TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií

ASOCIAČNÍ ALGORITMY V DATAMININGOVÝCH ÚLOHÁCH

Bakalářská práce

Studijní program: B2646 – Informační technologie
Studijní obor: 1802R007 – Informační technologie

Autor práce: **Milan Kováčik**
Vedoucí práce: Ing. Bc. Marián Lamr

Liberec 2014

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

ASSOCIATION ALGORITHMS IN DATAMINING TASKS

Bachelor thesis

Study programme: B2646 – Information technology
Study branch: 1802R007 – Information technology

Author: **Milan Kováčik**
Supervisor: Ing. Bc. Marián Lamr

Liberec 2014

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií
Akademický rok: **2013/2014**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Milan Kováčik**
Osobní číslo:	**M11000098**
Studijní program:	**B2646 Informační technologie**
Studijní obor:	**Informační technologie**
Název tématu:	**Asociační algoritmy v dataminingových úlohách**
Zadávající katedra:	**Ústav mechatroniky a technické informatiky**

Zásady pro vypracování:

1. Seznamte se s dataminingem a nástrojem IBM SPSS Modeler.
2. Připravte zpracování "nákupního košíku" v IBM SPSS Modeleru a pro výuku připravovaného předmětu Datamining vytvořte modelovou úlohu v tomto prostředí.
3. Zpracujte případovou studii pro modelovou úlohu.
4. Prostudujte asociační algoritmy a téma zpracujte jako e-learningovou podporu pro předmět Datamining.
5. Nastudujte algoritmus Apriori a naprogramujte jej v libovolném programovacím jazyce.

Rozsah grafických prací: **dle potřeby dokumentace**

Rozsah pracovní zprávy: **30–40 stran**

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

Seznam odborné literatury:

[1] **Yong Yin, Ikou Kaku, Jiafu Tang: Data Mining, Springer London Ltd, 2011**

[2] **Steve McConnell: Dokonalý kód, Computer Press, a.s., 2006**

[3] **Kotler Philip.: Marketing management, Grada, 2003**

[4] **Hendl J.: Přehled statistických metod zpracování dat, Portál, s.r.o. 2006**

[5] **Olivia Parr Rud: Datamining, Computer Press, a.s., 2006**

[6] **http://www.msps.cz/data-mining/**

[7] **http://www.spss.cz/pasw_modeler.htm**

[8] **tutorialy k SAP a další materiály na WWW stránkách**

Vedoucí bakalářské práce: **Ing. Marián Lamr**
Ústav mechatroniky a technické informatiky

Datum zadání bakalářské práce: **10. října 2013**

Termín odevzdání bakalářské práce: **16. května 2014**

prof. Ing. Václav Kopecký, CSc.
děkan

L.S.

doc. Ing. Milan Kolář, CSc.
vedoucí ústavu

V Liberci dne 10. října 2013

Prohlášení

Byl jsem seznámen s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Datum:

Podpis:

Abstrakt

Tato bakalářská práce se zabývá asociačními algoritmy v data miningových úlohách. V teoretické části je rozebrána metodologie CRISP-DM, podle které je zpracována případová studie pro modelovou úlohu analýza nákupního košíku. Jako e-learningová podpora pro předmět data mining byla naprogramována aplikace pro generování asociačních pravidel pomocí algoritmu Apriori.

Klíčová slova: Data mining, CRISP-DM, analýza nákupního košíku, asociační pravidla, Apriori

Abstract

This bachelor thesis talks about association algorithms in data mining tasks. There is the analyse of CRISP-DM methodology in the teoretical part, which is base for the case study of the model task: market basket analyse. The application for generating of association rules was programmed with using Apriori algorithm as the e-learning support for the Data mining course.

Keywords: Data mining, CRISP-DM, market basket analyse, association rules, Apriori

Poděkování

Tímto bych rád poděkoval svému vedoucímu práce Ing. Bc. Mariánu Lamrovi za konzultace a užitečné rady při řešení práce. Děkuji své rodině a přátelům za podporu a RNDr. Kláře Císařové, Ph.D. děkuji za konzultace.

Obsah

Úvod

Tématem bakalářské práce jsou asociační algoritmy v data miningových úlohách. V dnešní době, je ukládáno stále více dat, které jen bezcenně leží v databázích, a proto se ve světě informatiky stále více objevuje pojem data mining. Tento pojem se dá přeložit jako dolování dat, nebo vytěžování dat z databází. Pomocí dolování z dat je možné získat potencionálně užitečné informace. Tyto informace je nutné zpracovat a posoudit, jak s nimi naložit. Tato práce se zabývá analýzou nákupního košíku, z něhož lze získat spoustu informací použitelných v marketingu. Jen z jednoduchého průchodu dat lze zjistit strukturu, obsah, úplnost a kvalitu dat. Po podrobnějším zkoumání můžeme zjistit, jací uživatelé se vracejí a co si kupují, to je dobré především z obchodního hlediska, jelikož udržení zákazníka je méně nákladné než zisk nového. Získané znalosti je možné využít na cílenou marketingovou kampaň, což může ušetřit velké množství peněz. Hlavní je však zjistit, jaké zboží si zákazníci nejčastěji kupují v kombinaci s jiným zbožím a pomocí těchto informací zákazníkům nabízet produkty podle toho, co mají aktuálně ve svém nákupním košíku.

Hlavním cílem bakalářské práce je naprogramovat aplikaci na generování asociačních pravidel. Pro generování asociačních pravidel bude použit algoritmus Apriori. Pro aplikaci bude navrhnuto grafické rozhraní s možností volby vstupního souboru. Po načtení dat, si uživatel bude moci zvolit, z kterých atributů (druh zboží) má generovat asociační pravidla. V případě, kdy jsou k dispozici osobní data o zákazníkovi, lze generovat pravidla pouze na určité skupiny lidí. Nakonec uživatel nastaví citlivost algoritmu a zahájí generování pravidel. Výstupem programu je tabulka vygenerovaných asociačních pravidel. Přímo v aplikaci bude možné vyzkoušet si pravidla pomocí simulace nákupu a podle obsahu košíku se cíleně doporučí další zboží. Výsledné implikace bude možné uložit do souboru a dále použít v praxi.

Dílčím cílem je zpracování případové studie pro předmět v navazujícím stu-

diu. Studie zahrnuje celou analýzu všech košíků, zjištění a rozebrání důležitých faktorů a zpracování dat do podoby vhodné pro algoritmus Apriori. Analýza zahrnuje sestavení funkčního modelu a vyzkoušení vygenerovaných pravidel. Pro analýzu a sestavení modelu bude použit data miningový nástroj IBM SPSS Modeler a postupovat se bude pomocí metodologie CRISP-DM popsané v teoretické části.

K dispozici jsou však jen ideální data, vytvořená ke studijním účelům, proto z nich nelze zjistit některé faktory vyplývající z praxe. Získáním dat z reálného obchodu by celá analýza i výsledná pravidla mohla být použita v praxi. Celá analýza a program mohou posloužit dalším provozovatelům jako návod, jak se svojí databází naložit co nejlépe.

1. Data mining

Termín data mining v překladu znamená dolování dat nebo vytěžování dat, někdy chápána jako dobývání znalostí z databáze (Knowledge Discovery in Databases [5]). Data mining je proces vytěžování dat z rozsáhlých databází, k němuž se využívají metody umělé inteligence, strojového učení, statistik a databázových systémů. Obecně jde o vytěžování informací z databází a transformovaní do srozumitelné podoby použitelné k dalšímu použití.

Manuální získávání informací z dat je známé již několik století, mezi první používané metody patří Bayeovská věta (1700) a regresní analýza (1800). S rostoucím vlivem výpočetní techniky se zvýšilo shromažďování a složitost dat. Nebylo již možné ruční zpracování a analyzování dat, a proto začal vzestup automatického zpracování dat, který byl podporován dalšími objevy v informatice, jako jsou shlukové analýzy, neuronové sítě, rozhodovací stromy a genetické algoritmy. Data mining vlastně využívá tyto metody k zisku skrytých vzorců v rozsáhlých datech. [1]

1.1 Úlohy v data miningu

Dolování dat lze použít na velké množství nejrůznějších problémů. Jednotlivé problémy lze zařadit do kategorií, avšak rozdělení do kategorií není jasně stanovené. Zde je několik základních úloh řešených v data miningu.

- Predikce – na základě statických technik jsou předpovídány následující hodnoty z předešlých hodnot.
- Deskripce – nalezení skryté struktury nebo vazeb, které jsou použity k následnému vyhodnocování.
- Klasifikace – rozdělení objektů do tříd na základě společných charakteristických rysů. Třídy jsou dány předem, a každý objekt do nich lze jednoznačně zařadit.

- Regrese – na základě předešlých zkušeností předpovídá následující hodnotu. Jedná se o statickou metodu popisující vztahy mezi vstupními a výstupními hodnotami.
- Segmentace – jedná se o nejstarší techniku používanou v data miningu, založenou na rozdělení objektů do skupin (shluků), které jsou vytvářeny v analýze dat. Objekty jsou zařazeny do shluků podle podobnosti charakteristických rysů.
- Sumarizace – sumarizaci je vhodné použít u velkého počtu dat, k zjištění struktury těchto dat. Jsou zde použity základní aritmetické operace.

1.2 Metodologie

V současnosti se data miningem zabývá stále více firem, které chtějí výsledky rychle, levně a efektivně, v důsledku toho přicházejí na scénu metodologie. Standardizace postupů je jedním ze způsobů, jak šetřit prostředky a čas. Bylo vytvořeno několik metodologií popisující efektivní postup zpracování projektu, mezi nejznámější a nejpoužívanější patří metodologie SEMMA a CRISP-DM.

