**EKONOMICKÁ UNIVERZITA V BRATISLAVE**
**FAKULTA HOSPODÁRSKEJ INFORMATIKY**

Evidenčné číslo: 103004/I/2014/1541500730

# BUSINESS ANALÝZA V PRAXI

Diplomová práca

**Bratislava 2014** **Bóllová Monika Bc.**

**EKONOMICKÁ UNIVERZITA V BRATISLAVE**
**FAKULTA HOSPODÁRSKEJ INFORMATIKY**

# BUSINESS ANALÝZA V PRAXI

Diplomová práca

| | |
|---|---|
| **Študijný program:** | Manažérske rozhodovanie a informačné technológie |
| **Študijný odbor:** | 6258 Kvantitatívne metódy v ekonómii |
| **Školiace pracovisko:** | Katedra aplikovanej informatiky FHI |
| **Vedúci záverečnej práce:** | Dr. Ing. Miroslav Hudec |

**Bratislava 2014** **Bóllová Monika Bc.**

**Čestné vyhlásenie**

**Čestne vyhlasujem, že záverečnú prácu som vypracovala samostatne a že som uviedla všetku použitú literatúru.**

**Dátum: 25.4.2014** **…….....**

**(Podpis študenta)**

**ABSTRAKT**

Bóllová Monika: Business analýza v praxi.

Ekonomická univerzita v Bratislave: FHI, Katedra aplikovanej informatiky.

Vedúci záverečnej práce: Dr. Ing. Miroslav Hudec

Bratislava: FHI EU 2014 (67 s.)

Cieľom diplomovej práce je preukázať opodstatnenosť a užitočnosť nasadenia data miningu v prostredí malého, alebo stredného podniku, a použitie príslušnej aplikácie na analyzovanie konkrétnych podnikových dát. Predmetom diplomovej práce je analýza firemných údajov, konkrétne reálnych hotelových údajov, prostredníctvom dataminingových techník.

Práca je rozdelená do kapitol. Obsahuje základné teoretické vymedzenie najdôležitejších dataminingových techník, a prínosy použitých metód pre malý podnik. Práca obsahuje 2 tabulky, 3 grafy, 5 obrázkov a 4 prílohy.

Prvá kapitola je venovaná súčasnému stavu Business Intelligence a jeho prínosy pre organizácie ako aj rozdelenie nástrojov Business Inteligence. Druhá kapitola sa venuje popisu prostredia open source nástroja Weka v ktorom budeme realizovať praktickú časť práce. Tretia kapitola popisuje cieľ práce a štvrtá metodiku práce, teoretické východiská dataminingu. V ďalšej časti prakticky realizujeme datamining na konkrétnych hotelových dátach. Záverečná kapitola sa zaoberá zhodnotením práce, a jej prínosom pre malý podnik.

Výsledkom riešenia je zhodnotenie možností použitia open source nástroja Weka v praxi, a vytvorenie vhodných študijných podkladov pre výučbu základov data miningu.

Pri spracovaní diplomovej práce bola použitá dostupná odborná literatúra a internetové zdroje, pričom sme v teoretickej aj praktickej časti použili metódu deskripcie a analýzy.

Kľúčové slová: Business Inteligence, Data mining, podnikové dáta, deskripcia, analýza, open source, Weka

**Abstract in English**

The aim is to demonstrate the relevance and usefulness of thesis assignments in an environment of a small or medium-sized enterprise datamining, and use the appropriate application for analyzing specific corporate data. Subject of the thesis is the analysis of corporate data, namely data through datamining techniques, the hotel real.

The work is divided into chapters. It contains the basic theoretical definition of the most important techniques, and benefits of the datamining methods used for a small business. The work contains 2 tables, 3 charts, 5 pictures and 4 attachments.

The first chapter is devoted to the current state of Business Intelligence and its benefits for the Organization as well as the distribution of Business Intelligence tools. The second chapter is devoted to a description of the open source tools Weka environment in which we will carry out the practical part of the job. The third chapter describes the objective of the work and the methodology of work, theoretical bases of four dataminingu. In the next section we carry out datamining on specific hotel data virtually. The final chapter deals with the assessment of the work and its benefits for small business.

As a result, the assessment of options for solutions using open source tools Weka in practice, and the provision of appropriate learning materials for teaching the foundations of datamining.

When processing the thesis available literature and Internet sources have been used, while we in theoretical and practical part of the used method and analysis deskripcion.

Keywords: Business Intelligence, Datamining, enterprise data, deskripcion, analysis, open source, Weka

# Obsah

Úvod ..... 10

1 Súčasný stav problematiky ..... 12

1.1 Prínosy Business Intelligence pre organizáciu ..... 12

1.1.1 Štatistické zhodnotenie ..... 13

1.1.2 Budúcnosť Business Intelligence ..... 14

1.2 Analýza trhu Business Intelligence riešení ..... 15

1.3 Komerčné nástroje na dolovanie dát ..... 16

1.3.1 SAP ..... 16

1.3.2 ORACLE ..... 16

1.3.3 SAS ..... 16

1.3.4 SYBASE ..... 17

1.3.5 IBM ..... 17

1.3.6 MICROSOFT ..... 18

1.3.7 COGNOS ..... 18

1.4 Open source nástroje na dolovanie dát ..... 19

1.4.1 RapidMiner ..... 19

1.4.2 Weka ..... 20

1.5 Porovnanie komerčných a open source systémov ..... 20

1.6 Manažérske rozhranie ..... 21

2 Weka ..... 22

2.1 Grafické rozhranie Weka ..... 22

2.2 Filtre vo Weke ..... 24

2.3 Klasifikácia vo Weke ..... 25

2.3.1 Voľba klasifikátora ..... 25

2.3.2 Testovacie režimy ..... 25

2.3.3 Ďalšie možnosti testovania ..... 26

2.3.4 Výstup textovej klasifikácie ..... 27
3 Cieľ práce ..... 28
4 Metodika práce a metódy skúmania ..... 28
4.1 Využívanie dataminingových metód v praxi ..... 28
4.2 Rozhodovacie stromy ..... 29
4.2.1 Vytvorenie rozhodovacieho stromu ..... 30
4.2.2 Problémy rozhodovacích stromov pri získavaní dát ..... 31
4.2.3 Výhody a nevýhody rozhodovacích stromov ..... 32
4.2.4 Presnosť modelu ..... 32
4.3 Predikatívne techniky ktoré sa využívajú pri dolovaní dát ..... 33
4.3.1 Kategorická a numerická predikcia ..... 33
4.3.2 Priebeh klasifikácie ..... 33
4.3.3 Metriky ..... 34
4.3.4 Techniky trénovania dát ..... 35
4.3.5 Informačný zisk a entropia ..... 35
4.4 Algoritmy na tvorbu rozhodovacích stromov ..... 36
4.4.1 Algoritmy vytvárajúce jeden strom ..... 37
4.4.2 Algoritmy vytvárajúce súbory stromov ..... 37
4.4.3 Genetické algoritmi a rozhodovacie stromy ..... 38
4.5 Metodológia CRIPS-DM ..... 38
4.5.1 Jednotlivé fázy metodiky CRIPS-DM ..... 39
5 Praktická časť práce ..... 40
5.1 Pochopenie problému ..... 40
5.2 Definovanie problému ..... 40
5.3 Súčasná situácia skúmaného objektu ..... 41
5.3.1 Hotelový informačný systém ..... 42
5.4 Porozumenie dátam a spôsob zbierania údajov ..... 43

5.4.1 Vstupné údaje od zamestnancov ..... 44
5.5 Príprava dát ..... 45
5.6 Analýza dát ..... 46
5.7 Algoritmus modelu ..... 49
6 Výsledky práce ..... 49
6.1 Použitie modelu ..... 49
6.2 Načítanie dát do modelu ..... 50
6.3 Úprava dát pre model ..... 51
6.4 Ohodnotenie modelu ..... 54
6.5 Sumarizácia výsledkov pri tvorbe prediktívneho modelu ..... 58
6.6 Integrácia výsledku do manažérskeho rozhodnutia, prínosy práce ..... 60
Záver ..... 63
Zoznam použitej literatúry ..... 64
Prílohy ..... 66

## Zoznam tabuliek

**Tabuľka 1.** Použitie Dataminingu v priemyselných oblastiach v roku 2010 a 2011**. Zdroj:** http://www.kdnuggets.com/polls/, vlastné spracovanie
**Tabuľka 2.** Ukážka dát pred a po príprave**. Zdroj:** Vlastné spracovanie

## Zoznam obrázkov

**Obrázok 1.** Úvodné okno a KnowledgeFlow aplikácie Weka. **Zdroj**: Vlastné spracovanie
**Obrázok 2.** Údaje o návštevnosti vo formáte .xls. a .txt. **Zdroj**: Vlastné spracovanie
**Obrázok 3.** Okno predspracovania údajov v programe Weka. **Zdroj:** Vlastné spracovanie
**Obrázok 4**. Vizualizácia stromového grafu. **Zdroj**: Vlastné spracovanie
**Obrázok 5.** Vizualizácia štatistických charakteristík. **Zdroj:** Vlastné spracovanie

## Zoznam schém

**Schéma 1.** Grafické znázornenie metodológie CRIPS DM**. Zdroj:** neuron.tuke.sk/ zvada/ statnice/ II/08/index.html

## Zoznam grafov

**Graf 1.** Návštevnosť za mesiac marec v rokoch 2011-2012. **Zdroj**: Vlastné spracovanie
**Graf 2.** Tržba za rok 2011 a 2012 po mesiacoch. **Zdroj**: Vlastné spracovanie
**Graf 3.** Obsadenosť za rok 2011 a 2012 maximálna a minimálna početnosť. **Zdroj**: Vlastné spracovanie

## Zoznam príloh

**Príloha A** Grafy agregovaných tržieb a obsadenosti
**Príloha B** Textový výstup rozhodovacieho stromu
**Príloha C** Okno predspracovania údajov v programe Weka
**Príloha D** Vizualizácia modelu rozhodovacieho stromu

# Úvod

Základom každej firmy sú znalosti. Flexibilita firmy je závislá na znalostiach. Znalosti, ktorými firma disponuje, musia predstihnúť znalosti konkurentov, ale aj zákazníkov. Firma musí nielen reagovať na zákaznícke objednávky, ale nutne musí iniciatívne podporovať dopyt.

Firma si aktívne vytvára svoju budúcnosť a náskok pred ostatnými firmami. Efektívne pracovať s informáciami vo firme znamená aj vedieť, ako získať informácie, ktoré sú „roztrúsené“ vo firme. To však vôbec nemusí byť jednoduché. Databázy dnes ukrývajú nesmierne informačné bohatstvo.

Jednotlivec môže znalosti využívať vždy, pretože ich má stále k dispozícii, ale pre firmu s množstvom zamestnancov je to problém. Je možné nariadiť zdieľať znalosti, ale lepšie je riadiť prostredie, v ktorom znalosti vznikajú. Informácie musíme zachytávať, zdieľať, ale hlavne aplikovať. To je úlohou správnej firemnej stratégie. Vedenie firmy musí odstraňovať prekážky, aby znalosti boli využívané.

Výrobné aj ostatné firmy využívajú dnes pre svoje riadenie množstvo sofistikovaných systémov. Tieto sú často zložené z mnohých vzájomne prepojených modulov obsluhujúcich napríklad nákup, predaj alebo controlling.

Zložitá štruktúra systému generuje obrovské množstvo dát a informácií. Pre optimalizáciu a flexibilitu, potom firmy zavádzajú interné komunikačné technológie, ktoré by mali zaistiť konzistenciu a previazanosť dát. Takto koncipovaná architektúra je často zložitá a neposkytuje flexibilitu pri zmene výrobného procesu.

Aby bolo možné vybudovať otvorený a flexibilný systém, ktorý bude poskytovať rýchle a dynamické vykonávanie analýz, je vhodné zvoliť riešenie typu business intelligence. Je implementované priamo nad dátovým skladom a tým je zaistená jednotná interpretácia dát z rôznych systémov.

Každá firma má špecifické požiadavky a rôzne agendy. Busines intelligence dokáže odpovedať na otázky, ktoré majú priamy kladný dopad na procesy riadenia. Pomáhajú zostaviť dlhodobé plány z možnosťou simulovať scenáre, hľadá vzory v skutočných dátach a aplikuje ich na akúkoľvek dimenziu plánu.

Business intelligence je vybudované na jednotnej dátovej základni, zo schválených a preverených dát, bez ohľadu na ich pôvod. Do implementačného procesu sú zapojený všetci kľúčový užívatelia. Cieľovým stavom je komplexná podpora riadenia spoločnosti.[1]

Práca sa skladá z týchto kapitol: Prvá kapitola sa venuje súčasnému stavu Business Intelligence, jeho prínosom pre organizáciu, ako aj rozdeleniu nástrojov Business Inteligence. Analýza trhu BI riešení podrobnejšie opisuje existujúce systémy na trhu. SAP, ORACLE, SAS, SYBASE, IBM, MICROSOFT, COGNOS, Rapidminer, Weka.

Druhá kapitola popisuje najznámejšie open source nástroje na dolovanie dát. Weka a Rapidminer. Weka je prostredie v ktorom budeme realizovať praktickú časť práce.

Tretia kapitola popisuje cieľ práce a štvrtá metodiku práce, teoretické východiská dataminingu, ako aj prípravu dát na ďalšie skúmanie pomocou vybranej metodiky .

V ďalšej časti prakticky realizujeme datamining na konkrétnych hotelových dátach. Bližšie opíšeme hotel, ktorý budeme skúmať a spôsob akým boli zbierané a vybrané údaje. Použijeme metódy zvolené na skúmanie a interpretáciu výsledku.

Záverečná kapitola sa zaoberá zhodnotením práce, a jej prínosom pre malý podnik, zvýšenie efektivity organizácie. Táto kapitola rozvádza ako sa predikcia môže podieľať na manažérskom rozhodovaní. Výsledky pri dolovaní dát, presnosť modelu, vstupy, výstupy a integrácia výsledku do manažérskeho rozhodnutia.

