UNIVERZITA PARDUBICE

Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

KLASIFIKAČNÍ ÚLOHY PRO DATA MINING

Petra Jandová

Bakalářská práce
2013

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2012/2013**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Petra Jandová**
Osobní číslo:	**E09235**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Regionální a informační management**
Název tématu:	**Klasifikační úlohy pro Data Mining**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Závěrečná práce se bude týkat problematiky Data Mining-u (DM) a metod klasifikace. Na vybraných příkladech budou vysvětleny: základní pojmy z klasifikačních metod, vybraná problematika sběru a předzpracování dat pro tři klasifikační úlohy, vybrané zásady návrhu a analýzy modelů pro jednotlivé úlohy.

i. Vysvětlení základních pojmů DM a metod klasifikace
ii. Sběr dat pro tři klasifikační úlohy
iii. Modelování jednotlivých úloh
iv. Vyhodnocení výsledků

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 55 stran**

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

Seznam odborné literatury:

1) BERKA, Petr. Dobývání znalostí z databází. 1. vyd. Praha: Academica, 2003. ISBN 80-200-1062-9.
2) PETR, Pavel. Data Mining - Díl I. 3. vyd. Pardubice: Univerzita Pardubice, 2010. ISBN 978-80-7395-325-6.
3) RUD, Olivia Parr. Data Mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001. ISBN 80-7226-577-6.
4) BERRY, Michael J. A., LINOFF Gordon. Data mining techniques: for marketing, sales and customer support. New York: John Wiley & Sons, 1997. ISBN 0-471-17980-9.

Vedoucí bakalářské práce: **doc. Ing. Jiří Křupka, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **3. října 2012**

Termín odevzdání bakalářské práce: **30. dubna 2013**

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

L.S.

prof. Ing. Jan Čapek, CSc.
vedoucí ústavu

V Pardubicích dne 15. října 2012

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracovala samostatně. Veškeré prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

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

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 25. 6. 2013 Petra Jandová

PODĚKOVÁNÍ

Tímto bych ráda poděkovala svému vedoucímu doc. Ing. Jiřímu Křupkovi, PhD. za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování bakalářské práce.

ANOTACE

Bakalářská práce se zabývá klasifikačními úlohami v data miningu. Cílem této práce je vytvoření modelu, který bude schopen spolehlivě klasifikovat data do cílových atributů. Pro splnění cíle byl proveden sběr dat na tři klasifikační úlohy, které mají za úkol roztřídit data do skupin podle souvislostí mezi nimi. Z výsledků měření vzešly jako optimální tyto klasifikační techniky: rozhodovací stromy s algoritmy C5.0, CHAID, CART, QUEST a neuronové sítě. Data zpracovaná pomocí těchto technik jsou poté analyzována a v závěru je vyhodnocena nejpřesnější technika.

KLÍČOVÁ SLOVA

Data mining, dobývání znalostí, klasifikace, klasifikační metody, rozhodovací stromy, neuronové sítě, IBM SPSS Modeler

TITLE

The classification tasks for Data Mining

ANNOTATION

The bachelor thesis deals with the classification tasks for data mining. The goal of the thesis is to create the model that is able to qualify data to target attributes. For reach the aim data for three classification task was collected. Data was analyzed with these methods: decision trees with algorithms C5.0, CHAID, CART, QUEST and neural networks. Selected methods were compared on the basis of the results of testing and the precise method was chosen.

KEYWORDS

Data mining, knowledge discovery, classification, classification methods, decision trees, neural networks, IBM SPSS Modeler

OBSAH

SEZNAM TABULEK

SEZNAM ILUSTRACÍ

SEZNAM ZKRATEK A ZNAČEK

CART – Classification And Regression Tree

CRISP-DM – Cross Industry Standard Process for Data Mining

CRM – Customer Relationship Management

DM – Data Mining

ECML/PKDD – European Conference on Machine Learning / Practice of Knowledge Discovery in Databases

ESPRIT - European Strategic Program on Research in Information Technology

FAM – Fuzzy Associative Memory

GNU – GNU's Not Unix

CHAID – CHi-squared Automatic Interaction Detector

IBM – International Business Machines

ID3 – Iterative Dichotomizer

KDD – Knowledge Discovery in Databases

MATLAB – MATrix LABoratory

MPL – Multi Layer Percepton

PAKDD - Pacific-Asia Conference in Knowledge Discovery and Data mining

QUEST – Quick Unbiased and Efficient Statistical Tree

RBF – Radial Basis Function

SAS - Statistical Analysis System

SOM – Self Organizing Map

SPSS – Statistical Package for the Social Sciences

TDIDT – Top Down Induction of Decision Trees

UNICEF - United Nations International Children's Emergency Fund

WEKA – Waikato Environment for Knowledge Analysis

ÚVOD

Data mining je dnes hojně využívaným nástrojem pro získávání nových zajímavých znalostí z nejrůznějších databází. V dnešní době tržní konkurence a pokroků ve vědě se stalo získávání nových a zároveň cenných informací téměř nutností. Právě díky této nutnosti se data mining těší čím dál větší oblibě, ať už například v oblasti marketingu, bankovnictví nebo medicíny, kde data mining pomáhá uživatelům k efektivnějšímu rozhodování.

O vzrůstající data miningové popularitě v odborných kruzích svědčí i vznik odborných časopisů a množství různých světových konferencí zabývající se tímto tématem. Nejznámější a nejvýznamnější konference jsou americká KDD, asijská PAKDD a evropská ECML/PKDD. Shodou náhod připadlo konání letošní evropské konference na Českou republiku. V září se v Praze uskuteční 13. ročník. [3] [1]

Existuje mnoho důvodů proč a jak data mining využívat v praxi. Pro větší přehlednost můžeme proto rozdělit data miningové úlohy na několik typů, podle toho, jakou informaci potřebujeme z dat získat. Jednou z typových úloh pro data mining je klasifikace. Právě tou se tato práce bude zabývat podrobněji než jinými úlohami, které budou zmíněny pouze ve stručnosti. Popsán bude samotný pojmem data mining, jeho typové úlohy a softwarové systémy, pomocí nichž lze data mining provádět. Konkrétněji se práce zaměřuje na klasifikační úlohy a techniky pro jejich modelování.

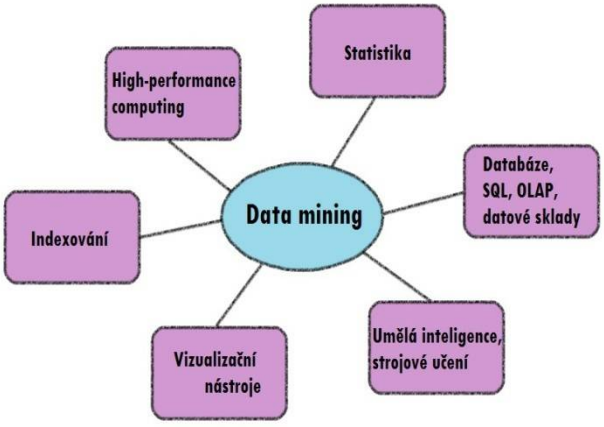
Cílem práce je vytvoření modelu, který bude schopen spolehlivě klasifikovat data do cílových atributů. Pro splnění cíle jsou stanoveny tři klasifikační úlohy, které mají za úkol roztřídit data do skupin podle souvislostí mezi nimi. Data budou zpracována pomocí různých klasifikačních technik, poté budou analyzována a v závěru bude vyhodnocena nejpřesnější technika klasifikace. Modelování je prováděno v softwaru IMB SPSS Modeler.

Další cíl je související s předchozím. Pro navrhování modelu je potřeba nejprve vybudovat teoretický aparát. Cílem je tedy seznámení s pojmem data mining, s úlohami a technikami, které jsou v této oblasti využívány, zejména pak ke klasifikaci dat.

1 DATA MINING

Na počátku 90. let 20. století, společně s rozvojem počítačové techniky, se z dosud samostatných vědních disciplín, zejména statistiky a umělé inteligence (respektive metod strojového učení), začala vyvíjet nová disciplína – data mining. Nejčastější překlad slova data mining je dobývání (či dolování) znalostí z dat (či databází). Přesná definice data miningu v současné době neexistuje, často se ale používá tato definice uznávaného výzkumníka Usamy Fayyada [1] : „*Dobývání znalostí z databází lze definovat jako netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat.*“ Zjednodušeně řečeno lze chápat data mining jako metodu, která získává skryté, ale užitečné informace z obrovského množství dat. Přitom je potřeba nezapomínat na vhodnou interpretaci výsledku koncovým uživatelům, protože pokud uživatel výsledkům data miningu nerozumí, je celý proces takřka zbytečný.

V data miningu se kromě již zmíněné statistiky a umělé inteligence prolínají různé směry vědeckého zkoumání – nejvýznamnější z nich jsou zobrazeny na obrázku č. 1.



Obrázek č. 1: Hlavní složky data miningu

Zdroj: upraveno podle [9] [7]

Rozdíl mezi použitím „dřívějších“ metod strojového učení a statistiky je v prvé řadě pohled na výstupní data – výsledek musí být pro koncového uživatele srozumitelný a pokud možno užitečný. Při používání data miningu získáváme nové skryté znalosti o zkoumané databázi. V neposlední řadě je v data miningu kladen důraz na přípravu (předzpracování) dat ještě před tím, než data budeme dále analyzovat. [1]

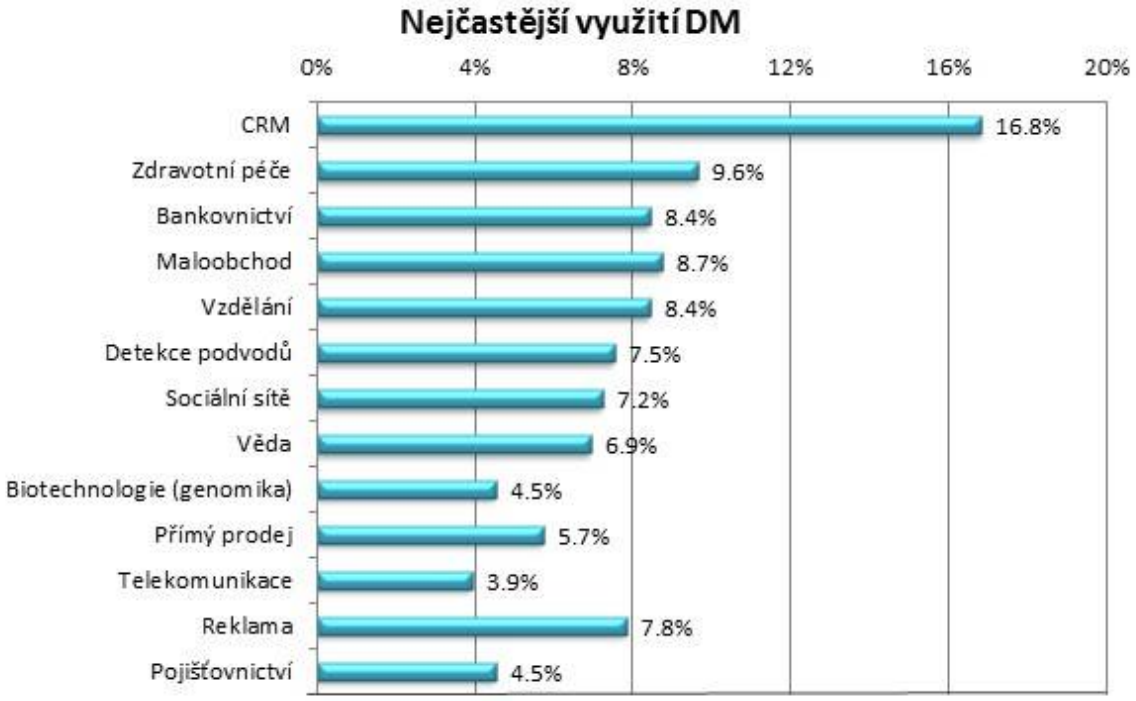
Účel data miningu je pomoc při jakémkoli typu rozhodování. Data mining obecně spočívá v prohledávání často i velmi rozsáhlých databází, analyzování dat nejrůznějšími

matematickými metodami, a nalezení co nejvíce relevantních informací vhodných k řešení problému. Toto souhrnné pojetí někteří autoři (například Fayyad [4]) označují pojmem Knowledge discovery in databases (KDD) neboli objevování znalostí z databází, zatímco pojmem Data mining označují pouze jednu z fází KDD. V této práci budou významy obou pojmů synonymní. [30]

1.1 Využití data miningu

Data mining má velmi širokou škálu využití. Od obecných oblastí – např. obchodu - CRM (řízení vztahů se zákazníky), obchodování po internetu, přes medicínu, reklamu, bankovnictví, různé prevence (terorismu, detekce podvodů) po specializované oblasti genomiky nebo astrofyziky. [33] Příkladem může být nalezení skupin obchodů pomocí shlukové analýzy na základě jejich obratu, sortimentu a typu zákazníků. Nalezené skupiny pak lze použít při vytváření specifické reklamní kampaně.

Podle průzkumu na internetovém portálu KD nuggets byly zjišťovány nejčastější oblasti využití data miningu v roce 2012. Bylo vybráno 13 položek, které si rozdělily 333 hlasů z celkového počtu 507 hlasů. Výsledky jsou patrné na grafu v obrázku č. 2.



Obrázek č. 2: Graf nejčastějšího využití data miningu

Zdroj: upraveno podle [24]

1.1.1 Customer Relationship Management (CRM)

Z grafu na obrázku č. 2 je patrné, že data mining je nejčastěji využíván při CRM - řízení vztahů se zákazníky. CRM spočívá v nalezení a pokud možno pochopení zákazníkova chování, zkrátka jde o vztah zákazníka a firmy. V dnešní době vyostřeného konkurenčního boje, se čím dál více upouští od tradičních přístupů marketingu, ve kterých se hledí především na to, jak nejlevněji vyrobit co nejvíce zboží a o samotný odbyt již není nutné se tolik starat. Doba, kdy byla konkurence v odvětví tak nízká, že zákazník v podstatě nakoupil to, co mu bylo na trhu nabízeno, přičemž měl jen minimální volbu při výběru produktu, je nenávratně pryč. V současné době již není zásadní problém v ekonomické výrobě, do popředí zájmu se dostává otázka, jak daný produkt co nejefektivněji prodat. [27]

V této situaci je tedy nutné pracovat se zákazníkem. Cílem je zaujmout a přilákat nové zákazníky a udržet stávající. Je tudíž nutností řídit vztahy se zákazníkem. Podniky staví zákazníka do popředí všech procesů, zohledňují se v první řadě přání a požadavky zákazníků. Tento přístup se také označuje pojmem one-to-one marketing.

CRM se dělí na dvě části [33] :

- *Operativní CRM*: Zahrnuje tvorbu marketingových kampaní, podporuje prodej, obsahuje všechny prostředky pro komunikaci se zákazníkem (telefonicky, e-mailem, pop-up nabídky po přihlášení do uživatelského účtu, atd.)
- *Analytický CRM:* Analyzuje zákaznická data, tvoří vzorce chování podobných zákazníků, slouží jako podklad pro operativní CRM. Podstatou je tzv. Bussines intelligence, což je [2] : „*souhrnný pojem pro procesy, technologie a nástroje potřebné k přetvoření dat do informací, informací do znalostí a znalostí do plánů, které umožní provést akce podporující splnění primárních cílů organizace*.“ Právě zde, v analytickém CRM, je pro složité analýzy zákaznických dat využíván data mining.

V následující tabulce jsou zobrazeny rozdíly mezi tradičním marketingem, který se v prvé řadě neorientuje na zákazníka a mezi marketingem s použitím CRM (a data miningu).