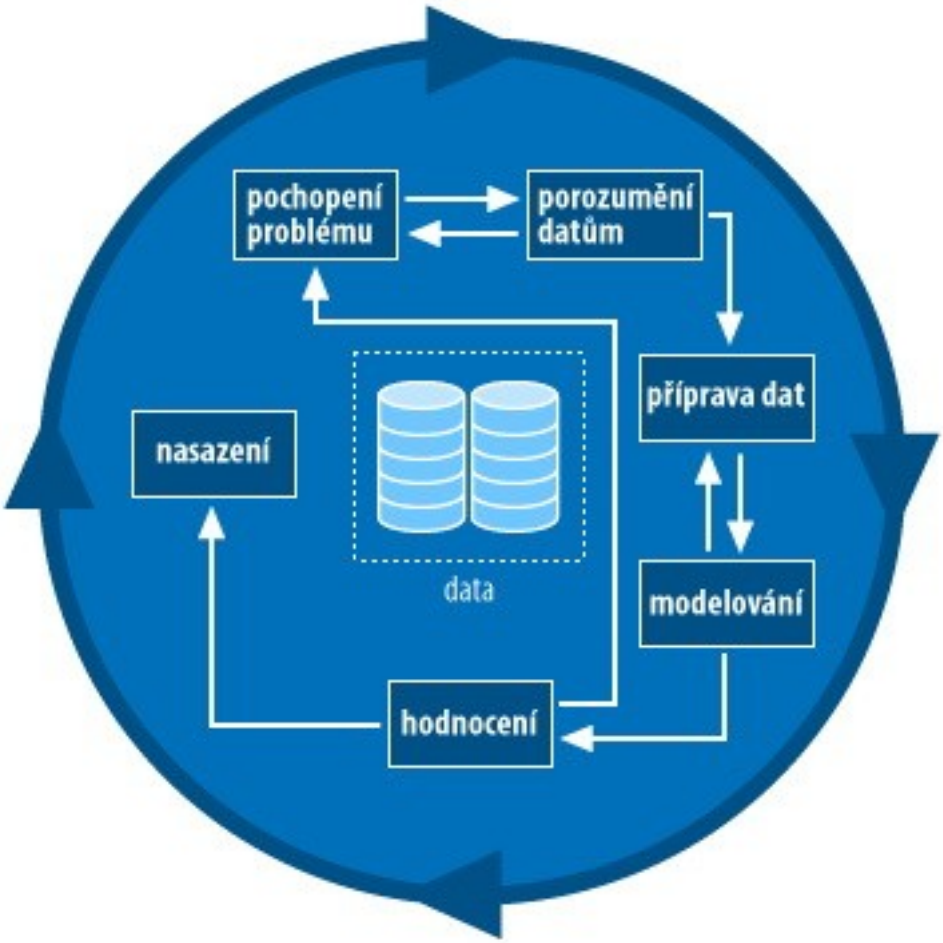
1.2.1 SEMMA

Jméno této metodologie je složením prvních písmen z jednotlivých fází vytěžování dat. SEMMA byla vyvinuta společností SAS Institute a je považována za obecnou metodiku dolování dat. Společnost SAS Institute tvrdí, že SEMMA je spíše logická organizace funkční sady nástrojů pro produkt SAS Enterprise Miner, a proto její používání mimo tento produkt může být dvojznačné. Je zaměřena především na modelování úloh a oproti CRISP-DM nezahrnuje obchodní stránku projektu. [6]

- Sample – výběr dat dostatečně velikých a zároveň dostatečně malých, aby byla data efektivně využita.
- Explore – porozumění datům, objevování souvislostí pomocí vizualizace.
- Modify – příprava dat pro modelování.
- Model – modelování na připravených datech k dosažení výsledku.
- Assess – zhodnocení výsledků.

1.2.2 CRISP-DM

Metodologie CRISP-DM (CRoss – Industry Standard Process for Data Mining) byla vyvinuta jako projekt Evropské komise standardizující postup vytváření data miningových projektů. CRISP-DM nabízí návod krok po kroku pro každou část projektu. Model pomáhá zpracovávat projekty rychleji, efektivněji, s nižšími náklady a bez běžných chyb. Metodologie je popsána v šesti krocích, to však neznamená, že musíme jít od prvního kroku k poslednímu, v rámci celého projektu se můžeme vracet k minulým krokům a měnit je tak, aby bylo dosaženo požadovaného cíle. V praxi je běžné vracet se i několikrát do stejného bodu. Na obrázku 1.1 je zobrazen průběh cyklu CRISP-DM. [4]



Obrázek 1.1: Průběh metodologie CRISP-DM [7]

- **Business understanding - (porozumnění problematice)**
  V první fázi je nutné pochopit, čeho chce zákazník dosáhnout z obchodního hlediska. Zákazníci mají občas protichůdné cíle a omezení, kterým musí analytik porozumět a navrhnout vyváženou cestu. Dalším úkolem analytika je odhalení důležitých faktorů, jež by mohly ovlivnit výsledek projektu. Nejdůležitější je stanovit správný cíl a to především z obchodního hlediska. Zde přichází i první plán projektu, jak dolovat data za správným účelem projektu, stanovení základních postupů a výběr nástrojů a technik. Je důležité stanovit kritéria pro úspěch z podnikatelského hlediska. Dalším bodem je podrobnější zjištění o zdrojích dat, vytvořit seznam dostupných zdrojů pro projekt, a to i lidských zdrojů a software. Zkoumaná vstupní data musí být zhodnocena, zda by neměla být doplněna nebo modifikována. Na základě známých faktů je nutné zhodnotit míru rizika, dostupnost zdrojů a výši nákladů. Na rozsáhlejších projektech, na kterých spolupracuje velký tým je dobré sestavit slovník termínů.

- **Data understanding - (porozumnění datům)**
  Zisk dat, nebo přístupu k datům z projektových zdrojů, je nutné tato data pochopit, k tomu je k dispozici velká řada nástrojů. Pomocí těchto nástrojů jsou data charakterizována a jsou popsány jejich vlastnosti, včetně formátu dat, množství dat, popis polí každé tabulky a další objevené vlastnosti, pomocí nichž je vyhodnoceno, zda data splňují požadavky. Pro analýzu se zde často využívá jednoduchých funkcí, jako nalezení minima a maxima, průměrné hodnoty, nebo četnosti jednotlivých hodnot. Poté jsou zkoumány data důkladněji a jsou vytipovány související množiny a jejich podmnožiny. Zhodnotí se první hypotézy a jejich vliv na výsledek projektu. Důležité je zhodnocení kvality dat, zda jsou data kompletní a neobsahují chyby. Pokud obsahují chyby, je nutné vědět, o jaké chyby se jedná, jak jsou časté a jestli mohou ovlivnit výsledek.

- **Data preparation - (příprava dat)**
  Na začátku je důležité zhodnotit technické omezení, jako objem dat nebo datových typů. Hlavní je selekce potřebných atributů (sloupců) a výběr záznamů (řádků). Musí být rozhodnuto jaké atributy a záznamy budou vybrány nebo vyloučeny. Pomocí vyloučení některých záznamů je zvýšena kvalita dat. Tím jsou vybrána pouze kompletní data, je možné chybějící data doplnit pomocí technik, odhadujících chybějící údaje. Při konstrukci dat mohou vznikat

nové atributy nebo generované záznamy, potřebné pro modelování. Často se slučuje více zdrojů do jednoho společného a vytváří se zcela nové tabulky.

- **Modeling - (modelování)**
  V prvním kroku byl vybrán nástroj pro modelování, ale v tomto kroku je nutné vybrat konkrétní modelovací techniku, která bude použita. Mezi používané techniky patří například rozhodovací stromy, neuronové sítě nebo asociační pravidla. Je možné použít více modelovacích technik a porovnávat výsledky, což zvýší pravděpodobnost správného výsledku. Některé modelovací techniky mají specifické požadavky na data, které musí být splněny. Před samotným modelováním je dobré sestavit mechanismus na testování kvality modelu. Následuje sestavení modelu a nastavení parametrů a citlivost jednotlivých modelovacích technik pro potřebný výsledek. Na závěr je posuzována přesnost a kvalita modelu.

- **Evaluation - (zhodnocení)**
  V minulém kroku byl model posuzován z hlediska přesnosti a obecnosti. V tomto kroku je však posuzován z obchodního hlediska, zda model splňuje cíle projektu, je hodnoceno jestli je model použitelný či nikoliv. K hodnocení kvality modelu slouží dvě množiny, první množina vstupních dat, na kterých se model naučí generovat pravidla, druhá množina slouží k otestování pravidel. Pomocí testovacích dat lze určit procentuální úspěšnost modelu. Výsledky mohou ukázat další možnosti směřování obchodní taktiky, odhalit nové výzvy a informace. V okamžiku, kdy se zdá, že jsou výsledky uspokojivé, je vhodné udělat přezkoumání a podívat se do minulých kroků, jestli nedošlo k přehlédnutí chyby nebo některého z významných faktorů. Na závěr je sestaven seznam kroků dalších možných akcí, jak model zlepšit nebo modifikovat. Nakonec je rozhodnuto zda se modul nasadí do praxe nebo jestli bude poslán zpět k přepracování některého z kroků.

- **Deployment - (uvedení do praxe)**
  Naplánování strategie pro nasazení do praxe popsána krok po kroku. Po nasazení do praxe je nutné monitorovat a udržovat. Správná strategie údržby pomáhá vyhnout se nesprávnému používání. Po zavedení do praxe je nutné zhodnotit zda se nevyskytli nějaké chyby a zjistit co se mělo stát ale nestalo.