V závere bude stručné zhodnotenie, porovnanie cieľa a výsledku. Okomentovanie skutočností a vlastné zhodnotenie.

[1] Radek Kafka, 2013, SystemOnline, **Implementace BI řešení ve výrobní firmě,** www.systemonline.cz/ business-intelligence/ implementace-bi-reseni-ve-vyrobni-firme-1.htm

# 1 Súčasný stav problematiky

## 1.1 Prínosy Business Intelligence pre organizáciu

Už z definície Business Intelligence je jasné, že cieľom užívateľa je získať z dostupných dát komplexné informácie pre riadenie, rozhodovanie a výkazníctvo. Odpovede na všetky otázky sú v dátach a systémoch, ktoré už existujú. Úlohou Business Intelligence je dáta získať, prepojiť a vytvoriť prostredie v ktorom bude možné jednoducho formulovať dotazy, na ktoré dostaneme v čo najkratšom čase odpovede.

Mnoho projektov Business Intelligence neprináša očakávané výsledky. Na vine je niekoľko faktorov. Patrí medzi ne hlavne nezladenie cieľov projektu a strategických cieľov spoločnosti. Z tohto pohľadu je dobré venovať pozornosť disciplíne Busines Intelligence governance. Je to disciplína, ktorá je súčasťou riadenia informačných technológií v organizácii.

Zaoberá sa definovaním a riadením Business Intelligence programov a projektov, aby sme zabránili vzniku niekoľkých útržkovitých Business Intelligence riešení v rôznych prostrediach. Toto priamo súvisí s finančnou náročnosťou a návratnosťou vložených finančných prostriedkov, ako aj zo sklamaním, že projekt nepriniesol očakávané výsledky.[2]

Náklady na zavedenie Business Intelligence sú zložené z niekoľkých položiek. Okrem ceny licencie príslušného softvéru, treba počítať s cenou za prostredie pre získanie a integráciu dát z rôznych dátových zdrojov, cenu za databázu, alebo dátový sklad. Náklady na implementáciu, rozvoj, údržbu a administráciu, školenie pracovníkov, prípadne náklady na hardware.

[2] Tomáš Třminek, 2012, SystemOnline, **Prínosy a náklady Business Intelligence,** www.systemonline.cz/business-intelligence/prinosy-a-naklady-business-intelligence.htm

### 1.1.1 Štatistické zhodnotenie

Medzi kritérium najviac ovplyvňujúce zavádzanie Business Intelligence do praxe patrí cena. Všeobecne sa uvádza, že celkové náklady na informačné služby predstavujú okolo 6 až 8 % z celkových firemných nákladov, pri organizáciách zameraných na vývoj, napr. pri farmaceutických firmách, je to až okolo 10 %.

| Priemyselné oblasti kde hlasujúci použili Analytics/ Data Mining v roku 2011 | | |
|---|---|---|
| Počet voličov 228 | 2011(% voličov) | 2010(% voličov) |
| CRM /analytics spotrebiteľov (57) | 25,00% | 26,80% |
| Bankovníctvo (43) | 18,90% | 19,20% |
| Zdravotná starostlivosť / HR (38) | 16,70% | 13,10% |
| Detekcia podvodov (32) | 14,00% | 12,70% |
| Veda (31) | 13,60% | 10,30% |

**Tabuľka 1.** Použitie Dataminingu v priemyselných oblastiach v roku 2011 a 2010**. Zdroj:** http://www.kdnuggets.com/polls/, vlastné spracovanie

Najväčší nárast zaznamenalo cestovanie a pohostinstvo až o 429%, sociálne siete o 100%, vzdelanie o 65%, biotechnológie o 64%, a credit scoring o 59%. Najväčší pokles zaznamenala výroba až o -34%, reklama o -29%, e-commerce o -25%.[3]

Gartner líder v oblasti výskumu informačných technológií očakáva, že okolo roku 2016 bude 30% podnikov využívať svoje informačné zdroje. Na predikatívne analýzy sa však bude orientovať menej ako 25% projektov. Tieto projekty však vyprodukujú viac ako 50% obchodnej hodnoty. Trh s analytickými riešeniami bude aj naďalej jedným z najrýchlejšie rastúcich softvérových trhov. Ročná miera rastu sa bude pohybovať okolo 7%.[4]

Práve preto, že najväčší rast použitia systémov dolovania dát sme zaznamenali práve v odvetví cestovania a pohostinstva, rozhodli sme sa aplikovať konkrétny dataminingový

[3] **KDnugets™**, www.kdnuggets.com/polls
[4] Lacko Ľ., Asseco, **BI: Trendy, prehľad riešení, poskytovateľov, možností a cien**, Lacko Ľ., Asseco, Asseco.com/ce/assets/Uploads/ attachments/news-items/NMinfoware2013060720.pd

systém na hotelové dáta. Predpokladáme, že analyzovanie dostupných informácií, okrem iného umožní zlepšenie služieb pre hostí.

Bratislava v roku 2013 zaznamenal najväčší počet návštevníkov z Ukrajiny (+100%), Ruska (+47,7%) a Nemecka (+27%). Najväčšiu skupinu zahraničných návštevníkov už dlho tvoria Česi. Minulý rok ich do Bratislavy prišlo 87.000. Nasledujú Nemci (70.000), Rakúšania (40.000) turistov. Počet prenocovaní v Bratislave medziročne vzrástol o 11%.

Ubytovacie zariadenia vykázali viac ako 1,9 mil. prenocovaní. Priemer je 1,66 dňa na jedného návštevníka mesta. Samotná Bratislava mala minulý rok 23,4% podielu na trhu s cestovným ruchom v rámci celej republiky. Priemerné denné náklady na pobyt v slovenskej metropole sa pohybujú na úrovni 100 eur, dve tretiny sumy turisti minú na ubytovanie a stravu, zvyšok na nákupy či na zábavu.[5]

### 1.1.2 Budúcnosť Business Intelligence

V súčasnosti sa ako veľmi zaujímavé javí využiť systém založený na princípe memory analýzy. Vďaka tejto analýze sa dá docieliť, že budú naplnené všetky požadované kritériá a zároveň priaznivý pomer cena, výkon a úspešnosť Business Intelligence[6]

Prevládajúcim trendom v Bussines Inteligence je úsilie správne zareagovať na hromadenie sa väčšinou neštruktorovaných dát, ktoré napriek tomu že sú nesmierne cenné nevieme využiť. Vo veľkých organizáciách kde je povinnosť dáta uchovávať tvoria náklady na ich uchovanie nemalé čiastky. Na druhej strane, firmy nemajú vhodné technológie na využitie týchto dát. Ak vie organizácia dáta využiť, predstavuje to skutočnú konkurenčnú výhodu.

Podľa analytických štúdií sa v najbližšom období očakáva presun analýz do mobilov. Najprv sa bude častejšie len pristupovať k existujúcim reportom, neskôr budú aj mobilné aplikácie schopné plniť konkrétne úlohy. Dnes už pomerne častou súčasťou implementácie BI je cloud

[5] Pacherová S., Pravda, 1.4.2014, **Milionárka na Dunaji? O Bratislave to neplatí**

[6] Dostál M. 2012/5, SystemOnline, **Věnujte pozornost BI governance**, www.systemonline.cz/business-intelligence/ venujte-pozornost-bi-governance.htm

computing, ktorý bol v minulosti cenovo nedostupný. Dnes je naopak pre malé a stredné firmy ideálnym riešením.

Ďalšou oblasťou, je analytické zapracovanie v pamäti (in-memory). Rastúce využívanie Bussines Intelligence funkcií urýchli aj vznik nových. Bude stále častejšie obsahovať text, data mining, predikácie, regresiu, optimalizáciu, alebo simulácie pomocou komplexného modelovania dát.[7]

Budúcnosť bude patriť logickým dátovým skladom, čo je spojenie tradičného dátového skladu s distribuovaným dát. Nové nástroje budú špecializované na analýzu neštruktúrovaných dát. Zaujimavá bude tá časť trhu ktorá predstavuje dáta ako služby (DaaS). Pôjde o odber dátovo špecifikovaných služieb, zameraných na úzku oblasť záujmu.[8]

## 1.2 Analýza trhu Business Intelligence riešení

Business Intelligence dnes charakterizuje dynamika a neustále zmeny. Existuje množstvo dodávateľov so širokou ponukou nástrojov z oblasti Business Intelligence. Firmy zaoberajúce sa Business Intelligence chcú svojím zákazníkom ponúknuť čo najkomplexnejšie služby. Podniky stále častejšie zavádzajú Business Intelligence pomocou externých firiem.

Takáto služby začínajú analýzou súčasného stavu v podniku a končia kompletným návrhom riešenia celého Business Intelligence projektu. Spravili sme prehľad najznámejších spoločností zaoberajúcimi sa Business Intelligence.[9]

---

[7] Komora M., Hečková G., SystemOnline, **Rychlejší analýza dat s in-memory BI,** www.systemonline.cz/ business-intelligence/ rychlejsi-analyza-dat-s-in-memory-bi-1.htm

[8] Lacko Ľ., Asseco, **BI: Trendy, prehľad riešení, poskytovateľov, možností a cien**, asseco.com/ce/assets/ Uploads/ attachments/news-items/NMinfoware2013060720.pd

[9] **SAP**, www.sap.com/sk/index.html

## 1.3 Komerčné nástroje na dolovanie dát

### 1.3.1 SAP

Obsadil tretie miesto ako dodávateľ softvéru, a je najväčšou firmou na svete dodávajúcou medzi podnikové softvérové riešenia. V oblasti Business Intelligence ponúka dve aplikácie: SAP Business Integration Warehouse (SAP BW) a SAP Stratégie Enterprise Manager (SAP SEM).

Hlavný cieľ aplikácie SAP BW je spojenie ERP systémov rôzneho pôvodu s dátovým skladom na báze jednotného modelu. Na základe tohto prepojenia môžeme pomocou OLAP funkcií analyzovať reportovacie dáta.

### 1.3.2 ORACLE

Oracle je líder v oblasti databázových technológií. Má Business Intelligence riešenie rozdelené do štyroch hlavných bodov. Začína Data Warehousing, pokračuje v ETL a OLAP a končí v Data Mining. V oblasti dátových skladov je hlavný produkt databázový server Oracle BI 10g. Pracuje priamo s dátovým skladom. Medzi hlavné funkcie patrí: návrh, implementácia a správa dátového skladu.

Výhodou nástroja je, že umožňuje načítať cez Gateway Oraclu dáta do dátového skladu z iných zdrojov (iné typy Oracle databáz, Informix, Sybase, DB2). Ďalšou výhodou je jeho použiteľnosť na ľubovoľnej platforme v prostredí Java.[10]

### 1.3.3 SAS

Lídrom v oblastí softvérových riešení pre oblasť Business Intelligence je aj spoločnosť SAS. Silné postavenie získala aj vďaka investíciám do vývoja. Spoločnosť má naozaj širokú škálu modulov. Rieši úlohy dátových skladov, ukladanie, popis, transformáciu, až sprístupnenie užívateľovi. Je aj nezávislé na použitom hardvéri a softvéri. Warehouse Administrator SAS je produkt obsahujúci hlavné procesy potrebné k vytvoreniu dátového skladu.[11]

[10] **ORACLE**, www.oracle.com/sk/index.html?ssSourceSiteId=ocomen
[11] **SAS,** www.sas.com/offices/europe/slovakia/

### 1.3.4 SYBASE

Spoločnosť má významne postavenie, hlavne v oblasti relačných databáz, vďaka svojmu produktu Sybase IQ server. Vychádza z potreby v prípade agregácie v oblasti Data Warehouse prejsť celou tabuľkou pre výpočet jedného stĺpca. V bežnom databázovom serveri sa dáta ukladajú po vetách.

Počet načítaní sa výrazne zníži, ak do databázy uložíme dáta tak, že uložíme jednotlivé stĺpce zvlášť. Tento princíp ukladania do stĺpcov namiesto riadkov sa nazýva Bitwise Technology. Týmto zabezpečujeme rýchlosť prístupu do databázy, a taktiež umožníme získať homogénne údaje.

Jednou z ďalších výhod je aj podpora metodológie UML a dvojúrovňový návrh databáz. Všetky riešenia od firmy SYBASE je možné použiť s nástrojmi inej firmy a to či už s iným databázovým serverom alebo analytickým nástrojom.[12]

### 1.3.5 IBM

IBM si prednú pozíciu medzi lídrami v oblasti Business Intelligence získalo vďaka kvalitnej realizácii systémov. Riešenia pokrývajú veľkú časť z nástrojov od dátových trhovísk až po podnikové dátové sklady.

Firma ponúka integrované riešenie DB2 Data Warehouse Editions (DB2 DWE). Najsilnejšou verziou je DB2 Enterprise Server Edition (DB2 ESE). Je navrhnutá tak aby poskytovala vysoký výkon, kvalitnú správu a uchovávanie dát, ako aj analytických časti z cieľom poskytnúť užívateľovi v reálnom čase potrebnú informáciu.[13]

[12] **SYBASE**, www.sybaseproducts.com

[13] **IBM**, www-03.ibm.com/software/products/en/category/business-intelligence

### 1.3.6 MICROSOFT

Microsoft je najvýznamnejšou svetovou firmou v oblasti softvéru, služieb a internetových technológií. Pre svoje operačné systémy ponúka radu SQL Server 2005 rozdelenú do 4 editácií. Express, Workgroup, Standard a Enterprise.

Ponúkajú množstvo funkcionalit, dostupnosť, robustnosť, škálovitosť cez nástroje pre Business Intelligence. Tiež vysokú bezpečnosť, spoľahlivosť a jednoduchú správu. Medzi hlavné prednosti patrí analýza a spracovanie dát cez internet. Prenos dát medzi voľne prepojenými systémami je na báze jazyka XML.[14]

### 1.3.7 COGNOS

Medzi najvýznamnejších svetových dodávateľov manažérskych informačných systémov patrí Cognos. Radí sa v oblasti Business Intelligence a plánovaní produkcie k svetovým lídrom. Má software na realizáciu kompletných procesov vo firmách, od plánovania a zostavovania rozpočtu, cez meranie a monitorovanie, až k reportom a analýzam.

