící položku. Tito zákazníci buď učinili velký nákup v jiném časovém horizontu, nebo u konkurence.

**5. segment** zahrnuje nejmenší koše s čerstvými potravinami. Jedná se o druhý nejčetnější segment. Z profilace vyplývá, že kromě produktů z kategorie FOOD2 (čerstvé potraviny) nezahrnuje téměř žádné další položky. Tento segment lze proto označit jako segment běžných denních nákupů.

Obr. č. 38: Profilace vytvořených segmentů metodou *k*-průměrů




Zdroj: Vlastní výpočty

### Grafické znázornění vytvořených segmentů pomocí hlavních komponent

Aby bylo možné znázornit vytvořené segmenty v trojrozměrném bodovém grafu, bylo nutné transformovat vstupní proměnné na hlavní komponenty. Komponentní skóre pak sloužilo k vynesení jednotek do grafu.

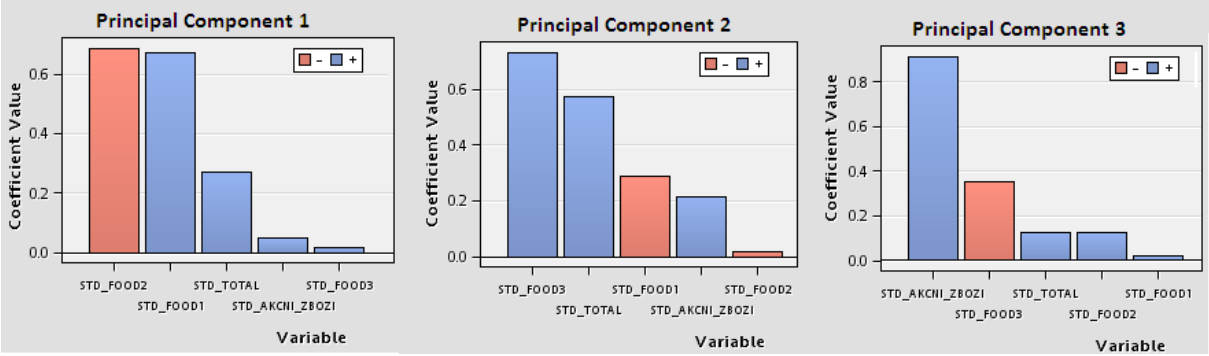
Realizovaná analýza hlavních komponent vycházela z korelační matice. Z níže uvedené tabulky (Tab. 16) je patrné, že první tři požadované hlavní komponenty vysvětlují 85,8 % informace obsažené v původních proměnných.

| PC | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 1,987 | 0,706 | 0,397 | 0,397 |
| 2 | 1,281 | 0,261 | 0,256 | 0,654 |
| 3 | 1,020 | 0,308 | 0,204 | 0,858 |
| 4 | 0,712 | 0,712 | 0,142 | 1,000 |
| 5 | 0,000 | 0,000 | 0,000 | 1,000 |

Tab. 16: Vlastní čísla korelační matice

Zdroj: Vlastní výpočty

První hlavní komponenta, vysvětlující téměř 40 % rozptylu původních proměnných, představuje komponentu středně velkých košů, obsahujících především produkty z kategorie FOOD1. Na vytvoření druhé komponenty se nejvíce podílela proměnná FOOD3 a hodnota nákupního koše. Tato komponenta, vysvětlující 26 % původní informace, představuje velké nákupní koše s výrazným zastoupením řeznických produktů. Třetí komponenta je komponenta akčního zboží. Vysvětluje 20 % původní informace a souvisí především s proměnnou podíl akčního zboží. Zastoupení jednotlivých proměnných v rámci hlavních komponent je uvedeno na schématu níže (Obr. č. 39).

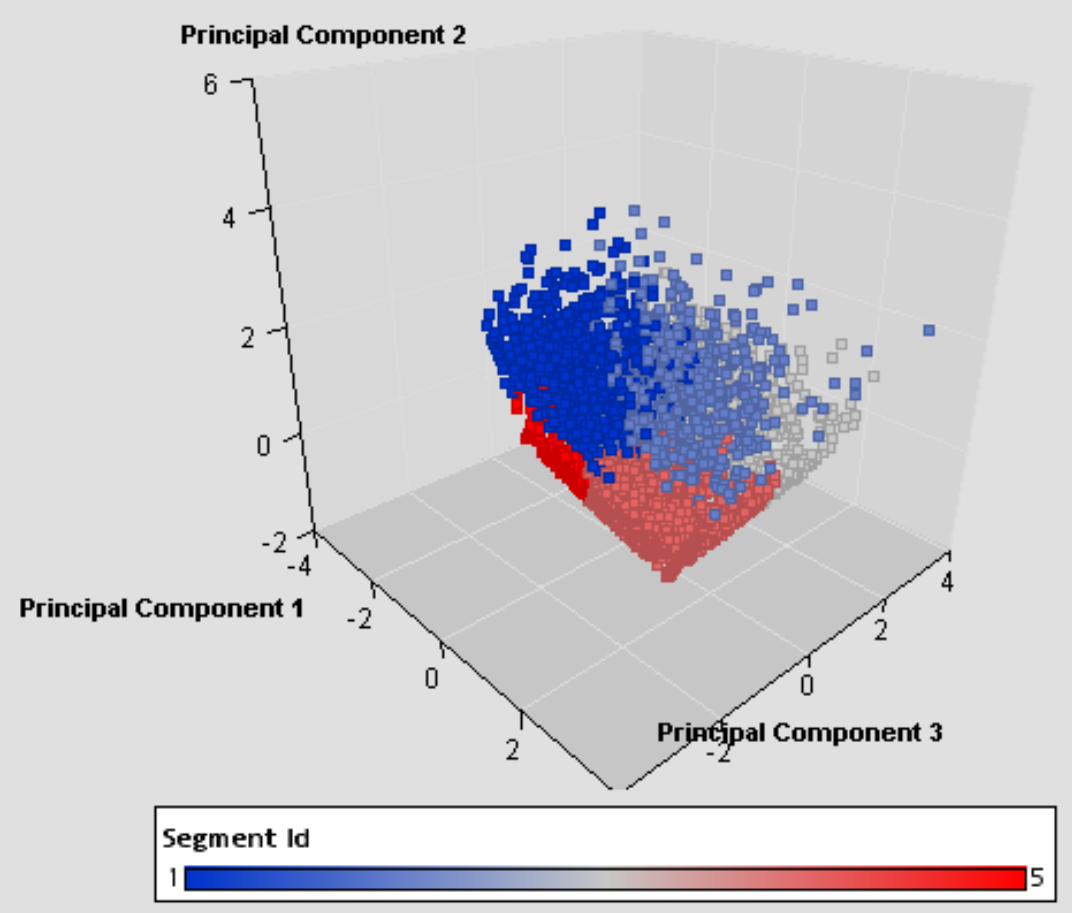



Obr. č. 39: Podíl vstupních proměnných na složení hlavních komponent

Zdroj: Vlastní výpočty

Po vynesení jednotlivých komponentních skóre do grafu získáme následující bodový graf zachycující rozdělení vstupního souboru do nově vytvořených segmentů. Modře jsou zachyceny segmenty rodinných nákupů. Tmavě modrá představuje velké rodinné nákupy, světle modrá naopak menší rodinné nákupy. Šedivě je znázorněn segment akčního zboží. Červená barva představuje malé nákupní koše, které se dělí na příležitostný nákup nápojů, drogerie či baleného zboží (světle červená) a denní nákup čerstvých potravin (tmavě červená).

Obr. č. 40: Vizualizace segmentů vytvořených metodou *k*-průměrů




Zdroj: Vlastní výpočty

### 6.2.7 Shlukování pomocí Kohonenových map

Kohonenovy mapy představují speciální shlukovací algoritmus vycházející z metodiky neuronových sítí. Jedná se o síť bez skryté vrstvy uspořádané do mapy ve tvaru obdélníku či čtverce. V rámci procesu segmentace byly využity zpravidla přednastavené parametry shukování, jelikož jejich upravení nevedlo ke zlepšení modelu.

Jednotlivé charakteristiky vytvořených shluků metodou Kohonenových samoorganizačních map jsou uveden v Tab. 16. Na základě hodnot kubického shlukovacího kritéria a pseudo F kritéria byla zvolena dimenze 2 x 2 se čtyřmi výslednými shluky.

| KRITÉRIUM | DIMENZE | | | |
|---|---|---|---|---|
| | 2x2 | 2x3 | 2x4 | 4x4 |
| PSEUDO_F | **1603,726** | 1266,348 | 990,451 | 662,693 |
| ERSQ | 0,363 | 0,513 | 0,566 | 0,672 |
| CCC | **75,183** | 36,180 | 19,064 | 7,425 |
| TOTAL_STD | 1,000 | 1,000 | 1,000 | 1,000 |
| WITHIN_STD | 0,698 | 0,647 | 0,630 | 0,561 |
| RSQ | 0,513 | 0,581 | 0,604 | 0,686 |
| RSQ_RATIO | 1,055 | 1,389 | 1,522 | 2,186 |