1.3 Cross-selling

Bez Cross-sellingu (křížový prodej) se v dnešní době neobejde žádný větší internetový obchod, právě tato metoda nejvíce souvisí s prodejem zboží na internetu. Cross-selling je marketingová metoda zajišťující zvýšení tržeb, pomocí cílené nabídky doplňků k zakoupenému zboží. K získání modelu této metody je potřeba znát, co si zákazníci koupili dříve v kombinaci s jiným zbožím. To lze zjistit z databází, které si každý internetový obchod ukládá. Z těchto dat jsou doslova vydolována pravidla určující nabídku k vybranému zboží. Pro tento druh dolování se nejčastěji používá algoritmus Apriori.

1.4 Software

Pro sestavování modelů je důležité mít software, který zná potřebné modely ke generování pravidel a zároveň umí pracovat s různými formáty vstupních dat. Výběr software závisí na prostředcích, můžeme si vybrat mezi open source a komerčním software. Pro tuto práci byl vybrán komerční nástroj IBM SPSS Modeler.

1.4.1 IBM SPSS Modeler

Původně se software jmenoval Clementine a byl vyvinut společností Integral Solutions Limited (ISL) ve Velké Británii. První verze vyšla roku 1994 pod označením Clementine 1.0. Tento nástroj se rychle stal oblíbeným v oblasti data miningu, hlavně díky použití ikon v grafickém prostředí, což uživatele osvobozovalo od ručního psaní kódu v programovacím jazyce. Ovšem první verze Modeleru vydaná společností IBM je až verze 14.2 z roku 2011.

IBM SPSS Modeler je data miningový nástroj určený na analýzu textu a dolování dat z databází. Vytváří prediktivní modely a provádí různé analytické úlohy pomocí grafického rozhraní, díky kterému může uživatel používat data miningové algoritmy. [3]

2. Modelování v data miningu pomocí asociačních pravidel

2.1 Asociační pravidla

V běžném jazyce se hojně využívá posuzování buď a nebo, tato syntaxe je základem aciační pravidla. Jelikož se jedná o jedno z nejstarších a nejjednodušších vyhodnocování, patří mezi nejpoužívanější prostředky pro reprezentaci znalostí. Asociační pravidla jsou spjata především s analýzou nákupního košíku. V analýze jde o hledání společných vztahů mezi jednotlivými atributy, přítomnost jedné položky implikuje jednu nebo více položek v jedné transakci.

U nalezených pravidel z dat je důležité najít vztahy mezi předpokladem a závěrem.

$$Ant \Rightarrow Con \tag{2.1}$$

Kde *Ant* (antecedent, předpoklad) implikuje *Con* (consequent, závěr). Kombinace kategorií pro n košíků znázorňuje kontingenční tabulka (Tabulka 2.1).

	Con	**-Con**	$\sum$
Ant	a	b	r
-Ant	c	d	s
$\sum$	k	l	n

Tabulka 2.1: Kontingenční tabulka

- $\mathbf{a = n(Ant \wedge Con)}$
  Počet případů kdy je splněn předpoklad a zároveň závěr.
- $\mathbf{b = n(Ant \wedge \text{-}Con)}$
  Počet případů kdy je splněn předpoklad a závěr není splněn.

- **c = n(-Ant $\wedge$ Con)**
  Počet případů kdy předpoklad není splněn a závěr je splněn.

- **d = n(-Ant $\wedge$ -Con)**
  Počet případů kdy není splněn předpoklad ani závěr.

Z těchto četností lze vypočítat charakteristiky vypovídající o kvalitě nalezeného pravidla. Základními charakteristikami jsou podpora (support) a spolehlivost (confidence). Podpora je v kolika procentech případů byl splněn předpoklad i závěr.

$$Support = \frac{a}{n} \tag{2.2}$$

Spolehlivost je pravděpodobnost splnění závěru, pokud je splněn předpoklad.

$$Confidence = \frac{a}{r} \tag{2.3}$$

2.2 Princip algoritmu Apriori

Algoritmus Apriori slouží k vyhledávání frekventovaných množin a k následnému generování asociačních pravidel. Snahou algoritmu je nalézt vazby mezi jednotlivými atributy v databázi, takové že přítomnost jednoho nebo více atributů implikuje přítomnost jiných atributů v jedné transakci. Hlavní snahou je získat co nejsilnější asociační pravidla. Pomocí následujících metrik jsou vybrána nejsilnější pravidla. [2]

- **Podpora (support)**
  Minimální práh četnosti množiny položek v celé databázi, vyjádřené v procentech. Pokud množina položek splňuje minimální podporu, je pro algoritmus zajímavá a bude s ní dále pracovat.

$$podpora = \frac{kosiky\ obsahujici\ mnozinu\ prvku}{vsechny\ kosiky} * 100\% \tag{2.4}$$

- **Spolehlivost (confidence)**
  Jak moc se lze spolehnout na výsledné pravidla. Spolehlivost je počítána pro každý prvek ve frekventované množině a jsou vybrána jen ta nejsilnější pravidla, tedy jen pravidla splňující zadanou spolehlivost. Každý prvek frekventované množiny je porovnáván se všemi jeho podmnožinami.

$$podpora = \frac{pocet\ vyskytu\ mnoziny}{pocet\ vyskytu\ podmnoziny} * 100\% \tag{2.5}$$

2.2.1 Generování frekventovaných množin

Generování začíná průchodem databáze a zjištěním všech dostupných atributů. Z těch je sestavena první jednopoložková množina kandidátů, která obsahuje všechny atributy.

Id objednávky	**Seznam zboží**
1	I1, I2, I4
2	I1, I5
3	I2, I4
4	I1, I2, I4, I5
5	I4, I5
6	I1, I2, I4
7	I2, I4, I5
8	I1, I2, I3, I4, I5

Tabulka 2.2: Databáze objednávek

V tabulce 2.2 je vidět 8 nákupních košíků, kde každý má svůj seznam zboží. Právě zboží budeme potřebovat ke generování frekventovaných množin. Předtím je však nutné si stanovit vstupní podmínky pro generování těchto množin. Pokud zvolíme minimální podporu 25%, je vypočítán minimální počet košíků, které musí obsahovat množinu zboží.

$$minimalni\ support = \frac{pocet\ vsech\ nakupu * zadany\ support}{100} = \frac{8 * 25}{100} = 2 \quad (2.6)$$

Pokud je znám minimální support, může být sestavena první množina kandidátů C1, která bude obsahovat všechny druhy zboží z nákupů. Následně bude sestavena frekventovaná množina L1 z kandidátů, kteří splňují minimální support. Pro zjištění četnosti výskytu je nutné v každém kroku projít celou databázi.

Zboží	Počet výskytů
I1	5
I2	6
I3	1
I4	7
I5	5

Tabulka 2.3: Množina kandidátů C1

Z množiny kandidátů C1 vybereme pouze prvky splňující minimální support, tak dostaneme frekventovanou množinu L1.

Množina	Počet výskytů
I1	5
I2	6
I4	7
I5	5

Tabulka 2.4: Frekventovaná množina L1

Z vygenerované frekventované množiny z tabulky 2.3 sestavíme novou množinu kandidátů C2 spojením množiny L1 s množinou L1. V tomto kroku je využita vlastnost algoritmu Apriori, ten kontroluje jestli každá podmnožina z množiny kandidátů C2 je frekventovanou množinou. Pokud některá z podmnožin není frekventovanou množinou je množina vyloučena z množiny kandidátů.

Množina	Počet výskytů
I1, I2	4
I1, I4	4
I1, I5	3
I2, I4	5
I2, I5	3
I4, I5	3

Tabulka 2.5: Množina kandidátů C2

Jelikož všechny množiny v množině kandidátů splňují minimální podporu je výsledná frekventovaná množina L2 rovna množině kandidátů C2. Následovalo

by další spojení množiny L2 s množinou L2. Takto by algoritmus pokračoval dokud by byly nalézány frekventované množiny.

2.2.2 Generování asociačních pravidel

Nalezení asociačních pravidel se provádí pomocí využití silných množin, odstraněna jsou pouze pravidla, jejichž confidence nesplňuje minimální confidenci. Minimální confidence se volí již na začátku, pro tento příklad je určena minimální confidence 75%. Například z frekventované množiny L2 je vybrána množina {I2, I4}, z níž jsou generována pravidla.

- **I2** $\rightarrow$ **I4**

$$confidence = \frac{support\{I2, I4\}}{support\{I2\}} * 100 = \frac{5}{6} * 100 = 83,3\% \tag{2.7}$$

- **I4** $\rightarrow$ **I2**

$$confidence = \frac{support\{I2, I4\}}{support\{I4\}} * 100 = \frac{5}{7} * 100 = 71,4\% \tag{2.8}$$

Odstraněno je druhé pravidlo, které nesplňuje minimální confidenci. Takto jsou generována všechna pravidla ze všech nalezených frekventovaných množin. Pro algoritmus je nalezení silných pravidel nenáročné oproti hledání frekventovaných množin.

3. Analýza nákupních košíků

Cílem této práce je zpracování úlohy analýza nákupního košíku pro předmět data mining. Pro tento účel byl vybrán nástroj IBM SPSS Modeler určen k realizaci celých projektů v oblasti data miningu. V tomto projektu bude použit především k analýze dat a následnému zpracování dat. Analytická část má za úkol zjistit vše potřebné o datech, jako jsou chybějící hodnoty, nalezení extrémních hodnot. Pomocí údajů z analýzy se budou data moci zpracovat do výsledné podoby, potřebné pro zpracování algoritmem Apriori. Dále bude z analýzy vyplývat, jak často se zákazníci vracejí, z čehož lze vyvodit závěry, jak k takovýmto zákazníkům přistupovat.