Je jedinou firmou, ktorá poskytuje všetky tieto nástroje vjednom riešení. Vďaka svojej univerzálnosti našla uplatneni vo všetkých sférach hospodárstva.

PowerPlay Enterprise Server je nástroj na distribúciu OLAP analýz a zostáv pre najširší okruhu používateľov v podniku aj mimo, cez web portál Cognos UpFront. Škálovateľnosť, centrálna správa a integracia, umožňuje Windows a web klientom využívať z jediného miesta analytické a reportovacie OLAP služby, zahŕňajúce zdieľanie, publikovanie a tímovú spoluprácu pri vytváraní analýz.

Upfront vytvára multidimenzionálne kocky, analýzy a zostavy podniku do prehľadnej štruktúry. Umožňuje používateľom prezerať kocky, prezerať a tlačiť zostavy, premeniť statickú zostavu na dynamickú a pokračovať v analýze priamo v prostredí webu, alebo na Windows klientovi.[15]

Okrem týchto spoločností, ktoré tvoria vrchol rebríčkav data miningu, existuje mnoho spoločností, možno nie tak známych, ale pre malé a stredné firmy hlavne cenovo

[14] **Microsoft**, www.microsoft.com/slovakia/sqlserver/default.mspx

[15] **Cognos**, www-03.ibm.com/software/products/sk

prístupnejších. Ako sme už spomínali v súčastnosti je pre malú firmu zaujímavé riešenie ako outsourcing a cloudové riešenia, ktoré ponúka napríklad firma ABRA softvare.

## 1.4 Open source nástroje na dolovanie dát

Datamining sa spravidla nevykonáva bez použitia softvérových nástrojov, komerčného alebo open source. Líšia sa hlavne v užívateľským rozhraním a funkcionalitou. Ako sme už uviedli na trhu existuje veľa kvalitných komerčných riešení, ktoré sa používajú na riešenie problémov dataminingu. Ku komerčným nástrojov som sa pri svojej práci nedostala.

Oveľa lepší prístup je k open source nástrojom. Medzi významné open source nástroje na dolovanie dát patrí Weka a RapidMiner.

### 1.4.1 RapidMiner

Vznikol na univerzite v Dortmunde v roku 2001 (predtým YALE). Počas niekoľkých rokov sa z neho stal jeden z najúspešnejších open source nástrojov na dolovanie dát. Softvér je napísaný v jazyku Java. Umožňuje integrovať sa do existujúcich aplikácií. V aktuálnej verzii 5.0 je dostupný ako open source nástroj pod licenciou GPL, alebo ako komerčný software, ktorý má užívateľskú podporu.

RapidMiner má dobre vypracované užívateľské rozhranie, v ktorom môžeme vytvárať proces dolovania dát ako súbor na seba naväzujúcich operátorov. RapidMiner má úplne integrovanú knižnicu Weka, preto sú všetky operátory zo systému Wekka prístupné, dajú sa kombinovať s operátormi z RapidMineru.

Proces je možné uložiť vo formáte XML, načítať ho z užívateľského rozhrania, ako aj cez API z vlastného kódu. Má to výhodu - procesy môžeme vytvoriť a odladiť v GUI. Pri integrácii do nášho systému ho načítame a aplikujeme na vstupné dáta. Takto môžeme dynamicky zmeniť napr. učiaci algoritmus v aplikácii.[16]

[16] **Rapidminer**, http://rapidminer.com/products/rapidminer-studio/

### 1.4.2 Weka

Weka (*Waikato Environment for Knowledge Analysis*) je softvér vyvinutý na novozélandskej univerzite Waikato. Je to komplexný nástroj napísaný v jazyku Java. Ponúka širokú škálu nástrojov a algoritmov na predspracovanie dát, výber atribútov, klasifikáciu, regresiu, clustering, vytváranie asociačných pravidiel a vizualizáciu. V súčasnosti je Weka voľne dostupná vo verzii 3.7 pod licenciou GNU GPL.

Weka poskytuje implementáciu učiacich algoritmov, a ich jednoduchú aplikáciu na akýkoľvek dataset. Výhodou je možnosť integrácie s vlastnou aplikáciou, jednoduchým použitím Weka ako knižnice. Algoritmy a nástroje sa dajú volať aj priamo z našej aplikácie.

Je konštruovaná tak, že používateľ môže rýchlo a flexibilne otestovať existujúce dataminingové metódy. Poskytuje rozsiahle zázemie pre celý proces dataminingu od prípravy vstupných dát, cez štatistické ohodnotenie učiacich schém, až po vizualizáciu získaného výsledku.

Zahŕňa všetky dataminingové metódy ako regresia, klasifikácia, zhnukovanie, asociačné pravidlá a výber atribútov. Podporuje formáty ARFF, XRFF, CSV a ďalšie, vstup sa dá načítať z flat-file súborov. Tiež je možné načítať dáta zo súborov, alebo získanie dát priamo z databáze pomocou SQL dotazu cez rozhranie JDBC.[17]

## 1.5 Porovnanie komerčných a open source systémov

Komerčná verzia je vždy obsiahlejšia ako open source. Grafické prostredie je prepracované a program obsahuje množstvo nástrojov. Je doplnená o zákaznícku podporu cez telefón, email a školenia. Neporovnateľná je cena, ktorá je pri komerčných nástrojoch vyššia.

Open source je slobodný software, dodávaný spolu so zdrojovým kódom. To znamená voľnosť v modifikácii, distribúcii a užívaní. Open source ponúka užívateľovi voľnosť

[17] **Weka,** http://www.cs.waikato.ac.nz/ml/weka/downloading.html

modifikovať program, ale kladie dôraz na to, aby tato modifikácia bola zdokumentovaná a modifikovaný zdrojový kód aby bol opäť prístupný.[18]

Produkt je vyvíjaný jednotlivcami, alebo skupinkami vývojárov. Od toho sa odvíja aj kvalita produktu. Preto sú aplikácie bez záruky a zákazníckej podpory. Nie je pravda, že open source produkty nie sú vhodné pre nasadenie do riadenia. V podniku však nie sú vždy na úrovni produktov výrobcov komerčne ponúkaných systémov.

Pomocou týchto produktov je možné realizovať skôr projekty menších a stredných organizácií. Rozdiely sú vidieť hlavne pri užívateľskom prostredí, ktoré sa však dá vďaka otvorenosti modifikovať. Podstatným nedostatkom je absencia niektorých funkcií, tým si výrobca komerčných produktov stále drží výhodu. Väčšinou sa jedná aj o nedostatočnú dokumentáciu, chýbajúcu podporu, alebo obmedzenú funkčnosť.

Sú však často zdrojom nových riešení, ktoré sa neskôr presúvajú do komerčnej sféry. Spoločnosti sponzorujúce open source projekty takto získavajú nové technológie, postupy a aj nové talenty. Všeobecne, najlepšie open surce produkty sú natoľko vyspelé, že môžu konkurovať komerčným produktom miliónmi inštalácií.[19]

## 1.6 Manažérske rozhranie

Weka sa zaoberá funkčnou špecifikáciou akéhokoľvek manažérskeho systému. Preto je dôležité, nezávisle na oblasti v ktorej predikáciu robíme venovať sa manažérskemu rozhraniu. Návrhu manažérskeho rozhrania predchádza zistenie funkčných a nefunkčných požiadaviek obsluhujúcich pracovníkov.

Existuje niekoľko manažérskych systémov, ktorých použitie je obmedzené len správnym nakonfigurovaným na ich aplikačný server. Je dobré ak je informačný systém úplne oddelený od manažérskeho systému ako na aplikačnej tak aj na dátovej vrstve.

[18] Arnošt P., 22.8.2001, ROOT.cz, **Co je to „Open Source software“**, www.root.cz/clanky/co-je-to-open-source-software

[19] Janů S., 18.4.2014, Zive.cz, **Open-source je poprvé kvalitnější než proprietární software**, www.zive.cz

Existuje niekoľko manažérskych prostredí. Medzi open source platformi patrí napríklad Pantheo. Obsahuje služby ako generovanie reportov, analýzy, alebo dátovú integráciu. Existuje možnosť využiť niektoré portálové riešenie, alebo vytvoriť vlastnú webovú aplikáciu.

Na implementáciu webového rozhrania je vhodný napríklad nástroj Google Web Tolkit. Pri menších projektoch je vhodnejšia druhá možnosť, pretože webové aplikácie majú oveľa viac funkcionalít ktoré by zostali nevyužité.

# 2 Weka

Weka je často používaná v akademickej sfére. Začiatočník, ktorý nepozná všetky operátory, si ich význam môže vyhľadať v dostupnej dokumentácii, alebo si môže zobraziť nápovedu priamo v programe. Dostupná je po výbere a kliknutí na daný operátor.

## 2.1 Grafické rozhranie Weka

Práca so systémom Weka prebieha cez rôzne grafické rozhrania:

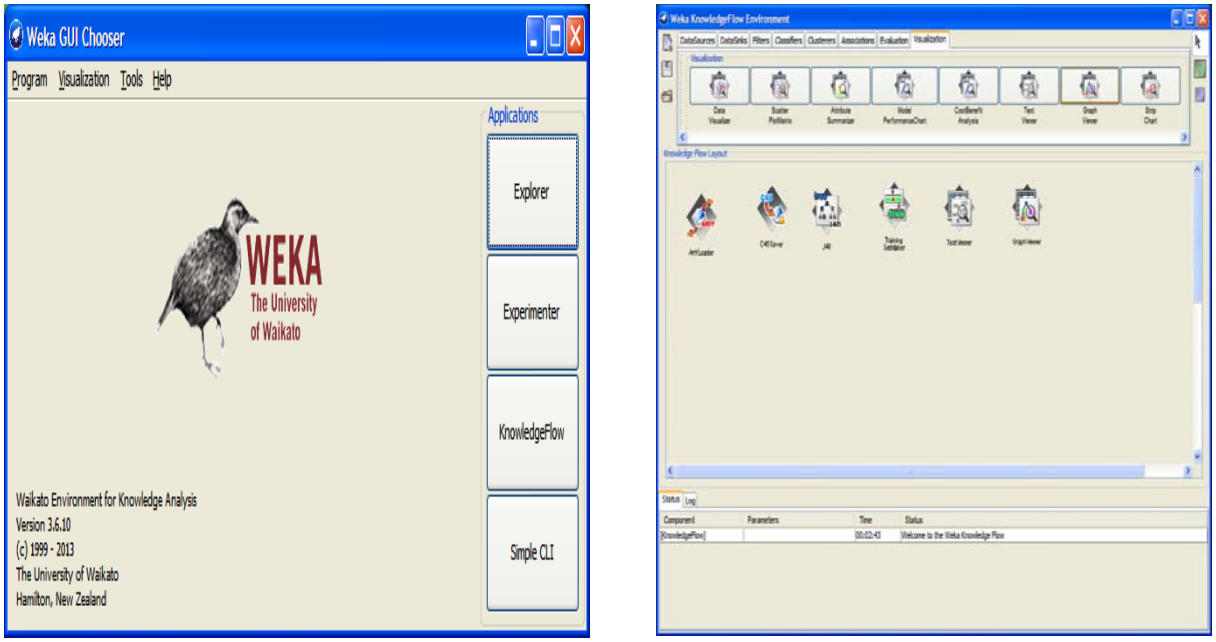
**Explorer** - hlavné užívateľské rozhranie. Umožňuje aplikovať niekoľko typov filtrov. Filtre na výber atribútov, transformáciu atribútov, postupne vybrať algoritmus pre danú úlohu (klasifikácia, regresia,...) a výsledky učenia zobraziť.

Je to najjednoduchší spôsob oboznámenia sa s používaním aplikácie, pretože je to grafické užívateľské prostredie. Prostredníctvom neho máme prístup ku kompletnému nástrojovému vybaveniu.

Vyberáme si z ponúkaných možností a zadávame úlohy. Možnosti ponúka vo forme menu, čím zabezpečuje plynulosť krokov v požadovanom poradí. Ďalšie kroky sú neaktívne pokiaľ nie sú ukončené predchádzajúce.

Pomocným nástrojom je „tool tips.“ Jeho úlohou je oboznámiť nás s funkciami jednotlivých ikon. Ďalšou pomôckou je prednastavenie hodnôt a filtrov čo nám zabezpečuje s minimálnym úsilím získať výsledky.

Nevýhodou Explorera je, že všetko uchováva v hlavnej pamäti. Po otvorení datasetu sa tento celý načíta, čo predstavuje obmedzenie. Môže byť použitý, len na skúmanie malých, alebo stredných problémov.




**Obrázok 1**. Úvodné okno a KnowledgeFlow aplikácie Weka . **Zdroj**: Vlastné spracovanie

**KnowledgeFlow** – komponentovo orientované užívateľské rozhranie. Má tú istú funkcionalitu ako Explorer, ale odlišný prístup k práci. Weka komponenty si môže používateľ vybrať, umiestniť ich na plátno a navzájom ich poprepájať tak, aby vytvárali tok znalostí pre spracovanie a analýzu dát.

**Experimenter** – je rozhranie, kde môžeme permanentne porovnať výsledky dosiahnuté zo zvolených algoritmov. Porovnáva rozličné učiace techniky, interaktívne práve cez rozhranie Experimenter. Experimenter automatizuje tento proces, zjednodušuje ho a vykonáva testy významnosti modelu.

**Simple CLI** – analógia príkazového riadku. Z neho môžeme volať ktorúkoľvek Weka triedu s vybratými vstupmi.

## 2.2 Filtre vo Weke

Weka ponúka pre úpravu dát niekoľko filtrov, ktoré upravujú dáta a uľahčujú tak prácu s datasetmi. Zistili sme že:

♦ **filter Normalize** – normuje všetky atribúty, ale neponúka možnosť výberu iba niektorého atribútu.

♦ **filter Discretize** – je vhodný na úpravu numerických atribútov. Rozdelí hodnoty do intervalov, počet intervalov definuje užívateľ. Štandardné je zvoliť 10 intervalov.