Tabulka č. 1: Rozdíl mezi marketingem bez použití DM a s použitím DM

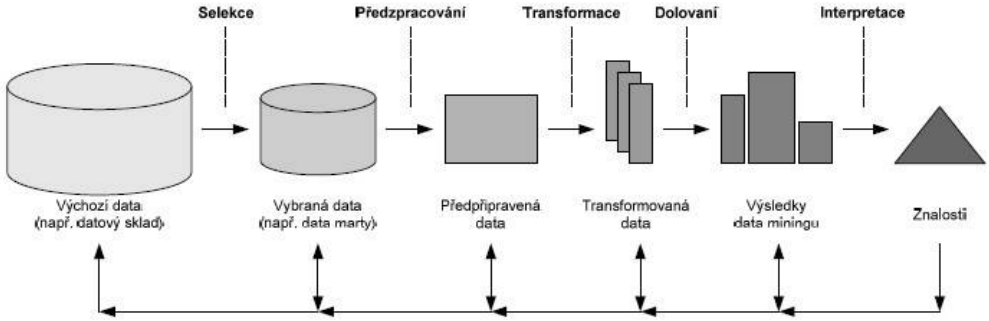
Tradiční marketing (bez použití DM)	One-to-one marketing (s použitím DM)
Anonymní zákazník	Individuální, určitý zákazník
Běžný produkt	Přizpůsobení produktů/služeb zákazníkovi
Sériová produkce	Výroba na míru
Masová reklama	Individuální zprávy
Jednostranná komunikace	Interaktivní komunikace
Úspěch z prodeje, vysokého zájmu	Rozvoj zákaznické věrnosti
Tržní podíl	Zákaznický podíl
Široké (obecné) cíle	Výnosný specializovaný segment na trhu (mezery na trhu)
Tradiční distribuce kanálů komunikace (vzájemně nesouvisejících)	Nové, vzájemně propojené kanály komunikace (internet, chytré mobilní telefony)
Produktově orientovaný marketing	Marketing orientovaný na zákazníka

Zdroj: [33]

1.2 Proces dobývání znalostí

Proces dobývání znalostí je interaktivní a iterativní proces. To znamená, že jednotlivé kroky se mohou několikrát opakovat a dle výsledků je možné se vracet zpět k předchozím krokům. Celý proces je zachycen na obrázku č. 3. Podobně celý proces při řešení problému popisuje i metodika CRISP-DM (podrobněji viz podkapitola 1.3). Hlavními kroky při dobývání znalostí podle Fayyada jsou [4] :

- *Selekce* – z výchozích dat vybereme pouze data, která souvisejí s cílem procesu
- *Předzpracování* – očištění dat (například od odlehlých hodnot)
- *Transformace* – generalizace nebo normalizace dat, podle zvolené metody dolování
- *Dolování (data mining)* – vytváření data miningových modelů pro úspěšné řešení problému
- *Interpretace* – vhodně zvolené vysvětlení výsledků dolování (například vizualizace), při správném pochopení výsledků jsou získány nové znalosti



Obrázek č. 3: Proces dobývání dat z databází

Zdroj: upraveno podle [4]

1.3 Metodiky data miningu

Pro lepší přehlednost a efektivnost při řešení data miningových úloh vznikly metodiky, které uživatelům udávají jednak jednotný rámcový postup pro vytvoření celého data miningového projektu a zároveň umožňují sdílet a předávat zkušenosti z jiných úspěšných projektů. V následujících bodech jsou popsány tři metodiky, z nichž nejrozšířenější je metodika CRISP-DM. [1]

- ***Metodika CRISP-DM***

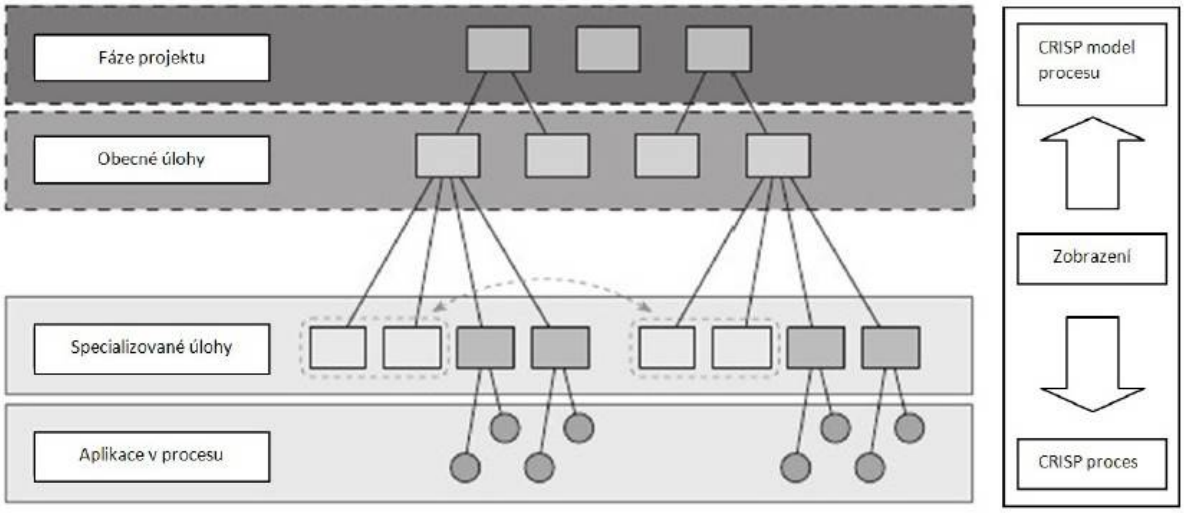
Metodika vznikla v roce 1996 v rámci projektu ESPRIT (výzkumný evropský program). Na vytvoření spolupracovaly firmy NCR (USA), Teradata (USA), Daimler AG (původně DaimlerChrysler - Německo), SPSS (USA) a OHRA (Holandsko), všechny mají letité zkušenosti s data miningovými projekty. [11]

Cílem metodiky je navrhnout univerzální postup při řešení projektů, použitelný v různých komerčních softwarových aplikacích, a návrh řešení problémů, které mohou během projektu nastat. CRISP-DM umožňuje řešit data miningové úlohy rychleji, efektivněji, spolehlivěji a s nižšími náklady.

Hierarchická struktura CRISP-DM (obrázek č. 4) je složena z těchto čtyř abstraktních úrovní [11] :

- *Fáze projektu* – šest fází (podrobněji viz níže)
- *Obecné úkoly* – rozdělení fází na obecné úlohy tak, aby pokryly všechny DM situace, které kdy mohou nastat, a zároveň aby úlohy byly stabilní – musí počítat s nepředvídatelným vývojem

- *Specializované úkoly* - převod obecných úkolů na konkrétní, podle potřeb daného problému se přiřazuje způsob jeho řešení
- *Aplikace v procesu* – technická realizace specializovaných úloh



Obrázek č. 4: Hierarchická struktura CRISP-DM

Zdroj: upraveno podle [22]

Životní cyklus projektu spadá pod první úroveň hierarchické struktury CRISP-DM – do úrovně Fáze projektu. Podle metodiky CRISP-DM je tento bod rozdělen na šest částí (fází). Výsledky jednotlivých fází se navzájem ovlivňují a na základě těchto výsledků je nezřídka potřeba se k předchozím fázím vracet. Fáze metodiky CRISP-DM jsou zobrazeny na obrázku č. 5. Jedná se o tyto fáze [11] :

- *Business understanding* (Porozumění problému)

Tato fáze vyžaduje pochopení cílů a požadavků projektu z obchodního hlediska, posuzuje se zde možnost rizika a přínos projektu. Tyto znalosti se poté přeformulují na požadavky z analytického hlediska a stanoví se předběžný plán prací.

- *Data understanding* (Porozumění datům)

Porozumění dat začíná sběrem dat, se kterými chceme dále pracovat. Následuje seznámení s daty například pomocí popisné statistiky (četnosti hodnot atributů, průměry, minima, maxima atd.).

- *Data preparation* (Příprava dat)

Fáze zahrnuje veškeré činnosti potřebné k vytvoření konečné datové sady, se kterou bude pracovat data miningový software. Provádí se zde selekce dat, čištění dat (např. od odlehlých

hodnot), transformace dat (např. generalizace, normalizace), odvozování dat atd. Tato fáze bývá nejpracnější a obvykle se provádí opakovaně.

- *Modeling* (Modelování)

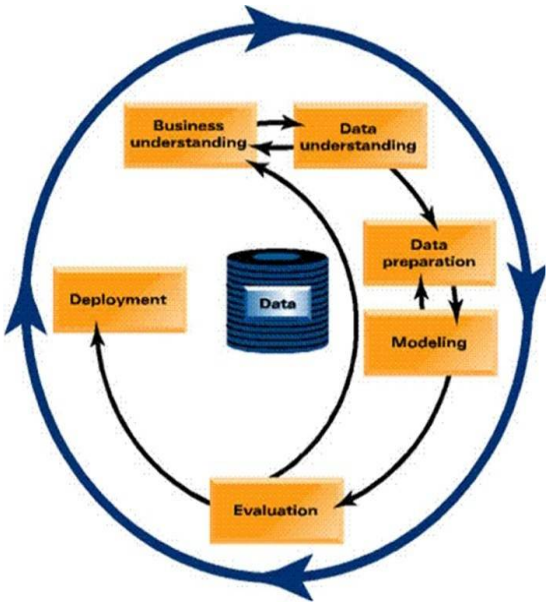
Zde dochází k výběru nejvhodnějšího algoritmu pro analýzu dat, nejčastěji pomocí data miningového softwaru. Pokud to problém dovoluje, doporučuje se vyzkoušet více různých metod s různým nastavením parametrů a výsledky pak porovnat, případně zkombinovat. Některé metody mají specifické požadavky na formu dat, často je tudíž potřeba vrátit se o krok zpět k Přípravě dat.

- *Evaluation* (Vyhodnocení výsledků)

V této fázi je již vytvořen kvalitní model. Na řadu přicházejí opět manažeři, kteří posuzují, zda bylo dosaženo všech předem zadaných cílů. Podstata je nalezení důležitých položek, které na začátku nemusely být zřejmé. Na konci této fáze by mělo padnout rozhodnutí o výběru a způsobu užití nejlepších výsledků data miningu.

- *Deployment* (Využití výsledků)

Vytvořením nejlepšího modelu proces nekončí. Je potřeba výsledky interpretovat v dostatečně srozumitelné podobě, kterou koncový uživatel bez problému pochopí a efektivně výstup procesu aplikuje do praxe. Výstupem celého procesu může být jak „pouhé“ sepsání závěrečné zprávy, tak složitější zavedení (hardwarové, softwarové, organizační) systému pro automatickou klasifikaci a predikci nových případů.



Obrázek č. 5: Metodika CRISP-DM

Zdroj: upraveno podle [1]

- ***Metodika 5A***

Metodika používaná firmou SPSS, tedy producentem data miningového softwaru. Kroky metodiky jsou [1] :

- *Assess* (posouzení) – posouzení potřeb projektu
- *Access* (získávání) – shromáždění dat potřebných k projektu
- *Analyze* (analyzování) – provedení datových analýz – přeměna dat na informace a znalosti
- *Act* (provedení) – přeměna znalostí na akční znalosti
- *Automate* (automatizace) – převedení výsledků analýz do praxe a následné užívání

- ***Metodika SEMMA***

Metodika používaná firmou SAS, tedy producentem data miningového softwaru. Kroky metodiky jsou [1] :

- *Sample* (vzorek) – výběr vhodných dat
- *Explore* (poznávání) – průzkum dat a jejich redukce a vizualizace
- *Modify* (úprava) – datové transformace, vytváření veličin
- *Model* (modelování) – analýza dat
- *Assess* (posouzení) – porovnávání modelů a jejich srozumitelná interpretace

2 TYPY ÚLOH PRO DATA MINING

Data mining řeší mnohdy zdánlivě nesouvisející problémy z mnoha různých oborů. Ještě předtím, než je vybrána konkrétní metoda pro modelování problému, je potřeba úloze přiřadit typ (skupinu), do které spadá. Pro splnění cílů data miningového problému lze využít více metod, které spadají do jedné skupiny – je dokonce žádoucí mít v závěru k dispozici porovnání více výsledků odlišných metod. K porovnávání výsledků je dobré znát výhody a nevýhody jednotlivých metod.

Jednotná podoba rozdělení typů úloh neexistuje, nejčastěji se však v literatuře uvádějí tři typy data miningových úloh. Jsou to [1] : predikce a klasifikace, deskripce a hledání nuggetů. Na obrázku č. 6 jsou pro představu znázorněny znalosti, které v rámci jednoho konceptu v jednotlivých typech úloh zkoumáme.

2.1 Predikce a klasifikace

Predikce znamená předpověď, odhad budoucích hodnot na základě znalosti hodnot jiných, většinou minulých. Cílem predikce a klasifikace je tedy nalézat znalosti, které pomáhají nová data roztřídit (klasifikovat) do předem určených skupin. Podstatou těchto úloh je výběr jednoho cílového atributu *A* a následné modelování vlivu ostatních atributů na atribut *A*. Je dána přednost většímu počtu znalostí, pokrývajících daný problém, někdy na úkor jejich srozumitelnosti. Nejčastějším příkladem z praxe bývá rozpoznání problémového nebo naopak bonitního zákazníka banky.

Metody a techniky nejčastěji používané pro predikci a klasifikaci: regrese, neuronové sítě, rozhodovací stromy. Podrobněji je o některých těchto metodách pojednáno v kapitole 3.

2.2 Deskripce

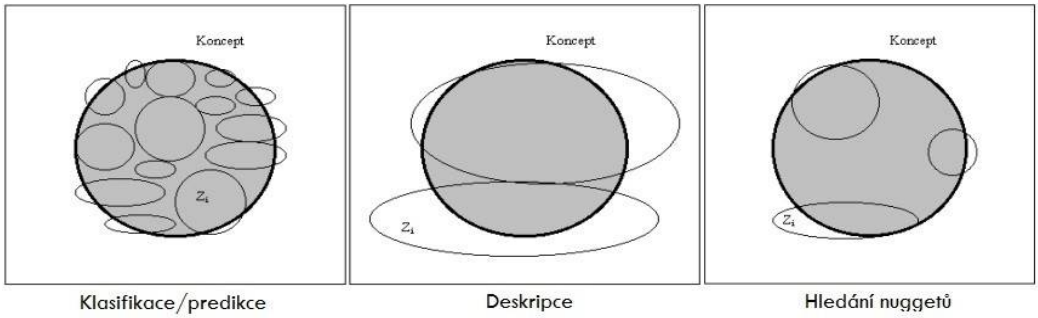
Deskripce, jak je již z názvu patrné, popisuje (charakterizuje) data jako celek. Odhaluje nové skryté zákonitosti (znalosti), které přispívají k lidskému poznání ve zkoumané oblasti. Cílem je získání menšího počtu srozumitelných znalostí, které pokrývají celý problém.

Metody a techniky nejčastěji používané pro deskripci: shluková analýza, vizualizace, sumarizace. [25]

2.3 Hledání nuggetů

Hledání nuggetů je podobné deskripci s tím rozdílem, že při hledání nuggetů získáváme nové srozumitelné znalosti, které nemusejí pokrývat celý problém, ale jen jeho zajímavé části. Slovo nugget můžeme přeložit jako střípek, což odpovídá principu tohoto typu úloh. Jelikož postupy při modelování těchto typů úloh jsou do značné míry podobné deskripci a jejich používané metody se prolínají, někteří autoři (například Fayyad [4]) hledání nuggetů řadí pod deskripci.

Metody a techniky nejčastěji používané při hledání nuggetů: asociační pravidla, vzorkování, segmentace, shluková analýza.



Obrázek č. 6: Znázornění typů data miningových úloh

Zdroj: [1]

2.4 Metody pro řešení data miningových úloh

K dosažení cílů data miningových můžeme použít různé metody, které se mnohdy ve všech typech úloh mohou prolínat. Jelikož metody nemůžeme přesně přiřadit jen k jedné skupině úloh, je zde stručný přehled metod, podle Usamy Fayyada, které pomáhají k dosažení cílů predikce, deskripce a hledání nuggetů [4] :

- *Regrese* – statistická metoda, kde známe řadu v minulosti naměřených hodnot a pomocí regresní analýzy předpovídáme vývoj hodnot do budoucna
- *Shlukování* – úkolem je rozdělení datového souboru do skupin (shluků) na základě jejich podobnosti, přičemž cílem shlukování je vytvoření, nalezení těchto skupin – skupiny nejsou předem známé
- *Klasifikace* – úkolem je rozdělení datového souboru do skupin, dle jejich podobnosti, které jsou předem známé

- *Sumarizace* – hledání uceleného popisu datového souboru (například záznam průměrů a směrodatných odchylek)
- *Modelování závislostí* – zachycování podstatných závislostí atributů. Existují dva typy závislostí – strukturální (graficky znázorňuje závislostní vztahy) a kvantitativní (určuje sílu závislosti na číselné škále)
- *Detekce změn a odchylek* – objevování nejpodstatnějších změn z dříve naměřených nebo normových hodnot

3 KLASIFIKACE A PREDIKCE

Mezi klíčové úlohy, které se mohou v data miningu aplikovat jsou klasifikace a predikce. Použití těchto pojmů v literatuře často splývá. Jako hlavní **rozdíl mezi klasifikací a predikcí** uvádějí autoři odborné literatury to, že klasifikace pracuje s diskrétními daty (zařazování proměnných do kategoriálních skupin), zatímco predikce se spojitými daty (například odhad vývoje časových řad). Holčík považuje za nejpřesnější tento výklad rozdílu mezi klasifikací a predikcí [8] : „*Pojem klasifikace používáme tehdy, pokud vybíráme identifikátor klasifikační třídy z určitého diskrétního konečného počtu možných identifikátorů. Pokud určujeme (predikujeme) spojitou hodnotu, například pomocí regrese, pak hovoříme o predikci.*“

Cílem klasifikace a predikce je vytvořit model, pomocí něhož bude možné co nejpřesněji klasifikovat (predikovat) nové případy, které v budoucnu nastanou, do vhodných tříd. K tomu musí klasifikační model umět rozpoznat podstatné rozdíly mezi prvky v kategoriích od rozdílů nepodstatných a podle toho kategorie odlišit. Podle Skalské jsou cíle klasifikačního procesu vymezeny těmito požadavky [30] :

- Nalezení přesného klasifikačního pravidla.
- Výběr proměnných, které klasifikaci umožňují.
- Možnost srozumitelně interpretovat výsledek klasifikace, porozumění struktuře klasifikačního modelu.
- Nalezení rozhodujícího parametru klasifikace.
- Logicky zdůvodnitelná kritéria zařazení prvku do dané skupiny, s ohledem na daný problém.