Když jsou známa všechna fakta o datech, je potřeba data transformovat na základě známých informací. Provedená transformace nemusí být konečná, jestliže další kroky ukáží, že jsou data nedostatečná nebo naopak obsahují více informací než je potřeba. Konečná transformovaná data budou uložena do nové datového souboru, který bude následně zpracován. Zpracování dat proběhne pomocí algoritmu Apriori, obsaženým v IBM SPSS Modeleru. V tomto kroku je důležité nastavení správné citlivosti algoritmu, jinak by výsledek mohl být znehodnocen. Získané implikace budou aplikovány a na základě výsledků budou zákazníkovi nabízeny další druhy zboží.

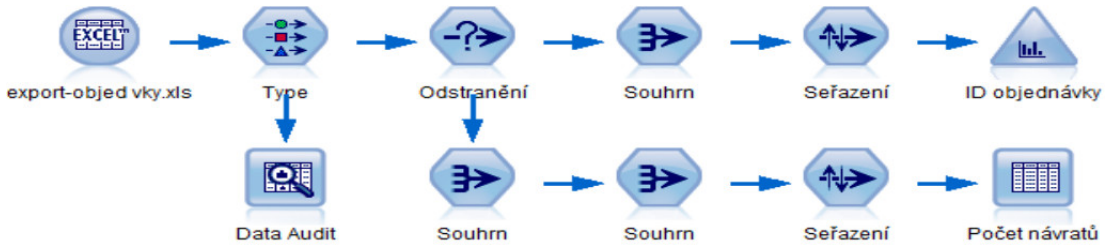
Pro analýzu nákupního košíku je rozhodnuto, pokud je potřeba efektivně prodávat související produkty. Pomocí analýzy se dozvíme, co si zákazníci nejčastěji kupují v kombinaci s jinými produkty. Toto zjištění je pro majitele internetového obchodu velmi zajímavé, jelikož může těchto znalostí využít k cílené nabídce produktu, které zákazník zatím nekoupil. Zákazníkovi, který má v košíku notebook, tak bude nabídnuto to, co si lidé nejčastěji kupují právě s notebookem, jako je taška na notebook, myš nebo chladící podložka. Cílem prodávajícího je tedy nabídnout zákazníkovi co nejzajímavější zboží tak, aby si ho zákazník koupil.

3.1 Analýza datového souboru

K dispozici byla pouze ideální data, ze kterých nelze získat informace z praxe. Tato data jsou vytvořena pouze k simulaci a jsou nastavena tak, aby z nich bylo možné vydolovat informace. Proto byla sehnána reálná data z internetového obchodu, která budou zkoumána.

Před začátkem generování pravidel je nutné zjistit podobu dat. Jelikož vstupní soubor pro tento projekt je ve formátu xsl, může probíhat analýza v excelu, vzhledem k množství dat, by analýza byla časově extrémně náročné. Ovšem v dnešní době kdy je data mining hojně využíván existuje spousta nástrojů určených právě pro dolování dat. Pro tento projekt byl zvolen IBM SPSS Modeler, který slouží jak k analýze dat tak i k modelování celých projektů. Pomocí IBM SPSS Modeleru budou data prozkoumána a na základě zkoumání modifikována do potřebné podoby.

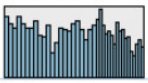
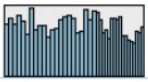
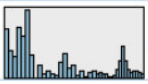
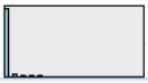
Na obrázku 3.1 je proud použitý pro analýzu dat. Vstupem do proudu je soubor export-objednavky.xsl, což jsou data určená pro tento projekt v nezměněné podobě z internetového obchodu. V uzlu Type se pouze načítají hodnoty jednotlivých atributů, popřípadě se zde mohou měnit jejich datové typy. Tento proud umožňuje prozkoumat data pomocí několika uzlů, které odhalí některá fakta o datech.



Obrázek 3.1: Proud pro analýzu dat

3.1.1 Data audit

Uzel Data Audit je často používán k prvotnímu zkoumání vstupních dat, poskytuje komplexní pohled na data. Zobrazuje souhrnné statistiky, histogram a distribuční grafy pro každé pole. Uzel má dvě karty použitelné pro zkoumání dat první z nich karta Audit zobrazuje již zmíněné statistiky a grafy. Na kartě quality je posuzována kvalita dat, zobrazuje informace o extrémních, odlehlých a chybějící hodnotách. Poskytuje také nástroj pro zpracování těchto dat.

Field	Sample Graph	Measurement	Min	Max	Mean	Unique	Valid
ID zákazníka		Continuous	7.000	14558.000	6723.961	--	49428
ID objednávky		Continuous	7.000	17047.000	8221.727	--	49428
ID produktu		Continuous	47.000	478238.000	150676.761	--	49428
ID kategorie produktu		Nominal	--	--	--	206	28700
Název kategorie produktu		Nominal	--	--	--	203	28700
Množství produktu		Continuous	1.000	20.000	1.028	--	49428

Obrázek 3.2: Zobrazení karty quality z uzlu Data Audit

Z auditu dat zobrazeném na obrázku 3.2 je patrná struktura souboru, který obsahuje šest atributů (sloupců) a 49 428 záznamů (řádků). Po průchodu souboru bylo zjištěno, že každý záznam obsahuje jeden produkt v košíku, takže více záznamů se společným ID objednávky tvoří jeden nákupní košík. Proto bude nutné v přípravě dat přetransformovat data do podoby, kde jeden záznam bude obsahovat jeden nákupní košík. Ze sloupce Unique je patrný počet kategorií obsažených v souboru. Ovšem 203 kategorií je pro sestavování modelu příliš mnoho. Bylo vyzkoušeno, že s takto velkým množstvím kategorií není možné najít společné množiny.Z průchodu internetových stránek je patrné, že se jedná o podkategorie. Proto je nutné tyto podkategorie sloučit do příslušných kategorií.

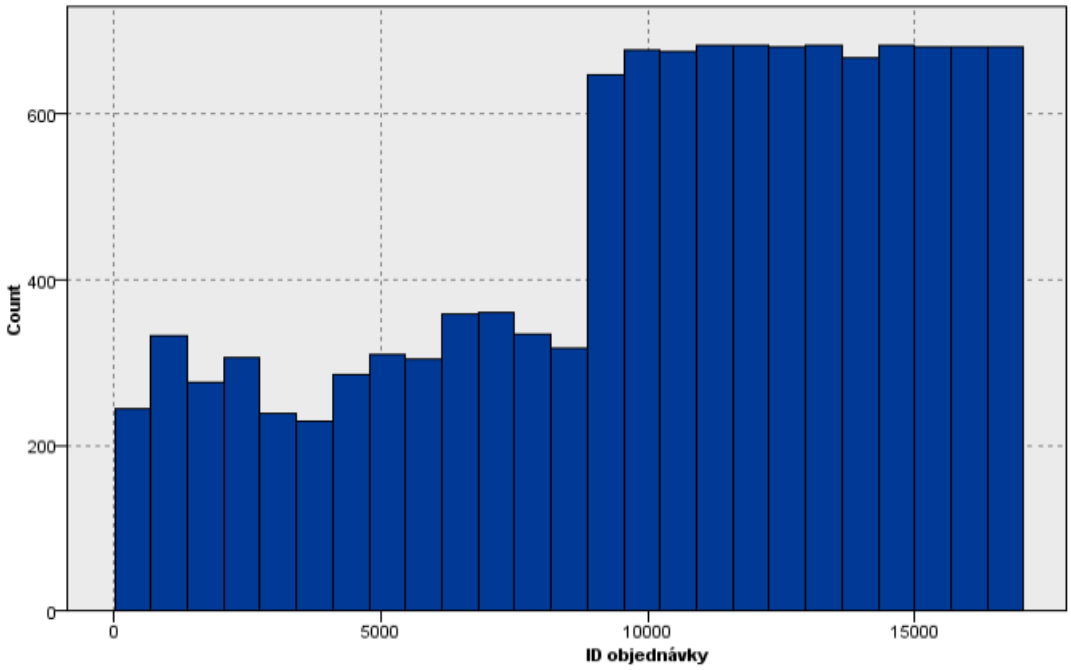
Field	% Complete	Valid Records	White Space	Extremes
ID zákazníka	100	49428	0	0
ID objednávky	100	49428	0	0
ID produktu	100	49428	0	0
ID kategorie produktu	58,06	28700	20728	--
Název kategorie produktu	58,06	28700	20728	--
Množství produktu	100	49428	0	178

Obrázek 3.3: Zobrazení karty quality z uzlu Data Audit

Z karty quality auditu dat bylo zjištěno, že 58% dat není kompletní což je vidět na Obrázku 3.3. V ID kategorie a Název kategorie produktu je spousta chybějících hodnot, což může být problém, jelikož právě atribut Název kategorie produktu bude pro projekt nejdůležitější a je nutné získat pouze data obsahující tento atribut.

3.1.2 Histogram ID objednávky

Do histogramu vstupují již kompletní data, v uzlu Odstranění došlo k selekci dat bez chybějících prvků. U vybraných dat bylo spočítáno, kolik záznamů obsahuje jeden nákupní košík, jednotlivé košíky byly setříděny podle ID objednávky tak, aby bylo patrné, jaký byl vývoj počtu nákupů na košík v čase.



Obrázek 3.4: Počet nákupů na ID objednávky

Z histogramu na obrázku 3.4 je patrný nárůst okolo hodnoty ID objednávky 9 000, takže na jeden košík připadá v průměru až dvakrát více zboží. Tento nárůst lze vysvětlit velkou obchodní kampaní, která nalákala zákazníky ke koupi zboží z obchodu. Po průchodu dat je však patrné, že důvod je zcela jiný. Do ID objednávky 8 914, nejsou kompletní data u názvů kategorií produktu.