♦ **filter Replace missing values** – vstupný dataset očistí od chýb a doplní chýbajúce hodnoty. Platí, že nominálne hodnoty nahradí najčastejšou hodnotou a numerické priemernou.

Nevýhoda:

- ♦ nie je možné nahradiť chýbajúce hodnoty iným spôsobom
- ♦ nahradí chýbajúce hodnoty pri všetkých atribútoch

♦ **filter Remove percentage** – často používaný filter, ktorý transformuje numerické hodnoty na nominálne. Weka ho často využíva rozdelenie datasetu na množiny.

Nevýhoda

- ♦ dataset môže byť rozdelený len na dve množiny
- ♦ neposkytuje možnosť nastaviť pomer rozdelenia cieľovej premennej[20]

[20] Witten I.H., Eibe F., Elsevier, 2005, **Data Mining: Practical Machine Learning Tools and Techniques**, ISBA 0-12-088407-0

## 2.3 Klasifikácia vo Weke

### 2.3.1 Voľba klasifikátora

Klasifikátori Weka sú navrhnuté tak, aby vedeli predpovedať iba jednu triedu, atribút, ktorý je cieľom predikcie. Niektoré klasifikátori môžu učiť nominálne triedy, iné môžu naučiť iba numerické triedy (regresná problémy).

Ak chceme naše údaje preskúmať pomocou rozhodovacích stromov, zvolíme v hornej časti box **Klasifikátor**. Tento box má textové pole, v ktorom je názov aktuálne vybraného triedenia, a jeho možnosti. Tlačidlom **Choose** si môžeme vybrať jeden z klasifikátorov, ktoré sú k dispozícii.

Box pod ním má meno **Test options** (Testovacie možnosti). Zvolený klasifikátor bude testovaný v súlade s možnosťami, ktoré sú nastavené v **Test options.** Ďalšie možnosti testovania možno nastaviť kliknutím na viac možností.

### 2.3.2 Testovacie režimy

K dispozícii sú štyri testovacie režimy:

**1. Use traning set** (*použite tréningového setu*)

Klasifikátor je hodnotený podľa toho, ako dobre predpovedá triedy inštancií na ktorých bol trénovaný čiže nám známe výsledky.

**2. Supplied test set** (*dodávaná testovacia sada*)

Klasifikátor je hodnotený podľa toho, ako dobre predpovedá triedy inštancií načítané zo súboru. Kliknutím na tlačidlo **Set** otvoríme dialóg umožňujúci zvoliť súbor, ktorý chceme vyskúšať.

**3. Cross-validation** (*krížové overenie*). Klasifikátor je hodnotený krížovým testom.

**4. Percentage split** (*percentuálne rozdelenie*). Klasifikátor je hodnotený podľa toho, ako dobre predpovedá určité percento z údajov, pre ktoré sa koná testovanie. Množstvo stiahnutých dát závisí na hodnote zadanej v poli %.

Bez ohľadu na to, ktorú metódu hodnotenia použijeme, model má na výstupe vždy iba jeden výstup zostavený zo všetkých trénovaných dát.

### 2.3.3 Ďalšie možnosti testovania

Ďalšie možnosti testovania možno nastaviť kliknutím na viac možností:

**1. Output Model**. *(výstup je klasický model)*. Model klasifikácie na celej trénovanej množine. Výstup môže byť vizualizovaný. Táto možnosť je prednastavená.

**2. Output per-class stats** (*výstup na štatistickej triede*). Presná štatistika pre každú triedu na výstupe.

**3. Output entropy evaluation measures** (*výstupné hodnotenie entropie*). Hodnotenie entropie bude zahrnuté vo výstupe. Táto možnosť nie je prednastavená.

**4. Output confusion matrix**. (*výstup je krížový test)*. Výstup je hodnotený krížovým testom, test je súčasťou výstupu. Možnosť je prednastavená.

**5. Store predictions for visualization** (*uloženie predpovede pre vizualizácie*) Klasifikátori sú v pamäti tak, aby mohli byť zobrazené. Možnosť je predvolená.

**6. Output predictions** (*výstupná predpoveď*). Predpoveď hodnotiacich dát je na výstupe.

**7. Output additional attributes** (*ďalšie atribúty na výstupe*). Ak musia byť ďalšie atribúty na výstupe spolu s predpoveďou, potom index tohto atribútu môže byť uvedený tu.

**8. Cost-sensitive evaluation** (*vyhodnotenie nákladov*). Chyby sa hodnotia s ohľadom na nákladovú maticu. Použijeme tlačidlo Set, umožňuje určiť a použiť nákladovú maticu.

**9. Random seed for xval % Split.** Určuje použitie súboru náhodných čísel, pred tým, než sú rozdelené na účely hodnotenia.

**10. Preserve order for % Split.** Zachováva poradie, potláča randomizáciu dát pred rozdelením.

**11. Output source code** (*zdrojový kód výstupu*). Ak je klasifikačný model postavený ako Java zdrojový kód, môžeme zadať názov triedy tu.

### 2.3.4 Výstup textovej klasifikácie

Výstup je rozdelený do niekoľkých sekcií:

**1. Run information** *(spustiť informácie)* - zoznam informácií, ukazuje schému učenia, názov relácie, inštancie, atribúty a testovací režim, ktorý bol zapojený do procesu.

**2. Classifier model** *(plný tréningový set)* - textová reprezentácia modelu klasifikácie, ktorá bola vytvorená z údajov celého školenia.

**3. Results** – *(výsledky)* výsledky zvoleného testovacieho režimu, členíme dalej takto:

**4. Summary** *(súhrn)* – zoznam, štatisticky sumarizuje ako presne klasifikátor bol schopný predpovedať skutočnú triedu inštancií v rámci zvoleného testovacieho režimu.

**5. Detailed Accuracy By Class** - presnosť triedy. Podrobnejšie zobrazenie presnosti predikcie klasifikátora.

**6. Confusion Matrix** *(krížová matica)* - ukazuje, koľko inštancií bolo pridelených pre každú triedu. Prvky ukazujú počet skúšobných príkladov, ktorých skutočná trieda je rad a ktorých predpokladané trieda je stĺpec.

**7. Source code** - zdrojový kód *(voliteľne*) - táto časť obsahuje zdrojový kód v jazyku Java, ak sa preň rozhodneme.[21]

[21] **Weka Tutoriál**, http://www.cs.waikato.ac.nz/ml/weka/index.html

# 3 Cieľ práce

Cieľom diplomovej práce je preukázať opodstatnenosť a užitočnosť nasadenia open source aplikácie na dolovanie dát v prostredí malého a stredného podniku. Zároveň popíšeme podrobnejšie tento nástroj pre potreby ďalšieho individuálneho vzdelávania.

Zosumarizujeme teoretické poznatky o metóde rozhodovacích stromov a niektorých jednoduchý algoritmoch pre tvorbu týchto stromov. Popíšeme výhody a nevýhody rozhodovacích stromov.

Predmetom diplomovej práce je analýza firemných údajov prostredníctvom data miningových metód. Práca obsahuje základné teoretické vymedzenia postupov pri použití data miningu. Vyhodnotenie výsledkov použitia dataminingu a prínos pre prax.

Pri spracúvaní diplomovej práce bola využitá dostupná odborná literatúra a internetové zdroje. V praktickej časti sme použili metódu deskripcie a metódu analýzy. Prínos spočíva aj vo vytvorení študijných podkladov pre individuálne vzdelávanie v data miningu, pomocou open source programu Weka.

# 4 Metodika práce a metódy skúmania

## 4.1 Využívanie dataminingových metód v praxi

*„Dolovanie dát je proces objavovania nových, vopred neznámych, zmysluplných vzorov a trendov, prostredníctvom preskúmania veľkých objemov dát, za použitia technológií vzorového rozpoznávania, štatistických a matematických metód.“(Dr. Daniel T. Larose, LAROSE 2005)*

Rozhodovacie procesy definujeme ako individuálne zhodnotenie situácie, na základe dostupných informácií a zdrojov. V bežnom živote sa s rozhodovacím procesom stretávame neustále. Rozhodovanie je stratégia nášho riešenia. Výsledok bude ovplyvnený veľkým počtom neznámych vplyvov, ktoré sťaží porovnanie výsledku.[22]

[22] Sarnovský J., Liguš J., Benko P., Košice 2001, **Kybernetika a manažment**, http://web.tuke.sk/kybernetika/kam

## 4.2 Rozhodovacie stromy

Rozhodovacie stromy zaraďujeme k jednému zo základných princípov symbolických metód strojového učenia. Sú jedinečným nástrojom používaným na klasifikáciu a predikciu, ktorú budeme realizovať. Záujem o ne spôsobuje fakt, že v porovnaní s inými napríklad neurónovými sieťami, rozhodovacie stromy sú o pravidlách. Každé pravidlo je možné pomerne ľahko, vyjadriť prirodzeným, alebo databázovým jazykom.[23]

Rozhodovací strom zobrazí logický vývoj nadväzujúcich možností z hľadiska alternatívnych rozhodnutí a potenciálnych výsledkov. Je to klasifikátor zo stromovou štruktúrou. Vnútorné uzly sú rozhodovacie. Množstvo algoritmov na učenie rozhodovacích stromov je variáciou základného algoritmu.

Rozhodovací strom definujeme ako strom, kde všetky nelistové uzly stromu budú obsahovať test na hodnotu atribútu. Vetvy vedúce z tohto uzlu budú výsledky testu. Listové uzly stromu budú jednotlivé triedy, výsledky našej klasifikácie. Atribúty ktoré budeme testovať v jednotlivých uzloch pri budovaní stromu, vyberáme pomocou metód ako je entropia a informačný zisk.

Rozhodovací strom reprezentuje procedúru pri ktorej bude dataset na základe jeho vlastností zaradený do jeden z vopred definovaných tried. Vlastnosti datasetu sú reprezentované spojitými alebo diskrétnymi atribútmi. Najrozšírenejší spôsobom pri tvorbe rozhodovacieho stromu je induktívne generovanie s využitím metódy rozdeľuj a panuj.

Na začiatku bude koreňovému uzlu priradený atribút. Na základe hodnôt atribútu budú naše tréningové dáta rozdelené na podmnožiny. Pre každú z týchto podmnožín je vytvorený uzol, ktorému je následne priradený ďalší atribút. Toto delenie pokračuje až kým nie je splnená ukončovacia podmienka.

[23] Kostík Ľ., Saloky T., 2006, Riadenie a identifikácia systémov, **Niektoré z problémov pri získavaní dát pomocou rozhodovacích stromov**, AT&P journal PLUS2, Riadenie a identifikácia systémov

Dolovanie dát je proces výberu vhodných, skrytých, alebo nedefinovaných informácií z rozsiahlej databázy. Datamining je analýza, ktorá nie je vopred daná užívateľom. Charakter odvodených informácií je predovšetkým predikatívny.[24]

Metóda rozhodovacích stromov sa uplatňuje hlavne pri rozhodnutiach spojených s inováciami. Umožňuje nám identifikovať vzťah medzi stratégiou ktorú uplatňujeme a stratégiou ktorú by sa mali uplatniť v budúcnosti.

Existuje niekoľko typov rozhodovacích stromov. Sú určené na klasifikáciu do tried, alebo umožňujú predikovať numerické atribúty, regresné stromy. Prednosťou rozhodovacích stromov je ich ľahká interpretovateľnosť a prehľadnosť.

Vieme pomocou nich rýchlo a ľahko vyhodnotiť získané výsledky, identifikovať kľúčové položky. Rozhodovacie stromy nevyžadujú normalizáciu dát, alebo dodatočné atribúty. Vedia pracovať s kategorickými, aj numerickými znakmi.

### 4.2.1 Vytvorenie rozhodovacieho stromu

Najprv nájdeme atribút, ktorý obsahuje maximálne množstvo informácií. Tento atribút bude koreň stromu. Potom rozdelíme množinu príkladov na také množstvo podmnožín koľko je hodnôt koreňového atribútu. V jednotlivých podmnožinách sú príklady s jednou hodnotu tohto atribútu.

V každej podmnožine vyhľadáme ďalší najvýznamnejší atribút. Takto pokračujeme pokiaľ nespotrebujeme všetky atribúty Pri výbere najvýznamnejšieho atribútu poznáme niekoľko kritérií.

Meranie vhodnosti atribútu pre zistenie najvhodnejšieho nazývame informačný zisk. Udáva mieru rozdelenia atribútov do tréningových príkladov v ich cieľovej klasifikácii. Aby sme správne stanovili informačný zisk, musíme určiť entropiu, ktorá charakterizuje nečistoty v ľubovoľnej skupine príkladov.

[24] Novotný O., Pour J., Slánsky D., 2005, **Business Intelligence –** Jak využít bohatství ve vaších dátech, Grada Publishing, ISBN 80-247-1094-3

Množina bude obsahovať pozitívne a negatívne príklady nejakého cieľového plánu. Bude to binárna klasifikácia. Ako kritérium na výber najvhodnejšieho atribútu sa používa entropia, informačný zisk, pomerný informačný zisk alebo giny index.[25]

### 4.2.2 Problémy rozhodovacích stromov pri získavaní dát

V každom systéme sa stretávame s problémami. Najčastejší a najzávažnejší problém je určenie hĺbky, do akej bude rozhodovací strom rásť, spracovanie spojitých atribútov, výber správneho výberu atribútu, spracovanie cvičných dát ak chýbajú hodnoty atribútov, alebo spracovanie atribútov s odlišným ohodnotením a zvyšovanie efektívnosti výpočtu.

Vytvorenie rozhodovacieho stromu je pomerne rozsiahly proces, pri ktorom môže nastať niekoľko problémov. Tieto problémy môžu znížiť zrozumiteľnosť nášho modelu.

Zostrojené rozhodovacie stromy nemusia mať vždy vyhovujúci tvar, a potom nie sú vhodné pre rýchlu a správnu klasifikáciu. Ak strom dostatočne nezovšeobecňuje trénovacie dáta hovoríme, že došlo k preučeniu (*overfitting*) stromu, to znamená že sme docielili neúmernú presnosť stromu. Strom môžeme zjednodušiť tak, že príklady jednej triedy budú v listovom uzle len prevažovať.