Příkladem **využití klasifikace a predikce v praxi** mohou být systémy v bankách nebo telekomunikačních společnostech, které detekují výskyt tzv. „fraudů“, neboli podvodníků, na základě platební morálky zákazníků. Dalším příkladem může být pomoc při určování lékařských diagnóz na základě příznaků, které pacient uvede. Pomocí rozhodovacích stromů lze efektivněji diagnózu rozpoznat.

Vzhledem k zaměření práce se bude dále pojednávat především o klasifikaci, nikoli predikci. Klasifikace se opírá o strojové učení, tedy o techniky, které umožňují počítačovému systému učit se. Jde o oblast umělé inteligence. Učením v tomto kontextu rozumíme takovou činnost, při níž dochází ke změnám vnitřního stavu systému, a která zefektivňuje schopnost přizpůsobení se změnám v okolním prostředí. Pro klasifikaci je nejčastěji využíván algoritmus

učení s učitelem, kde je pro vstupní data určen správný výstup, který slouží jako kontrola správnosti. [19]

Proces klasifikace a tedy sestavení klasifikačního modelu se podle Skalské skládá ze dvou hlavních kroků [30] :

1) *Diskriminace*:

   Diskriminace vyžaduje množinu znaků charakterizující každý prvek v datovém souboru, a která později umožňují učinit rozhodnutí o příslušnosti prvku k určité skupině (= klasifikace). Hlavním procesem diskriminace je učení algoritmu s učitelem. To je prováděno na takzvaných trénovacích datech. To jsou vzorky dat, která obsahují informaci („učitele“) o příslušnosti ke skupině (tj. třídu do které patří). Pomocí těchto dat se systém naučí klasifikovat data do tříd a sestaví tak diskriminační model, se kterým v dalším kroku dále pracuje vlastní klasifikace.

2) *Vlastní klasifikace*:

   Vlastní klasifikace je proces, při němž se definují klasifikační pravidla (klasifikátory) na základě předchozího učícího souboru a diskriminačního modelu, a která slouží při zařazování nových prvků do skupin.

Ještě před procesem klasifikace je potřeba provést předzpracování dat. To spočívá především v úpravách dat typu [28] :

- *Čištění dat* - doplňuje chybějící hodnoty nebo odstraňuje z dat přebytečný šum
- *Určení relevantnosti dat* - nevýznamná data jsou buď odstraňována nebo je jim přirazena váha jejich významu
- *Transformace dat* – zahrnuje zjednodušení (například převedení spojitých atributů na atributy diskrétní) a normalizaci dat za účelem přizpůsobení klasifikačnímu modelu

Existuje mnoho způsobů, jak provádět procesy klasifikace. Kritéria pro porovnání metod jsou zejména přesnost, výpočetní složitost, odolnost vůči chybám (robustnost), škálovatelnost a snadná interpretace. Pravděpodobně nejsledovanějším faktorem metod klasifikace a nejvýznamnější vlastností konkrétních klasifikačních modelů je jejich přesnost. V následujících podkapitolách jsou představeny některé techniky využívané při tvorbě

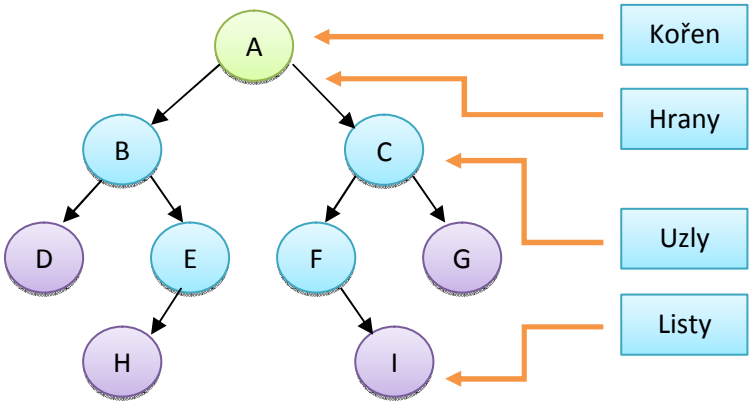
klasifikačních modelů. Jsou to: rozhodovací stromy, rozhodovací (klasifikační) pravidla a neuronové sítě.

3.1 Rozhodovací stromy

Rozhodovací stromy jsou pravděpodobně nejužívanějším způsobem interpretace data miningových výsledků. To je dáno především grafickou podobou, díky které jsou stromy jednou z nejpřehlednějších technik. Známým příkladem rozhodovacího stromu z praxe je zařazování rostlin a živočichů do biologických skupin. Od kořene (třída/řád/říše organismu) se po větvi stromu postupuje na základě otázek na vlastnosti organismu hlouběji do stromu, až k cílovému listu (konkrétní organismus).

Rozhodovací stromy jsou acyklické grafy, skládají se z hran a uzlů. Na obrázku č. 7 jsou zobrazeny všechny části stromu. Jsou to:

- *Kořen:* uzel, který je na vrcholu stromu, nevstupuje do něj žádná hrana
- *Hrany:* spojnice mezi uzly, ohodnocená hodnotou příslušného atributu
- *Uzly:* vnitřní uzly, které mají další potomky (vedou z nich hrany)
- *Listy:* konečné uzly bez potomků (nevede z nich žádná hrana), každý list představuje cílovou klasifikační třídu
- *Větve:* postup stromem od kořene až do cílového listu (v obr. č. 7 například ACFI)



Obrázek č. 7: Části rozhodovacího stromu

Zdroj: vlastní zpracování

Rozhodovací stromy se podle topologie dělí na [14]:

- *Binární stromy*: z uzlu vystupují pouze dvě větve
- *Nebinární (vícecestné) stromy*: z uzlu vystupuje více než dvě větve

Při tvorbě rozhodovacích stromů je využívána metoda „rozděl a panuj“. Princip této metody je následující: Celá trénovací množina s daty se rozděluje na menší podmnožiny – uzly stromu tak, aby v jednom uzlu byly navzájem co nejpodobnější příklady. Z těchto uzlů se rekurzivně postupuje, do té doby, dokud se atributy v uzlech dají dělit do skupin. Tento postup dělení používá obecný algoritmus TDIDT, který postupuje takzvaně „shora dolů“ – od kořene po listy stromu. Postup algoritmu je popsán v tabulce č. 2. [15], [1]

Tabulka č. 2: Obecný algoritmus pro tvorbu rozhodovacích stromů (TDIDT)

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

Zdroj: [1]

3.1.1 Problémy při tvorbě stromů

Při tvorbě rozhodovacích stromů může dojít ke dvěma zásadním problémům - strom může mít příliš velký rozsah, snižující jeho srozumitelnost nebo může dojít k „přeučení“, což znamená přílišné přizpůsobení systému trénovací množině dat – tedy snaha o bezchybnou klasifikaci trénovačích dat, která jsou zatížena šumem. Systém pak nemá schopnost generalizace a selhává. Strom je proto možné zjednodušit tak, že v jednom listu stromu nejsou příklady výhradně jedné třídy, ale že příklady jedné třídy pouze převažují. Toto redukovaného stromu se dá dostáhnout dvěma způsoby: [1]

- *Modifikací původního algoritmu* – růst stromu se předčasně zastaví a přidáním nového kritéria, které určuje, zda má uzel dále expandovat, se redukovaný strom se vytvoří přímo.
- *Prořezáváním stromu* – jedná se o následné procházení již hotového stromu „zdola nahoru“ a u každého podstromu se podle nějakého kritéria rozhoduje, zda tento podstrom nahradit konečným listem. Tento způsob je častěji využíván, zejména proto, že je obtížné poznat, kdy růst stromu předčasně zastavit. V tabulce č. 3 je popsán algoritmus prořezávání stromu.

Tabulka č. 3: Algoritmus prořezávání stromu

	Prořezávání stromu
1)	Převeď strom na pravidla
2)	Generalizuj pravidlo odstraněním podmínky z předpokladu, pokud dojde ke zlepšení odhadované správnosti
3)	Uspořádej prořezaná pravidla podle odhadované správnosti; v tomto pořadí budou použita pravidla pro klasifikaci

Zdroj: [1]

3.1.2 Klasifikační stromy

Pomocí klasifikačních stromů je možné provádět pouze klasifikaci diskrétních, tedy nespojitých, hodnot. Pokud mají původní data spojitý charakter, je potřeba je pomocí generalizace převést na charakter diskrétní.

Klasifikační stromy fungují podle algoritmu TDIDT, popsaného v tabulce č. 2. Aby mohl být vykonán první krok algoritmu, tedy volba kořene a jeho rozvětvení, je potřeba najít nejvhodnější atribut pro toto větvení tak, aby co od sebe nejlépe odlišoval příklady různých tříd. K tomu slouží několik kritérií podle charakteristik atributu. Tato kritéria jsou používána v algoritmech pro rozhodovací stromy. Pro rozhodování o vhodném atributu pro větvení jsou představena tato kritéria: Entropie, Informační a poměrný informační zisk, Gini index a χ^2 test. [1] [6]

- ***Entropie***

Entropie je veličina udávající míru neuspořádanosti zkoumaného systému nebo také neurčitosti daného procesu. Pro vyjádření entropie se využívá vztahu:

$$H = -\sum_{t=1}^{T}(p_{t1} * \log_2 p_{t2}) \quad (1)$$

t … třída

T … počet tříd

p … pravděpodobnost výskytu třídy t

Výpočet entropie pro jeden atribut, podle kterého by se mohl rozhodovací strom větvit, se provádí tímto způsobem [1] : Pro každou hodnotu v, kterou může nabýt uvažovaný atribut A, se spočítá podle vzorce (1) entropie $H(A(v))$ na skupině příkladů, které jsou pokryty kategorií $A(v)$.

$$H\big(A(v)\big) = -\sum_{t=1}^{T} \frac{n_t\big(A(v)\big)}{n\big(A(v)\big)} * \log_2 \frac{n_t(A(v))}{n(A(v))} \quad (2)$$

Dále se spočítá střední entropie *H(A)* jako vážený součet entropií *H(A(v))*, přičemž váhy v součtu jsou relativní četností kategorií *A(v)* v trénovacích datech. Na základě nejmenší střední entropie *H(A)* se rozhoduje o vhodnosti atributu pro větvení.

$$H(A) = -\sum_{v \in Val(A)} \frac{n(A(v))}{n} * H(A(v)) \quad (3)$$

- ***Informační zisk / Poměrný informační zisk***

Informační zisk je definován jako rozdíl entropie pro cílový atribut (pro celá data) a pro atribut o kterém se rozhoduje. Informační zisk měří redukci entropie způsobenou volbou atributu *A*.

$$Zisk\ (A) = H(C) - H(A) \quad (4)$$

kde se *H(A)* spočte podle vzorce (3) a *H(C):*

$$H(C) = -\sum_{t=1}^{T} \frac{n_t}{n} \log_2 \frac{n_t}{n} \quad (5)$$

Při hledání nejvhodnějšího atributu pro větvení hledáme atribut s největší hodnotu *Zisk (A)*.

Protože informační zisk nebere v potaz počet hodnot zvoleného atributu, a to vede k některým chybám (nevhodnost pro klasifikaci nových případů), používá se jako kritérium poměrný informační zisk:

$$Poměrný\ zisk\ (A) = \frac{Zisk\ (A)}{-\sum_{v \in Val(A)} (\frac{n(A(v))}{n} * \log_2 \frac{n(A(v))}{n})} \quad (6)$$

Pro nejvhodnější atribut opět hledáme největší hodnotu *Poměrného zisku (A).*

- ***Gini index***

Tak jako entropie hraje podobnou úlohu i Gini index *GI*. Pro výpočet Gini indexu se používá vztah:

$$GI = 1 - \sum_{t=1}^{T} (p_t^2) \quad (7)$$

p_t … relativní počet příkladů *t*-té třídy

Hodnotu *GI* pro jeden atribut spočítáme stejně, jako tomu bylo u entropie. Pro každou hodnotu *v*, kterou může nabýt uvažovaný atribut *A*, se spočítá podle vzorce (7) Gini index *GI(A(v))* na skupině příkladů, které jsou pokryty kategorií *A(v).*

$$GI\big(A(v)\big) = 1 - \sum_{t=1}^{T}\left(\frac{n_t(A(v))}{n(A(v))}\right)^2 \quad (8)$$

Dále se spočítá Gini index *GI(A)* jako vážený součet Gini indexu *GI(A(v))*, přičemž váhy v součtu jsou relativní četností kategorií *A(v)* v trénovacích datech.

Jestliže je *GI(A)* rovno nule, v konečném uzlu je pouze jedna třída – to je žádoucí, tudíž jako nejvhodnější atribut pro větvení hledáme co nejmenší hodnotu *GI(A)*.

$$GI(A) = \sum_{v \in Val(A)} \frac{n(A(v))}{n} * Gini(A(v)) \quad (9)$$

- ***χ^2 (chí kvadrát test)***

Posledním kritériem je χ^2, který umí najít vzájemnou souvislost mezi dvěma atributy v kontingenční tabulce. Pokud je X a Y v tabulce nezávislé (nemají souvislost), má test přibližně Pearsonovo χ^2 rozdělení se stupni volnosti $v = (r\text{-}1)(s\text{-}1)$, kde r je počet řádků v tabulce a s počet sloupců v tabulce. Výpočet testu χ^2 je [14]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(p_{ij}-o_{ij})^2}{o_{ij}} \quad (10)$$

kde o_{ij} jsou očekávané četnosti dané vztahem:

$$o_{ij} = \frac{R_i S_j}{n} \quad (11)$$

i,j … označení řádků a sloupců tabulky

$p_{i,j}$ … pozorovaná četnost

n … celkový počet pozorování

R_i … počet pozorování v řádku i

S_j … počet pozorování ve sloupci j

Ten atribut, který nejvíce souvisí s cílovým atributem, má tedy největší hodnotu χ^2, je vhodným kandidátem na atribut pro větvení stromu.

3.1.3 Regresní stromy

Kromě klasifikačních stromů, existují i stromy regresní, kde na rozdíl od klasifikačních stromů, které zařazují data do tříd, regresní stromy odhadují hodnotu numerických atributů. To znamená, že v koncovém listu může být místo názvu třídy konkrétní hodnota (konstanta), která udává například průměrnou hodnotu.

Dalším rozdílem mezi klasifikačním a regresním stromem je způsob volby atributu pro větvení. Oba stromy fungují na principu algoritmu TDIDT, ale způsob větvení

v regresním stromu vychází z měření směrodatné odchylky hodnot cílového atributu. Redukce směrodatné odchylky se počítá podle následujícího vzorce [1] :

$$S_y - \sum_{v \in Val(A)} \frac{n(A(v))}{n} * S_y(A(v)) \tag{12}$$

S_y^2 … rozptyl hodnot cílového atributu pro celá trénovací data

$S_y^2(A(v))$ … rozptyl hodnot cílového atributu pro příklady pokryté kategorií $A(v)$.

Atribut, který maximalizuje toto kritérium je vybrán pro větvení stromu. Větvení končí, pokud je $S_y(A(v))$ menší než 5% S_y , nebo pokud je v uzlu málo příkladů.