3.1.3 Počet návratů zákazníka

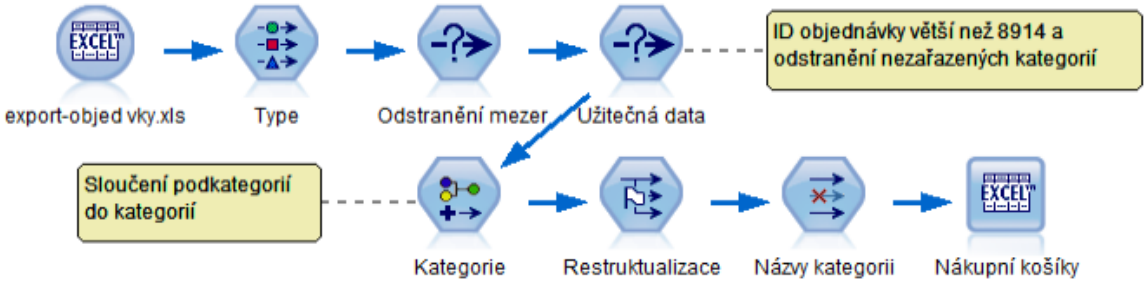
Z obchodního hlediska je dobré vědět jací zákazníci se vrací a nakupují pravidelně, o takové zákazníky je potřeba se starat, aby neutekli ke konkurenci. Věrným zákazníkům jsou nabízeny výhody ve formě bonusů, slev nebo dárkových poukazů. Proč se ale starat o několik zákazníků, kteří se vrací? Odpověď je jednoduchá, protože je známo, že náklady na udržení stávajících uživatelů jsou několikanásobně menší, než na zisk nových zákazníků. Nejde však jen o peněžní stránku, ale i o čas strávený nad reklamní kampaní a administrativou, která však nakonec nemusí mít žádnou odezvu. Z Obrázku 3.5 je patrné, kteří zákazníci se nejčastěji vracejí do tohoto obchodu.

ID zákazníka	Počet návratů
2878.000	16
2738.000	12
5563.000	12
6529.000	11
8212.000	11
11053.000	10
6267.000	10
6559.000	10
12601.000	9
4108.000	9
6754.000	8
13420.000	8
5539.000	8
4153.000	8
5540.000	8

Obrázek 3.5: Počet návratů zákazníků

3.2 Příprava dat

Z analýzy dat je zřejmé, že data musí být restrukturalizována a některé záznamy musí být odstraněny. Průběh přípravy dat je zobrazen na obrázku 3.6.



Obrázek 3.6: Proud pro přípravu dat

Vstupem do proudu jsou původní data, která pomocí několika uzlů budou změněna do potřebné podoby. Jednotlivé uzly jsou popsány na další stránce.

- **Odstranění mezer a užitečná data**
  V uzlu Odstranění mezer jsou odstraněna prázdná místa z atributu název kategorie produktu. V následujícím uzlu Užitečná data jsou vybrána pouze data s ID objednávky vyšší než 8 914, protože právě do této objednávky je častý výskyt prázdných míst. U vyšších ID objednávky je počet výskytů prázdných míst zanedbatelný a nemá velký vliv na výsledná pravidla.

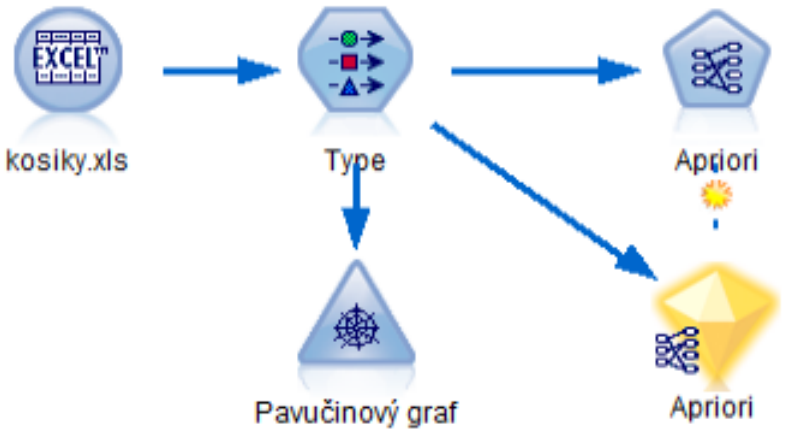
- **Kategorie**
  Zde se provádí sloučení podkategorií do kategorií podle internetového obchodu. Jelikož se nejčastěji kupují trička a mikiny, bylo rozhodnuto, že tyto dvě kategorie se rozdělí do podkategorií a to pánská, dámská a dětská. Výsledných implikací je s takto rozdělenými kategoriemi více, než v případě kdy jsou pouze kategorie trička a mikiny.

- **Restrukturalizace**
  Konečné sloučení jednotlivých řádků do jednotlivých košíků podle ID objednávky. Výstupem z tohoto uzlu jsou data, kde v prvním sloupci je ID objednávky a v dalších kategorie. Pro každý košík jsou v jednotlivých sloupcích hodnoty T nebo F, podle toho, zda v košíku byl produkt z dané kategorie (T) či nikoliv (F).

- **Název kategorií a nákupní košíky**
  Pouze kosmetická úprava jednotlivých atributů, jelikož uzel restrukturalizace automaticky přidá prefix každé kategorii. Nakonec jsou data exportována do souboru v uzlu nákupní košíky.

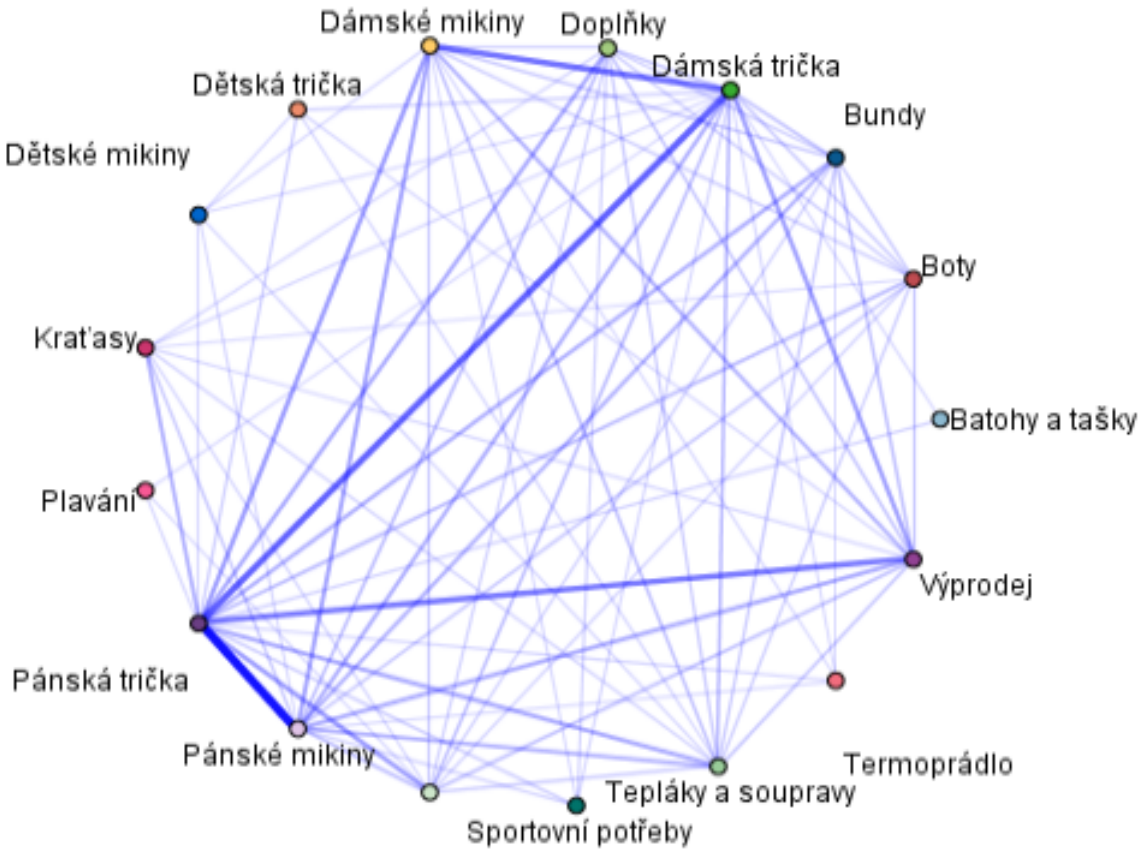
3.3 Modelování

V této fázi projektu dochází k získání původních požadavků, v tomto případě k vygenerování asociačních pravidel pomocí algoritmu Apriori. Je nutné nastavit správné vstupní podmínky algoritmu, abychom nalezli implikace. Optimálními vstupními podmínkami jsou confidence 30% a minimální support 3%. Je zřejmé, že vstupní podmínky jsou nízké, to je zapříčiněno především velkou spoustou jednopoložkových nákupů.



Obrázek 3.7: Proud pro generování asociačních pravidel

Před samotným generováním pravidel pomocí Apriori se podíváme na vzájemné vztahy všech druhů zboží. Pavučinový graf z obrázku 3.8 zobrazuje četnost společných výskytů jednotlivých položek v jednom nákupním košíku, síla čáry určuje četnost společných výskytů. Nejčastěji zákazníci kupují pánská trička s pánskými mikinami, tento vztah by měl být zřejmý i ve vygenerovaných asociačních pravidlech.