Zjednodušenie, redukciu stromu môžeme urobiť dvoma spôsobmi.

1. Algoritmus budeme modifikovať doplnením nejakého kritéria, ktoré indikuje či má uzol ďalej expandovať. Týmto spôsobom sa redukovaný strom vytvorí priamo.

2. Vytvoríme úplný strom a následne urobíme jeho prerezávanie (*post-prunning*). Postupujeme zdola nahor a pri každom podstrome sa podľa nejakého kritéria rozhodneme, či sa má podstrom nahradiť listovým uzlom. Tento spôsob sa používa častejšie.[26]

[25] Berka P., Academia, Praha 2003, **Dobývaní znalostí z databází**, ISBN 80-200-1026–9

[26] Kostík Ľ., Saloky T., 2006, Riadenie a identifikácia systémov, **Niektoré z problémov pri získavaní dát pomocou rozhodovacích stromov**, AT&P journal PLUS2

### 4.2.3 Výhody a nevýhody rozhodovacích stromov

Medzi výhody rozhodovacích stromov patrí

- schopnosť generovať pochopiteľné dáta
- klasifikácia bez veľkého počítania
- schopnosť pracovať zo spojitými aj kategorickými premennými
- jasná indikácia najdôležitejších oblastí pre predikáciu, alebo klasifikáciu

Nevýhody rozhodovacích stromov

- nie sú vhodné ak cieľ predikácie je spojitý atribút
- rast rozhodovacieho stromu je náročný na výpočet
- je potrebné určiť limit do akej hĺbky má strom rasť
- sú náchylné k chybám pri klasifikácii problémov s veľa triedami a relatívne malým počtom tréningových príkladov.

### 4.2.4 Presnosť modelu

Po zostavení rozhodovacieho stromu, musíme odhadnúť jeho kvalitu. Rozhodovací strom je presnejší vtedy ak lepšie klasifikuje prvky z testovacej množiny. Medzi hlavné kritériá pre hodnotenie kvality stromu patrí miera chybovosti. Je to pomer nesprávne klasifikovaných príkladov a všetkých príkladov v našej testovacej množine.

Dôležitým krokom pri vytváraní modelu je ladenie. Prebehne po zhodnotení presnosti nášho modelu. Presnosť predikčného modelu hovorí o miere schopnosti modelu predikovať nad neznámymi dátami. To sú dátach, nad ktorými prebehne vlastná klasifikácia.

Trénovacie dáta použijeme k tvorbe klasifikátora. Nie sú vhodné k výpočtu presnosti modelu mohlo by to viesť k chybným, neprimerane optimistickým výsledkom. Presnosť modelu odhadujeme pomocou rôznych technik. Základom je výpočet chyby podľa určitej metriky.

## 4.3 Predikatívne techniky ktoré sa využívajú pri dolovaní dát

### 4.3.1 Kategorická a numerická predikcia

Kategorická predikcia predpokladá kategóriu, hodnotu na základe znalosti iných veličín. Je to v podstate klasifikácia. Cieľom klasifikácie je vytvoriť model s prívlastkom klasifikátor, ktorý bude predikovať nespojité hodnoty. Záznam v našom datasete bude zaradený len do jednej triedy. Predikovaný atribút je „label", ktorý môže mať charakter kategórie alebo triedy.

Numerická predikcia predikuje spojitú funkciu, predikcia časovej rady ako napríklad hodnota výšky úveru pre klienta ktorá je pre banku ešte bezpečná.

### 4.3.2 Priebeh klasifikácie

Priebeh klasifikácie sa skladá z dvoch krokov:

**Učenie** - vytvoríme klasifikačný model pomocou našich trénovacích dát. Trénovacie dáta sú dáta, pri ktorých vieme výsledky klasifikácie, čiže triedu do ktorej patria. Toto je učenie s učiteľom (*supervised learning)*. Nájdené mapovanie, alebo funkcia nám rozdelí niekoľko rozmerný priestor rozdelí na jednotlivé triedy.

Ide o vyhľadanie mapovania, alebo funkcie, pre ktorú bude platiť:

**Vzorec 1** $y = f(X)$

- **X** je vektor známych hodnôt atribútov vzorky
- **y** je predikovaný atribút (label).

**Vlastná klasifikácia -** použijeme vytvorený model na klasifikáciu nových dát, tu už label nie je známy. Základným krokom pred použitím modelu bude určenie presnosti modelu. Presnosť s akou predikuje jednotlivé triedy. Výpočet presnosti prebehne na testovacej množine. Prvky sú náhodne vybrané z celej množiny príkladov. Na klasifikáciu príkladov do tried môžeme použiť rôzne techniky. My sme si už vybrali rozhodovacie stromy.

### 4.3.3 Metriky

Pri kategorickej predikcie vypočítame presnosť výsledného klasifikátora ako pomer počtu správne klasifikovaných záznamov ku všetkým záznamom, vyjadrený v percentách. Môžeme uviesť aj chybovosť (*error rate*) klasifikátora. Vyjadríme ju ako:

**Vzorec 2** l-Acc(M)

- **Acc** je presnosť klasifikátora M.

Potrebným nástrojom pri porovnávaní presnosti jednotlivých tried je matica zámien (*confusion matrix*). Je to kontingenčná tabuľka medzi reálnym zatriedením príkladu a predikciami klasifikátora. Najlepšie je ak bude väčšina príkladov ležať na diagonále matice, klasifikátor ich správne zaradil do triedy.

Pri numerickej predikcii poznáme viac metrík ako pri kategorickej predikcii. Ak vytvorený model predikuje hodnotu spojitej funkcie, metriky sa sústredia na určenie vzdialenosti predikovanej hodnoty $y_i'$ od skutočnej hodnoty $\mathbf{y_i}$.

Medzi najčastejšie používané metriky patria:

Stredná absolútna chyba MAE (*mean absolute error*)

**Vzorec 3** $$MAE = \frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$$

Stredná štvorcová chyba MSE (*mean squared error*)

**Vzorec 4** $$MSE = \frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$$

- **d** vyjadruje veľkosť testovacej množiny

Odmocnina zo strednej štvorcovej chyby RMSE (*root mean squared error*) prednosťou tejto metriky je, že má rovnakú magnitúdu ako predikovaný atribút.

**Vzorec 5** **RMSE=** $\sqrt{MSE}$

Ak nás zaujíma relatívna chyba voči predikcii, ktorú by sme spravili na základe strednej hodnoty predikovaného atribútu, potom vypočítame relatívnu absolútna a štvorcovú chybu. Vo vzorci pre výpočet MAE a MSE nahradíme **d** výpočtom príslušnej strednej hodnoty.[27]

### 4.3.4 Techniky trénovania dát

Poznáme niekoľko techník, ktorými zistíme s akými dátami trénovať a potom vyhodnocovať presnosť modelu ktorý vznikol. Metóda holdout a krížová validácia patria medzi najpoužívanejšie.

Holdout je metóda pri ktorej náš trénovací model rozdelíme na dve náhodné množiny, trenovaciu a testovaciu množinu. Na určenie presnosti použijeme naše testovacie dáta. Ak je metóda holdout zopakovaná niekoľkokrát a presnosť je braná ako priemer presností jednotlivých iterácií, ide o náhodné vzorkovanie (*random subsampling*).

Krížová validácia alebo aj k-fold cross-validation je druhá technika. Dáta sú náhodne rozdelené a to do **k** disjunktných množín $\mathbf{D_1}$ až $\mathbf{D_k}$**.** Sú asi rovnakej veľkosti. Množina $D_i$ je pri i-tej iterácii braná ako testovacia. Ostatné množiny sú použité na trénovanie modelu.

Od holdout metódy sa odlišuje tým, že každý príklad je použitý rovnaký počet krát na učenie a raz na testovanie. Konečná presnosť pri klasifikácii je vypočítaná ako pomer počtu správnych klasifikácií zo všetkých iterácií k počtu všetkých záznamov.

Pri numerickej predikcii sa odhad chyby počíta ako pomer celkovej chyby zo všetkých iterácií k počtu všetkých záznamov.

### 4.3.5 Informačný zisk a entropia

V každom rozhodovacom uzle musíme zvoliť vhodný klasifikátor. Na kvantitatívne meranie hodnosti atribútu, využívame vlastnosť s názvom informačný zisk. Táto vlastnosť udáva mieru rozdelenia atribútov do cieľovej klasifikácie. Meranie sa vykonáva pri výbere atribútov v každom kroku rastu stromu. Na to aby sme mohli stanoviť informačný zisk musíme vypočítať entropiu.[28]

[27] **Principles of Data Mining,** Hand D.,Mannila H.,Smyth P., MIT Press, 2001, IBAN 0-261-08290-X

Entropia charakterizuje stupeň nečistoty v skupine príkladov.

**Vzorec 6** $$E(S) = -\sum_{t=1}^{T}(p_t \log_2(p_t)$$

- **S** je množina obsahujúca pozitívne aj negatívne príklady cieľového konceptu
- **E(S)** je entropia
- **$p_t$** je podiel pozitívnych príkladov v **S**
- **($p_t$)** je miera negatívnych príkladov v **S**

Pokiaľ všetky členy spadajú do jednej skupiny entropia sa rovná nule. Ak skupina obsahuje rovnaký počet pozitívnych aj negatívnych príkladov potom je entropia 1 čiže maximum. pre entropiu **E(S)** je možne definovať mieru efektivity atribútu čiže informačný zisk.[29]

**Vzorec 7** $$G(S,A) = E'(S) - \sum_{veV(A)} \frac{|S_v|}{|S|} E(S_v)$$

- **G(S,A)** je informačný zisk atribútu **A**
- **$S_v$** je podmnožina **S**

## 4.4 Algoritmy na tvorbu rozhodovacích stromov

Rozhodli sme sa že modelovanie našich dát zrealizujeme v programe Weka. Zozbierané a pripravené dáta budeme modelovať až keď vyberieme vhodný algoritmus. Vo Weke je k dispozícii spektrum algoritmov. My hľadáme algoritmus, ktorý bude najlepšie popisovať hľadané vzťahy v našich dátach.

Na to aký algoritmus vyberieme vplýva niekoľko faktorov. Patrí medzi ne aj to ako sme sformulovali problém, a aký máme typ zozbieraných dát. Je prirodzené že sa budeme skúšať niekoľko variant a štatisticky ich hodnotiť. Na tvorbu rozhodovacích stromov existuje mnoho algoritmov. Základné rozdelenie pozná tri druhy.

[29] Han J., Kamber M.; Diane C., **Data Mining Concepts and Technigues**; 2006; ISBN 978-1-55860-901

### 4.4.1 Algoritmy vytvárajúce jeden strom

Rozhodovacie stromy je oblasť umelej inteligencie ktorá má v súčasnosti veľké uplatnenie, a to nielen staršie typy algoritmov rozhodovacích stromov ( ID3, ID5R, C4.5 ...), ale aj novšie v oblasti umelej inteligencie, ako učenie neurónových sietí či genetické algoritmy.

Medzi jednoduchšie algoritmy patrí algoritmus ID3 (*Iterative Dichotomiser*). Algoritmus pri budovaní rozhodovacieho stromu vychádza z princípu známeho ako Occamova britva. To znamená že ak existuje viacero vysvetlení, použijeme to najmenej komplikované. Pri rozhodovacích stromoch je to najkratšia cesta od koreňa k listu. Pri výber atribútov používa ako rozhodovacie kritérium informačný zisk. Dokáže však pracovať iba s diskrétnymi atribútmi. Nástupcom algoritmu ID3 je algoritmus C4.5.

Algoritmus C4.5 pracuje na rovnakom princípe ako ID3, ale na rozdiel od neho používa ako rozhodovacie kritérium pomerný informačný zisk. Vie pracovať aj so spojitými atribútmi a učiť sa z príkladov v ktorých chýbajú hodnoty niektorých atribútov. J48 je open source Java implementácia algoritmu C4.5 dolovania dát vo Weke.

J48 je založený na myšlienke že pôvodná heterogénna množina príkladov sa dá postupne rozdeliť na homogénnejšie podmnožiny. Homogénnejšia množina má nižšiu entropiu než heterogénnejšia. Množina obsahujúca prvky výhradne jednej triedy je dokonale homogénna, má nulovú entropiu. Algoritmus J48 teda hľadá taký atribút ktorý rozdelí heterogénnu množinu na podmnožiny s najnižšou možnou entropiou.

Rekurzívne sa postup opakuje tak dlho pokiaľ sa dá entropia znižovať. V najhoršom prípade vzniknú ako listy stromu podmnožiny obsahujúce iba jeden prvok nastane preučenie.

Najpoužívanejším algoritmom dnes je CART (*Classification and Regresion Trees*). Popisuje vytváranie binárnych rozhodovacích stromov. CART delí príklady v každom uzle na dve skupiny. Ako kritérium pre výber atribútov je použitý Gini index.

### 4.4.2 Algoritmy vytvárajúce súbory stromov

Jeden model často nepopisuje vzťahy dostatočne. V snahe vytvoriť presnejší a róbusnejší klasifikátor vznikla myšlienka učiť viac modelov a potom ich použiť súhrnne. V každom

modeli je kladený dôraz na inú časť dát. Týmto sa zaoberajú ansámblové učiace systémy (enable learners). Medzi najznámejšie ansámblové učiace systémy patrí bagging, boosting a random Forests.[30]

### 4.4.3 Genetické algoritmi a rozhodovacie stromy

Použitie techniky rozhodovacích stromou a genetických algoritmov prináša zo sebou zásadný problém. Reprezentácia stromov vo forme chromozómou. Stromy nemajú fixne danú štruktúru ani dlžku. Medzi dva najznámejšie patri algoritmus GA Tree a GA-ID3.

## 4.5 Metodológia CRIPS-DM

Proces dataminingu zahrňuje niekoľko metód a postupov práce. Nie je jednoznačný návod ako postupovať. V 90. rokoch sa vykryštalizovali dve všeobecne vhodné metodológie: metodológie *SEMMA*, od firmy SAS, a *CRISP-DM*, vyvinutá firmami medzi ktoré patrila aj firma SPSS.