Tab. 17: Statistiky vytvořených shluků pomocí algoritmu Kohonenových map

Zdroj: Vlastní výpočty

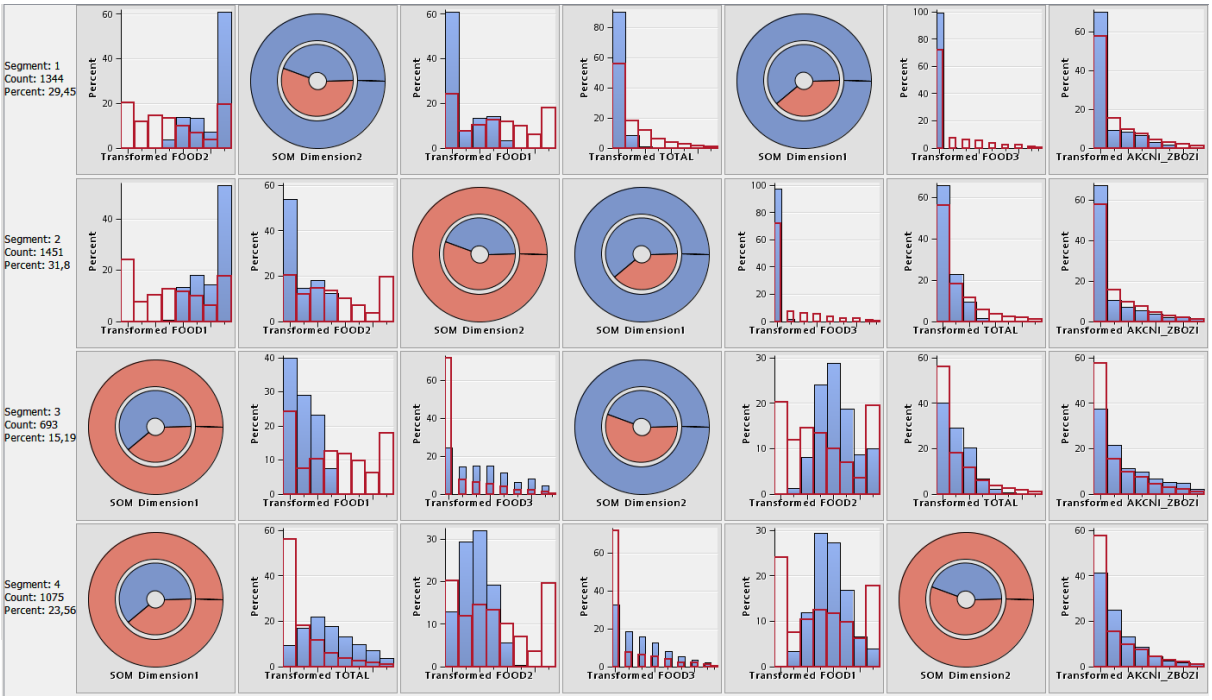
**Profilace vytvořených segmentů metodou samoorganizující se map**

Pomocí metody Kohonenových map byly získány čtyři výsledné segmenty, které profilaci přibližně odpovídají segmentům vytvořených metodou *k*-průměrů. Vlastnosti jednotlivých segmentů lze odvodit z četnostního rozložení, které je zachyceno na Obr. č. 41.

Dle uvedeného schématu (Obr. č. 41) lze vytvořené segmenty popsat následovně:

- **1. segment** je tvořen spíše menšími nákupními koši, které obsahují zvýšený podíl položek z řeznictví.
- **2. segment** představují malé koše obsahující nápoje, drogerii či balené zboží.
- **3. segment** tvoří středně velké koše obsahující čerstvé potraviny a produkty řeznictví.
- **4. segment** zahrnuje drahé nákupní koše, které značí velké rodinné nákupy.

Obr. č. 41: Profilace segmentů získaných pomocí algoritmu Kohonenovy mapy



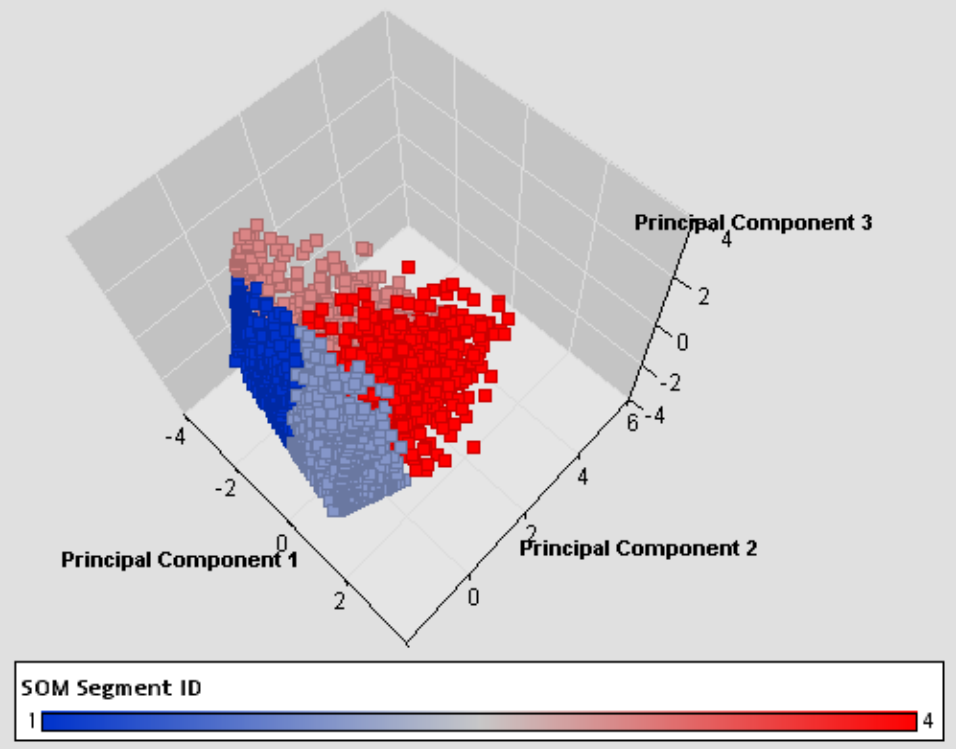

Zdroj: Vlastní výpočty

Grafické znázornění vytvořených segmentů pomocí již definovaných hlavních komponent zachycuje graf níže (

Obr. č. 42). Z grafu je patrné, že v tomto případě algoritmus nebral v úvahu existující třetí dimenzi, tedy zastoupení akčního zboží.

Přestože Kohonenovy mapy jsou oproti metodě *k*-průměrů relativně novou moderní technikou, v tomto případě nevykázaly lepší výsledky a nejsou k modelování nákupních košů vhodným řešením. Výhody těchto map jsou spatřovány především v případě většího množství vytvářených shluků, což je v praxi méně často využívané.

Obr. č. 42: Vizualizace segmentů vytvořených metodou Kohonenových map




Zdroj: Vlastní výpočty

## 6.3 Porovnání výsledných shluků získaných metodou *k*-průměrů

Pro porovnání výsledků získaných v prostředí nástrojů Modeler a Enterprise Miner byla zvolena jednoduchá frekvenční tabulka. Pomocí nástroje Modeler bylo na základě hodnoty průměrné siluety získáno šest výsledných shluků, oproti tomu v Enterprise Mineru bylo dle kubického kritéria zvoleno pět segmentů.

Z výsledné četnostní tabulky je patrné, že vytvořené segmenty v obou nástrojích zpravidla odpovídají, odlišně zařazené jednotky se pohybují vždy v maximální četnosti do 2 % záznamů. Tmavě modrý segment představuje segment menších rodinných nákupů. Jedná se o druhý nejhodnotnější segment. Světle modře je označen nejhodnotnější segment velkých rodinných nákupů. Šedivé shluky odpovídají segmentu akčního zboží. Modeler tento shluk rozdělil na dva segmenty: segment akčních čerstvých potravin a segment nápojů a drogérie

v akci. Světle červený segment představuje segment malých náhodných nákupů. Poslední tmavě červený segment lze označit segmentem běžných denních nákupů.

| Shluky vytvořené v nástroji EM | Shluky vytvořené nástrojem Modeler | | | | | |
|---|---|---|---|---|---|---|
| | 1. segment | 2. segment | 3. segment | 4. segment | 5. segment | 6. segment |
| 1. segment | 0% | 0% | 13% | 0% | 0% | 1% |
| 2. segment | 0% | 0% | 0% | 0% | 0% | 14% |
| 3. segment | 1% | 7% | 0% | 6% | 0% | 1% |
| 4. segment | 27% | 0% | 0% | 1% | 0% | 2% |
| 5. segment | 0% | 2% | 0% | 0% | 24% | 0% |

Tab. 18: Relativní četnost shluků vytvořených nástroji Modeler a Enterprise Miner

Zdroj: Vlastní výpočty

## 6.4 Výhody a nevýhody využitých dataminingových nástrojů

V rámci disertační práce byla hodnocena práce ve dvou nejčastěji využívaných dataminingových nástrojích. Prvním nástrojem je IBM SPSS Modeler, druhým Enterprise Miner společnosti SAS. Porovnání bylo provedeno z pohledu výzkumného pracovníka, který připravuje podporu pro marketingové kampaně. Určité vlastnosti mohou být pro pokročilou skupinu uživatelů výhodou, pro začátečníky naopak nevýhodou.

Z pohledu výzkumného pracovníka jsou první kroky s modelovacím nástrojem Modeler výrazně jednodušší. Snadná instalace systému a uživatelsky přívětivé a intuitivní rozhraní jsou velkým přínosem tohoto nástroje. Naopak Enterprise Miner z tohoto pohledu působí poněkud těžkopádně. Jeho tvůrci pravděpodobně nepředpokládají, že by byl tento nástroj využíván začátečníky samostatně bez zkušeností s prací v jiných modulech společnosti SAS. Lze předpokládat, že zkušenější uživatele jiných nástrojů společnosti SAS komplikovanější instalace a nepříliš uživatelsky přívětivé prostředí nepřekvapí.

Zdánlivá jednoduchost Modeleru však nemusí být vždy k užitku. Například nastavení formátů vytvořených proměnných je v Enterprise Mineru skryto v jednotlivých uzlech, v Modeleru je ale nutné po jakékoliv manipulaci s daty nastavovat typ proměnných v samostatném uzlu, což nejenže představuje činnost navíc, ale s každým novým uzlem se modelovací prostředí zaplňuje a stává nepřehlednější.

Oba nástroje mají velmi dobře propracovanou nápovědu včetně řešených příkladů a ukázkových datových souborů. Taktéž uživatelská podpora je v obou případech na velmi vysoké úrovni. Oba nástroje jsou podpořeny množstvím školení v mateřských organizacích. Velká výhoda Modeleru však může být spatřována v možnosti semestrálního dataminingového kurzu, který pořádá společnost ACREA ČR, výhradní distributor softwaru IBM SPSS. Naopak konkurenční SAS pořádá pouze 3denní školení, což je z hlediska obsáhlosti této tematiky nedostačující.

Jednotlivé zaznamenané výhody a nevýhody využitých dataminingových nástrojů jsou shrnuty níže.