Obrázek 3.8: Pavučinový graf vztahů mezi položkami

Datový tok vstupující do uzlu Apriori vygeneruje krystal, v němž jsou zobrazeny nalezené implikace. Vygenerované implikace podle vstupních podmínek jsou zobrazeny na obrázku 3.9. Implikace jsou znázorněny dle následujícího vztahu.

$$X \Leftarrow A \ \& \ B \ \& \ C \ \& \ ... \tag{3.1}$$

X znázorňuje sloupec consequent (závěr), předpoklady (sloupec antecedent) jsou znázorněny pomocí hodnot A, B, C, atd. Aby bylo možné nabídnout zákazníkovi závěr, musí být splněny všechny předpoklady. Ostatní sloupce popisují statistické hodnoty jednotlivých implikací.

Consequent	Antecedent	Instances	Support %	Confidence %	Rule Support %	Lift	Deployability
Pánská trička	Kraťasy	325	4,053	53,846	2,183	1,712	1,871
Dámská trička	Dámské mikiny Pánská trička	247	3,081	51,012	1,571	2,979	1,509
Pánské mikiny	Dámské mikiny Pánská trička	247	3,081	43,32	1,334	1,849	1,746
Pánská trička	Pánské mikiny	1 879	23,435	39,063	9,154	1,242	14,28
Pánská trička	Spodní prádlo	667	8,319	37,931	3,155	1,206	5,163
Pánská trička	Doplňky	534	6,66	37,64	2,507	1,197	4,153
Pánské mikiny	Spodní prádlo Pánská trička	253	3,155	33,597	1,06	1,434	2,095
Pánská trička	Dámská trička	1 373	17,124	33,503	5,737	1,065	11,387
Dámská trička	Dámské mikiny	1 236	15,415	33,172	5,113	1,937	10,302
Pánská trička	Tepláky a soupravy	705	8,793	30,78	2,706	0,979	6,086
Pánská trička	Dámské mikiny Dámská trička	410	5,113	30,732	1,571	0,977	3,542

Obrázek 3.9: Nalezené implikace

- **Instances**

  Suma košíků splňujících předpoklad.

- **Support**

  Kolik procent ze všech košíků splnilo předpoklad.

- **Confidence**

  Procento případů, kdy byl splněn předpoklad a zároveň závěr. Výpočet je prováděn pouze z košíků, které splnily předpoklad. Tento atribut určuje míru spolehlivosti pravidla.

- **Rule support**

  V kolika procentech všech košíků se objevil předpoklad i závěr.

- **Lift**

  Zlepšení pravidla, kolikrát je pravidlo lepší při použití předpokladu, než při náhodném výběru zboží bez ohledu na ostatní zboží.

$$Lift = \frac{confidence}{support\ antecedentu\ (\%)} \tag{3.2}$$

- **Deployability** Procento případů, kdy byl splněn předpoklad, ale závěr ne. Pro tyto případy je pravidlo použito a je nabídnut zákazníkovi závěr.

$$Deployability = support - rule\ support \tag{3.3}$$

3.4 Nasazení

Uvedení do praxe v Modeleru lze nasimulovat. Do košíku je vložen zákazníkův nákup a na základě vygenerovaných pravidel rozhodne uzel Apriori co zákazníkovi nabídnout. Například zákazník má v košíku dámskou mikinu a pánské tričko, v uzlu nabídka bude doporučení dalšího zboží.

Obrázek 3.10: Doporučení nákupu

Po průchodu uzlem Apriori budou zákazníkovi nabídnuty dva produkty, které splňují předpoklad.

```
Zákazníkovi 123.000 nabídni:
1) Dámská trička
2) Pánské mikiny
```

Obrázek 3.11: Doporučené produkty

Nabídku lze ověřit, na obrázku 3.9 je předpoklad dámská mikina s pánským tričkem dvakrát a jako své závěry má právě dámské tričko a pánskou mikinu.

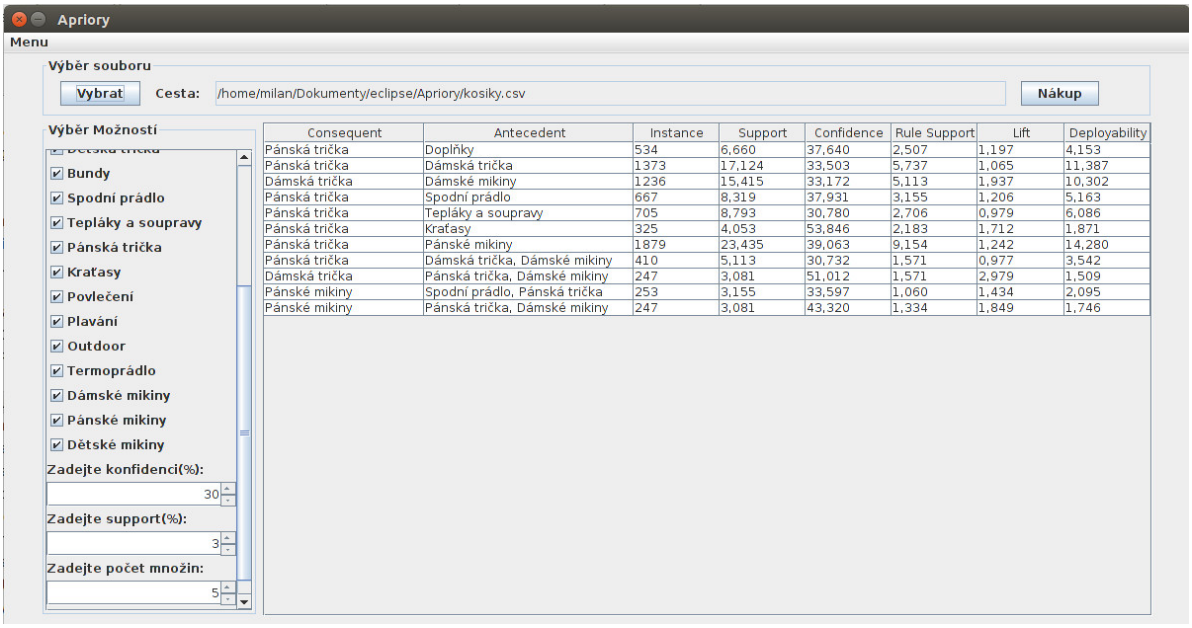
4. Implementace aplikace

Hlavním cílem je naprogramovat aplikaci umožňující zpracování připravených dat algoritmem Apriori. Spustitelná aplikace bude načítat data ve formátu csv [8]. P výběru dat lez nastavit atributy a citlivost pro algoritmus. Pokud jsou dostupné osobní informace o zákaznících, lze generovat pravidla pouze pro určité skupiny zákazníků. Získané implikace budou zobrazeny v tabulce společně se statistickými údaji o implikacích. Uživatel používající aplikaci může simulovat nákup a pomocí získaných implikací nabídnout další produkty.

Aplikace byla navržena v jazyce Java, jako podpora pro předmět data mining. Pro jazyk Java bylo rozhodnuto především kvůli tomu, že je multiplatformní. Proto jej studenti budou moci používat bez ohledu na vlastní operační systém. Aplikace byla navržena tak, aby simulovala chování algoritmu Apriori v IBM SPSS Modeleru a navíc mohla generovat pravidla pouze na cílené skupiny.

4.1 Návrh grafického rozhraní

Grafické rozhraní slouží především ke snadnější manipulaci s algoritmem. Jeho podoba je znázorněna na obrázku 4.1 s již vygenerovanými pravidly ze souboru z přípravy dat. Do aplikace je možné nahrávat pouze csv soubory v horní části rozhraní pomocí tlačítka vybrat. Po výběru se celý soubor projde a zjistí se jeho atributy, které mohou být, buď jednotlivé druhy zboží nebo osobní informace o zákazníkovi, ty jsou zobrazeny v levé části aplikace. Jako poslední možnost je zvolení citlivosti algoritmu pomocí nastavení minimálního supportu a minimální confidence, kde obě hodnoty jsou zadávány procentuálně. Po vygenerování pravidel se v pravé části zobrazí tabulka s nalezenými implikacemi, které mohou být vyzkoušeny pomocí tlačítka nákup. To vytvoří nové okno s položkami a dle vybraného zboží se zobrazí nabídka.



Obrázek 4.1: Vzhled grafického rozhraní

4.2 Hierarchie tříd

V příloze A je znázorněn diagram tříd, kde jsou vidět vazby mezi třídami. Je zřejmé, že třída *Data* je jádrem této aplikace a řídí téměř celý program, poskytuje data pouze pro třídu *GUI*, která řídí veškerou komunikaci s uživatelem a zobrazuje výsledná data. K uchování atributů a jejich proměnných je zde třída *HeadItems*. Z třídy *GUI* je možné spustit pouze *GUIBuy*, ta slouží k ověření vygenerovaných implikací. V třídě *FrequentItemsL* jsou generovány frekventované množiny a je zde prováděno generování kandidátů a odstranění množin nesplňujících support. Položky ve frekventovaných množinách jsou reprezentovány instancemi z třídy *ItemSet*. Abstraktní třída *Reader* slouží pouze k dědění a jejím potomkem je *CSVReader*, ten poskytuje data z vybraného souboru. Podrobnější popis tříd je popsán níže.