3.1.4 Algoritmy rozhodovacích stromů

Jak již bylo popsáno výše, možností jak určit atribut, který bude strom podle nějakého pravidla dále dělit, je vícero. Proto byly vyvinuty algoritmy, které na základě výše zmíněných kritérií hledají nejvhodnější atribut pro větvení. V tabulce č. 4 jsou některé známé algoritmy porovnány podle svých vlastností. Zde jsou představeny podrobněji:

- ***ID3, C4.5 a C5.0***

Základním algoritmem je algoritmus ID3, vyvinut australským vědcem J. R. Quinlanem v roce 1986. Kritériem pro větvení je v tomto algoritmu informační zisk. V roce 1993 rozšířil Quinlan algoritmus ID3 na algoritmus C4.5 tak, aby mohl pracovat s numerickými atributy, chybějícími hodnotami a brát v potaz cenu za chybná rozhodnutí. Za dalších pět let, tedy v roce 1998 se C4.5 dostalo dalšího rozšíření, zejména přátelštějšího uživatelského ovládání. Tato nová modifikace nese název C5.0, v některých platformách, například Windows, je pojmenován See5. C5.0 je vhodný pro všechny typy proměnných, na výstupu však vykazuje pouze kategoriální hodnoty. Kritériem pro větvení je informační zisk a entropie. [1] [33]

Silné stránky: C5.0 má dobré využití v případě v případě velkého množství vstupních polí. Je schopen rychle odhadnout klasifikační model. Nabízí také posilovací metodu (boost) pro zvýšení přesnosti klasifikace. [13]

- ***CART***

Algoritmus CART byl vyvinut v roce 1984 statistiky Barkleyské a Stanfordské univerzity. Kromě názvu CART je možné setkat se s různými modifikacemi jako C&RT nebo CRT. CART vytváří pouze binární stromy, což znamená, že z jednoho uzlu vedou pouze dvě větve.

Binární stromy jsou obvykle přesnější než nebinární. Jako kritérium pro větvení je zde používán Gini index. [33]

Silné stránky: CART má podobné výhody jako C5.0, na rozdíl od něj však dokáže pracovat jak s diskrétními, tak se spojitými výstupy. [12]

- ***CHAID***

Princip algoritmu CHAID byl zkoumán už od roku 1975 J. A. Hartiganem, samotný algoritmus byl vyvinut v roce 1980 G. V. Kassem. Je často využíván v komerční sféře, zejména v marketingu (například výběr cílových zákazníků). K výběru nejvhodnějšího atributu k větvení používá CHAID χ^2 test, z čehož vyplývá, že vstupní proměnné musí být pouze kategoriální. S tímto problémem si ale většina softwaru umí poradit a spojitá data převede automaticky na diskrétní. [33]

Silné stránky: CHAID dokáže vytvářet nebinární stromy, což znamená, že uzly mohou mít při dělení více než dvě větve. Stromy proto rostou více do šířky, než je tomu u binárních stromů (QUEST a CART). CHAID je vhodnější pro větší datové soubory. [12]

- ***QUEST***

QUEST je statistický algoritmus, který vybírá proměnné bez zkreslení a zajišťuje rychlé a přesné sestavení binárního stromu. Podobně jako C5.0 má na výstupu pouze kategoriální proměnné. Kritérium pro větvení funguje pomocí statistických metod.

Silné stránky: QUEST tvoří binární stromy, které jsou obvykle více přesné než stromy nebinární. Mimo to je algoritmus tvořící binární strom rychlejší. [12]

Tabulka č. 4: Porovnání algoritmů pro tvorbu stromů

	C5.0	CART	CHAID	QUEST
Větvení	Vícenásobné	Binární	Vícenásobné	Binární
Spojitý vstup	Ano	Ano	Ne	Ano
Spojitý výstup	Ne	Ano	Ano	Ne
Kritérium výběru atributu	Informační zisk, entropie	Gini index	Chí kvadrát	Statistické metody
Interaktivní tvorba stromu	Ne	Ano	Ano	Ano

Zdroj: [15]

3.2 Rozhodovací (klasifikační) pravidla

Pro začátek je dobré objasnit si rozdíl mezi asociačními pravidly, která se v data miningu také běžně používají, a rozhodovacími, neboli klasifikačními, pravidly. Asociační pravidla hledají zajímavé souvislosti mezi hodnotami atributů a jejich kombinací bez toho aniž by byl v závěru pravidla předem vyhrazen cílový atribut určující zařazení do třídy, zatím co rozhodovací pravidla se, jak už název napovídá, používají při klasifikaci a v závěru pravidla je vyhrazen cílový atribut, určující příslušnost ke třídě. Rozhodovací pravidla jsou lineárním přepisem rozhodovacích stromů. [1]

Syntaxe rozhodovacího pravidla:

IF *Předpoklad* THEN *Třída,*

kde podmínka pravidla IF znamená logický součin všech podmínek testů atributů po cestě z kořene do listu rozhodovacího stromu a závěr pravidla THEN je přidělení hodnoty do třídy cílového atributu (cílového listu stromu). [28]

Jednou z metod, jak vytvořit rozhodovací pravidla, je přepis rozhodovacího stromu do pravidel. Postupuje se postupně od kořene k listům, po všech větvích stromu. Další možností jsou algoritmy, tvořící rozhodovací pravidla. Ty jsou následující [1] :

- ***Pokrývání množin***

  Autorem algoritmu pokrývání množin, též známým jako AQ, je polsko-americký vědec Ryszard Michalski. Na rozdíl od rozhodovacích stromů, kde je používána metoda „rozděl a panuj“ je v tomto algoritmu použita metoda „odděl a panuj“, což je snaha nalézt pravidla, která pokrývají příklady téže třídy a oddělit je od třídy jiné. Rozhodovací pravidla, která jsou určena tímto algoritmem, mají stejnou vyjadřovací sílu jako rozhodovací stromy. Zatímco však u rozhodovacích stromů postupujeme v prostoru hypotéz pouze „shora dolů“, v pokrývání množin lze použít jak postup „shora dolů“ (specializace – od jednodušších případů po složitější), tak postup „zdola nahoru“ (generalizace – zjednodušování složitějších případů). Při postupu „zdola nahoru“ je využíván právě algoritmus AQ, který vytváří takzvaný neuspořádaný soubor IF-THEN pravidel. Pro postup „shora dolů“ je využíván algoritmus CN4, který je podrobněji popsán v následujícím bodu - Rozhodovací seznam. Obecný postup při pokrývání množin je v tabulce č. 5. [1]

Tabulka č. 5: Algoritmus pokrývání množin

	Algoritmus pokrývání množin
1)	Najdi pravidlo, které pokrývá pozitivní příklady a žádný negativní
2)	Odstraň pokryté příklady z trénovací množiny
3)	Pokud v trénovací množině zbývají nějaké nepokryté pozitivní příklady, vrať se k bodu 1, jinak skonči

Zdroj: [1]

- ***Rozhodovací seznam***

Opakem pro neuspořádaný soubor IF-THEN pravidel je rozhodovací seznam neboli uspořádaný soubor pravidel. Zde je syntaxe pravidel následující:

IF *Předpoklad_1* THEN *Třída_i*

ELSE IF *Předpoklad_2* THEN *Třída_j*

ELSE IF *Předpoklad_3* THEN *Třída_k* ….

Každá podmínka ELSE IF propojuje všechny pravidla navzájem, pravidla v seznamu jsou tudíž navzájem závislá. Předpoklad za ELSE IF v sobě implicitně skrývá negaci podmínek všech předešlých pravidel.

Pro vytváření jak uspořádaného, tak neuspořádaného souboru pravidel můžeme použit algoritmus CN2 a jeho rozšíření CN4. Je založen na podobném principu jako obecný algoritmus pro pokrytí množin, jeho postup je ale „shora dolů“. Rozdíl mezi tvorbou uspořádaných a neuspořádaných pravidel v hlavním cyklu algoritmu CN4 je v tabulce č. 6.

Tabulka č. 6: Hlavní algoritmus CN4 pro neuspořádaná a uspořádaná pravidla

	CN4 pro neuspořádaná pravidla		CN4 pro uspořádaná pravidla
1)	Nechť *Seznam_pravidel* je prázdný seznam	1)	Nechť *Seznam_pravidel* je prázdný seznam
2)	Pro každou třídu *C*	2)	Dokud trénovací množina D není prázdná
2a)	Dokud množina *D* (trénovací množina) pozitivních příkladů této třídy není prázdná	2a)	Pomocí funkce Search(*Předpoklad, D*) nalezni nejlepší kombinaci Předpokladů
2aa)	Pomocí funkce Search(*Předpoklad, D*) nalezni nejlepší kombinaci Předpokladů	2b)	Přiřaď D:= D – D(*Předpoklad*), kde D(*Předpoklad*) jsou příklady pokryté kombinací *Předpokladů*
2ab)	Přiřaď D:= D – D(*Předpoklad*), kde D(*Předpoklad*) jsou příklady pokryté kombinací *Předpokladů*	2c)	Do *Seznam_pravidel* přidej pravidlo IF *Předpoklad* THEN *Class*, kde *Class* je majoritní třída příkladů v D(*Předpoklad*)
2ac)	Do *Seznam_pravidel* přidej pravidlo IF *Předpoklad* THEN *C*		

Zdroj: [1]

- ***Pravděpodobnostní pravidla (algoritmy ITRule a ESOD)***

Pravděpodobnostní pravidla jsou založena na hodnotách (vahách, pravděpodobnostech), které jsou pravidlům přidělovány. Na principu ohodnocování a následné klasifikace nových případů jsou založeny expertní systémy, což jsou počítačové programy, které dokáží řešit velmi složité úlohy, podobně jako by je řešil lidský expert – tedy specialista na danou problematiku. Algoritmy pro tvoření tohoto typu pravidel jsou ITRule a ESOD. [1] [19]

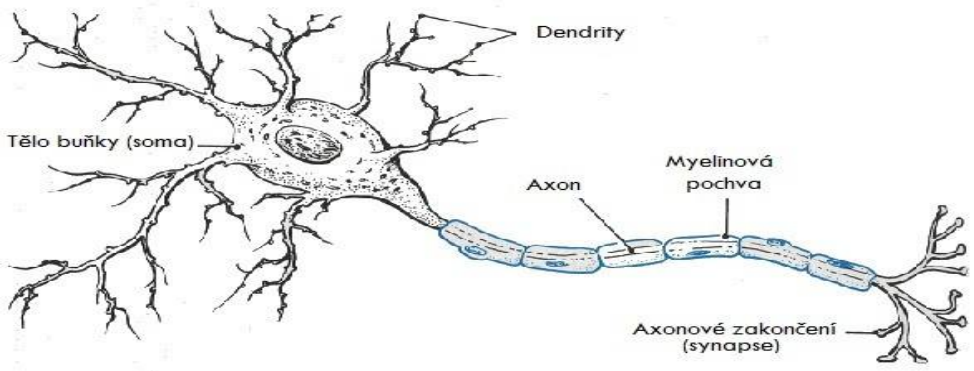
3.3 Neuronové sítě

Další technikou pro tvorbu automatických systémů pro klasifikaci a predikci jsou neuronové sítě. V případech, kdy primárně nezáleží na srozumitelnosti výstupu, mohou být použity jako náhrada místo rozhodovacích stromů a pravidel. Na rozdíl od rozhodovacích stromů, kde je lepší použití kategoriálních dat, se však v neuronových sítích více uplatní spojitá numerická data. Podobně jako bylo nutné v rozhodovacích stromech použít diskretizaci spojitých proměnných, v neuronových sítích je potřeba provést tzv. „binarizaci“ kategoriálních proměnných. Pro každý kategoriální atribut se vytvoří tolik nových binárních atributů, kolik měl původní atribut různých hodnot. Stejně jako u rozhodovacích stromů je tento proces ve většině softwarů prováděn automaticky. [1]

Cílem neuronových sítí je napodobování chování lidského mozku. Zejména pak ve třech aspektech:

- Umět uložit znalosti pomocí synapsí
- Aplikovat znalosti na řešení problémů - uvažování
- Získávání nové znalosti v průběhu učení neuronové sítě

Základním stavebním kamenem nervové soustavy člověka je nervová buňka – neuron. Ten se skládá z těla buňky, ze kterého vybíhá mnoho větvících se výběžků - dendritů a jeden dlouhý výběžek, zvaný axon. Dendrity přijímají vzruchy z okolí, vedou vzruchy k buňce a tvoří tak vstupy do neuronu. Pomocí axonu je neuron schopen sám vyslat další signál do svého koncového rozvětvení a pomocí synapse, tedy spojení axonového rozvětvení s dendrity jiného neuronu, předat informaci (výstup) do okolních neuronů. Spojením neuronů pomocí synapsí vzniká biologická neuronová síť. Tímto biologickým systémem jsou informatické neuronové sítě inspirovány. Obrázek č. 8 znázorňuje stavbu biologického neuronu. [20]



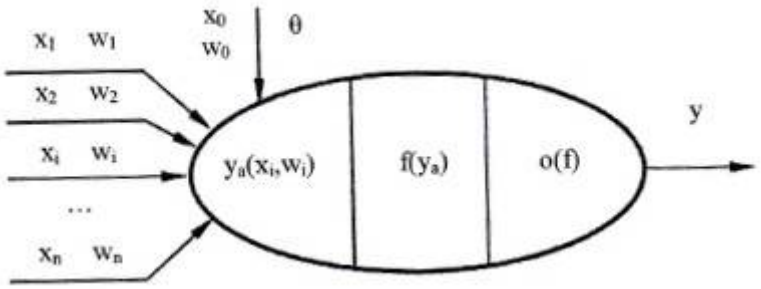
Obrázek č. 8: Biologický neuron

Zdroj: [10]

Pomocí počítačů lze simulovat mnohem rychlejší neuron než biologický neuron, problémem však je obrovské množství biologických neuronů a množství synapsí v mozku. Velikost a struktura takto masivní mozkové sítě zatím není možné pomocí umělé inteligence napodobit.

K tomu, aby mohl neuron vyslat signál je potřeba, aby vstupní signály překročily určitou prahovou hodnotu. Pokud tuto hodnotu signály nepřekročí, v neuronu nevznikne žádná reakce. Z toho vycházejí matematické modely umělých neuronů. Nejznámějšími modely

umělých neuronů jsou McCulloch-Pittsův „logický neuron“ a Wodrowův „adaptivní lineární neuron Adaline“. Na obrázku č. 9 je popsaná struktura McCulloch-Pittsova neuronu.



Obrázek č. 9: Model umělého neuronu

Zdroj: [20]

- x_i – vstupy neuronu, i=1,2,…,n
- n – počet vstupů
- w_i – synaptické váhy
- y_a – vstupní potenciál neuronu $y_a = \sum_{i=1} x_i w_i$
- $f(y_a)$ – aktivační funkce neuronu
- $o(f)$ – výstupní funkce neuronu
- θ – práh neuronu (x_0, w_0)
- y – výstup neuronu

Činnost neuronu lze popsat matematicky:

- neuron vyšle signál: y = 1 pro $\sum_{i=1} x_i w_i > w_0$
- neuron nevyšle signál: y = 0 pro $\sum_{i=1} x_i w_i < w_0$

Podmínku $y_a > w_0$ lze přepsat pomocí aktivační funkce $f(y_a)$. Jejím úkolem je převést hodnotu vstupního potenciálu na výstupní hodnotu z neuronu. Výběr vhodné aktivační funkce závisí na konkrétním typu řešené úlohy. [20], [10]

Neurony mají schopnost učit se, neboli adaptovat se. Činnost neuronových sítí lze rozdělit do dvou fází [20]:

- ***Fáze učení:***

Znalosti se ukládají do synaptických vah neuronové sítě. Ty se během učení mění na základě pravidel daných typem učení neuronové sítě. Pro učení platí daný vztah:

$$\frac{\partial w}{\partial t} \neq 0 \qquad (13)$$

kde W je matice všech synaptických vah

- ***Fáze života:***

Znalosti se využívají pro řešení konkrétních problémů (klasifikace, predikce). Synaptické váhy se během fáze života nemění. Platí vztah:

$$\frac{\partial w}{\partial t} = 0 \tag{14}$$

Množina vzorů O, podle kterých se síť učí, je obvykle rozdělena na trénovací množinu, která se využívá ve fázi učení a testovací množinu, která se využívá ve fázi života. Pokud má systém k dispozici toto rozdělení množin, mluvíme o **učení s učitelem**, které poskytuje správnou informaci o požadovaném výstupu sítě. Pokud systém tuto informaci nemá, odvozuje si ji sám pomocí zpětné vazby, tento postup se nazývá **učení bez učitele** (například Kohonenovy samoorganizující se mapy). [20]

Nejznámějším algoritmem pro učení dopředných neuronových sítí je algoritmus Backpropagation, tedy metoda zpětného šíření chyby. Chyba se šíří přes všechny vrstvy až k první vrstvě sítě. Cílem je minimalizovat chybu založenou na druhé mocnině rozdílu mezi skutečným a očekávaným výstupem. Chybová funkce *E* je dána vztahem:

$$E = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

y_i … hodnota cílového atributu (očekávaný výstup)