**Výhody nástroje IBM SPSS Modeler 14.2**:

- Snadná instalace systému
- Jednoduchá orientace v programu
- Přátelské uživatelské rozhraní
- Snadná orientace v generovaných výstupech
- Možnost tvořit super uzly, které umožňují skrýt část kódu do jednoho samostatného uzlu (zapouzdření).
- Obsahuje automatizované uzly (např. Auto Cluster), které umožňují vytvoření většího počtu modelů s různými parametry v jednom kroku
- Velké množství názorných příkladů v nápovědě, včetně ukázkových souborů a připravených datových toků
- Kontextová nápověda
- Kvalitní uživatelská podpora
- Podpora produktů formou školení (možnost semestrálního kurzu na téma Data Mining)

**Nevýhody nástroje IBM SPSS Modeler 14.2**:

- Obtížná orientace na pracovní ploše projektu po jejím zaplnění – projekt se stává nepřehledný (většina algoritmů má samostatný uzel; neustále je nutné nastavovat formáty proměnných v samostatných uzlech; připojování výstupních uzlů apod.)
- Nedostatečná parametrizace jednotlivých algoritmů
- Nízká odbornost nápovědy – chybí podrobná metodika k užívaným metodám
- Výstupy jsou zjednodušené a často neobsahují důležité informace (např. nelze zjistit hodnotu BIC ani jiného obdobného kritéria ve výsledcích shlukové analýzy)
- Vysoká náročnost na využívanou paměť a procesorový čas (např. u grafických úloh, složitých modelů…)
- Pády systému při zpracování hardwarově náročných úloh

**Výhody nástroje SAS Enterprise Miner 12.1**:

- Jednotlivé uzly umožňují vyšší parametrizaci, což ocení především pokročilejší uživatelé
- Přehlednější pracovní plocha (není třeba využívat velkého počtu uzlů)
- Možnost měnit rozlišení (velikost) pracovní plochy
- Možnost vygenerovat kód, který může být využit i v dalších modulech, např. SAS 9.3, nebo naopak vložit už vytvořený kód do datového toku
- Interaktivní práce s daty v průběhu explorační analýzy i při hodnocení výsledků analýz
- Velké množství výstupů z jednotlivých analýz
- I při náročných úlohách systém nepadá
- Propracovaná nápověda s detailní metodologií
- Kvalitní, ochotná a rychlá uživatelská podpora
- Podpora nástroje formou školení (možnost 3denního kurzu Data Miningu)

**Nevýhody nástroje SAS Enterprise Miner 12.1**:

- Složitá instalace systému
- Časová náročnost složitějších výpočtů
- Není user-friendly, složitější orientace v pracovním prostředí i v generovaných výstupech
- Rozsáhlost nápovědy způsobuje její částečnou nepřehlednost
- Chybí kontextová nápověda

Porovnání obou využívaných systémů z pohledu výzkumného či marketingového pracovníka je uvedeno v souhrnu níže (Tab. 19). Oba systémy byly hodnoceny na základě realizovaného dataminingového procesu, kterému předcházela instalace obou systémů. Hodnocení vychází z výše uvedených výhod a nevýhod obou nástrojů. Výsledné porovnání spočívalo v bodovém hodnocení na škále 1 – 5, kde 1 znamená naprostou spokojenost s realizací dané funkcionality/vlastnosti systému a 5 naprostou nespokojenost s danou funkcionalitou.

Z uvedeného srovnání obou nástrojů vychází jako mírný favorit systém SAS Enterprise Miner.

| Kritéria hodnocení | IBM SPSS Modeler | SAS Enterprise Miner |
|---|---|---|
| Instalace systému | 1 | 4 |
| Uživatelském rozhraní | 1 | 2 |
| Ovladatelnost | 2 | 4 |
| Přehlednost generovaných výstupů | 2 | 3 |
| Úplnost generovaných výstupů | 3 | 1 |
| Přehlednost procesního diagramu (pracovní plochy) | 3 | 1 |
| Automatizace modelů | 1 | 3 |
| Parametrizace modelů | 4 | 1 |
| Přehlednost a úplnost nápovědy | 1 | 2 |
| Odbornost nápovědy | 2 | 1 |
| Řešené příklady v nápovědě | 1 | 1 |
| Uživatelská podpora | 2 | 1 |
| Náročnost na využívanou paměť a procesorový čas, pády systému | 3 | 1 |

Tab. 19: Zhodnocení využitých softwarových systémů dle zvolených kritérií

Zdroj: Vlastní zpracování

# 7 SHRNUTÍ REALIZOVANÉHO POSTUPU SEGMENTACE

V rámci vlastní části práce byl analyzován datový soubor obsahující transakční údaje konkrétního hypermarketu. Úvodní fáze modelování se zaměřila na přípravu datového souboru. Jedná se o jednu z nejdůležitějších a časově nejnáročnějších fází dataminingového procesu, ve které byly řešeny možnosti úprav vstupních proměnných pro potřeby segmentace. Vstupní datový soubor obsahoval následující vstupní proměnné:

- identifikační číslo transakce (EAN),
- název zboží (Name),
- jednotková cena uvedeného zboží (Price),
- množství zboží v nákupního koši (PCount),
- celková cena (Total),
- produktová kategorie (WGR),
- typ platby (Payment),
- identifikační číslo nákupního koše (BasketN),
- typ probíhající akce (AKTIONSNR),
- podkategorie zboží (WGR01).

K porozumění vstupní datové matici byly využity možnosti datového auditu, který zahrnuje základní popisné charakteristiky a grafické znázornění rozdělení četností daných proměnných. Pro vytvoření modelu byly zvoleny pouze relevantní proměnné, které byly dále upraveny za účelem zkvalitnění výsledného segmentačního modelu. Důležitým krokem byl výběr vhodné úrovně produktových kategorií z produktového listu, který je uveden v příloze č. 1. Prokázalo se, že vhodným úvodním dělením jsou tři produktové kategorie. Toto rozdělení dosáhlo výrazně vyšších hodnot Silhouetovy míry než volba sedmi produktových kategorií. Jde o:

- kategorii FOOD1, která obsahuje alkoholické a nealkoholické nápoje, balené potraviny (konzervy, slané a sladké pečivo) a drogérii,
- kategorii FOOD2, jež zahrnuje čerstvé potraviny (samoobslužný prodej např. balené uzeniny a masné výrobky, mléčné výrobky apod.), pečivo, ovoce a zeleninu,

- kategorii FOOD3, která představuje produkty řeznictví.

Kromě těchto tří produktových kategorií do modelu vstoupila také proměnná *celková cena nákupního koše*, dále *počet položek* v koši a *akční zboží*. Tyto proměnné jsou výstupem realizované restrukturalizace transakcí v datovém souboru. Pomocí následné agregace vstupních proměnných byla datová matice uspořádána tak, aby každý řádek matice odpovídal jednomu nákupnímu koši. Úprava datového souboru po restrukturalizaci zahrnovala vypořádání se s vytvořenými chybějícími údaji, které ve skutečnosti představovaly nulové hodnoty, a vytvoření nové podílové proměnné: *podíl akčního zboží v koši*. Relativní hodnota této proměnné umožňuje vzájemně porovnat podíl akčního zboží v jednotlivých nákupních koších.

Následný datový audit vykázal výrazně asymetrické rozdělení vstupních proměnných. Z tohoto důvodu byly zvažovány možnosti transformace či diskretizace daných veličin. Logaritmická transformace vstupních proměnných nebyla aplikována, jelikož by došlo k zásadní ztrátě informace a ke snížení interpretovatelnosti výsledků. Obdobně bylo upuštěno také od možnosti diskretizace proměnných. Nejenže v takovém případě dochází ke ztrátě potenciálně důležité informace, výsledné dataminingové modely navíc vykázaly výrazně horší segmentační vlastnosti. V případě analýzy transakčních dat lze tedy doporučit z dostupných transformací především datovou normalizaci, neboli standardizaci proměnných na stejnou škálu.

V případě existence odlehlých pozorování v datové matici se potvrdilo jako žádoucí využít některou z technik na vypořádání se s extrémními hodnotami. V rámci modelování byla využita winsorizace, tedy metoda nahrazení. Existující anomálie byly upraveny především z důvodu citlivosti dále využívaných segmentačních algoritmů na extrémní pozorování.

Při vícerozměrné analýze transakčních dat zpravidla dochází k výskytu mulitikolinearity mezi vstupními veličinami. Konkrétně jde o vzájemnou závislost mezi cenou nákupního koše, počtem položek v koši a hodnotou potravin v jednotlivých produktových kategoriích. Jednou z ověřovaných možností na redukci existující dimenzionality v datovém souboru je analýza hlavních komponent. Výstupem této analýzy jsou ortogonální hlavní komponenty, které představují lineární kombinace původních vstupních proměnných. Využití těchto komponent jako vstupních proměnných k modelování vedlo ke kvalitnímu modelu, jenž vykázal velmi dobré segmentační vlastnosti. Nevýhodou tohoto přístupu je, podobně jako u logaritmické

transformace, ztráta interpretovatelnosti nalezeného řešení. Výsledný model není modelem vstupních proměnných, ale jejich lineárních kombinací. Z tohoto důvodu nebylo dané řešení zvoleno jako finální.

Nejvhodnějším přístupem k volbě vstupních proměnných byl zvolen expertní výběr s následnou úpravou analyzovaných veličin. Produktové kategorie byly převedeny na podíly těchto kategorií na hodnotě celého nákupního koše, čímž došlo k redukci vzájemné závislosti mezi kategoriemi a celkovou cenou. Proměnná udávající počet položek v nákupním koši byla z modelování vyloučena, jelikož nesla obdobnou informaci jako celková cena. Výsledný dataminingový model byl tedy sestaven na základě následujících vstupních proměnných:

- celková cena nákupního koše (Total),
- podíl hodnoty produktů první kategorie na celkové hodnotě nákupního koše (FOOD1),
- podíl hodnoty produktů druhé kategorie na celkové hodnotě nákupního koše (FOOD2),
- podíl hodnoty produktů třetí kategorie na celkové hodnotě nákupního koše (FOOD3),
- podíl akčního zboží v nákupním koši (Akcni_zbozi).

Následující fáze modelování zahrnuje aplikaci zvolených algoritmů shlukování na vstupní datovou matici. Samotný proces modelování byl realizován ve dvou dataminingových nástrojích – IBM SPSS Modeler a SAS Enterprise Miner.

K realizaci segmentace v nástroji Modeler byly využity tři dostupné shlukovací algoritmy. Jedná se o metodu *k*-průměru, metodu dvoustupňového shlukování a Kohonenovy mapy. Kvalita nalezených řešení byla hodnocena pomocí Silhouetovy míry a grafického znázornění získaných shluků. Přestože nástroj Modeler poskytuje možnost otestovat významnost rozdílů mezi jednotlivými shluky pomocí klasického F-testu, výsledky testů nebyly brány v potaz, jelikož v případě velkých datových souborů dochází k porušení stanovených předpokladů testování.