4.2.1 Zisk dat

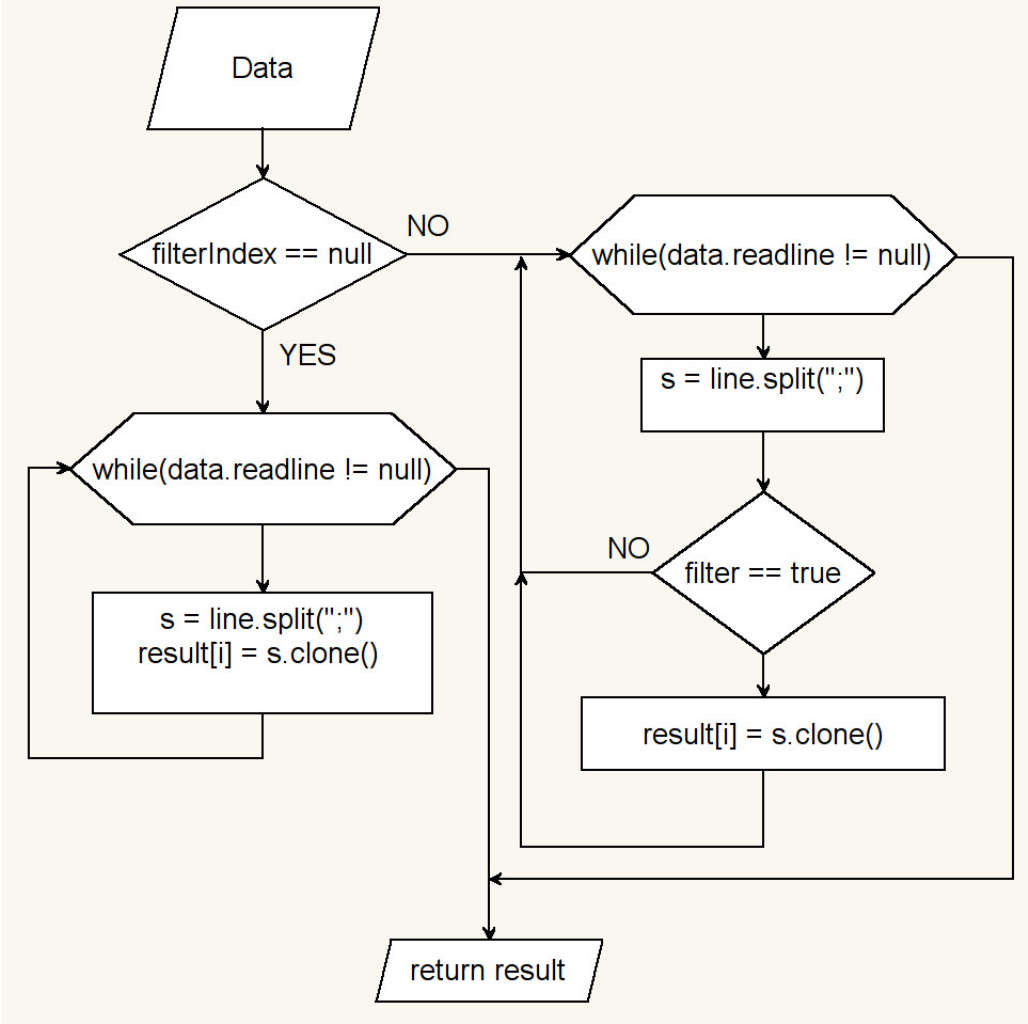
Pro zisk dat slouží abstraktní třída *Reader* sloužící pouze k dědění, byla použita především kvůli možnosti načítat další vstupní formáty souboru. Jednotné zpracování všech vstupních souborů není možné, jelikož každý formát má charakteristickou strukturu. Prozatím jsou načítány pouze csv soubory, které zpracovává potomek této třídy *CSVReader*.

První volaná metoda, bez ohledu na to, jaký je formát vstupu, je metoda *getHeadItems*. Jako parametr je předávána cesta k souboru. Metoda vrací pole stringů s názvy atributů a zároveň je v ní nastaveno počet transakcí obsažených ve vstupním souboru (*setNumberOfShoping*). Počet transakcí je důležitý pro výpočet minimálního supportu a confidence.

```
public String[] getHeadItems(String path){
  this.path = path;
  this.filterIndex = null;
  setNumberOfShoping();
  return getHead();
}
```

4.2.2 Čtení z csv souboru

CSVReader je potomek třídy *Reader*, zpracovává vstupní soubory ve formátu csv. Tato třída poskytuje data ze souboru a informace o něm ostatním třídám, probíhá zde i filtrování dat podle osobních informací zákazníka. Pokud je nastaven filtr, tak metoda *getData* vrací pouze vyfiltrovaná data. Na obrázku 4.2 je znázorněn vývojový diagram metody *getData*.



Obrázek 4.2: Zisk dat ze souboru

V cyklech while se čte celý soubor a je načítán po řádcích, ty jsou rozděleny podle středníku do pole stringů. Je-li nastaven filtr, data jsou filtrována pomocí indexů filtrovaných atributů a k nim přiřazených hodnot, které mají být vyfiltrovány.

Metoda *getHead* vrací pouze hlavičku souboru, tedy všechny atributy obsažené v souboru. Pomocí pozic těchto atributů vrací metoda *getFlags* hodnoty svých atributů, to je nutné abychom mohli nastavit filtrování v grafickém rozhraní. Jestliže je zvoleno filtrování podle některého atributu, volá se funkce *setFilter*. Ta nastaví indexy a s nimi spjaté hodnoty určené pro filtrování a také přepočítá hodnotu *numberOfShoping* tak, aby odpovídala počtu vyfiltrovaných košíků.

4.2.3 Uchování atributů

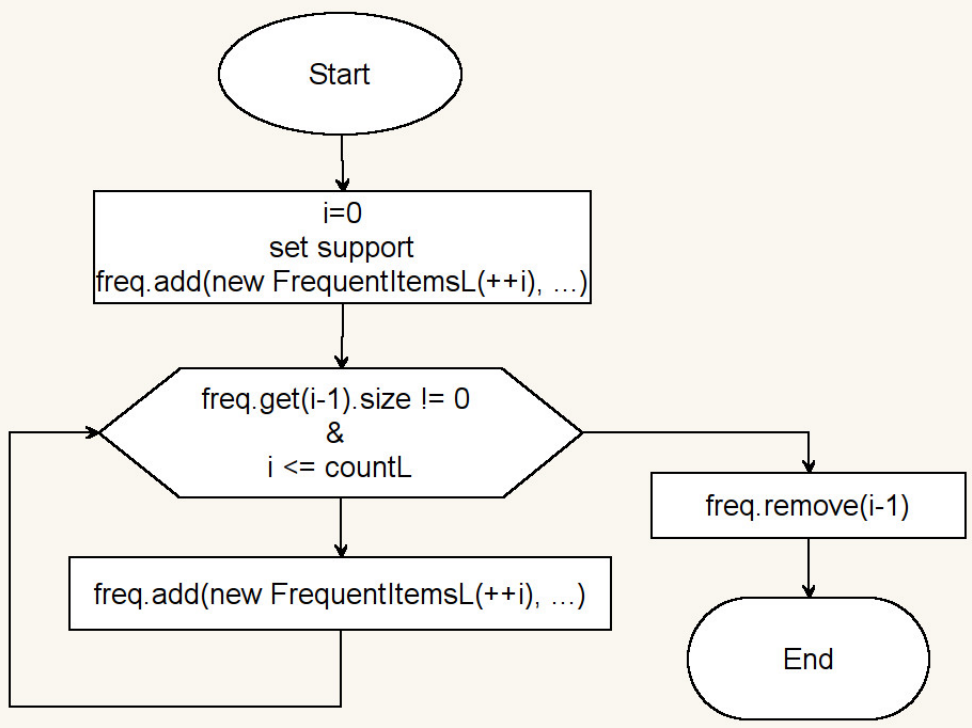
Pomocí instance třídy *HeadItem* jsou uchováván potřebná data o atributech. Pro aplikaci jsou uchovány hodnoty jméno, index, možné hodnoty atributu. Pokud je nastaven filtr uloží se vybraná položka.

```
public HeadItem(String name, int index) {
  this.setName(name);
  this.setIndex(index); }
```

Již v konstruktoru musí být uvedeny proměnné name a index, jelikož jsou po celou dobu běhu programu spolu spjaty. Ostatní metody v této třídě jsou pouze gettery a settery použitých proměnných.

4.2.4 Zprostředkování informací

Jádrem celé aplikace je třída Data, která provádí veškerou komunikace s grafickým rozhraním, poskytuje k dispozici nejen výsledné implikace, ale už od začátku programu předává důležitá data potřebná k zobrazení uživateli tak, aby mohl uživatel nastavovat filtry a atributy, které budou ovlivňovat výsledek. Uchovává v sobě potřebné informace o všech atributech, jejichž informace jsou uloženy v poli *HeadItem*. Toto pole je naplněno hned v konstruktoru, který také přijímá cestu k vybranému souboru, po nastavení cesty jsou vygenerovány informace o atributech ze souboru. Většina metod zajišťuje různou komunikaci, takže jenom přijímají nebo odesílají informace, některé metody zpracovávají přijatá data do potřebné podoby. Pro samotné generování frekventovaných množin slouží metoda *run*, která přijímá vybrané atributy, support, confidenci a maximální n-položkovou množinu kandidátů.