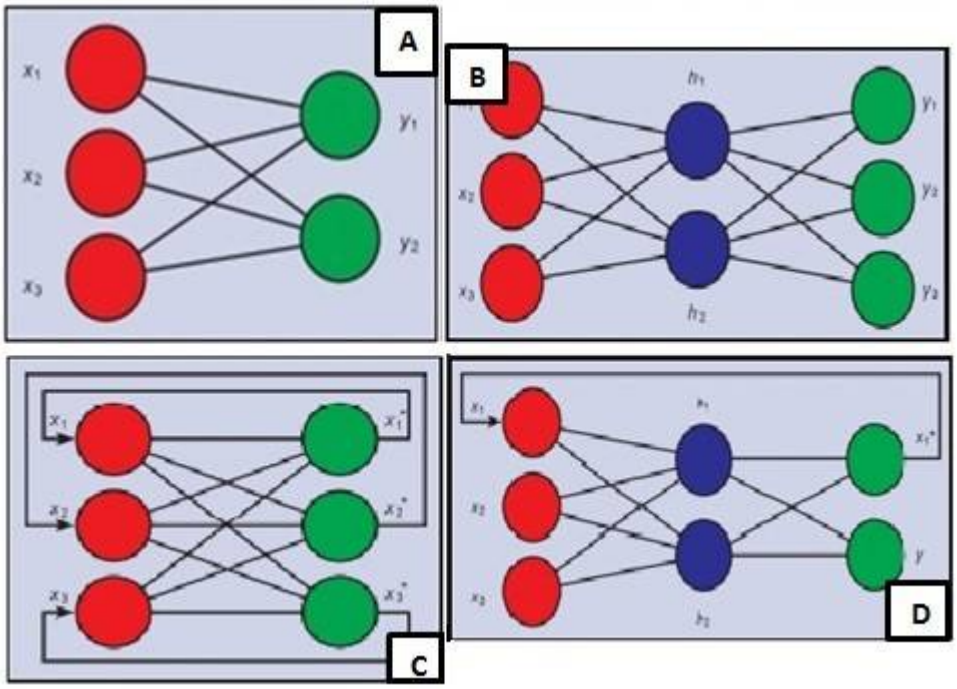
$\hat{y}_i$ … výsledek zařazení sítě (skutečný výstup)

3.3.1 Topologie neuronových sítí

Z hlediska průchodu informací neuronovou sítí rozlišujeme dva základní typy sítí:

- *Dopředné sítě* – signál se šíří jedním směrem od vstupu k výstupu
- *Rekurentní sítě* – díky zpětným vazbám (zapojení sítě v kruhu) se signál může šířit i opačným směrem

Struktura neuronové sítě se znázorňuje pomocí teorie grafů. Na obrázku č. 10 jsou zobrazeny některé typy sítí. Červeně jsou znázorněny vstupy, zeleně výstupy, modře skryté vrstvy. Prvním z nich (A) je základní dvouvrstvá síť se třemi vstupy a dvěma výstupy. Druhá síť (B) je hierarchická síť s jednou skrytou vrstvou, třetí síť (C) je autoasociativní síť, která má stejný počet vstupů i výstupů, přičemž z každého výstupu vede zpětná vazba, která vyvolává zpřesnění. Autoasociativní proces nemusí nikdy skončit nebo skončí v jednom z několika dovolených stavů. Poslední sítí (D) je rekurentní síť s jednou zpětnou vazbou. [16]



Obrázek č. 10: Struktury neuronových sítí

Zdroj: [16]

Existuje celá řada typů neuronových sítí. Každý typ je vhodný pro odlišný typ úlohy, přičemž některé neuronové sítě se mohou doplňovat. Některé typy sítí:

- *Vícevrstvé sítě (MPL)*

  Tato dopředná síť se skládá z tří a více vrstev neutronů, z čehož minimálně jedna vrstva je vrstva skrytá (viz obrázek č. 10B) Jsou to vrstvy:

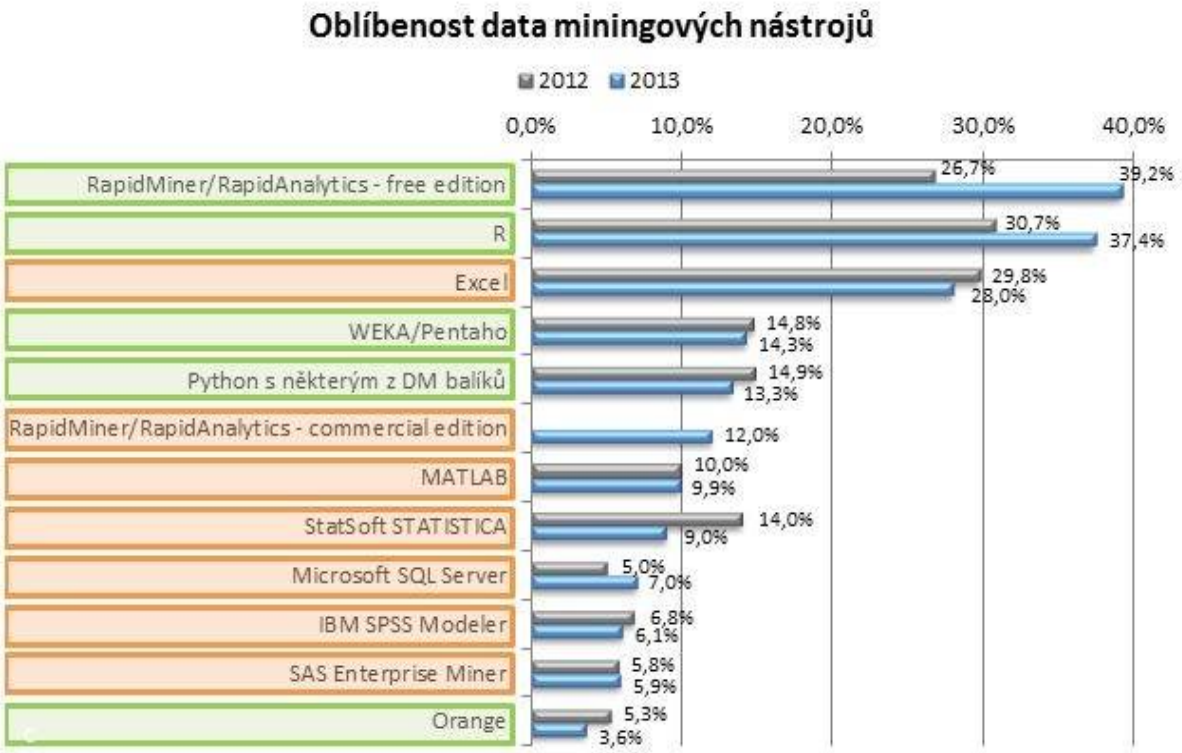
  - *Vstupní* – nejsou vzájemně propojeny, slouží jako vstupy pro další vrstvy
  - *Skryté* – bez skryté vrstvy nelze modelovat spojité funkce, čím větší je počet skrytých vrstev, tím obtížnější je učení sítě
  - *Výstupní* – převádí výsledky skryté vrstvy na výstupy sítě

- *Kohonenovy samoorganizující se mapy (SOM)* – k učení nepotřebují trénovací množinu (učení bez učitele), využívány pro roztřídění velkého počtu neznámých dat
- *Sítě s radiální bází (RBF)* – třívrstvé sítě s aktivační funkcí typu RBF (funkce určeny svým středem a pro argumenty se stejnou vzdáleností od tohoto středu dávají stejné funkční hodnoty)
- *Fuzzy neuronové sítě (FAM)* – určují řešení pomocí fuzzy množin

4 SOFTWAROVÉ NÁSTROJE PRO DATA MINING

Softwarových systémů pro řešení data miningových úloh je dnes již celá řada a neustále se vyvíjejí. Existují jak volně šiřitelné, tzv. open source systémy, které si uživatel může zdarma opatřit z internetu, tak i placené komerční systémy, které povětšinou nabízejí rozmanitější nabídku data miningových metod. Na obrázku č. 11 je znázorněn graf oblíbenosti produktů. Na otázku: „Který analytický data miningový nástroj jste využili v posledních dvanácti měsících pro tvorbu reálného projektu (květen 2012 - 2013)“, hlasovalo celkem 1880 uživatelů. Vybráno je 13 produktů, které se umístily v první dvacítce, přičemž výsledky jsou porovnávány s tím samým výzkumem, provedeným o rok dříve – tedy květen 2011 – 2012. Zeleně jsou označeny open source systémy, oranžově pak placené komerční systémy. [23]

Pro další usnadnění orientace v produktech lze kromě licence rozdělit systémy do několika skupin podle společných rysů. Jedním z těchto rysů je účel, se kterým je software vytvářen. Producenti na jedné straně vyvíjejí software, který slouží primárně pro data mining a na straně druhé je vývoj softwaru s jiným primárním využitím (například statistickým, matematickým atd.) umožňujícím data mining až v druhé řadě. [9]



Obrázek č. 11: Graf oblíbenosti data miningových nástrojů

Zdroj: upraveno podle [23]

V systémech, které byly vyvinuty speciálně pro potřeby data miningu, lze využívat jak jednoduché výpočty, tak výpočetně náročné metody, kvůli čemuž může proces dobývání znalostí trvat delší dobu. Uživatelské prostředí těchto nástrojů bývá grafické, na principu „drag and drop“ – v překladu „táhni a pusť“, který je dnes již znám z mnoha jiných aplikací či systémů (například MS Windows). Software je tudíž díky „drag and drop“ uživatelsky příjemný. V následujících dvou podkapitolách jsou uvedeny tři nejznámější komerční systémy a tři nejznámější open source systémy.

Kromě systémů, které jsou vytvářeny přímo za účelem podpory data miningu, se pro dobývání znalostí používají zhruba ve 30-40% i systémy, jejichž primární určení je jiné a data mining nabízejí pouze jako doplňkové rozšíření. To ostatně dokazuje i obrázek č. 11, podle kterého jsou na předních příčkách v oblíbenosti softwarových nástrojů i programy, které nejsou čistě „data miningové“. Jsou to především tyto skupiny softwaru: [9]

- *Databázové systémy*: Microsoft SQL Server, Oracle
- *Statistické systémy:* systémy SAS, SPSS, StatSoft Statistica, projekt R
- *Matematické systémy:* MATLAB

4.1 Nejznámější komerční systémy

- ***IBM SPSS Modeler (Clementine)***

V současné době je IBM SPSS Modeler jeden z nejvíce rozšířených systémů pro data mining. Původně byl vyvinut firmou Integral Solutions, pod názvem Clementine. V roce 1999 došlo ke sloučení Integral Solutions s věhlasnou firmou SPSS, zabývající se vývojem statistického softwaru. Od té doby probíhá vývoj softwaru pod hlavičkou SPSS. V roce 2009 firmu SPSS koupila světoznámá IBM, díky této změně se dostalo softwaru Clementine nového názvu používaného dodnes - IBM SPSS Modeler. [9] , [18]

Modeler nabízí širokou škálu vstupů dat do programu, umožňuje použití mnoha data miningových metod modelování, výsledné modely si uživatel může interaktivně zobrazovat pomocí tabulek, databází a různých vizualizačních prostředků. Systém důsledně vychází z metodiky CRISP-DM.

V této bakalářské práci byl pro modelování využit právě SPSS Modeler verze 14.2.

- ***SAS Enterprise Miner***

Další z data miningových systémů je vyvíjen od roku 1998 firmou SAS, tedy jednou z vůdčích firem na trhu se statistickým softwarem. Enterprise Miner je pouze jedním z mnoha modulů, které lze v rozsáhlém softwaru SAS zakoupit. Je však plně kompatibilní a snadno integrovatelný s dalšími produkty této firmy. Nabízí široké množství implementovaných metod, včetně možnosti využití vnitřního programovacího jazyka. Uživatelské prostředí je grafické, tím pádem uživatelsky příjemné. Celý systém je založen na metodologii SEMMA. [9] [29]

- ***StatSoft Statistica Data Miner***

Data Miner uzavírá trojici nejrozšířenějších nástrojů specializovaných pro data mining. Podobně jako předešlé dva produkty, je Statistica Data Miner příkladem systému, který vyvinula firma zabývající se původně statistikou. Uživatelské prostředí je zde opět grafické a přizpůsobivé uživateli. Systém nabízí pestrou škálu výběru data miningových metod. Výsledné modely lze vygenerovat jako spustitelný kód v různých programovacích jazycích, například VisualBasic, C++ (C#) nebo Java. Výhodou Data Mineru je možnost pracovat s pomocí systému WebStatistica, který umožňuje navrhovat a spravovat data miningové projekty na jiném počítači v prostředí internetového prohlížeče. [31]

4.2 Nejznámější volně šiřitelné systémy

- ***Rapid-I RapidMiner***

RapidMiner firmy Rapid-I je v současné době jedním z nejvyhledávanějších open source softwarů pro data mining. To ostatně dokládají i výsledky výzkumu na obrázku č. 11, kde se umístil na prvním místě v oblibě mezi uživateli. První verze programu vznikla na Fakultě umělé inteligence Dortmundské univerzity. Systém je díky svému úspěchu šířen i v placené verzi, která ho rozšiřuje o další možnosti. Systém je napsán kompletně v jazyce Java, což umožňuje spouštění téměř na jakémkoli operačním systému. [26]

- ***WEKA***

Systém WEKA byl původně vyvinut na novozélandské univerzitě Waikato pro akademické účely. Později byl přepsán do jazyka Java, byl upraven pro všeobecné použití a je šířen pomocí GNU licencí. Program nabízí nástroje pro předzpracování, klasifikaci, regresi, shlukování asociační pravidla a vizualizaci. V nedávné době se WEKA stala součástí

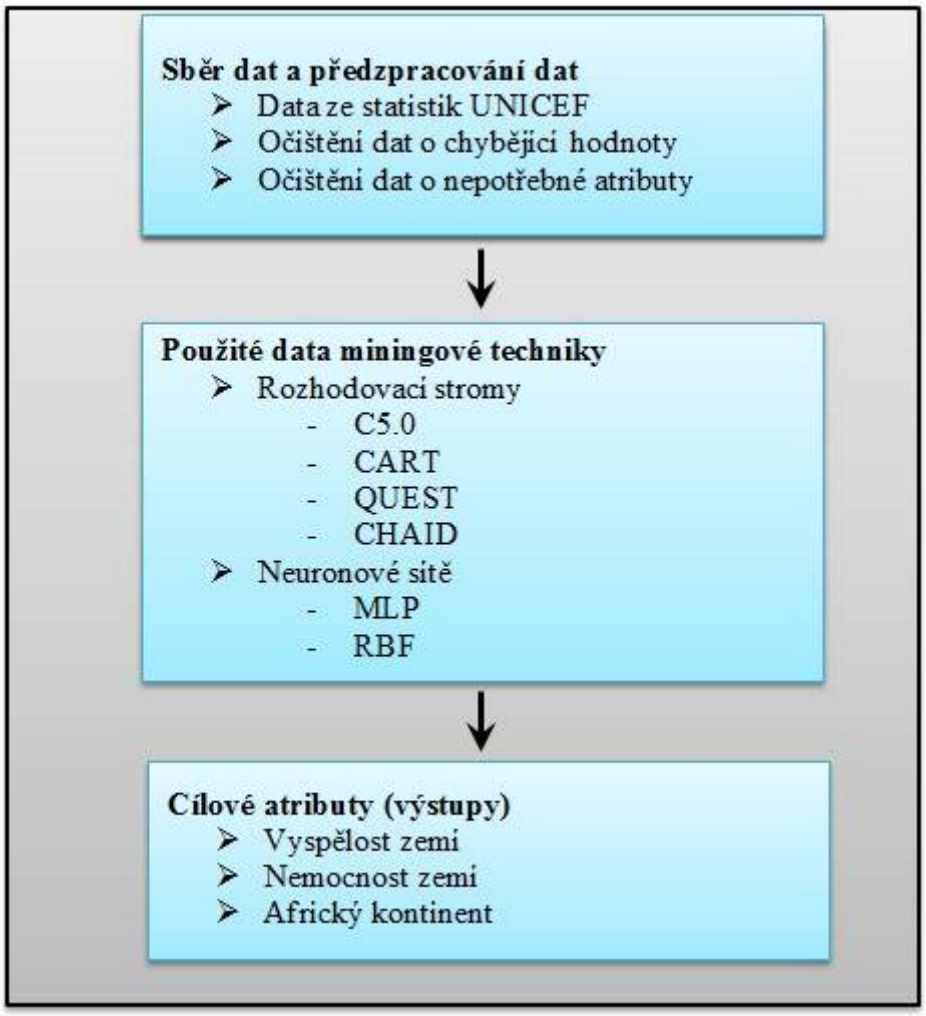
svobodného softwaru Pentaho, který se zaměřuje na business inteligenci a slučuje několik původně samostatných produktů. [34], [21]

- ***Orange***

Podobně jako dva předchozí open source programy byl Orange vyvinut na akademické půdě – na Fakultě počítačových a informačních věd slovinské univerzity v Lublani. Funguje pomocí vizuálního programování nebo skriptovacího jazyka Python. Nové funkce lze naprogramovat i pomocí jazyka C++. [5]

5 NÁVRH MODELU

Hlavním cílem modelování je vytvořit model tří klasifikačních úloh, jejichž přesné znění bude upřesněno v následující kapitole. Klasifikace bude provedena pomocí rozhodovacích stromů a neuronových sítí. Na obrázku č. 12 je zobrazen návrh modelu. Tento návrh klasifikačního modelu bude prováděn v softwaru IBM SPSS Modeler verze 14.2.