Nástroj Enterprise Miner poskytuje pro účely segmentace pouze dvě modelovací techniky: metodu *k*-průměru a Kohonenovy mapy. Oproti Modeleru však nabízí větší výběr kritérií shlukování. Kvalita nalezených řešení byla hodnocena pomocí kubického shlukovacího kritéria (CCC), pseudo F statistiky a grafického znázornění získaných shluků. K inicializaci zárodečných středů metody *k*-průměrů byla využita metoda plného nahrazení (Full

Replacement), která poskytla lepší výsledky jednotlivých shlukovacích statistik než metoda MacQueenova. Ke grafickému zobrazení výsledných shluků byly opět využity možnosti analýzy hlavních komponent, která posloužila k redukci dimenzionality vstupní datové matice.

V průběhu modelovací fáze bylo zjišťováno, zda má na výsledný model vliv zvolený počet iterací. Pomocí testování se potvrdilo, že automaticky přednastavená hranice 20 iterací je dostačující. Výsledky modelů s vyšším počtem iterací (50, 100, 200) se neliší. Výsledný model *k*-průměrů se šesti shluky byl sestaven na základě 12 kroků iteračního procesu.

V rámci procesu modelování byly využity různé přístupy k tvorbě výsledného segmentačního modelu. Kromě již zmiňované úpravy datové matice využitím expertního výběru a hlavních komponent byly ověřovány taktéž možnosti diskretizace asymetricky rozdělených vstupních proměnných. Výsledné ordinální proměnné byly využity pouze v případě metody dvoustupňového shlukování, jež využívá kromě euklidovské vzdálenosti také věrohodnostní funkci. Je tedy vhodná pro segmentaci diskrétních veličin. Ve výsledcích se však negativně projevila vlastnost této techniky vytvářet samostatný shluk nezařazených údajů. V případě modelu s nejvyšší hodnotou Silhouetovy míry (siluety) obsáhl tento shluk nezatříděných objektů přibližně 60 % všech vstupních jednotek (v dalších modelech bylo procento ještě vyšší). Případný přínos daného modelu v praxi je proto velmi nejistý.

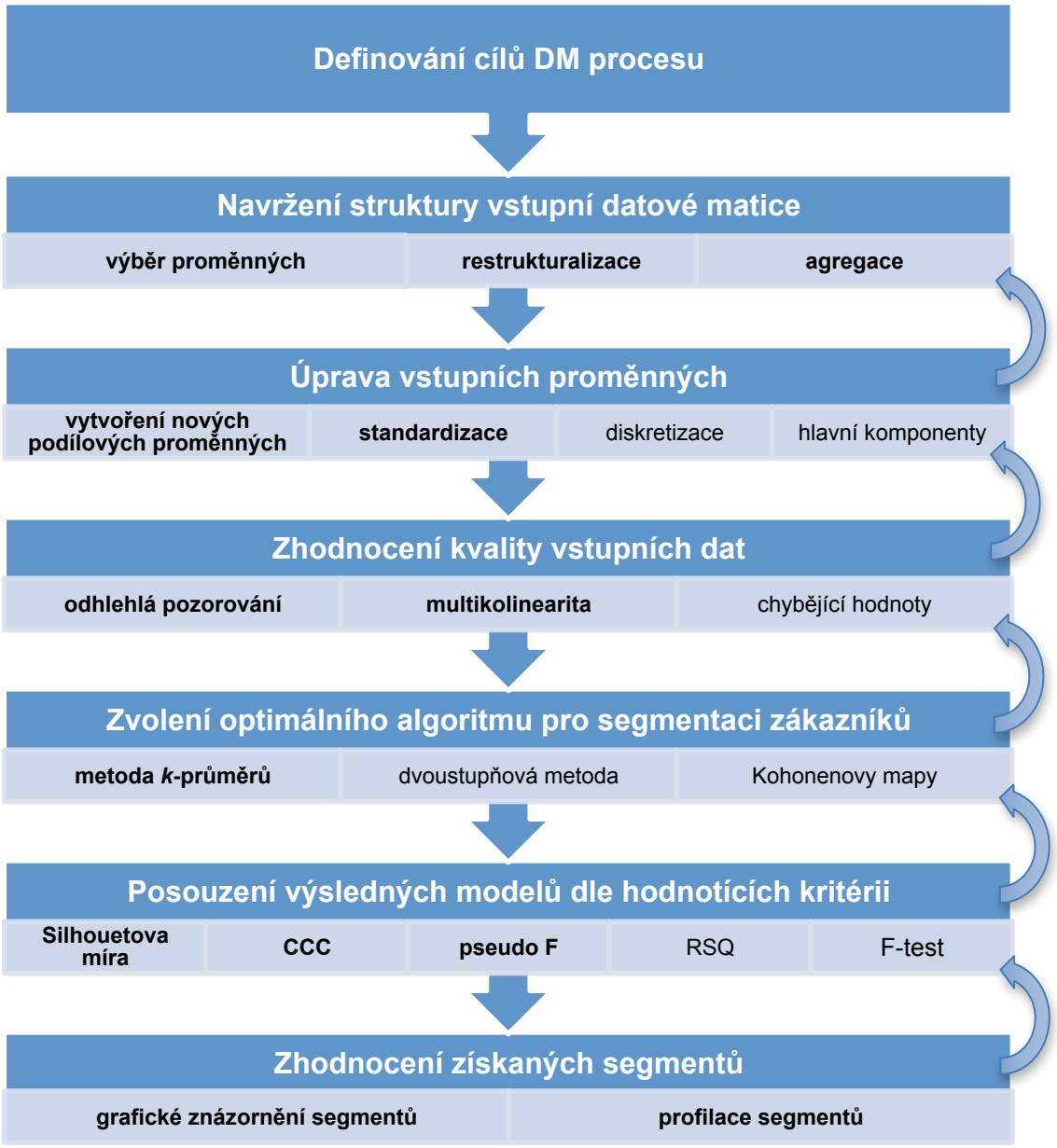
Segmentačním modelem, jenž využívá předností neuronových sítí, je algoritmus Kohonenových map. Tento algoritmus zpravidla vykazoval nejslabší výsledky segmentace. Jedním z důvodů je jeho vysoká parametrizovatelnost. Nalezení ideálního nastavení jednotlivých parametrů modelu je proto velmi náročné a závislé na zkušenostech analytika. Dalším možným důvodem je vlastnost algoritmu uspořádávat jednotlivé objekty do pravoúhle mřížky. Výsledný počet segmentů může být tedy např. čtyři (mřížka 2x2), šest (mřížka 2x3), devět (mřížka 3x3) atd. Tato rozdělení však nemusí korespondovat se skutečnou strukturou existující ve zkoumaných datech. Model s nejvyšší hodnotou siluety získaný pomocí Modeleru nevykázal vlastnosti využitelné z praktického hlediska, rozdělení shluků nebylo logické a počet shluků příliš vysoký (9). Naopak model získaný nástrojem Enterprise Miner vykazoval relativně dobré segmentační vlastnosti. Výsledné čtyři segementy se složením příliš nelišily od segmentů získaných metodou *k*-průměrů. Závažným nedostatkem však byla chybějící dimenze, která by zachycovala podíl akčního zboží.

Jako finální model s nejlepšími segmentačními vlastnostmi byl zvolen model realizovaný na expertně zvolených a upravených podílových vstupních proměnných. Jako nejlepší shlukovací algoritmus se projevila metoda *k*-průměrů, která dosáhla nejvyšší hodnoty Silhouetovy míry. Hodnota s = 0,51 poukazuje na skutečnost, že jde o dobře separované shluky, což potvrzuje i nízký počet iterací potřebný k dosažení výsledného řešení. Obdobně v případě Enterprise Mineru vykázal algoritmus metody *k*-průměrů kvalitní řešení segmentačního problému. Dle kubického shlukovacího kritéria a pseudo F statistiky byl zvolen jako nejlepší možný model s pěti výslednými shluky. Grafické znázornění finálních řešení bylo získáno využitím komponentních skóre. Výsledný trojrozměrný graf zachycuje rozdělení jednotlivých nákupních košů do šesti, respektive pěti, vytvořených homogenní shluků.

Na základě uvedeného postupu realizace segmentace zákazníků byl vytvořen následující procesní diagram (Obr. č. 43).

Proces výběru vhodné struktury datové matice, realizace shlukování a následné hodnocení získaných segmentů lze označit za expertní modelování. Existuje nespočet různých možností a nastavení, která je nutné v průběhu modelování respektovat. Jelikož výsledky segmentace není možné ověřit klasickými testovacími postupy, vychází výsledný model především ze zkušeností výzkumníka. Tato práce si neklade za cíl předložit vyčerpávající výčet všech možností, ale doporučit obecný metodický postup segmentace v dataminingovém softwarovém prostředí. Výše popsaný postup proto může sloužit jako určitý návod k realizaci segmentace transakčních údajů pro podporu marketingového rozhodování.

Obr. č. 43: Jednotlivé fáze realizovaného dataminingového procesu.




Zdroj: Vlastní zpracování

# 8 DISKUZE A ZÁVĚR

Shlukování či segmentace zákazníků představuje do jisté míry subjektivní proces dataminingového modelování. Jednoznačný postup, který by byl na konci modelování validován či testován, v rámci segmentace neexistuje. Proces shlukování hledá existující, ale zatím neznámou, strukturu v datovém souboru. Tedy něco, co není možné empiricky ověřit či otestovat. Přínosem předkládané práce je proto především její metodický charakter. Při tvorbě obecného postupu segmentace byl kladen důraz na jeho význam při praktickém využití. Vytvořený postup je přínosný především pro výzkumné a marketingové pracovníky.

Hlavním cílem disertační práce bylo vytvoření obecného metodického rámce pro segmentaci zákazníků. V rámci práce byly také zohledněny dílčí cíle dataminingového procesu definované v úvodní části práce. Jde o:

- navržení vhodné struktury vstupní datové matice,
- zvolení vhodné úpravy proměnných vstupujících do modelování,
- zhodnocení kvality vstupních dat dle předpokladů využitých modelů,
- zvolení optimálního algoritmu pro segmentaci zákazníků dle dostupných kritérií shlukování,
- provedení profilace a logického zhodnocení získaných segmentů,
- posouzení vhodnosti využitých softwarových nástrojů.

V první fázi DM procesu se potvrdilo, že prvotní výběr vhodné struktury datové matice z velkého datového skladu je klíčový pro další průběh shlukování. V rámci práce byla zvolena struktura produktového stromu se třemi hlavními skupinami potravin, jde o kategorii nápojů a drogérie (FOOD1), čerstvé potraviny (FOOD2) a produkty řeznictví (FOOD3). Jak bylo demonstrováno, v případě využití sedmi produktových kategorií dosáhly segmentační techniky výrazně horších výsledků shlukování dle segmentačních kritérií. Pro úvodní segmentaci zákazníků je nezbytné vytvořit jasnou a přehlednou strukturu, proto byly vybrány pouze tři hlavní produktové kategorie.