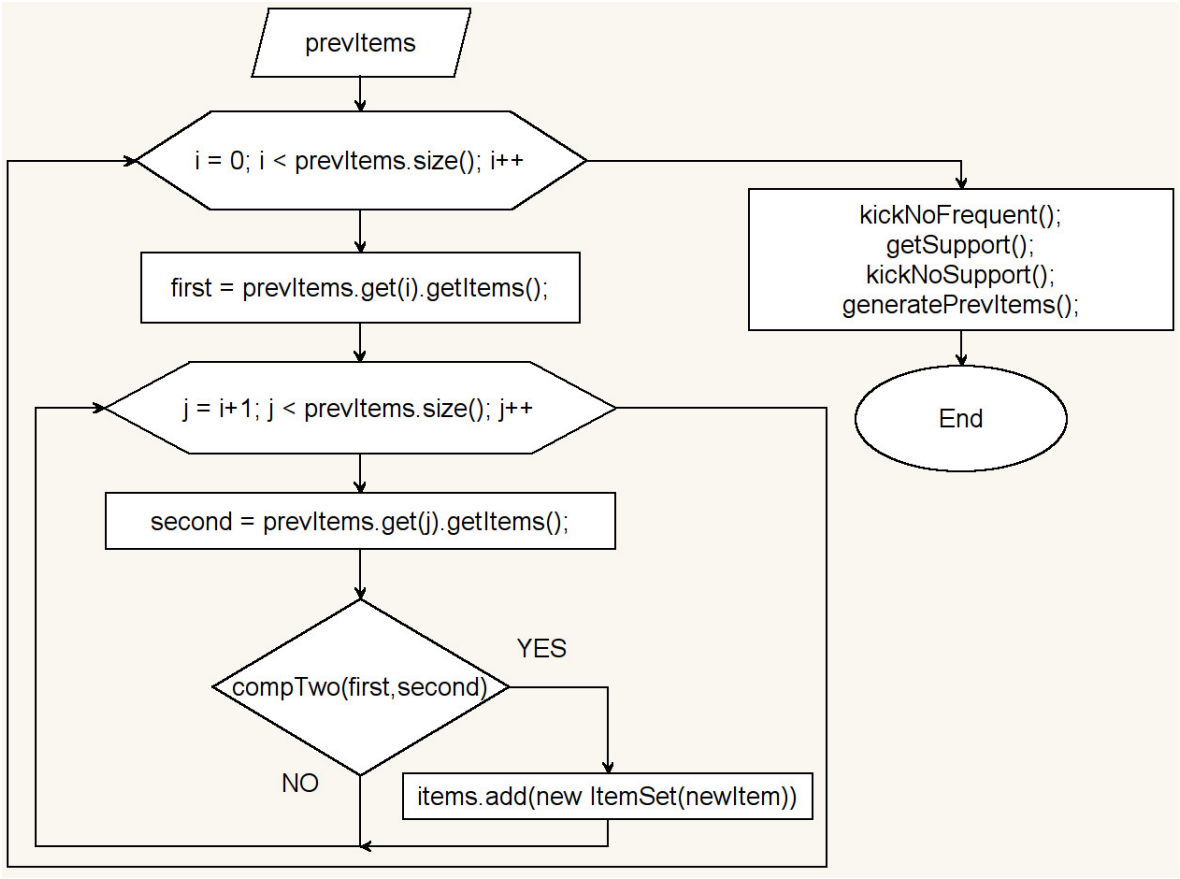


Obrázek 4.3: Zisk dat ze souboru

Vstupem do metody jsou parametry ovlivňující generování frekventovaných množin v každém kroku cyklu se vygeneruje nová frekventovaná množina, pokud je však množina prázdná nebo překročí maximální úroveň (Ln), je poslední množina smazána ze seznamu.

4.2.5 Frekventované množiny

V programu jsou reprezentovány instancí třídy *FrequentItemsetsL*, která se stará o vygenerování množiny kandidátů. Po vygenerování kandidátů jsou odstraněny množiny, jež nesplňují pravidlo o frekventovaných množinách (každá podmnožina musí být zároveň frekventovanou množinou). Průchodem souboru jsou zjištěny supporty množin a odstraněny ty, jež nesplňují minimální support. Pro výpočet pravidel jsou zde vygenerovány všechny podmnožiny z těchto množin.



Obrázek 4.4: Generování frekventované množiny

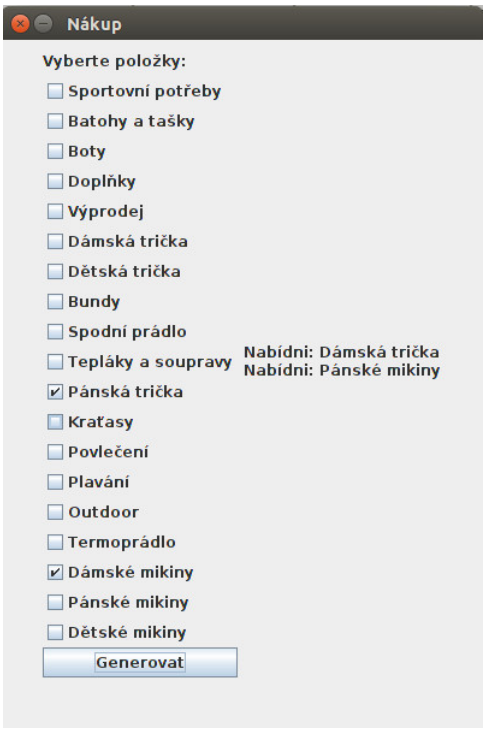
Vstupem jsou položky z minulé frekventované množiny, ty jsou pomocí dvou for cyklů spojeny, tak se vytvoří množina kandidátů pro tvorbu je důležitá metoda *compTwo*, která rozhodne zda, jsou obě položky vhodné ke spojení, pokud ano je vytvořena nová položka. Poté probíhá odstranění nefrekventovaných množin a vzniká nová frekventovaná množina.

4.2.6 Nalezení implikací

Položky ve frekventovaných množinách reprezentují instance třídy *ItemSet*, ve kterých dochází k nalezení implikací. V metodě *getImplication* se posuzuje, jaká pravidla splňují zadané vstupní podmínky a pro ty vypočítává statistické údaje. Všechny nalezené implikace jsou volány do frekventované množiny, tam se odstraňují duplicitní záznamy, po odstranění jsou nalezené implikace předány grafickému rozhraní.

4.2.7 Ověření nalezených implikací

Nalezené implikace lze lehce otestovat přímo v aplikaci pomocí tlačítka nákup, to otevře nové okno, kde je možné zadat, co má zákazník v nákupním košíku a na základě obsahu jeho košíku jsou vypsány druhy zboží, které mají být nabídnuty.



Obrázek 4.5: Ověření nalezených implikací

5. Závěr

Cílem bakalářské práce je provést analýzu nákupního košíku, nalézt důležité faktory v datech, zjistit vztahy mezi daty a nalézt v nich implikace pomocí asociačního algoritmu Apriori. Naprogramovat aplikaci určenou pro generování asociačních pravidel a vysvětlení algoritmu. Dále pak program a analýzu zpracovat jako e-learningovou podporu pro předmět data mining. V teoretické části je rozebírána problematika spjatá s tímto projektem, především metodologie CRISP-DM, podle které probíhala celá analýza nákupního košíku.

Pro zpracování bakalářské práce byla použita data z internetového obchodu. Analýza dat byla provedena data miningovým nástrojem IBM SPSS Modeler. Na základě analýzy se odstranily nepotřebné nebo zkreslující záznamy a byly zjištěny některé faktory důležité pro využití v marketingu. Data bylo nutné přetransformovat do podoby vhodné pro algoritmus Apriori sloučením záznamů, které obsahují jednu objednávku. Na závěr proběhlo modelování, nalezení a ověření implikací.

Aplikace byla naprogramována v jazyce java tak, aby byla nezávislá na operačním systému. V aplikaci je možné prohlédnout si, jak algoritmus Apriori funguje. Z předzpracovaných dat z analýzy lze vygenerovat implikace, které se zobrazí v tabulce. Pokud jsou dostupná data o zákaznících, je možné generovat pravidla pouze pro určitou skupinu lidí. Získané implikace pomocí programu se shodují s nalezenými implikacemi díky IBM SPSS Modeleru, při stejném nastavení vstupních podmínek. Tyto implikace je možné vyzkoušet v aplikaci, nebo je uložit do souboru a použít v praxi.

K nalezení implikací by bylo možné do aplikace přidat další algoritmy, výsledky by se porovnávaly a vybrány by byly jen ty nejlepší. Také je možné program rozšířit o možnost získání implikací v časovém rozmezí.

Požitá literatura

[1] DOUG, Alexander. Data Mining. *www.laits.utexas.edu* [online]. [cit. 2014-05-10]. Dostupné z: http://www.laits.utexas.edu/ anorman/BUS.FOR/course.mat/Alex/

[2] BERKA, Petr. *Dobývání znalostí z databází.* Praha: Academia, 2003. 366s. ISBN 80-200-1062-9

[3] *IBM SPSS Modeler.* [online]. [cit. 2014-05-16]. Dostupné z: http://www-01.ibm.com/software/analytics/spss/products/modeler/

[4] *What is the CRISP-DM methodology?*[online]. [cit. 2014-05-10]. Dostupné z: http://www.sv-europe.com/crisp-dm-methodology/

[5] *Knowledge Discovery in Databases (KDD).* [online]. [cit. 2014-05-10]. Dostupné z: http://www.usc.edu/dept/ancntr/Paris-in-LA/Analysis/discovery.html

[6] *SAS Enterprise Miner.* [online]. [cit. 2014-05-16]. Dostupné z: http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

[7] PROCHÁZKA, Michal. *Data mining: jiný pohled na problém.* [online]. [cit. 2014-05-16]. Dostupné z: http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem

[8] Shafranovich, Y. Common Format and MIME Type for Comma-Separated Values (CSV) Files RFC 4180, IETF, October 2005. Dostupné z: http://www.ietf.org/rfc/rfc4180.txt

[9] MCCONNELL, Steve. *Dokonalý kód: Umění programování a techniky tvorby software.* Computer press, 2006. ISBN 978-80-251-0849-9.

Seznam příloh

Příloha A: Diagram tříd