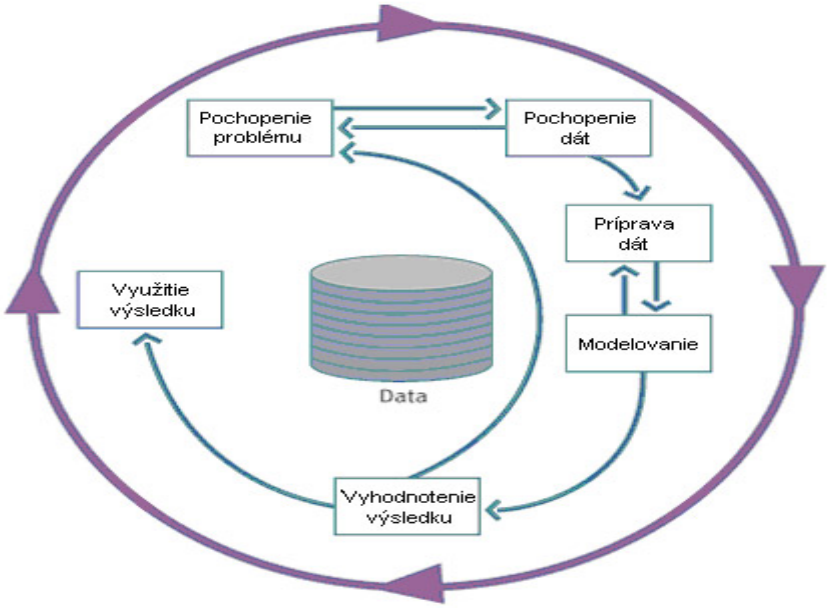



**Schéma 1.** Grafické znázornenie metodológie CRIPS DM. **Zdroj:** neuron.tuke.sk/ zvada/statnice/ II/08/index.html

[30] Dietterich T., **Ensemble Learning,** The MIT Press 2002, www.eecs.wsu.edu/~holder/ courses/ CptS570/fall07/ papers/Dietterich00.pdf

Metodológia CRISP-DM (Cross Industry Standard Process for Data Mining) realizuje projekt rýchlejšie, účinnejšie a lacnejšie. Sú to doporučené postupy obsahujúce celý životný cyklus projektu dolovania dát. Je rozdelený do 6 etáp. Poradie nie je pevne dané, ktorá fáza nasleduj záleží na výsledku predchádzajúcej fázy.

Hrany v grafe na obrázku zachytávajú najčastejšie závislosti medzi fázami. Ak vo fáze modelovania nevznikne model s vyhovujúcou presnosťou, vrátime sa späť do fázy prípravy dát, alebo až do fázy pochopenia problému.

Externý cyklus zobrazuje cyklickú povahu procesu dolovania dát. Po použití modelu, môže proces dolovania dát pokračovať. Môžu sa objaviť ďalšie problémy, ktoré súvisia s primárne riešenou úlohou. Ďalšie procesy môžu použiť poznatky získané pri riešení primárnej úlohy.

### 4.5.1 Jednotlivé fázy metodiky CRIPS-DM

**Pochopenie problému** - vstupná fáza, je orientovaná na pochopenie problému. Stanovíme ciele a požiadavky z business pohľadu. Poznatky ďalej transformujeme do popisu úloh dolovania dát a približného návrhu plánu, ako úlohu riešiť.

**Pochopenie dát** - fáza začína získaním dát. Potom je potrebné zoznámiť sa s dátami, určiť prípadné problémy s kvalitou dát. Je možné, že zistíme určité vzory, ktoré nám pomôžu pri zostavení prvých hypotéz.

**Príprava dát** - vytvoríme množinu dát z pôvodných dát, ktorú použijeme ako vstup modelu. Táto činnosť sa opakuje a obsahuje výber záznamov, atribútov, čistenie a transformáciu dát, podľa špecifických požiadaviek na formát dát.

**Modelovanie -** je fáza v ktorej požijeme rozličné metódy a techniky na vytvorenie modelu, pričom dochádza k optimalizácii parametrov. Jeden typ problému sa dá riešiť niekoľkými technikami a často majú tieto techniky rôzne nároky na dáta. Preto je v tejto fáze časté, že sa musíme vrátiť do fázy prípravy dát

**Hodnotenie modelu** - táto fáza hodnotí úroveň s akou model dosahuje definované obchodné ciele. Ak výsledky nie sú uspokojivé, snaží sa určiť dôvod, prečo je model neuspokojivý. Ak sú dosiahnuté výsledky vyhovujúce, urobíme konečnú kontrolu celej úlohy. Zistíme či sme neprehliadli dôležitý faktor. Rozhodneme či sa výsledný model použije, alebo sa vrátime do niektorej z predchádzajúcich fází.

**Nasadenie modelu** - ak model prináša novú znalosť, pokračujeme tak že túto znalosť prezentujeme zákazníkovi tak, aby ju mohol použiť. Prezentácia môže mať rôzne formy, väčšinou sa jedná o jednoduchý report, alebo iný nástroj na prezentovanie.

# 5 Praktická časť práce

## 5.1 Pochopenie problému

Ako praktickú časť práce sme zvolili predikáciu návštevnosti a ceny izby v hotely pomocou rozhodovacích stromov v súlade s metodikov CRIPS-DM. Použitím predikcie na tento konkrétny problém prispejeme ku kvalitnejšiemu a efektívnejšiemu rozhodovaniu manažmentu a zlepšenie služieb zákazníkom.

## 5.2 Definovanie problému

Budeme skúmať jeden z najpodstatnejších ukazovateľov úspešnosti hotela a to je návštevnosť a cena služieb. Od tohto ukazovateľa sa samozrejme odvíja tržba. Predikácia návštevnosti prinesie manažérovi niekoľko výhod medzi iným aj schopnosť plánovať zdroje.

S plánovaním zdrojov a tržieb súvisia niektoré výhody ako napríklad odvodenie tržieb. Ak dokážeme predikovať návštevnosť, môžeme si z toho odvodiť výšku tržieb, následné vieme plánovať cashflow, čo znamená že vieme rozhodnúť kedy a koľko môžeme investovať.

Ak vieme predikovať návštevnosť, môžeme zlepšiť management ľudských zdrojov. V prípade vysokej návštevnosti dokážeme včas zabezpečiť na jednotlivé úseky personál, prípadne plánovať presun, alebo voľno zamestnancom v pri nižšej návštevnosti.

V prípade ak nám predikcia dlhodobo vykazuje nižšiu návštevnosť, môže manažér rozhodnúť o použití reklamy, alebo použiť iné spôsoby na prilákanie hostí. Reklama v čase predikácie vysokej návštevnosti je neefektívna, z dôvodu obmedzených kapacít hotela.

## 5.3 Súčasná situácia skúmaného objektu

Vybrali sme spoločnosť Hotel SET s.r.o, Bratislava[31]. Spoločnosť podniká v oblasti poskytovania ubytovacích a stravovacích služieb tuzemským aj zahraničným zákazníkom a firmám.

Tiež ponúka prenájom konferenčných priestorov, a technickú podporu pri ich usporiadaní. Spoločnosť je vlastníkom ubytovacích priestorov a reštauračného zariadenia. Má kapacitu 26 izieb v rôznych kategóriách v ktorých je možné ubytovať asi 55 hostí. Kapacita reštaurácie je asi 70 hostí.

Výhoda je že hotel sa nachádza hneď vedľa Zimného štadióna Ondreja Nepelu, Futbalového štadióna AŠK Inter Bratislava, je prepojený s Národným tenisovým centrom NTC. A v dohľadnej budúcnosti bude dostavaný Futbalový štadión Slovan, Tehelné pole. Všetky tieto športové centrá sú v tesnej blízkosti hotela. Niekoľko krát do týždňa sú v týchto priestoroch organizované kultúrne podujatia.

Nevýhodou je, že v okolí pribúdajú kapacitne väčšie a novšie ubytovacie zariadenia, a v dôsledku poklesu vyťaženosti hotela spôsobeného aj hospodárskou krízou je nutné znižovať ceny ubytovania. Hotel je aktívny v predaji svojich voľných kapacít cez niekoľko ubytovacích portálov. Tu je však nutné znižovať ceny na hranicu rentability.

Je to malý podnik do 20 zamestnancov, z ktorých väčšinu tvoria recepčný, obsluha reštaurácie, chyžné, administratíva a riaditeľ spoločnosti.

[31] **Hotel SET** homepage, www.hotelset.sk

### 5.3.1 Hotelový informačný systém

Pre rezervácie hotel používa komerčný rezervačný systém Horec. Je to automatizovaný systém, ktorý odstraňuje náročnú evidenciu objednávok a materiálnych zásob. Umožňuje generovanie jednoduchých reportov, prípadne tvorbu jednoduchých analýz ako je napríklad analýza vyťaženosti.

Bol použitý ako prvý, ale nespĺňa všetky požiadavky na funkcionalitu, hlavne neumožňuje získať všetky informácie a v požadovanej podobe o zákazníkoch a partneroch. Tiež neumožňuje získavať a zaznamenať externé informácie napríklad o spoločenských podujatiach.

Informačný systém hotela preto čerpá dáta ktoré potrebuje z rôznych prameňov. Napríklad: základné údaje o zákazníkovi doplní recepčný z dokladu totožnosti pri príchode hosťa do systému Horec, ale ak je hosť zároveň zamestnancom firmy, nie je v tomto programe možnosť uviesť podrobnejšie informácie o firme a ani záznam znovu vygenerovať napríklad ako fakturačná adresa ak ho potrebujeme.

Na zaznamenanie tejto informácie je potrebné vytvárať extra tabuľky v Exceli, alebo v programe Outlook a potom ich pracne aktualizovať. Ak potrebujeme fakturovať, musíme kopírovať, alebo dopisovať tieto údaje ručne. Údaje o dochádzke zamestnancov sú niekoľko krát ručne prepisované podľa potreby manažéra.

Informácie o tržbách musí účtovník zozbierať osobne, každé ráno na prevádzke a potom si ich zaviesť do účtovného programu, ktorý nie je s programom Horec kompatibilný. Aj keď je systém pre prehľad tržieb najkomplexnejší, nie je prenos dát z jedného systému do druhého automatický .

Horec umožňuje vytvorenie účtov pre jednotlivých hostí, denný prehľad uzávierok. Tento systém umožňuje prepojenie na ďalšie podsystémami na príklad reštauračné zariadenie a sklad.

Dáta o návštevnosti tržbe alebo osobné údaje zákazníkov z programu Horec sa ukladajú do databázy. Reporty sa negenerujú z dátového skladu, ale z relačnej databázy pomocou SQL dotazov.

Aj z týchto dôvodov trvá vytvorenie reportu dlho. Čas závisí od počtu záznamov a úrovne agregácie. Reportovanie je nedostatočne flexibilné. Report neposkytuje požadovanú dimenziu ani úroveň agregácie.

V prípade ak je potrebný nový report, iný pohľad na aktuálny stav, iná dimenzia, alebo úroveň agregácie, nie je možné si ho nastaviť cez užívateľské prostredie a vygenerovať, ale je nutné ho do Horca doprogramovať pracovníkmi firmy Horec. Aj preto je práca s analytickými a reportovacími nástrojmi v tomto systéme používateľsky neprívetivá.

Tieto problémy by vyriešilo nasadenie technológie business intelligence z databázových a analytických komponentov, čo by znamenalo rýchlejšiu a efektívnejšiu prácu. To prinesie okrem pohodlnejšej práce aj ďalšie pridané hodnoty.

## 5.4 Porozumenie dátam a spôsob zbierania údajov

Definovali sme problém ktorý chceme riešiť a to ako nastaviť ceny v dňoch keď návštevnosť pravidelne klesá. Na základe zozbieraných údajov od zamestnancov a z dostupných zdrojov, dokážeme vysloviť prvé hypotézy, prvotnú analýzu dát. Zistili sme hlavné faktory ktoré ovplyvňujú návštevnosť na základe rozhovoru zo zamestnancami a zároveň sme zo systému získali informácie o denných tržbách a obsadenosti izieb za niekoľko rokov dozadu.

Zdrojom je tabuľkový procesor Excel kde je niekoľko rokov zaznamenávaná tržba a obsadenosť. Ako tréningový dataset použijeme dáta za rok 2011 a 2012 a pokiaľ by presnosť modelu nebola dostatočné pridáme rok 2010. Ako predikačné dáta použijeme rok 2013.

### 5.4.1 Vstupné údaje od zamestnancov

Manažéri ktorý na danom pracovisku pracujú, majú osobné skúsenosti a budú reporty a analýzy využívať ku svojej práci. Odpovede pracovníkov sa väčšinou zhodujú. Spoločne označili niekoľko hlavých faktorov zníženie návštevnosti v hotely.

**Deň v týždni** – hotel navštevuje prevažne stabilná skupina hostí. Počas pracovných dní sú to zástupcovia firiem, ktorý prichádzajú do mesta pracovne. Počas dňa pracujú, prichádzajú prevažne na 1-3 dni. V čase prázdnin a víkendov alebo pracovného voľna je ich návštevnosť nižšia, a teda je nižšia aj návštevnosť hotela.

**Mesiac v roku** – z predchádzajúceho vyplýva, že podstatný vplyv na návštevnosť má aj mesiac v roku. Návštevnosť teda vykazuje sezónnosť. Výrazne slabšie sú mesiace v ktorých firmy končia, alebo začínajú svoju činnosť, alebo mesiace keď zamestnanci čerpajú svoje dovolenky.

**Reklama** – uverejnenie reklamy v médiách významne vplývajú na návštevnosť. Väčšinou sa však jedná o krátkodobé zvýšenie návštevnosti. Prichádzajú noví návštevníci a pokiaľ sú spokojný vracajú sa znovu. Rozhodujúca je aj cena, ktorá je väčšinou v reklame nižšia ako sú ceny bežné.