Přístup k explorační analýze dat v rámci dataminingu se částečně liší od klasické přípravy dat pro statistickou analýzu. K těmto účelům zpravidla slouží speciální uzly, které zahrnují klasické jednorozměrné charakteristiky a jednoduché grafy a které poskytují přehled o rozložení vstupních proměnných. V případě výrazně jednostranně zešikmených

proměnných zahrnují tyto nástroje možnosti logaritmické či jiné transformace. Otázka, zda transformaci využít, není jednoduše zodpověditelná. Na jednu stranu vhodně využitá logaritmická transformace může výrazně zlepšit výsledky určitého modelovacího algoritmu, na druhou stranu pak dané výsledky neodpovídají zadání úlohy a jejich interpretace je náročná. Právě z důvodu interpretačního hlediska bylo od logaritmické transformace v rámci práce upuštěno. Nejednalo by se o model hodnoty nákupních košů a jejich složení, ale o model logaritmů těchto hodnot a jednotlivých produktových kategorií. Jedinou použitou transformací v modelu tak zůstala standardizace proměnných na stejnou škálu.

V přípravné fázi dat byla dále řešena problematika multikolinearity neboli neviditelného procesu vážení. Existence multikolinearity mezi vstupními proměnnými může zkreslit výsledky získaného modelu. Tento problém byl řešen dvěma způsoby: upravením datové matice a využitím analýzy hlavních komponent. Zvolený expertní způsob úpravy datové matice spočíval jednak ve vyřazení proměnné počet položek, jež velmi silně korelovala s celkovou hodnotou nákupního koše, jednak v převedení jednotlivých produktových kategorií, které byly vyjádřeny v Kč, na procentuální podíl těchto položek v nákupním koši.

Další možností, která byla řešena za účelem zkvalitnění modelu shlukování, byla diskretizace vstupních proměnných. Tato transformace, přestože způsobuje ztrátu určité informace obsažené v modelu, je velmi často používaným krokem v rámci dataminingového modelování. V DM nástrojích je zpravidla vyčleněn samostatný uzel či parametr pro ne/supervizovanou diskretizaci. Z hlediska analyzovaného datového souboru se tato možnost neprojevila jako žádoucí. V případě využití techniky *k*-průměrů je tento krok nelogický, jelikož algoritmus převádí kategorické proměnné zpět na číselné pomocí tzv. dummy proměnných (0/1). Využití získaných ordinálních proměnných přicházelo v úvahu pouze v nástroji Modeler, konkrétně při využití algoritmu dvoustupňového shlukování. Tato technika dosáhla dle hodnoty průměrné siluety obdobných výsledků jako ostatní modely, avšak vytvořila samostatný shluk nezatříděných údajů, který tvořil bezmála 60 % všech jednotek v modelu. Tento přístup byl tedy shledán pro zkoumanou problematiku jako nevhodný.

Jako protipól ke klasické metodě *k*-průměrů byl využit algoritmus Kohonenovy mapy. Tento algoritmus představuje relativně nový přístup ke shlukování. V tomto konkrétním případě shlukování nákupních košů však příliš neobstál. Nejlepší model velikosti mřížky 3 x 3, jenž

byl získán pomocí nástroje Modeler, dosáhl jednak výrazně nižší hodnoty průměrné siluety v porovnání s modelem *k*-průměrů, jednak je výsledných devět segmentů pro praktické využití neefektivní. Získané čtyři segmenty v nástroji Enterprise Miner by mohly být v praxi využitelné, avšak nereflektují existující třetí dimenzi, tedy podíl akčního zboží v nákupním koši.

Optimálním algoritmem pro segmentaci zákazníků tak byl zvolen postup využívající metodu *k*-průměrů aplikovanou na produktové kategorie vyjádřené podílem k celkové ceně nákupního koše. Dle zvolených kritérií na výběr požadovaného počtu shluků (silueta, kubické shlukovací kritérium) bylo vybráno šest výsledných shluků v případě nástroje Modeler a pět shluků při modelování v nástroji Enterprise Miner. Výsledné složení shluků z obou nástrojů téměř odpovídá, až na třetí shluk vytvořený systémem Enterprise Miner, který sdružuje jednotky ze dvou výsledných shluků nástroje Modeler.

**Hodnocení získaných segmentů**

Výsledné segmenty jsou jednoduše strukturované a logicky popsatelné, tudíž z praktického hlediska efektivně využitelné. Pomocí shlukovací techniky *k*-průměrů byl soubor zákazníků rozdělen do následujících skupin:

- **Segment velkých rodinných nákupů** - vyznačuje se drahými nákupními koši, obsahující všechny produktové kategorie.
- **Segment menších rodinných nákupů** - jde o zákazníky se středně velkými nákupními koši, které obsahují především produkty řeznictví, doplněny jsou čerstvými potravinami a v menším množství i nápoji či drogérií.
- **Segment akčního zboží** – tento segment je charakteristický malými nákupními koši s nejvyšším podílem akčního zboží. Jde o zákazníky kupující převážně čerstvé potraviny v akci nebo zákazníky kupující alkoholické a nealkoholické nápoje či drogérii v akci.
- **Segment běžných denních nákupů** - zahrnuje nejmenší nákupní koše s čerstvými potravinami.

- **Segment malých náhodných nákupů** - představuje malé nákupní koše obsahující především alkoholické a nealkoholické nápoje, balené potraviny nebo drogerii.

**Zhodnocení využitých DM nástrojů**

V rámci disertační práce byly také porovnány možnosti využitých dataminingových nástrojů a byla zhodnocena jejich využitelnost při řešení problému segmentace zákazníků.

Při hodnocení jednotlivých dataminingových nástrojů je patrný odlišný cíl tvůrců obou systémů. Enterprise Miner obsahuje nižší počet vysoce parametrizovaných uzlů, konkurenční Modeler disponuje rozsáhlým množstvím z velké části automatizovaných uzlů. Nástroj Modeler je vhodný pro začínající uživatele nebo uživatele nevyžadující vysoce parametrizované modely s detailními výstupy. Naopak tvůrci Enterprise Mineru, kteří pochází z akademického prostředí, vytvořili nástroj pro náročné uživatele, kteří jsou s problematikou dataminingu velmi dobře obeznámeni. Zatímco snahou tvůrců Modeleru je dosáhnout co nejvyšší srozumitelnosti systému z hlediska uživatelů, tvůrci Enterprise Mineru se orientují především na funkčnost a parametrizaci úloh.

Otázkou je využitelnost těchto systémů v praxi. Zatímco Modeler může intuitivně využít v podstatě každý, složitost Enterprise Mineru potenciální samouky zpravidla odradí. Pokud ale budeme vycházet z předpokladu, že kvalitní a v praxi „fungující“ model lze vytvořit pouze za předpokladu dobré znalosti dané problematiky a využitých algoritmů, pak je systém Enterprise Miner favoritem. Poskytuje totiž uživateli nejen podstatně rozšířenější možnosti vstupního nastavení modelů, ale především detailní výstupy zahrnující různorodá kritéria pro výběr vhodného modelu k implementaci. Tvorba rozhodnutí na základě získaných výsledů dataminingového procesu vyžaduje, aby měl uživatel k dispozici veškerá podpůrná kritéria a byl si vědom možných rizik spojených např. s nereflektováním předpokladů využitých algoritmů. Pro nejnáročnější uživatele obsahuje tento nástroj možnost vložit vlastní kód výpočtu, jehož syntaxe je stejná jako v jiných modulech systému SAS. Cenou za funkčnost, kterou poskytuje Enterprise Miner, je však náročnější instalace a jeho složitější ovladatelnost.

Taktéž z realizovaného bodového hodnocení obou systémů vychází jako mírný favorit systém Enterprise Miner, který získal lepší hodnocení především v odborněji definovaných kritériích. Naopak nástroj Modeler dominoval převážně ve srozumitelnosti a jednoduchosti ovládání.

**Obecná metodologie segmentace**

Na základě realizované segmentační analýzy lze marketingovým a výzkumným pracovníkům doporučit obecný postup segmentace zákazníků. Tento postup včetně grafického znázornění je podrobně popsán v kapitole 7.

Výběr dataminingového nástroje vychází z předchozích zkušeností analytika. Pro začínající uživatele je jednoznačně přínosnější nástroj IBM SPSS Modeler, pro pokročilé uživatele s vyššími nároky na parametrizaci modelů je pak vhodnější systém SAS Enterprise Miner.

Při přípravě datového souboru k modelování by nemělo docházet k podcenění této fáze procesu. Jedná se o nejdůležitější krok analýzy. Vlastní práce potvrdila, že kvalita výsledků segmentačních modelů je přímo úměrná kvalitě vstupních dat.

Do datové matice vstupují pouze relevantní proměnné, které mají logický význam pro plnění stanovených cílů segmentace. V úvodní fázi DM postupu je nutné provést restrukturalizaci a následnou agregaci transakčních dat tak, aby každý řádek vstupní matice představoval právě jeden nákupní koš (jednoho zákazníka). K odstranění nežádoucí multikolinearity vyskytující se v datech, lze doporučit úpravu vstupních proměnných, např. vytvořením podílových proměnných. Z dalších úprav datové matice je žádoucí využít standardizaci proměnných na stejnou škálu a upravení existujících odlehlých pozorování. Naopak se neprojevilo jako efektivní využit diskretizaci vstupních proměnných, ani jejich logaritmickou transformaci. Využití hlavních komponent, jako vstupních proměnných do shlukové analýzy, lze doporučit v případě většího množství vstupních proměnných. Nevýhodou tohoto přístupu je náročnost následné interpretace výstupů a hledání logických souvislostí.

K realizaci segmentace je žádoucí využít metodu *k*-průměrů, která prokázala ve všech realizovaných přístupech nejlepší segmentační vlastnosti, z čehož vyplývá, že ji lze obecně doporučit k realizaci segmentace zákazníků dle jejich nákupního chování.

Ke zhodnocení získaných výsledků a ke zvolení ideálního počtu segmentů je vhodné použít kritéria CCC nebo pseudo F statistiku. V nástroji Modeler je pak dostupná pouze Silhouettova míra. Logické posouzení segmentů vyplývá z realizované profilace.