Obrázek č. 12: Návrh modelu

Zdroj: vlastní zpracování

5.1 Formulace problému a sběr dat

Pro splnění cílů modelování byl nejprve proveden sběr dat. Data pro modelování pocházejí ze statistik uvedených na webových stránkách Dětského fondu OSN – UNICEF [32]. UNICEF je hlavní světovou organizací, zabývající se ochranou a zlepšováním životních podmínek dětí a matek. Fond působí ve 157 zemích světa a jeho prostřednictvím poskytují

pracovníci i dobrovolníci humanitní, materiální a finanční podporu státům, které ji potřebují nejvíce. [17]

Hlavní cíle modelování jsou tři klasifikační úlohy:

- *Rozdělení zemí v atributu „Vyspělost zemí“ se třemi skupinami podle vyspělosti*
- *Rozdělení zemí v atributu „Nemocnost zemí“ se třemi skupinami podle nemocnosti*
- *Oddělení afrických zemí od zemí jiných kontinentů (atribut „Africký kontinent“)*

Toto rozdělení může sloužit jako podklad při rozhodování pro rozdělování humanitních, materiálních a finančních prostředků do nejpotřebnějších oblastí. Jelikož je zde předpoklad, že Afrika bude potřebovat nejvíce pomoci, protože zde pravděpodobně bude nejvíce nerozvinutých a nemocných zemí, jsou africké státy odděleny od zemí ostatních kontinentů. V práci bude proto znázorněno, které atributy souvisí s příslušností států k africkému kontinentu a ke zbytku světa. Každá úloha bude mít v Modeleru svůj vlastní stream.

5.2 Předzpracování

V prostředí MS Excel byla původní data (198 údajů - zemí s 44 atributy) předzpracována do požadované podoby a poté na novém listu očištěna o chybějící hodnoty vymazáním některých zemí a atributů. Tyto chybějící hodnoty by se sice později v Modeleru dala nahradit (např. pomocí průměru, mediánu, modu, atd.), ale jelikož jsou v souboru dat hodnoty od vyspělých zemí až po velmi chudé země, mohly by být nahrazené hodnoty zkreslené a znehodnocovaly by celý datový soubor. Po očištění vzniklo konečných 123 údajů s 38 atributy. K bližšímu seznámení s datovým souborem byla do nového listu provedena ještě popisná statistika. List s názvem očištěná data byl převeden do formátu *.csv. Tento soubor bude sloužit jako zdroj pro práci v Modeleru. Datový slovník k atributům je v tabulce č. 7.

Tabulka č. 7: Datový slovník – 1. část

Název atributu	Typ dat	Rozsah hodnot	Popis dat
Název	Nominal	[Albánie; Zimbabwe]	Jednotlivé státy, které UNICEF monitoruje
Úmrtnost dětí mladších 5 let	Continuous	[3; 178]	Úmrtnost do 5 let života - na 1000 dětí
Kojenecká úmrtnost	Continuous	[2; 114]	Úmrtnost do 1. roku života - na 1000 dětí
Novorozenecká úmrtnost	Continuous	[2; 48]	Úmrtnost do 28 dní života – na 1000 živě narozených dětí
Celková populace (tisíce)	Continuous	[104,06; 1341335,15]	Celková populace státu – v tisících
Počet narozených dětí	Continuous	[2,82; 27165,24]	Počet narozených dětí v tisících za rok
Počet mrtvých dětí do 5 let	Continuous	[0; 1696]	Počet mrtvých dětí do 5 let v tisících za rok
Průměrná ročnímíra růstu HDP (%)	Continuous	[-3,2; 16,3]	Průměrná roční míra růstu HDP na osobu – v %
Hrubý národní důchod na obyvatele (v US $)	Continuous	[160; 73060]	Hrubý národní důchod na osobu – v dolarech
Celkem: % populace využívající kvalitní zdroje pitné vody	Continuous	[38; 100]	Celkové procento populace mající přístup ke kvalitnímu zdroji pitné vody
Města: % populace využívající kvalitní zdroje pitné vody	Continuous	[52; 100]	Procento populace žijící ve městech, mající přístup ke kvalitnímu zdroji pitné vody
Venkov: % populace využívající kvalitní zdroje pitné vody	Continuous	[26; 100]	Procento populace žijící na venkově, mající přístup ke kvalitnímu zdroji pitné vody
Celkem: % populace využívající kvalitní hygienické prostředky a zařízení	Continuous	[9; 100]	Celkové procento populace využívající kvalitní hygienické prostředky a zařízení
Města: % populace využívající kvalitní hygienické prostředky a zařízení	Continuous	[15; 100]	Procento populace žijící ve městech, využívající kvalitní hygienické prostředky a zařízení
Venkov: % populace využívající kvalitní hygienické prostředky a zařízení	Continuous	[3; 100]	Procento populace žijící na venkově, využívající kvalitní hygienické prostředky a zařízení
Odhadovaný počet osob žijících s HIV	Continuous	[0,1; 5600]	Odhadovaný počet osob žijících s HIV ve všech věkových kategoriích – v tisících
Muži: Míra gramotnosti mládeže (%)	Continuous	[47; 100]	Míra gramotností mladých mužů (15-24 let) v % na danou skupinu lidí
Ženy: Míra gramotnosti mládeže (%)	Continuous	[23; 100]	Míra gramotností mladých žen (15-24 let) v % na danou skupinu lidí
Celková míra gramotnosti dospělých	Continuous	[26; 100]	Celková míra gramotnosti dospělých v %
Míra zapsaných dětí na základní školu	Continuous	[31; 100]	Míra zapsaných dětí na základní školu – v % na 100 dětí ve školním věku
Používání moderních technologií - internet	Continuous	[0; 78]	Využívání internetu – počet na 100 obyvatel
Používání moderních technologií - mobily	Continuous	[1; 185]	Využívání mobilních telefonů – počet na 100 obyvatel

Zdroj: vlastní zpracování

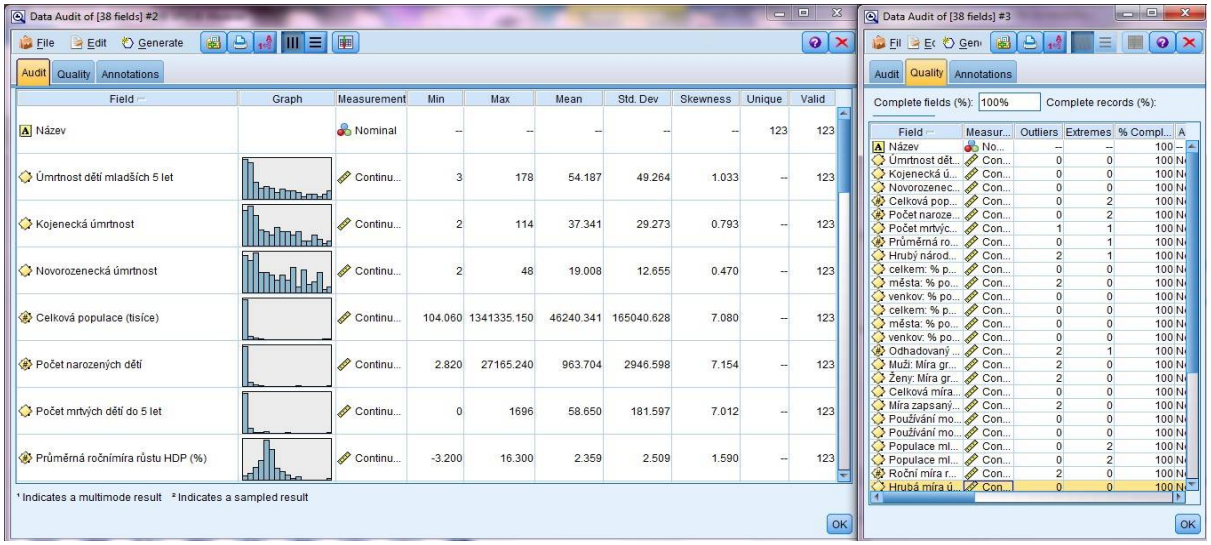
Tabulka č. 7: Datový slovník – 2. část

Název atributu	Typ dat	Rozsah hodnot	Popis dat
Populace mladší 18 let (v tis.)	Continuous	[46; 447309]	Populace mladší 18 let – v tisících
Populace mladší 5 let (v tis.)	Continuous	[14; 127979]	Populace mladší 5 let – v tisících
Roční míra růstu obyvatelstva (v %)	Continuous	[-1,1; 7,1]	Procentuální vyjádření poměru mezi počtem narozených a zemřelých, může být i záporná.
Hrubá míra úmrtnosti	Continuous	[1; 17]	Počet zemřelých v populaci během 1 roku vydělený celkovým počtem obyvatel a vynásobený tisícem - výsledek je v tisících
Hrubá míra porodnosti	Continuous	[9; 49]	Počet narozených v populaci během 1 roku vydělený celkovým počtem obyvatel a vynásobený tisícem - výsledek je v tisících
Očekávaná délka života	Continuous	[47; 81]	Naděje na dožití při narození (v letech), průměr pro ženy a muže celkem
Úhrnná plodnost	Continuous	[1,1; 7,1]	Průměrný počet dětí, které se narodí jedné ženě během jejího života
% populace žijící v urbanizovaných oblastech (městech)	Continuous	[11; 98]	Procento populace žijící v urbanizovaných oblastech (městech)
Průměrná roční míra růstu obyvatelstva žijících ve městech (%)	Continuous	[-1,3; 8,2]	Průměrná roční míra růstu obyvatelstva žijících ve městech – v %
Počet narozených dětí na 1000 dívek ve věku 15-19 let	Continuous	[4; 199]	Počet narozených dětí dívkám ve věku 15-19 let – na 1000 dívek
Podíl mladistvých na celkové populaci	Continuous	[8; 26]	Podíl mladistvých na celkové populaci
Prenatální péče	Continuous	[28; 100]	Procento žen ve věku 15-49 let, které se zúčastnili alespoň jednou během těhotenství kvalifikované zdravotní prohlídky
Porodní péče	Continuous	[6; 100]	Procento porodů za účasti kvalifikovaných zdravotních pracovníků - lékaři, zdravotní sestry, porodní asistentky, apod.
Vytváření imunity dětského organismu před DPT (%)	Continuous	[45; 99]	Procento 12-23 měsíčních dětí, u kterých se vytváří imunita organismu před DPT (= záškrt, černý kašel a tetanus)
Výskyt tuberkulózy	Continuous	[3; 1287]	Výskyt tuberkulózy na 100 000 obyvatel
ID kontinentu	Nominal	[1,2,3,4,5,6]	Kontinent, na kterém stát leží (1=Evropa, 2=Asie, 3=Afrika, 4=S.Amerika, 5=J.Amerika, 6=Austrálie)

Zdroj: vlastní zpracování

5.3 Modelování

V prostředí softwaru IBM SPSS Modeler byl po načtení dat pomocí uzlu *Var file* proveden datový audit a kontrola kvality dat viz obrázek č. 13.



Obrázek č. 13: Datový audit

Zdroj: vlastní zpracování

Aby byla zřejmá určitá představa o závislostech jednotlivých atributů, pomocí uzlu *Statistics* byla spočítána korelace mezi atributy. Některé korelace jsou uvedeny v Tabulce č. 8. Korelace se pohybuje v intervalu <-1; 1>. Čím více se blíží hodnota od nuly k jedničce, tím větší mezi sebou mají atributy závislost – roste-li hodnota prvního atributu, roste i hodnota druhého atributu. Čím více se hodnota korelace blíží od nuly k minus jedničce, tím větší je závislost mezi atributy – roste-li hodnota prvního atributu, klesá hodnota druhého. Korelace v uzlu statistics se zdají být celkem logické, tudíž lze předpokládat správnost dat.

Tabulka č. 8: Korelace některých atributů

Názvy atributů		**Korelační koeficient**
Úmrtnost dětí mladších 5 let	% populace používající kvalitní hygienické prostředky	-0,806
Porodní péče	Úmrtnost dětí mladších 5 let	-0,735
% populace používající kvalitní zdroje pitné vody	Hrubá míra porodnosti	-0,755
Ženy: míra gramotnosti mládeže	Úhrnná plodnost	-0,780
Podíl mladistvých na celkové populaci	Požívání moderních technologií – internet	-0,817

Zdroj: vlastní zpracování

5.3.1 Vytváření nových atributů

Pomocí uzlu *Derive* bylo vytvořeno několik nových atributů, potřebných pro další zpracování. Jedná se o cílové atributy pro klasifikaci.

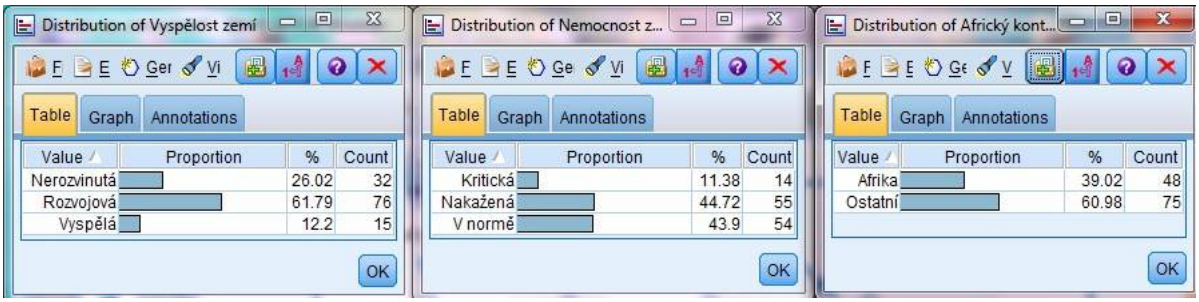
Prvním z nich byl atribut „Vyspělost zemí“. Pro potřebu modelování se tento atribut bude porovnávat podle atributu Hrubý národní důchod tak, že země s HNP na hlavu nižším než 1 000 dolarů byly zařazeny do skupiny „Nerozvinuté“, země s HNP do 10 000 dolarů do skupiny „Rozvojové“ a země s více jak 10 000 dolary HNP do skupiny „Vyspělé“. Konkrétní meze byly navrhnuty nahodile, pouze pro potřeby tohoto modelování.

Dalším atributem byla „Nemocnost zemí“. Skupiny „Kritická“, „Nakažená“ a „V normě“ byly stanoveny následovně: Jestliže výskyt tuberkulózy bude vyšší než 350, země bude ve skupině „Kritická“. Jestliže výskyt tuberkulózy bude vyšší než 80, země bude v kategorii „Nakažená“. Pokud budou hodnoty nižší, země bude ve skupině „V normě“. Meze jsou opět stanoveny bez ověřených podkladů, pouze pro potřeby modelování.

S předešlým atributem souvisí další nový atribut „HIV na 100 000 obyvatel“. Byl vytvořen proto, aby data byla v přesněji měřitelných relativních hodnotách.

Další atribut s názvem „Africký kontinent“ je rozdělení zemí podle toho zda se nacházejí v Africe, ty pak mají příznak „Afrika“, ostatní země mají příznak „Ostatní“. Příznak je udělen na základě atributu „ID kontinentu“. Příslušnost ke kontinentům je všeobecně známá.

Na obrázku č. 14 je vidět jak se země rozřadily do skupin.