**Spoločenské podujatia v okolí** – výrazne zvyšujú návštevnosť. Ubytovanie je však len na jednu noc, čiže má len krátkodobí charakter. Problém je že hostia preferujú lacné ubytovanie, a prevažne rezervujú cez rezervačná portály. Mnohokrát dlho dopredu, keď manažment ešte netuší že nejaké podujatie bude. Čiže ak manažér plošne nastaví ceny na slabšie víkendy nižšie, hotel prerába.

**Cena** – zákazníci preferujú lacné ubytovanie. Voľné ubytovanie hľadajú cez ubytovacie portály, kde je prvoradým kritériom pri výbere ubytovania cena, až potom úroveň ubytovania a ďalšie dôvody . Aj hotely v okolí využívajú znižovanie cien na prilákanie návštevníkov.

**Ostatné** – sem treba zaradiť čistotu prostredie, služby poskytované hosťom, úroveň stravovania, ale aj dostupnosť a vzdialenosť od cieľa kam má hosť namierené. Tiež bola často uvádzaná propagácia mesta a podujatí. Tieto faktory majú subjektívny charakter.

## 5.5 Príprava dát

Hotelové údaje boli pre naše účely čiastočne modifikované, aby nebola porušené diskrétnosť interných informácií. Táto časť práce nám zabrala najviac času. Bolo potrebné aby sme sa k tejto časti niekoľkokrát vrátili a údaje upravili, tak aby boli vhodná na ďalšie spracovanie a mali dostatočnú vypovedaciu schopnosť.

Všetky údaje sú kompletne nahraté na priloženom CD, alebo čiastočne uvádzané v texte a v prílohe na konci práce, ak to ich rozsah dovoľuje. V texte diplomovej práce sú priebežne odkazy na tieto materiály.

Miera kvality a relevantnosti údajov má priamy vplyv na správnosť výsledku. Keďže pracujeme s dátami ktoré boli určené na iné účely, bude nevyhnutné pred použitím dáta pripraviť.

Premenné v pôvodnej tabuľke boli dátum a tržba, ktoré sú menej dôležité a ďalej sú to premenné deň v týždni, mesiac, obsadenosť a podujatia v okolí hotela, ktoré sú pre našu predikciu podstatné.

Dataset obsahuje asi 700 riadkov údajov. Príprava dát sa skladá z vyhladzovania, generalizácie, diskretizácia, normalizácia, konštrukcia atribútov. Chýbajúce údaje vzhľadom k tomu že je to malý dataset sme nezaznamenali. Vyhladzovanie dát nebolo nutné.

Vzhľadom k tomu, že premenná dátum patrí medzi menej dôležité, a mi chceme predikovať návštevnosť a cenu počas dní keď v okolí prebieha podujatie, technikou kategorickej predikcie, urobili sme diskretizáciu a generalizáciu. Atribúty musia byť nominálne. Dátum sme zamenili za slovné názvy mesiacov - mesiac.

Premennú tržba sme použili len na prvotnú analýzu, potom sme ju vyradili z datasetu, vzhľadom k tomu, že ceny za izbu a noc nie sú jednotné, pevne stanovené. Na základnú pultovú cenu môže byť kedykoľvek poskytnutá zľava, ktorá nie je zaznamenaná. Tento údaj nemaná preto pri predikácii dostatočnú vypovedaciu schopnosť.

Premennú obsadenosť sme generalizovali do troch skupín. Nízka obsadenosť 0-8 (vrátane 8), stredná obsadenosť 9-16 (vrátane 16) a vysoká obsadenosť 17- 26. Týmto chceme podporiť proces dolovania dát, zvýšiť presnosť a porozumenie mnohorozmerným dátam v rozhodovacom strome.

Premennú podujatie sme skonštruovali a pridali ako nový atribút. Podporili sme tak proces dolovania dát, pretože táto premenná vysvetľuje prečo sa návštevnosť zvyšuje aj mimo už známych dní, kedy je návštevnosť vyššia v pravidelných intervaloch, ako to vidno na grafoch prvotnej analýzy.

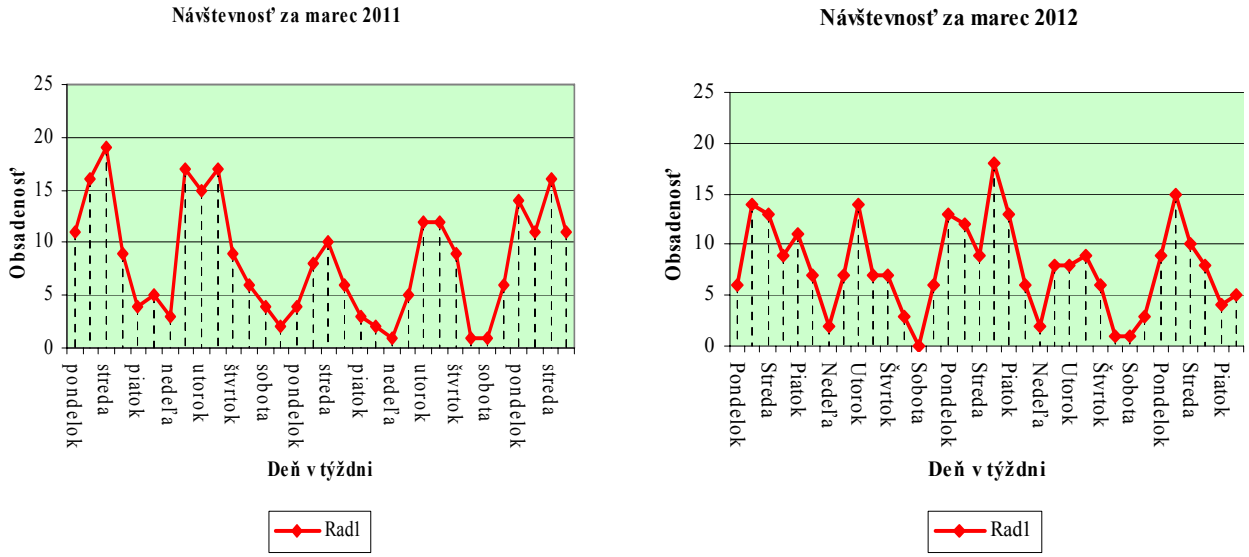
| Deň | Dátum | Tržba (€) | Obsadenosť | | Deň | Dátum | Obsadenosť | Akcia | Cena |
|---|---|---|---|---|---|---|---|---|---|
| sobota | 1.1.2011 | 297.00 | 5 | | sobota | Január | nízka | nie | znížiť |
| nedeľa | 2.1.2011 | 0.00 | 0 | | nedeľa | Január | nízka | nie | znížiť |
| pondelok | 3.1.2011 | 80.00 | 1 | | pondelok | Január | nízka | nie | znížiť |
| utorok | 4.1.2011 | 80.00 | 1 | | utorok | Január | nízka | nie | znížiť |
| streda | 5.1.2011 | 0.00 | 0 | | streda | Január | nízka | nie | znížiť |
| štvrtok | 6.1.2011 | 75.00 | 1 | | štvrtok | Január | nízka | nie | znížiť |
| piatok | 7.1.2011 | 150.00 | 5 | | piatok | Január | nízka | nie | znížiť |
| sobota | 8.1.2011 | 150.00 | 2 | | sobota | Január | nízka | nie | znížiť |
| nedeľa | 9.1.2011 | 85.00 | 1 | | nedeľa | Január | nízka | nie | znížiť |

**Tabuľka 2.** Ukážka dát pred a po príprave. **Zdroj**: Vlastné spracovanie

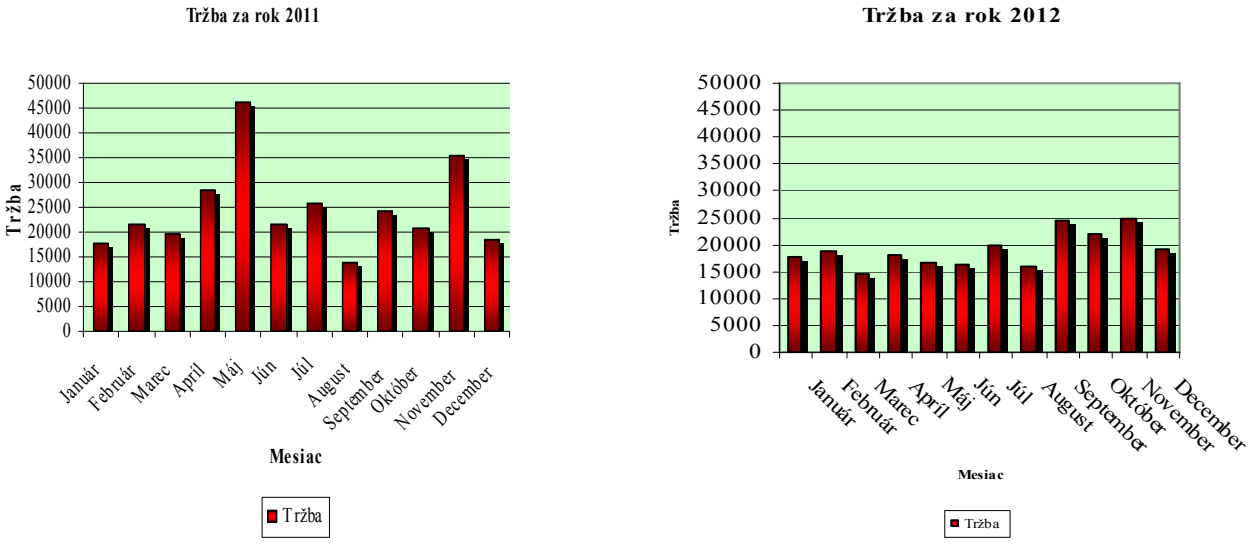
## 5.6 Analýza dát

Je to fáza zoznámenia sa s dátami. V tejto časti je možné formulovať prvé hypotézy, nájsť vzory ktoré sa opakujú. Za vzorovú množinu sme vybrali údaje o tržbe a návštevnosti v rokoch 2011-2012.

Vo fáze oboznamovania s dátami zakreslíme časť údajov ako časovú radu do grafu Návštevnosť podľa dní v týždni. Najvhodnejší spôsob interpretácie je vizualizácia. Zobrazenie pomocou grafov je rýchle a efektívne.




**Graf 1.** Návštevnosť za mesiac marec v rokoch 2011-2012. Zdroj: Vlastné spracovanie



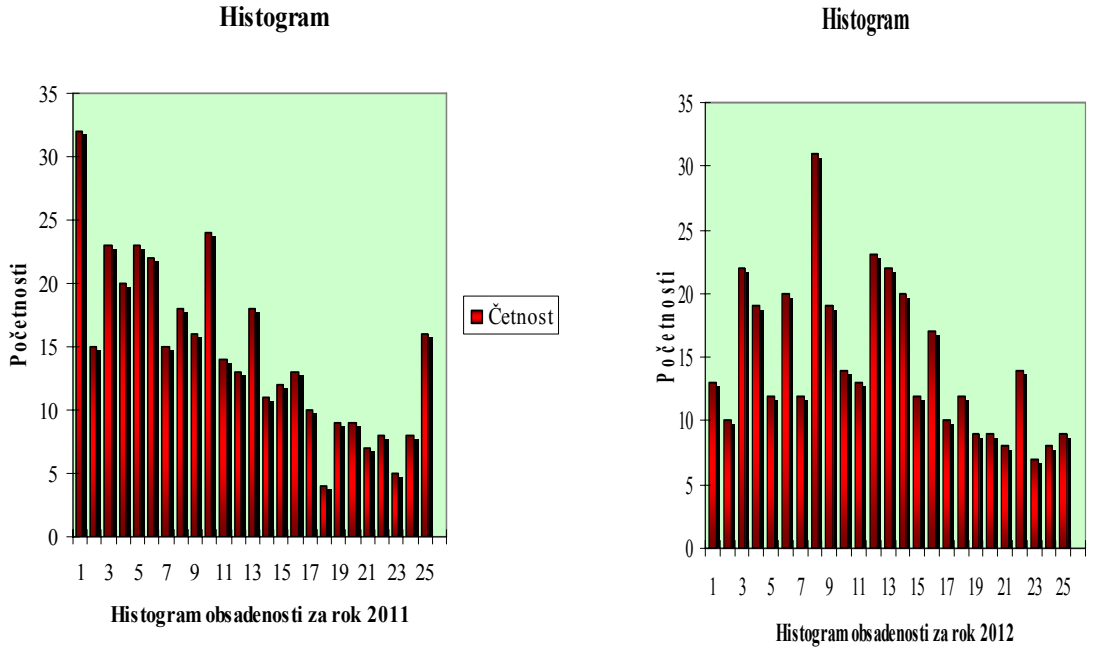

**Graf 2.** Tržba za rok 2011 a 2012 po mesiacoch. **Zdroj:** Vlastné spracovanie

Grafy návštevnosti majú približne rovnaký priebeh, pozorujeme tu značnú periodicitu. Graf začína vždy v pondelok, aj vtedy ak sme museli porušiť predpoklad, že údaje sú za mesiac marec. Návštevnosť sa od pondelka do stredy zvyšuje a ku koncu týždňa sa znižuje.

Grafy tržieb sa tiež zhodujú. Je tu vidieť sezónnosť. Tržby klesajú v letných mesiacoch počas prázdnin a znovu v decembri. Lepšie je to vidieť za rok 2011, ale grafy agregovaných tržieb z roku 2010 a 2009 ktoré sú v prílohe to potvrdzujú. Pri podrobnejšom skúmaní by sme zistili aj na pohľad že návštevnosť a tržby klesajú najviac v mesiacoch kde je veľa štátnych sviatkov, alebo si ľudia čerpajú dovolenky.

Z toho by sme mohli usúdiť že základ hotelových hostí tvoria zástupcovia firiem. Lokálne minimá sú zaznamenané v čase prázdni a v decembri. Lokálne maximá sú zaznamenané v marci, v októbri a novembri.

Pre ešte lepšiu prehľadnosť návštevnosti sme zostrojili histogram návštevnosti za roky 2011 a 2012. Priemerná početnosť obsadenosti je 3-15 izieb.