**Praktický přínos práce**

Z hlediska marketingových a obchodních cílů lze konstatovat, že získané segmenty poskytují kvalitní podklady pro účely plánování marketingových kampaní. Cílem analýzy bylo poznat strukturu zákazníků a jejich nákupní chování. Díky shlukové analýze mají marketingoví pracovníci nástroj na rozpoznání existujících segmentů zákazníků a jejich relativní zastoupení. Následně tyto informace mohou využít jako podporu k marketingovým rozhodnutím. Je zřejmé, že zvýšením podílu zákazníků v segmentu rodinných nákupů dojde ke zvýšení tržeb. Úkolem marketingu je tedy na základě získaných údajů naplánovat marketingovou kampaň s cílem přesunout zákazníky z malých segmentů do segmentů rodinných nákupů. Následné analýzy by měly prokázat, jak se mění struktura zákazníků v souvislosti se změnou marketingových kampaní.

Tato segmentace představuje první krok k poznání zákazníků. Pro praktické účely segmentace lze marketingovým pracovníkům doporučit personifikaci jednotlivých zákazníků například zavedením věrnostních karet, které poskytnou kontinuální informace o nákupním chování zákazníků. Dojde tím k získání základních socio-ekonomických informací o zákaznících, které lze taktéž využít pro účely segmentace. Sběr těchto dat musí probíhat v souladu s dodržováním pravidel o ochraně osobních údajů.

# 9 SEZNAM POUŽITÉ LITERATURY


1. Arnold, S. J. A Test for Clusters, *Journal of Marketing Research*, 1979, Vol. 16, pp. 545–551.

2. Awad, M. et al. *Design and implementation of data mining tools*. Boca Raton, FL: Auerbach Publications, 2009. ISBN 978-1-4200-4590-1.

3. Azevedo, A., Santos, M. F. KDD, SEMMA and CRISP-DM: A parallel overview. *In Proceedings of the IADIS European Conference on Data Mining* 2008, s. 182–185. ISBN 978-972-8924-63-8.

4. Ball, G. H., Hall, D. J. *A novel method of data analysis and pattern classification*. Technical report, Stanford, Research Insitute, Menlo Park, 1965.

5. Beh, E. J. Elliptical confidence regions for simple correspondence analysis. *Journal of Statistical Planning and Inference*, September 2010, vol. 140, iss. 9, s. 2582–2588.

6. Berry, D., A., Moore, D., S., Albert, J. Teaching Elementary Bayesian statistics with Real Application in Sciences. *The American Statistician*, August 1997, vol. 51, no. 3, s. 241–268.

7. Berry, M. J. A., Linoff, G. S. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. Second Edition. Indianapolis: Wiley Publishing, 2004. ISBN 0-471-47064-3.

8. Cerrito, P. *Introduction to Data Mining Using SAS Enterprise Miner*. Cary, NC: SAS Institute, 2006. ISNB 978-59049-829-5.

9. Collica, R. S. *CRM Segmentation and Clustering Using SAS Enterprise Miner*. Cary, NC: SAS Institute, 2007. ISNB 978-1-59047-508-9.

10. Duda, R. O., Hart, P. E., Stork, D. G. *Pattern Classification*. 2nd edition. New York: Wiley Publishing, 2001.

11. Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well Separated Clusters, *Journal of Cybernetics*, 1973, vol. 3, s. 32–57.

12. Drucker, P. F. *Výzvy managementu pro 21. století*. 1. vyd. Praha: Management Press, 2000. 187 s. ISBN 80-7261-021-X.

13. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press, 1996. 560 pages. ISBN 978-0262560979.

14. Field, A. *Discovering Statistics Using SPSS*. 2nd edition. London: SAGE Publications, 2005. 816 pages. ISBN 0-7619-4451-6.

15. Forgy, E. W. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 1965, vol. 21, s. 768–769.

16. Frawley, W., Piatetsky-Shapiro, G., Matheus, C. Knowledge discovery in databases – An overview. *AI Magazine*, 1992, vol. 13, no. 3, s. 57–70.

17. Giudici, P., Figini, S. *Applied Data Mining for Business and Industry*. 2nd edition. Cornwall: Wiley Publishing, 2009. ISBN 978-0-470-05886-2.

18. Hair, J. F., Anderson, R. E. *Multivariate data analysis*. 6th edition. London: Prentice Hall, 2010. 785 pages. ISBN 978-0-138-13263-7.

19. Han, J., Kamber, M. *Data Mining: Concepts and Techniques*. 1st edition. London: Morgan Kaufmann Publisher, 2000. ISBN 1-55860-498-8.

20. Hand, D. J. *What You Get Is What You Want? Some Dangers of Black Box Data Mining*. M2005 Conference Proceedings. Cary, NC SAS Institute, 2005.

21. Hand, D., J., Mannila, H., Smyth, P. *Principles of Data Mining*. Cambridge, MA: The MIT Press, 2001. 425 pages. ISBN 978-0262082907.

22. Hartigan, J. A. Asymptotic distribution for clustering criteria, *Annals of statistics*, 1978, vol. 6, s. 117–131.

23. Hartigan, J. A. *Clustering Algorithms (Probability & Mathematical Statistics)*. New York: Willey, 1975. 366 pages. ISBN 978-0471356455.

24. Hartigan, J. A., Wong, M. A. Algorithm AS 136. A k-means clustering algorithm. *Applied Statistics*, 1979, vol. 28, no. 1, s. 100–108.

25. Hastie, T., Tibshirani, R., Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*. Corrected edition. Springer, 2003. 552 pages. ISBN 978-0387952840.

26. Hebák, P. et al. *Vícerozměrné statistické metody 3*. 2. vyd. Praha: Informatorium, 2007. 256 s. ISBN 978-80-7333-001-9.

27. Hebák, P. Výuka statistiky 2007. *Forum Statisticum Slovacum*. Slovenská štatistická a demografická spoločnosť, 2007, vol. 5, s. 41–59, ISSN 1336-7420.

28. Hecht-Nielsen, R. *Neurocomputing.* Reading, MA: Addison-Wesley, 1990.

29. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. Chicago, IL: SPSS, 2000.

30. Chiu, S., Tavella, D. *Data Mining and Market Intelligence for Optimal Marketing Returns*. 1st edition. Oxford: Elsevier, 2008. ISBN 978-0-7506-8234-3.

31. Jain, A. K. *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, vol. 31, no. 8, s. 651–666, 2010.

32. Kohonen, T., Schroeder, M. R., Huang, T. S. *Self-Organizing Maps*. Secaucus, NJ: Springer-Verlag New York, 2001. 426 pages. ISBN 3540679219.

33. Kotler, P. *Marketing Management*. 12. vyd. Praha: Grada, 2007. ISBN 80-247-1359-4.

34. Kukal, J. Úvod do neuronových sítí. *Automa*, 2005, vol. 6, no. 01, s. 20.

35. Kůrková, V. Učení neuronových sítí se schopností generalizace, *Celostátní seminář a workshop Lázně Bohdaneč*, Seminář Analýza dat 2008, vol. 2, s. 9–19. ISBN 978-80-904-053-1-8.

36. Lavine, K. B. Clustering and Classification of Analytical Data. *Encyclopedia of Analytical Chemistry*, Chichester: Wiley Publishing, 2000. ISBN 9780470027318.

37. Lletí, R., Ortiz, M. C., Sarabia, L. A., Sánchez, M. S. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 2004, vol. 515, no. 1, s. 87–100.

38. Lloyd, S. P. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982, vol. IT-28, no. 2, s. 129–137.

39. MacQueen, J. B. *Some Methods for Classfication and Analysis of Multivariate Observations*. Proceedings of the Berkley Symposium on Mathematical Statistics and Probability. Berkley: University of California Press, s. 281–297.

40. Masters, T. *Neural, Novel & Hybrid Algorithms for Time-Series Predictions*. New York: Wiley Publishing, 1995. 514 pages. ISBN 978-0471130413.

41. Meloun, M., Militký, J., Hill, M. *Počítačová analýza vícerozměrných dat v příkladech*. Praha: Academia, 2005. ISBN 80-200-1335-0.

42. Meloun, M., Militký, J. Přednosti analýzy shluků ve vícerozměrné statistické analýze, *Sborník přednášek z konference: Zajištění kvality analytických výsledků*, s. 29–46, Medlov, 22. – 24. 3. 2004, ISBN 80-86380-22-X.

43. Nisbet, R., Elder, J., Miner, G. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009. 864 Pages. ISBN 978-0-12-374765-5.

44. Olej, V., Hájek, P. Modelování bonity obcí pomocí kohonenových samoorganizujících se map a LVQ neuronových sítí. *Scientific Papers of the University of Pardubice*, Series D, 2007, vol. 12, s. 142–149. ISSN 1211 – 555X.

45. Oslon, D., Delen, D. *Advanced Data Mining Techniques*. Berlin: Springer-Verlag, 2008. ISBN 978-3-540-76916-3.

46. Rahman, H. *Data Mining Applications for Empowering Knowledge Societies*. 1st edition. Bangladesh: IGI Global, 2008. 356 pages. ISBN 978-1599046570.

47. Ramos, M., Carvalho, H. Perceptions of quantitative methods in higher education: mapping student profiles. *Higher Education*, 2010, vol. 61, no. 6, s. 629–647. ISSN 0018-1560.

48. Rencher, A. *Methods of Multivariate Analysis*. Second Edition. New York: Wiley Publishing, 2002. ISBN 978-0471418894.

49. Rousseeuw, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, 1987, vol. 20, s. 53–65.

50. Řezanková, H. *Analýza dat z dotazníkových šetření*. Příbram: Professional Publishing, 2007. 217 str. ISBN 978-80-86946-49-8.

51. Řezanková, H., Húsek, D., Snášel, V. *Shluková analýza dat*. Příbram: Professional Publishing, 2007. 220 str. ISBN 978-80-86946-26-9.

52. Schwartz, D. *Encyclopedia of Knowledge Management*. London: Idea Group Publishing, 2006. 902 pages. ISBN 1-59140-574-2.

53. Swingler, K., Cairns, D. Making Decisions with Data: Using Computational Intelligence Within a Business Environment. *Data Mining Applications for Empowering Knowledge Societies*, Bangladesh: Idea Group Publishing, 2008. ISBN 978-1599046570.

54. Šály, M. CRM a segmentace dat: Cesta k rozhodnutí o CRM. *IT Systems*, 2003, vol. 11. ISSN 1802-615.