Obrázek č. 14: Graf Distribution pro nové atributy

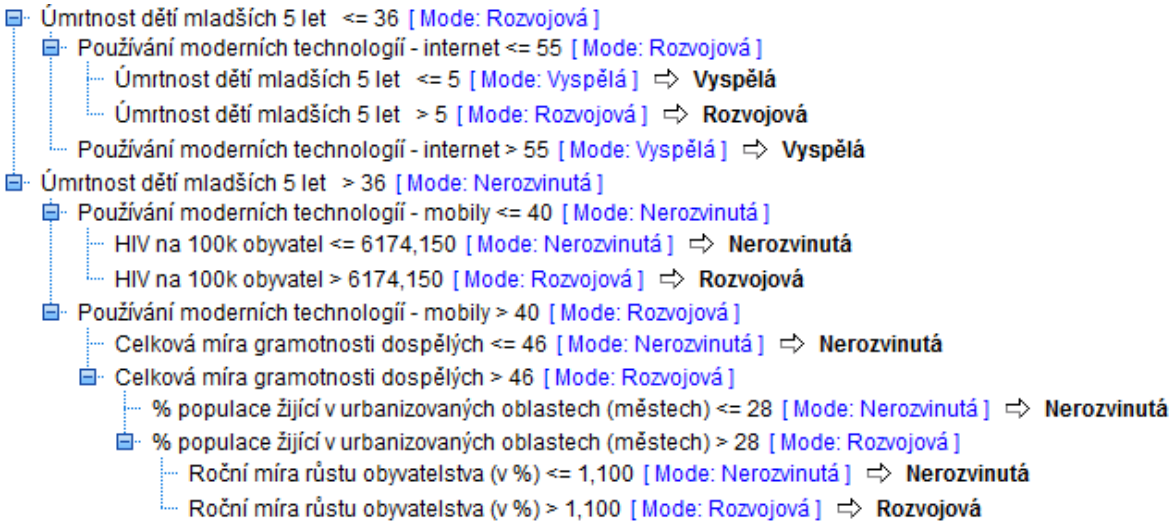
Zdroj: vlastní zpracování

5.3.2 Úloha „Vyspělost země“

První úlohou je klasifikace podle vyspělosti země. Pomocí uzlu *Filter* byly zvoleny atributy, které dále budou figurovat v analýze. Uzel *Partition* rozdělil datový soubor

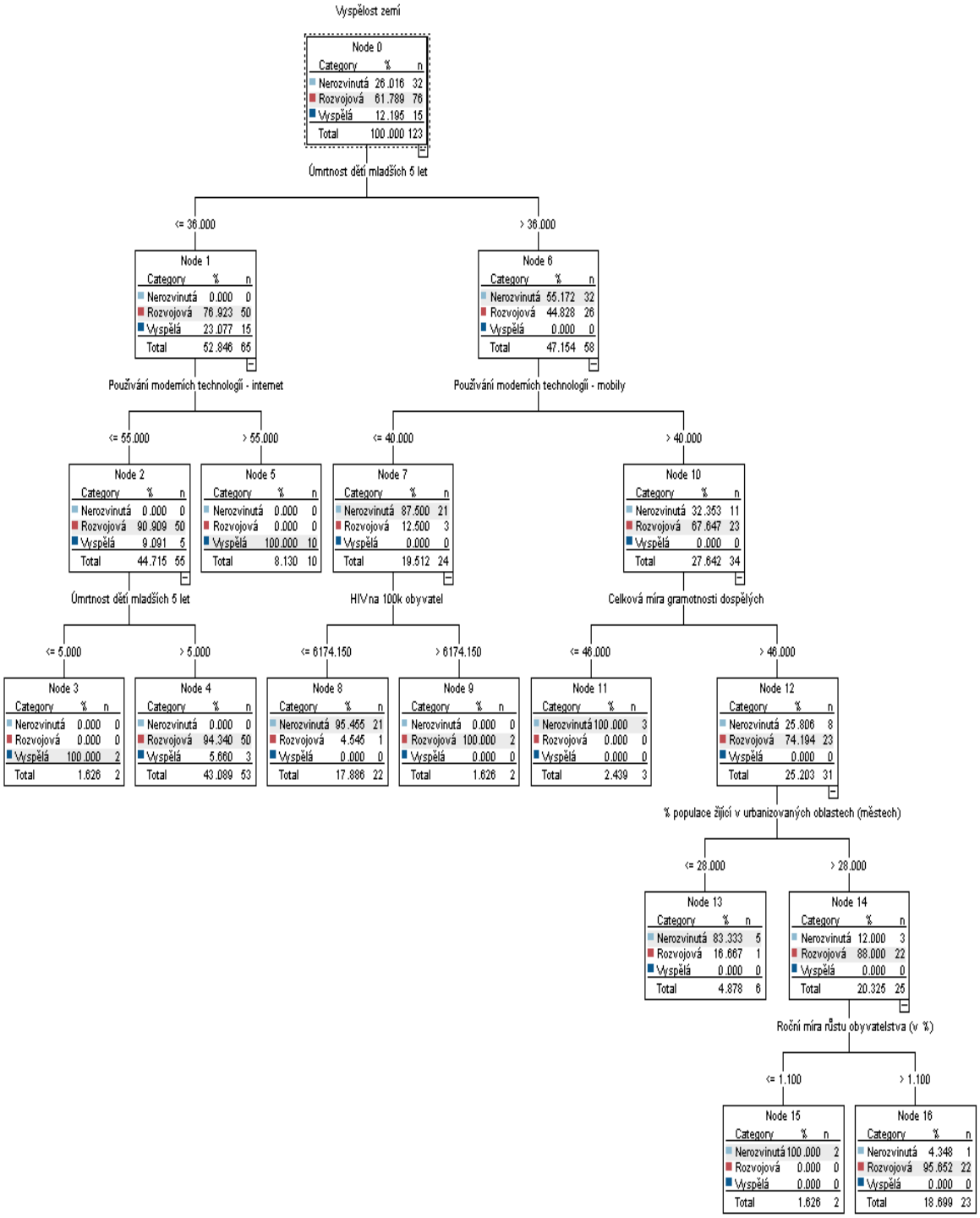
na trénovací a testovací množinu v poměru 60:40. Poté byl pomocí uzlu *Type* zvolen cíl klasifikace (atribut Vyspělost zemí). Při modelování bylo použito zjednodušení, které SPSS Modeler nabízí, a to automatické určení nejvhodnějších technik pro danou úlohu (uzel *Auto clasiffier*). Ten vyhodnotil jako nejvhodnější použití rozhodovacího stromu s algoritmem C5.0 a neuronovou síť. Tyto metody tedy byly uplatněny pro vytvoření modelu.

Algoritmus stromu ***C5.0*** rozhodl, že nejdůležitější atributy jsou: úmrtnost dětí do 5 let, používání moderních technologií – internet, používání moderních technologií – mobily, HIV na 100 tisíc obyvatel, celková míra gramotnosti dospělých, procento populace žijící v urbanizovaných oblastech a roční míra růstu obyvatelstva. Rozhodovací pravidla tohoto stromu jsou na obrázku č. 15 a struktura stromu je na obrázku č. 16.



Obrázek č. 15: Úloha „Vyspělost země“ - rozhodovací pravidla algoritmu C5.0

Zdroj: vlastní zpracování



Obrázek č. 16: Úloha „Vyspělost země“ - struktura rozhodovacího stromu C5.0

Zdroj: vlastní zpracování

Pro větší přehlednost o výsledcích klasifikace byl použit uzel *Analysis*, pomocí něhož je možné udělat si představu o správnosti klasifikačního modelu. Analysis porovnává výsledek

algoritmu C5.0 a původního rozdělení vyspělosti zemí jak na trénovacíh, tak na testovacích datech. Rozhodovací strom C5.0 je podle závěrečné analýzy přesný ve 100% případů testovacích dat. Analýza výsledků je na obrázku č. 17. Nejprve se porovnává přesnost celkové klasifikace, poté klasifikace po třídách.

Results for output field Vyspělost zemí

Overall Results

Comparing $C-Vyspělost zemí with Vyspělost zemí

'Partition'	1_Training		2_Testing	
Correct	66	91,67%	51	100%
Wrong	6	8,33%	0	0%
Total	72		51	

Output field Vyspělost zemí, splitting by field Vyspělost zemí

Vyspělost zemí = Nerozvinutá

Comparing $C-Vyspělost zemí with Vyspělost zemí

'Partition'	1_Training		2_Testing	
Correct	14	93,33%	17	100%
Wrong	1	6,67%	0	0%
Total	15		17	

Vyspělost zemí = Rozvojová

Comparing $C-Vyspělost zemí with Vyspělost zemí

'Partition'	1_Training		2_Testing	
Correct	43	95,56%	31	100%
Wrong	2	4,44%	0	0%
Total	45		31	

Vyspělost zemí = Vyspělá

Comparing $C-Vyspělost zemí with Vyspělost zemí

'Partition'	1_Training		2_Testing	
Correct	9	75%	3	100%
Wrong	3	25%	0	0%
Total	12		3	

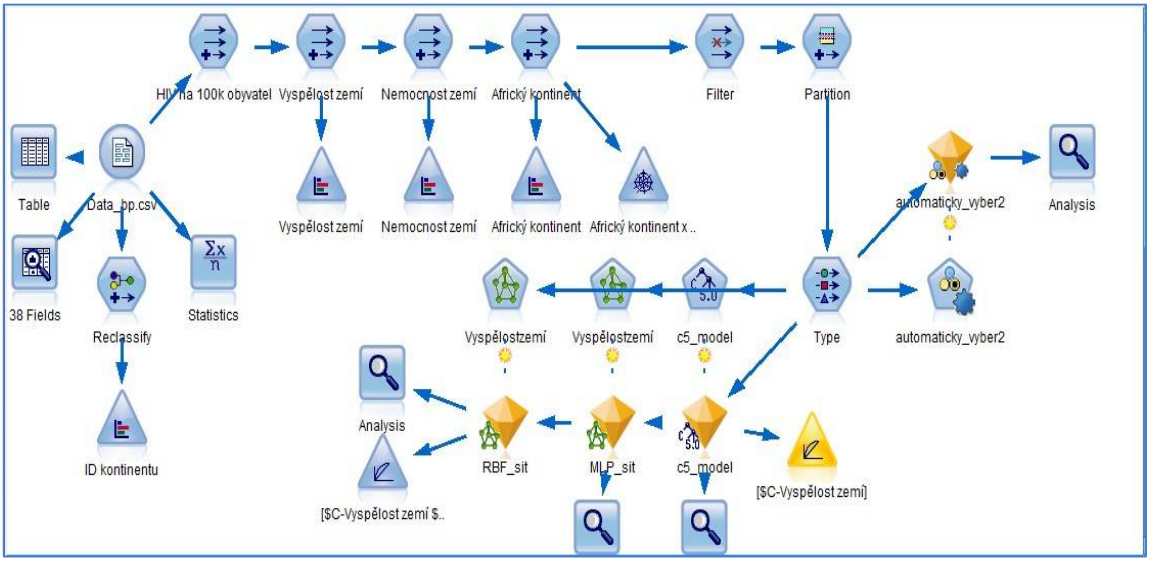
Obrázek č. 17: Úloha „Vyspělost země“ - výsledky algoritmu C5.0

Zdroj: vlastní zpracování

Jako další byly použity **neuronové sítě**, konkrétně typ sítě MLP a RBF (vícevrstvá a s radiální bází). Z porovnání výsledků těchto dvou typů vychází lépe síť MLP, avšak ani jedna ze sítí nedosahuje podle uzlu *Analysis* tak přesných výsledků jako algoritmus C5.0 (MLP 80,39% a RBF 76,47%).

Výsledek úlohy „Vyspělost země“: optimální je použití rozhodovacího stromu s algoritmem C5.0, s přesností 100%.

Celý stream k této úloze je na obrázku č. 18.



Obrázek č. 18: Stream k úloze „Vyspělost země“

Zdroj: vlastní zpracování

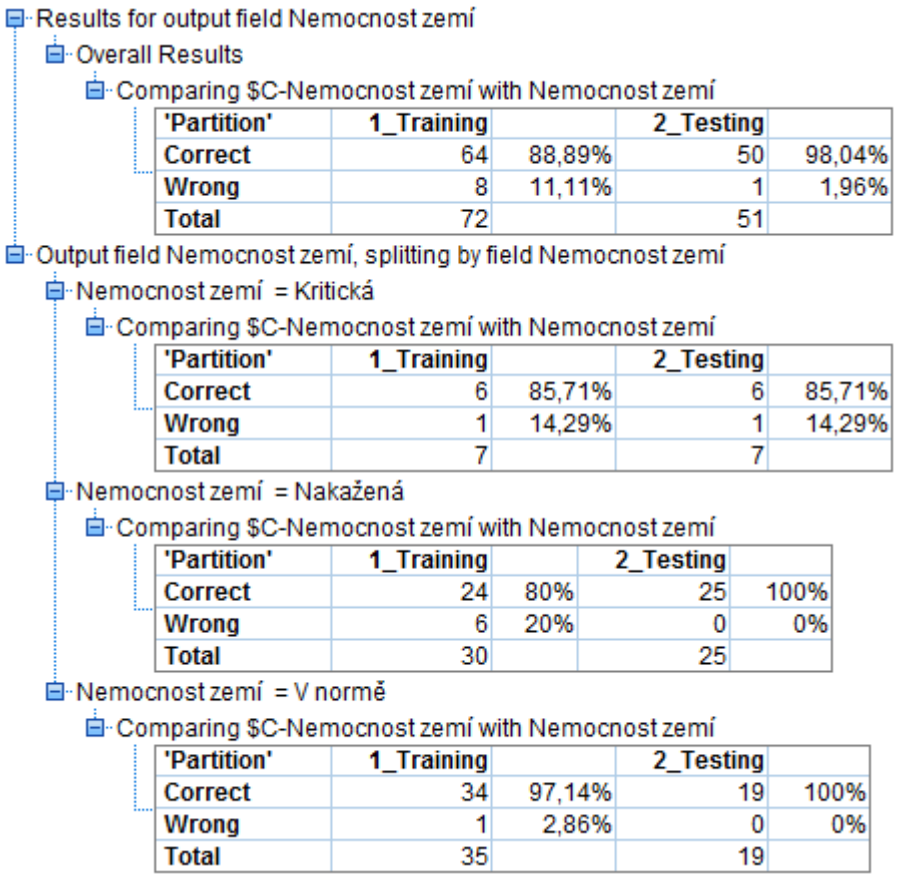
5.3.3 Úloha „Nemocnost země“

Postup dalších úloh je analogický s první úlohou „Vyspělost země“. Tentokrát byl cíl klasifikace rozřadit země podle nemocnosti. Opět byl použit uzel *Auto classiffier* k detekci nejvhodnějších metod řešení. Podle něj byl opět použit rozhodovací stromy s algoritmem ***C5.0***, dále pak strom s algoritmem ***CART*** a pro porovnání se stromy i **neuronová síť** typu MLP. Rozhodovací pravidla algoritmu C5.0 jsou na obrázku č. 19.

Obrázek č. 19: Úloha „Nemocnost země“ – rozhodovací pravidla algoritmu C5.0

Zdroj: vlastní zpracování

Dle výsledků uzlu *Analysis* byl nejlepším možným řešením opět algoritmus C5.0 s přesností 98,04% na testovacích datech. Algoritmus CART a neuronová síť MLP dosáhli nepříliš přesvědčivých výsledků, jejich přesnost byla pouze 75,55% (CART) a 68,62% (MLP). Výsledky algoritmu C5.0 v porovnání s původním atributem Nemocnost zemí je na obrázku č. 20.