55. Vapnik, V.N. *Statistical Learning Theory*. 1st edition, AT&T Research Laboratories. New York: Wiley Publishing, 1998. 736 pages. ISBN 0-471-03003-1.

56. Vlach, P. Data mining v malých a středních organizacích; Small Business Solutions. *IT Systems*, 2006. ISSN 1802-615.

57. Weinlichová, J., Fejfar, J. Usage of self-organizing neural networks in evaluation of consumer behaviour. *Acta univ. agric. et silvic. Mendel. Brun*., 2010, vol. 58, no. 6, s. 625–632.

58. Witten, I. H., Frank, E., Hall, M. A. *Data Mining Practical Machine Learning Tools and Techniques*. 3th edition. Morgan Kaufmann, 2011, 664 pages. ISBN 978-0123748560.

**Internetové zdroje:**

59. Katolický, A. *Knowledge Management* [online]. ZCU - FEK - katedra inovací a projektů, 2000, [cit. 27. 11. 2009]. Dostupné z: <http://www.volny.cz/akatolicky >.

60. KD Nuggets [online]. *Data Mining Community's Top Resource*, [cit. 11. 9. 2011]. Dostupné z: < http://www.kdnuggets.com>.

61. IBM SPSS Modeler Help, *IBM SPSS*, [online], 2011, [cit. 11. 9. 2012]. Dostupné z: <help/index.jsp?topic=/com.ibm.spss.modeler.help/clem_intro.htm>.

62. *Iuridicum Remedium, o. s.*, Nevládní nezisková organizace na ochranu lidských práv, 2012, [cit. 2. 10. 2012]. Dostupné z: <http://www.iure.org>.

63. Taboada Jimenez, H. A. *Multi-objective Optimization Algorithms Considering Objective Preferences and Solution Clusters*, [online], [cit. 2. 2. 2013]. ProQuest, 2007. 247 pages. ISBN 9781109045321. Dostupné z: <http://books.google.cz>.

64. Pejčoch, D. Metody řešení problematiky neúplných dat, *Data Quality Tutorial* [online], 2011, [cit. 2. 9. 2011]. Dostupné z: <http://www.dataquality.cz/tutorial/tutorial_04.pdf>.

65. SAS Enterprise Miner 12.1 Reference Help, *SAS Institute Inc.*, 2011, [cit. 5. 12. 2012].

66. SAS Enterprise Intelligence Platform [online]. *SAS Slovakia, s.r.o.*, 2006, [cit. 1. 6. 2011]. Dostupné z: <http://www.sas.com/offices/europe/slovakia/press/newsletters/SNLnovember2006.html>.

67. SAS OnlineDoc®, Version 8 [online]. *SAS Institute Inc.*, 2000, [cit. 8/2011]. Dostupné z: <http://v8doc.sas.com/sashtml/>.

68. SAS Product Documentation [online]. *SAS Institute Inc.*, 2011, [cit. 5. 9. 2011]. Dostupné z: <http://support.sas.com/documentation>.

69. SAS/STAT 9.2 User's Guide, Introduction to Clustering Procedures [online]. *SAS Institute Inc.*, 2008, [cit. 5. 12. 2012]. Dostupné z: <http://support.sas.com/documentation/cdl/en/statugclustering/61759/PDF/default/statugclustering.pdf>.

70. SAS Technical Report A-108. Cubic Clustering Criterion. *SAS Institute Inc.*, Cary, 1983, [cit. 15. 12. 2012].

71. Statistician and Chemist Jack Youden's Graphics. *Department of Statistics, The University of Chicago*, [cit. 9. 10. 2010]. Dostupné z: <http://www.stat.uchicago.edu/events/normal/youden.html>.

# 10 PŘÍLOHY

**Seznam uvedených příloh:**

Příloha č. 1: Hierarchie produktových kategorií

Příloha č. 2: Pracovní prostředí nástroje IBM SPSS Modeler 14.2

Příloha č. 3: Pracovní prostředí nástroje SAS Enterprise Miner 12.1

Příloha č. 4: Procesní diagram modelování v nástroji IBM SPPS Modeler

Příloha č. 5: Proces restrukturalizace a následné agregace vstupních záznamů

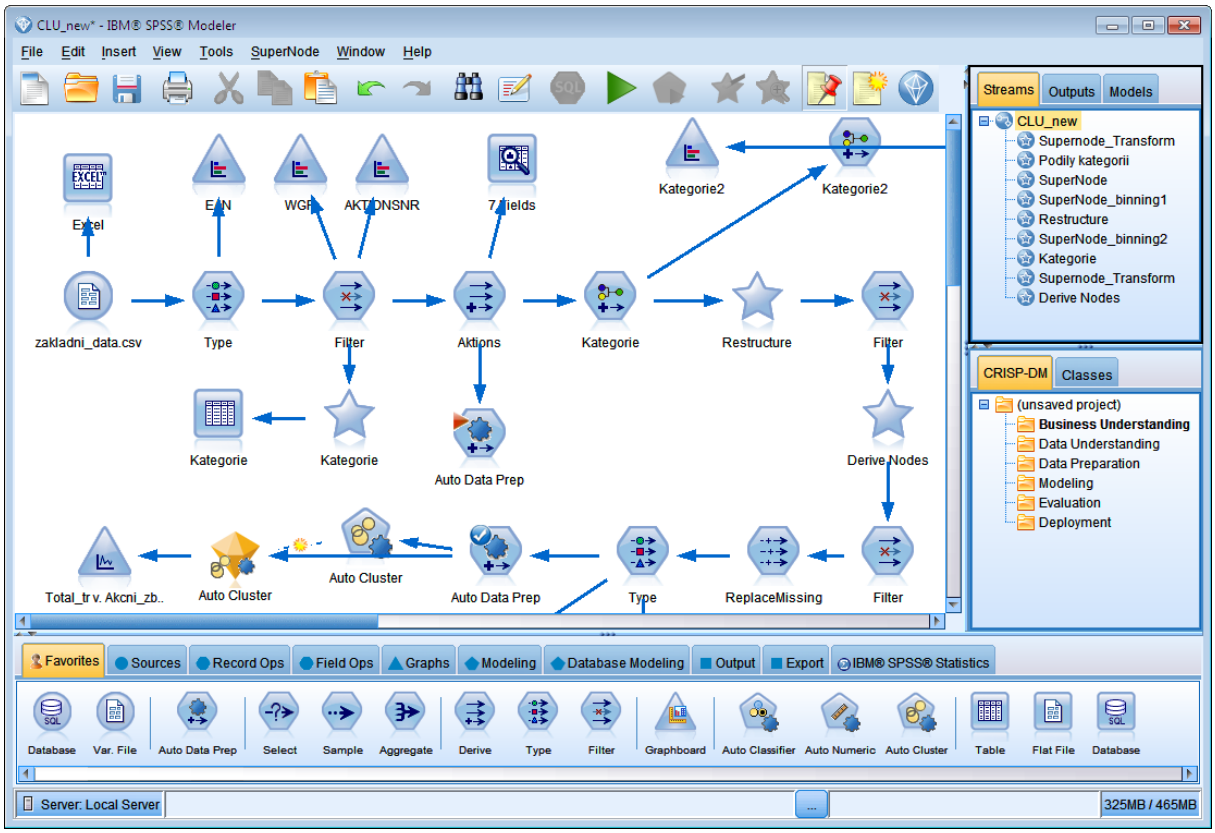
Příloha č. 6: Proces vytvoření nových podílových kategorií

**Příloha č. 1: Hierarchie produktových kategorií**

| KATEGORIE | PODKATEGORIE I | ČÍSLO PODKATEGORIE | PODKATEGORIE II |
|---|---|---|---|
| FOOD I | BALENÉ POTRAVINY | 610 | Konzervy |
| FOOD I | BALENÉ POTRAVINY | 615 | Slad., slan. pečivo, chrup. sortiment |
| FOOD I | BALENÉ POTRAVINY | 620 | Potraviny 14 % DPH |
| FOOD I | BALENÉ POTRAVINY | 630 | Potraviny 20 % DPH |
| FOOD I | ALKO A NEALKO NÁPOJE | 740 | Pivo (nápojové centrum) |
| FOOD I | ALKO A NEALKO NÁPOJE | 621 | Čaj, práškové nápoje |
| FOOD I | ALKO A NEALKO NÁPOJE | 627 | Káva, kakao a čok. v prášku |
| FOOD I | ALKO A NEALKO NÁPOJE | 641 | Lihoviny |
| FOOD I | ALKO A NEALKO NÁPOJE | 642 | Víno a sekt |
| FOOD I | ALKO A NEALKO NÁPOJE | 741 | Nealko – nápojové centrum |
| FOOD I | DROGERIE | 629 | Potrava pro psy |
| FOOD I | DROGERIE | 644 | Hygienické potřeby |
| FOOD I | DROGERIE | 645 | Kosmetika a výr. tel. kosmetiky |
| FOOD I | DROGERIE | 648 | Ostatní výr. |
| FOOD I | DROGERIE | 650 | Čistící a prací prostř. |
| FOOD I | DROGERIE | 689 | Zoo – artikl |
| FOOD I | DROGERIE | 726 | Dětská výživa |
| FOOD II | ČERSTVÉ POTRAVINY | 624 | Balené uzeniny a masné výrobky |
| FOOD II | ČERSTVÉ POTRAVINY | 628 | Chléb, pečivo |
| FOOD II | ČERSTVÉ POTRAVINY | 632 | Mražené výrobky |
| FOOD II | ČERSTVÉ POTRAVINY | 635 | Mléko, mléčné výrobky, tuky |
| FOOD II | ČERSTVÉ POTRAVINY | 746 | Drůbež, zvěřina |
| FOOD II | ČERSTVÉ POTRAVINY | 748 | Sýry |
| FOOD II | ČERSTVÉ POTRAVINY | 756 | Ryby + lahůdky – obslužný úsek |
| FOOD II | OVOCE A ZELENINA | 736 | Ovoce a zelenina |
| FOOD II | PEČIVO | 754 | Chléb – pečivo |
| FOOD III | ŘEZNICTVÍ | 600 | Maso/uzeniny prodej |
| FOOD III | ŘEZNICTVÍ | 602 | Uzeniny samoobslužný úsek |