'Partition'	1_Training		2_Testing	
Correct	64	88,89%	50	98,04%
Wrong	8	11,11%	1	1,96%
Total	72		51	

'Partition'	1_Training		2_Testing	
Correct	6	85,71%	6	85,71%
Wrong	1	14,29%	1	14,29%
Total	7		7	

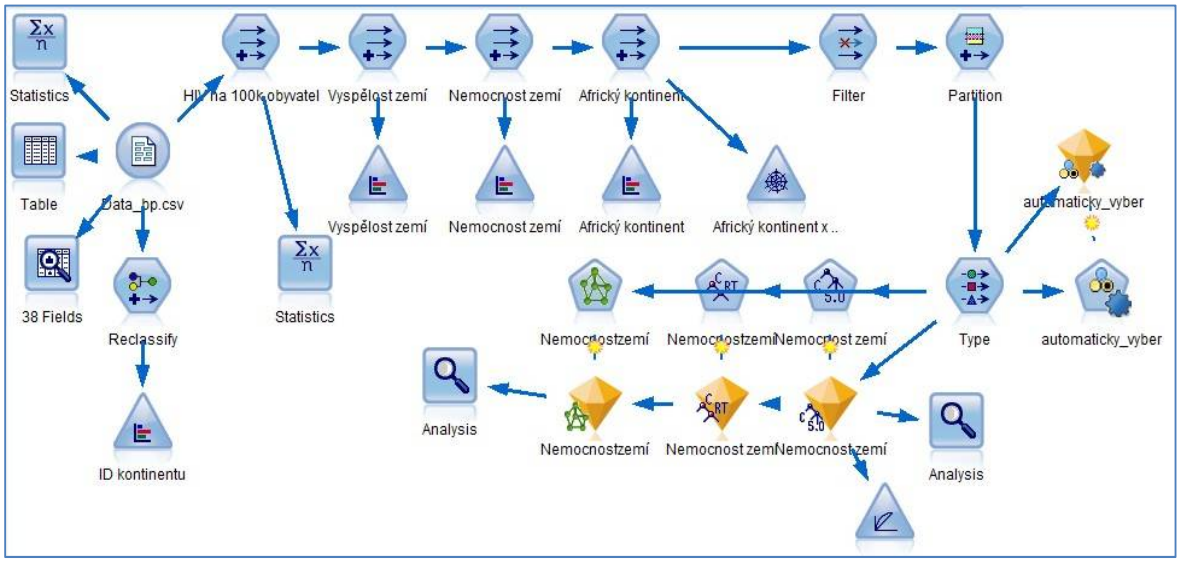
'Partition'	1_Training		2_Testing	
Correct	24	80%	25	100%
Wrong	6	20%	0	0%
Total	30		25	

'Partition'	1_Training		2_Testing	
Correct	34	97,14%	19	100%
Wrong	1	2,86%	0	0%
Total	35		19	

Obrázek č. 20: Úloha „Nemocnost země“ – výsledky algoritmu C5.0

Výsledek úlohy „Nemocnost země“: optimální je použití rozhodovacího stromu s algoritmem C5.0, s přesností 98,04%.

Výsledný stream k úloze „Nemocnost země“ je na obrázku č. 21.

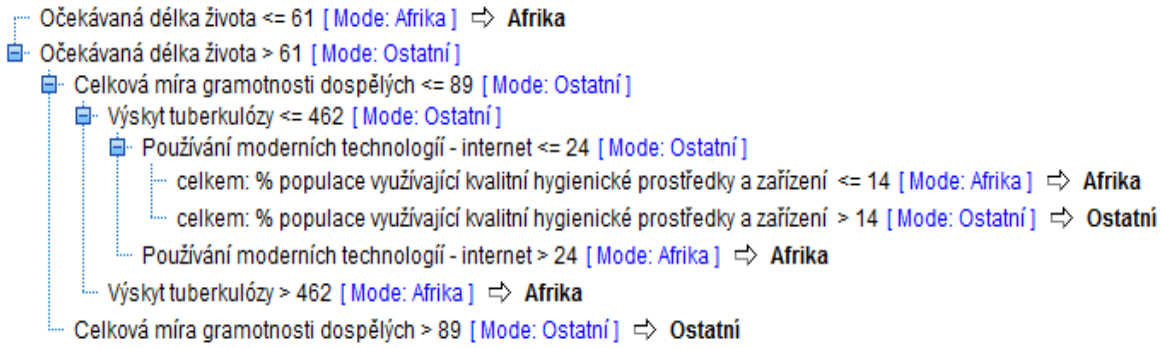


Obrázek č. 21: Stream k úloze „Nemocnost země“

Zdroj: vlastní zpracování

5.3.4 Úloha „Afrika“

V poslední úloze s názvem Afrika se opět postupuje stejným způsobem jako v předešlých úlohách. Uzlem *Auto classifier* byly vybrány rozhodovací stromy s algoritmem ***C5.0, CHAID*** a ***QUEST***. Opět byla pro porovnání použita i ***neuronová síť*** typu MLP. Pravidla rozhodovacího stromu C5.0 jsou na obrázku č. 22.



Obrázek č. 22: Úloha „Afrika“ – rozhodovací pravidla algoritmu C5.0

Zdroj: vlastní zpracování

Ve výstupovém uzlu *Analysis* je nepřesnější metodou algoritmus C5.0, kde je 96,08% shoda s původním atributem Africký kontinent, dále pak algoritmus CHAID s 90,2%, algoritmus QUEST s 84,31% a neuronová síť s 80,39%. Výsledné porovnání s původním atributem Africký kontinent je na obrázku č. 23.

Results for output field Africký kontinent

Overall Results

Comparing $C-Africký kontinent with Africký kontinent

'Partition'	1_Training		2_Testing	
Correct	70	97,22%	49	96,08%
Wrong	2	2,78%	2	3,92%
Total	72		51	

Output field Africký kontinent, splitting by field Africký kontinent

Africký kontinent = Afrika

Comparing $C-Africký kontinent with Africký kontinent

'Partition'	1_Training		2_Testing	
Correct	25	96,15%	21	95,45%
Wrong	1	3,85%	1	4,55%
Total	26		22	

Africký kontinent = Ostatní

Comparing $C-Africký kontinent with Africký kontinent

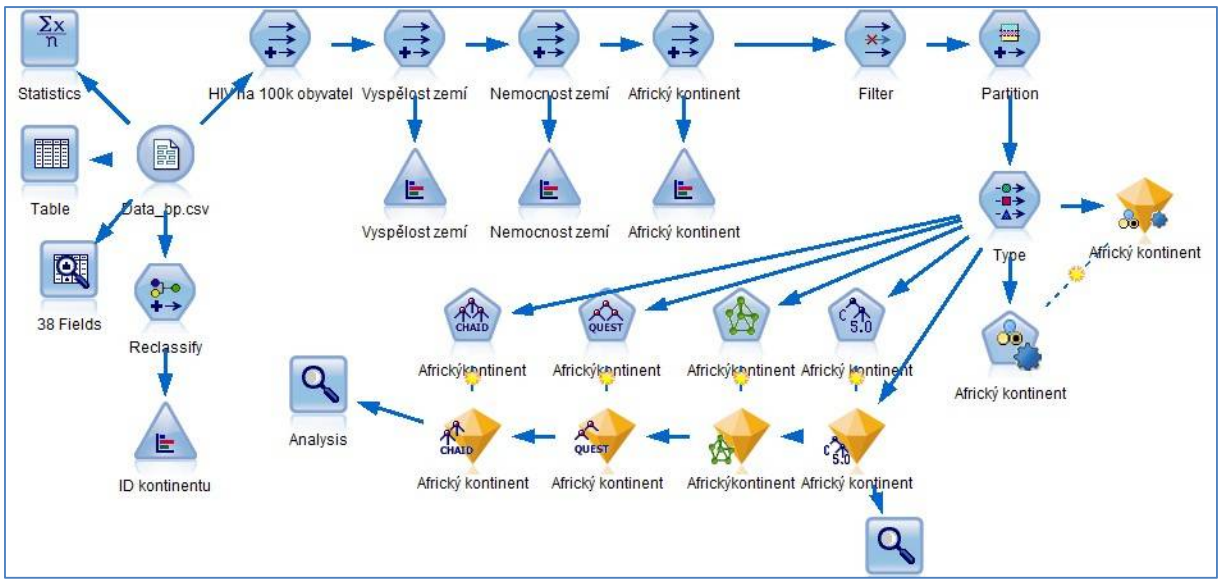
'Partition'	1_Training		2_Testing	
Correct	45	97,83%	28	96,55%
Wrong	1	2,17%	1	3,45%
Total	46		29	

Obrázek č. 23: Úloha „Afrika“ – rozhodovací pravidla algoritmu C5.0

Zdroj: vlastní zpracování

Výsledek úlohy „Afrika“: optimální je použití rozhodovacího stromu s algoritmem C5.0, s přesností 96,08%.

Celý stream třetí úlohy je na obrázku č. 24.



Obrázek č. 24: Stream k úloze „Afrika“

Zdroj: vlastní zpracování

ZÁVĚR

Cílem bakalářské práce bylo vytvořit model pro tři klasifikační úlohy, který bude schopen spolehlivě klasifikovat data do cílových atributů. Pro splnění tohoto cíle byl proveden sběr dat ze statistik světové organizace UNICEF, zabývající se ochranou a zlepšováním životních podmínek dětí a matek po celém světě. Na základě těchto dat byly vytyčeny tři úlohy, které měly za úkol rozdělit data do skupin podle souvislostí mezi nimi. Tyto atributy byly nejprve vytvořeny, přiřazeny k datovému souboru a následně zvoleny jako cílové. Sloužily také jako kontrola správnosti pro trénovací množinu dat. Pro samotné modelování byly zvoleny rozhodovací stromy s algoritmy C5.0, CART, CHAID a QUEST a neuronové sítě. Po sérii experimentů byla v každé úloze vyhodnocena nejpřesnější technika, tj. ta, která roztřídila data na testovací množině co nejpodobněji původním atributům (shoda na nejvíce procent). Ve všech třech úlohách vynikl jako nejlepší rozhodovací strom s algoritmem C5.0, který vykazoval přesnost na více než 95%. U úlohy „Vyspělost země“ měl tento algoritmus dokonce 100% přesnost. Data byla modelována pomocí softwaru IBM SPSS Modeler v. 14.2.

Dalším cílem bylo seznámení s pojmem data mining, jeho úlohami a s technikami, které slouží k analýze dat, především pak ke klasifikaci. Tohoto cíle bylo dosaženo v kapitolách 1 - 3, kde jsou uvedeny informace potřebné k vytvoření teoretického aparátu.

Závěrem lze říci, že díky těmto výsledkům byl cíl o vytvoření co nejlepšího klasifikačního modelu splněn a v případě aktualizace statistik UNICEFu a přidání nových případů by byla jejich klasifikace poměrně přesná.

POUŽITÁ LITERATURA

[1] BERKA, Petr. *Dobývání znalostí z databází*. 1. vyd. Praha: Academia, 2003. 365 s. ISBN 80-200-1062-9.

[2] CRM a Business Intelligence. *CRM portál: zpravodaj z oblasti CRM* [online]. 2012 [cit. 2013-04-03]. Dostupné z: http://www.crmportal.cz/redakcni/crm-a-business-intelligence

[3] ECML PKDD*: European Conference on Machine Learning and Practice of Knowledge Discovery in Databases* [online]. 2013 [cit. 2013-03-19]. Dostupné z: http://www.ecmlpkdd2013.org/

[4] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory. From data mining and knowledge discovery in databases. *AI magazine*. [online]. 1996, vol. 17. s. 37-54 [cit. 2013-03-15]. ISSN 0738-4602. Dostupné z: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

[5] Google Summer of Code. *Orange*. [online]. [cit. 2013-06-12]. Dostupné z: http://orange.biolab.si/trac/wiki/GSoC

[6] HAN, Jiawei. *Data mining concepts and techniques: Classification and Prediction*. [online]. Urbana and Champaign: University of Illinois. 2006 [cit. 2013-03-18]. Dostupné z: http://134.208.3.165/course/2006/Fall/Data_mining/06.pdf

[7] HAN, Jiawei. *Data mining concepts and techniques: Introduction*. [online]. Urbana and Champaign: University of Illinois. 2006 [cit. 2013-03-18]. Dostupné z: http://134.208.3.165/course/2006/Fall/Data_mining/01.pdf

[8] HOLČÍK, Jiří. *Analýza a klasifikace dat*. [online] 1. vyd. Brno: Cerm, 2012. 111 s. ISBN 978-80-7204-793-2. Dostupné z: http://www.iba.muni.cz/res/file/ucebnice/holcik-analyza-klasifikace-dat.pdf

[9] HOLEŇA, Martin. *Statistické aspekty dobývání znalostí z dat*. 1. vyd. Praha: Karolinum, 2006. 106 s. ISBN 80-246-1186-4.

[10] CHALUPNÍK, Vitalij. Biologické algoritmy – neuronové sítě. *Root.cz: Informace nejen ze světa Linuxu*. 25.4.2012 [cit. 2013-06-22]. Dostupné z: http://www.root.cz/clanky/biologicke-algoritmy-4-neuronove-site/

[11] CHAPMAN, Pete a kolektiv. *CRISP-DM 1.0: Steb-by-step data mining guide*. [online]. 2000 [cit. 2013-04-14]. Dostupné z:

ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[12] IBM SPSS Modeler Help. *Decision tree nodes*.

[13] IBM SPSS Modeler Help. *C5.0 node*.

[14] KOMPRDOVÁ, Klára. *Rozhodovací stromy a lesy*. [online]. 1. vyd. Brno: CERM, 2012 [cit. 2013-06-27]. 98 s. ISBN 978-80-7204-785-7. Dostupné z: http://www.iba.muni.cz/res/file/ucebnice/komprdova-rozhodovaci-stromy-lesy.pdf

[15] KORDÍK, Pavel. *Rozhodovací stromy a jejich regresní varianty*. 2011 [online]. [cit. 2013-06-20]. Dostupné z: http://old.avc-cvut.cz/?id=9688

[16] KUKAL, Jaromír. Úvod do neuronových sítí. *Automa: časopis pro automatizační techniku*. [online] ročník 2005, číslo 1 [cit 2013-06-22]. ISSN: 1210-9592. Dostupné z: http://www.odbornecasopisy.cz/index.php?id_document=30255

[17] Naše mise. UNICEF [online]. 2012 [cit. 2013-06-22]. Dostupné z: http://www.unicef.cz/co-delame/nase-mise

[18] Novým vlastníkem SPSS je IBM. *CIO Business World*. 2009 [cit. 2013-06-14]. Dostupné z: http://businessworld.cz/aktuality/novym-vlastnikem-spolecnosti-spss-je-ibm-5234

[19] OLEJ, Vladimír; HÁJEK, Petr. *Úvod do umělé inteligence*: *klasická umělá inteligence*. Pardubice: Univerzita Pardubice, 2009. 112 s. ISBN 978-80-7395-241-9.

[20] OLEJ, Vladimír; HÁJEK, Petr. *Úvod do umělé inteligence: moderní přístupy*. 1. vyd. Pardubice: Univerzita Pardubice, 2010. 98 s. ISBN 978-80-7395-307-2. [20]

[21] Our story. *Pentaho*. [online]. [cit. 2013-06-12]. Dostupné z: http://www.pentaho.com/about/story/

[22] PETR, Pavel. *Data mining – díl I*. 3. vyd. Pardubice: Univerzita Pardubice, 2010. ISBN 978-80-7395-325-6.

[23] Poll: What analytics, big data, data mining, data science software you used in the past 12 months? *KD nuggets* [online]. 2013 [cit. 2013-06-11]. Dostupné z: http://www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html

[24] Poll: Where did you apply Analytics/Data Mining in 2012? *KD nuggets* [online]. 2012 [cit. 2013-04-03]. Dostupné z: http://www.kdnuggets.com/polls/2012/where-applied-analytics-data-mining.html

[25] POSPÍŠIL, Jaromír; NEMRAVA, Michal. Dolování dat a jeho aplikace. *Slezská univerzita v Opavě: Filozoficko-přírodovědecká fakulta* [online]. 2006 [cit. 2013-06-05]. Dostupné z: http://axpsu.fpf.slu.cz/~sos10um/trendy/DM.pdf

[26] RapidMiner. *Rapid-I: Report the future.* [online]. [cit. 2013-06-12]. Dostupné z: http://rapid-i.com/content/view/181/190/

[27] RUD, Olivia Parr. *Data mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Přeložili I. MAGERA, M. DANĚK. Praha: Computer Press, 2001. 329 s. ISBN 80-7226-577-6.

[28] RYCHLÝ, Marek. Klasifikace a predikce. *VUT: Fakulta informačních technologií* [online]. Podzim 2005. [cit. 2013-03-15]. Dostupné z: http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/xhtml/classification-and-prediction.xhtml

[29] SAS Enterprise Miner. *SAS: The power to know*. [online]. [cit. 2013-06-12]. Dostupné z: http://www.sas.com/technologies/analytics/datamining/miner/#section=1

[30] SKALSKÁ, Hana. *Data mining a klasifikační modely*. 1. vyd. Hradec Králové: Gaudeamus, 2010. 154 s. ISBN 978-80-7435-088-7.

[31] Statistica Data Miner. *StatSoft*. [online]. 2013 [cit. 2013-06-12]. Dostupné z: http://www.statsoft.cz/produkty/5-dataminingove-nastroje/21-statistica-data-miner/detail/

[32] The State of the World's Children 2012 - Statistics – Tables. *UNICEF* [online]. 2012 [cit. 2013-06-22]. Dostupné z: http://www.unicef.org/sowc2012/statistics.php

[33] TUFFÉRY, Stéphane. *Data mining and statistics for decision making*. Hoboken: Wiley, 2011. 689 s. ISBN 978-0-470-68829-8.

[34] WEKA 3: Data mining software in Java. *WEKA: The university of Waikato*. [online]. [cit. 2013-06-12]. Dostupné z: http://www.cs.waikato.ac.nz/ml/weka/