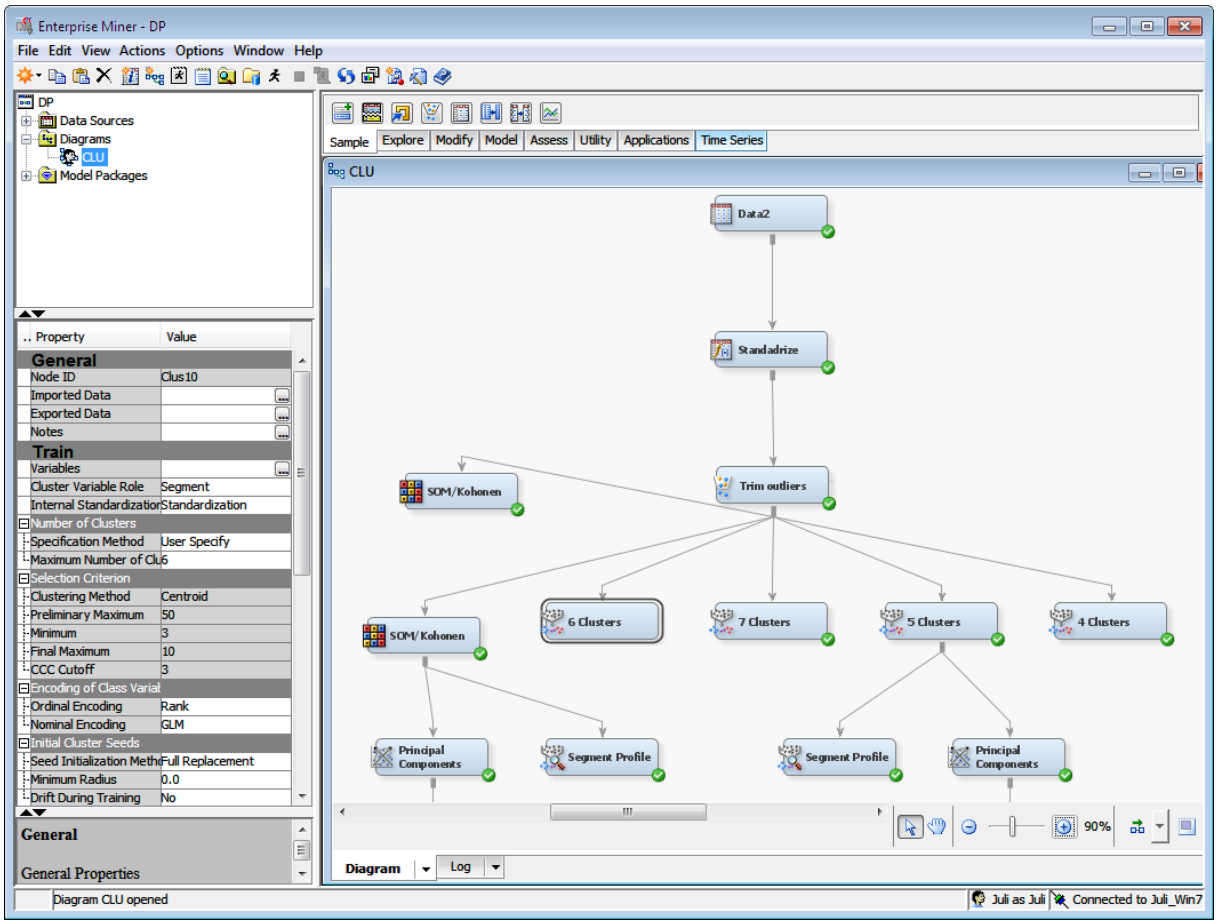
Zdroj: Interní dokument zvoleného hypermarketu, vlastní zpracování

**Příloha č. 2: Pracovní prostředí nástroje IBM SPSS Modeler 14.2**



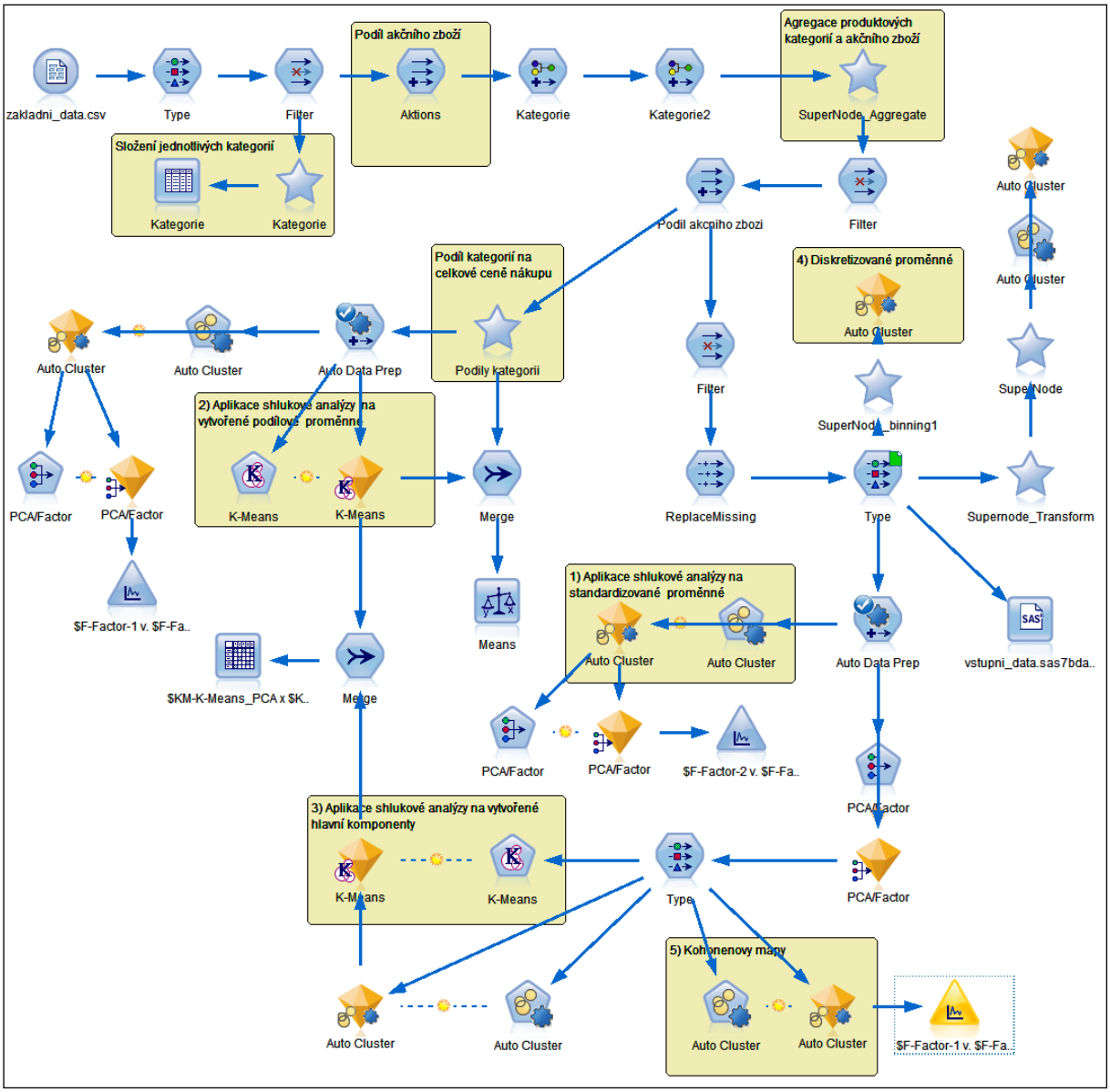

Zdroj: Vlastní zpracování

**Příloha č. 3: Pracovní prostředí nástroje SAS Enterprise Miner 12.1**



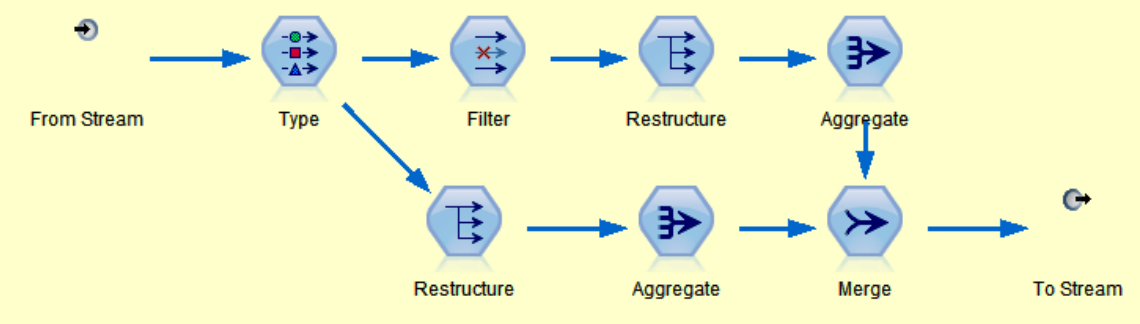

Zdroj: Vlastní zpracování

**Příloha č. 4: Procesní diagram modelování v nástroji IBM SPPS Modeler**



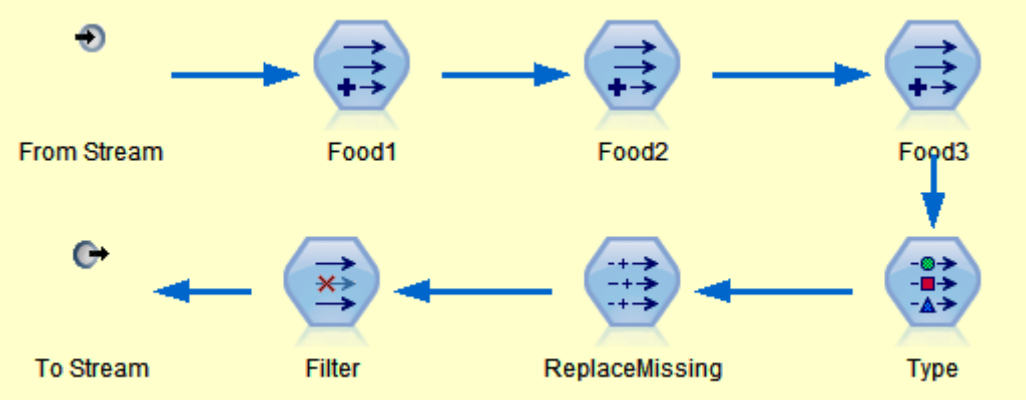

Zdroj: Vlastní zpracování

**Příloha č. 5: Proces restrukturalizace a následné agregace vstupních záznamů**




Zdroj: Vlastní zpracování

**Příloha č. 6: Proces vytvoření nových podílových kategorií**




Zdroj: Vlastní zpracování