**Graf 3.** Obsadenosť za rok 2011 a 2012 maximálna a minimálna početnosť. Zdroj: Vlastné spracovanie

## 5.7 Algoritmus modelu

Zozbierané a pripravené dáta budeme modelovať. Vybrali sme algoritmus J48, ktorý sme už popísali v metódach práce a použijeme ho na naše údaje. Weka ponúkne len algoritmy vhodné pre náš typ dát, ostatné algoritmy sú neaktívne. Hľadáme algoritmus, ktorý bude v našom modeli najlepšie popisovať hľadané vzťahy v dátach.

Na to aký algoritmus vyberieme vplýva mnoho faktorov. Patrí medzi ne aj to ako je formulovaný problém. Náš problém je pre väčšiu názornosť a pochopenie formulovaný slovne. Chceme vedieť či ceny za ubytovanie znížiť, ponechať alebo zvýšiť v prípade že v okolí podniku je organizované podujatie aj keď všeobecne v tento deň, časť týždňa je návštevnosť nízka.

Veľký vplyv na výber algoritmu má typ zozbieraných dát. Naše dáta sú nominálne. Je možné, že budeme musieť vyskúšať niekoľko variant modelov a potom vybrať ten, ktorý bude vyhodnotený najlepšie. To zistíme ak model, učiace schémy, ohodnotíme matematickými a štatisticky metódami.

Pri klasifikácii je cieľom vytvoriť funkciu, ktorá zaradí náš príklad do jednej z preddefinovaných tried. V našom prípade sú tieto triedy zníženie, ponechanie, zvýšenie ceny.

# 6 Výsledky práce

## 6.1 Použitie modelu

Najjednoduchší spôsob použitia Weky je cez grafické užívateľské rozhranie Explorer. Cez toto rozhranie máme prístup ku všetkým nástrojom. Data mining začína prípravou dát, a ich prevedením do formátu .arff. Formát .arff je prirodzená metóda ukladania údajov.

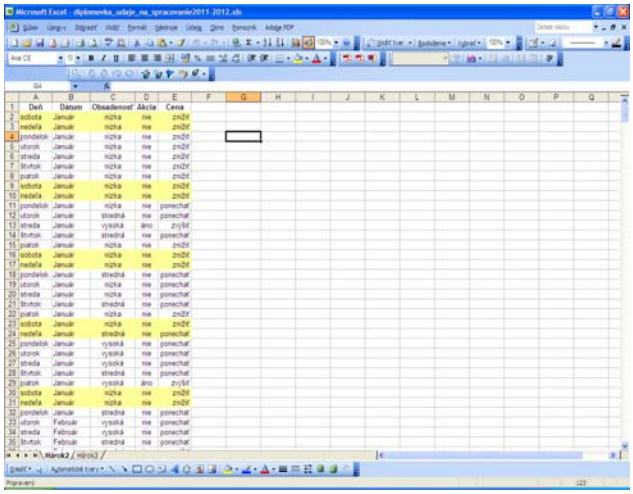
Súbor .arff je rad prípadov a hodnôt atribútov. Každý prípad je oddelený čiarkou. Väčšina tabuľkových a databázových programov umožňuje export dát vo formáte s čiarkami .csv. (*coma separated value*), list záznamov oddelených čiarkami.

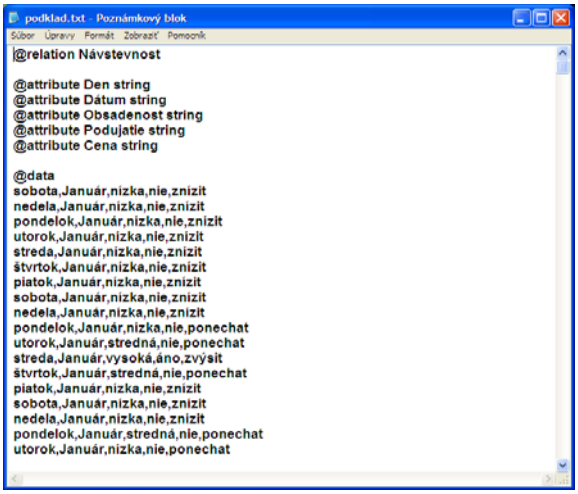
Dáta z roku 2011a 2012 prevedieme do jedného hárku v programe Excel. Údaje musia byť uložené iba na jednom hárku. Názvy stĺpcov ponecháme. Budeme potrebovať premenné deň, mesiac, obsadenosť, podujatie, cena. Radšej píšeme bez diakritiky. Weka ponúka možnosť v prípade chýbajúcich údajov doplniť miesto toho otáznik. Naše údaje boli kompletné.

## 6.2 Načítanie dát do modelu

Dataset je potrebné previesť do podoby, ktorú analytický software akceptuje. V hárku Excel nastavíme na karte **nástroje** → **možnosti** oddeľovač desatinných miest – bodku, ak je to nutné, pretože údaje sú numerické. Nastavíme oddeľovač stĺpcov a to cez **údaje** → **text na stĺpce** → ako oddelené, čiarka, formát údajov všeobecné. Takto upravíme všetky stĺpce nášho datasetu navštevnosť.xls.

Dáta z Excelu uložíme vo formáte .csv. V zošite musí byť len jeden hárok. **Súbor** → **uložiť ako**, vyberieme formát .csv oddeľovač s čiarkami. Excel nás upozorní že hárok obsahuje funkcie, ktoré nie sú kompaktibilná s .csv, nie je možné ich uložiť v tomto formáte a po odsúhlasený budú tieto odstránené. Odsúhlasíme tento postup.






**Obrázok 2** Údaje o návštevnosti vo formáte .xls. a .txt Zdroj: vlastné spracovanie

Originál tabuľky a textový dokument sú nahrané na CD nosiči. Pôvodný zošit Excel zatvoríme. Otvoríme v Exceli súbor **návštevnosť.csv** a skontrolujeme či je všetko v poriadku. Súbor zatvoríme a otvoríme ho vo formáte .txt. V našom prípade sme si ho otvorili v Notapede. Pred údaje pridáme názov, menovky a typy atribútu.

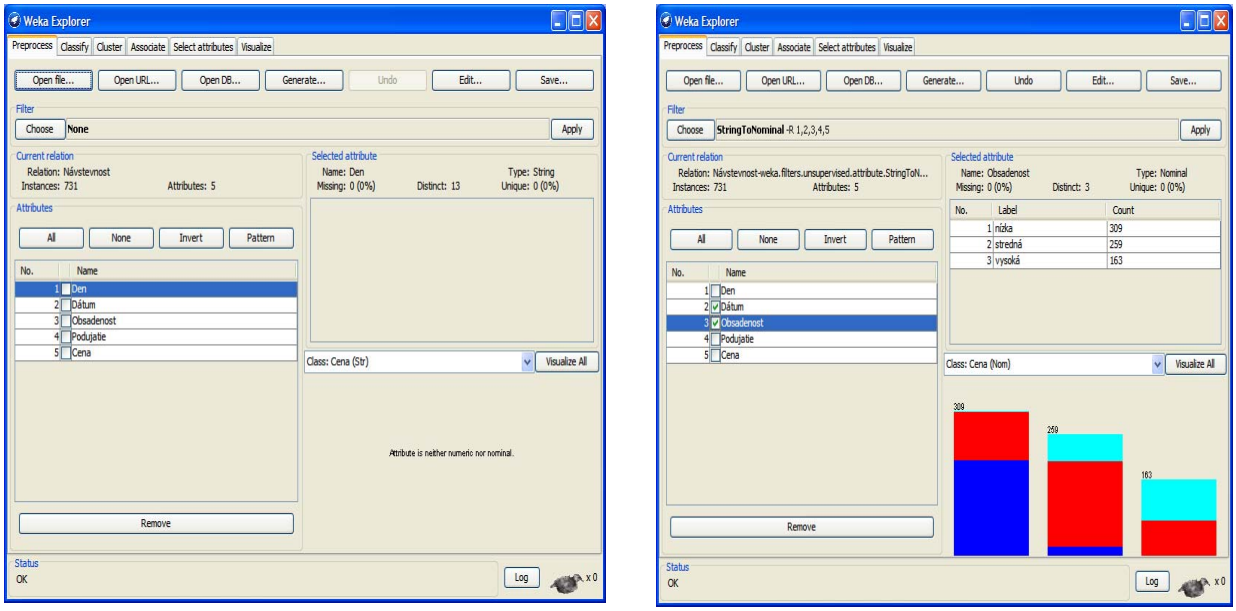
Data vo formáte .txt, v textovom editore upravíme. Pridáme dopredu názov datasetu s menovkou **@relation → navštevnosť @relation,** pridáme informácie o atribútoch s menovkou a s označením typu atributu, **@attribute → deň@attribute string.** Typ atribútu môže biť nominal, numeric, string a data. Pred samotné dáta pripojíme menovku **@data.** Všetky naše atribúty sú typ string.

Otvoríme aplikáciu Weka a načítame dáta. Karta **Preproces → Open file** a nájdeme náš dataset **navstevnost.csv**. Pri pokuse o načítanie nemusí ísť všetko hladko. Je dôležité dodržať všetky pravidlá pre načítanie dát vo formáte .arff. V prípade ak ani potom Weka nevie načítať náš súbor, ponúkne nám automatickú pomoc cez filtre, ktoré dokážu dáta zmeniť alebo automaticky upraviť tak aby boli pre aplikáciu zrozumiteľné. Krátky zoznam filtrov sme opísali pri predstavovaní prostredia programu Weka.

## 6.3 Úprava dát pre model

Pokiaľ nám Weka oznámi že nevie rozoznať dáta a ponúkne nám automatický konvertor, vo filtroch vyberieme **Choose** a potom náš nahrávač dát **Arffloader**. V základnom okne na ľavej strane sa objavia názvy atribútov, stĺpcov datasetu. Tiež môžeme postupovať takto: pozatvárame všetky súbory a v zošite v Exceli prepíšeme príponu . csv na .arff. Automaticky sa súbor zmení na súbor v programe Weka.

Pokiaľ sú v našom datasete dáta iné ako numerické, tie sú prednastavené aplikáciou, musíme znovu pomocou filtra určiť v ktorom atribúte, stĺpci sú dáta akého typu. V boxe **Filter** otvoríme ponuku **Choose → Filters → unsupervised → attributes**, a zvolíme **stringtonominal**. To sa objaví v okne Choose. Klikneme na názov filtra a do riadku attributerange doplníme čísla stĺpcov v ktorých máme dáta typu string, oddeľujeme čiarkami v našom prípade stĺpce 1,2,3,4,5. Zadáme aplikovať. Okno predspracovania údajov je vo väčšom formáte v prílohe.




**Obrázok 3** Okno predspracovania údajov v programe Weka. Zdroj: vlastné spracovanie

Ako náhle sú dáta načítané panel **Preprocess** ukazuje v boxe **Current relation** (aktuálny stav) tieto informácie:

**Relation** – je to názov relácie, ktorý bol priradený z názvu súboru. V našom prípade je to relácia Návstevnost.csv. Pomocou filtra je možné zmeniť názov relácie.
**Instance** – je počet riadkov záznamov. V našom prípade 731.
**Atribúty** - sú stĺpce tabuľky, v našom prípade 5.

Pod boxom Current relation je box s názvom **Atributes**. K dispozícii sú štyri tlačidlá, a pod nimi je zoznam aktuálnych atribútov. Naše atribúty sú deň, dátum, obsadenosť, podujatie, cena. Atribúty sú označené číslom, je možné ich vyberať pomocou výberového okienka a uvádza sa meno atribútu.

Ak klikneme na niektorí z riadkov v zozname atribútov, v poli **Selecte attribute** na ľavej strane sa zobrazia vlastnosti aktuálne označeného atribútu v zozname.

**Name -** je rovnaké ako v atribútoch navstevnost.arff
**Type** – naše atribúty sú typu string
**Missing** – počet (podiel) chýbajúcich riadkov, údajov je 0 (0%).

**Distinct** – je počet rôznych hodnôt na ktoré sa delia dáta cena, nízka, stredná, vysoká
**Unique** – je počet jedinečných inštancií, tých ktorých hodnotu nemá žiadna iná inštancia, čo je v našom prípade 0 (0%)

Pod týmito štatistikami je zoznam zobrazujúci viac informácií o hodnotách uložených v atribúte. Líšia sa v závislosti od typu. Ak je atribút číselný, zoznam uvádza štyri štatistické údaje opisujúce hodnotu dát – minimálna, maximálna a priemerná hodnota a tiež smerodajná odchýlka. Ak je atribút nominálny, ako v našom prípade „cena“ zoznam sa skladá z hodnôt nízka s početnosťou 309, stredná – 259, a vysoká – 163.

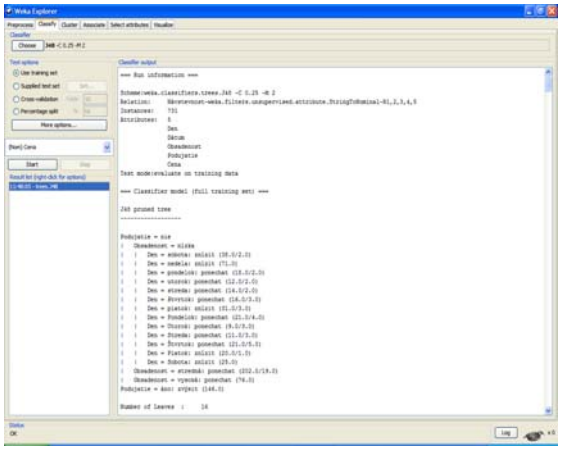
Pod takouto štatistikou je farebný histogram. Farebne rozlíšenie je podľa rozdelenia hodnôt v atribúte. V našom prípade vidíme atribút cena ktorý je rozdelený na tri hodnoty nízka, stredná, vysoká. Iba nominálne atribúty majú farebné kódovanie. Po stlačení tlačidla vizualizovať všetky atribúty, otvorí sa spoločné okno pre histogrami všetkých atribútov.

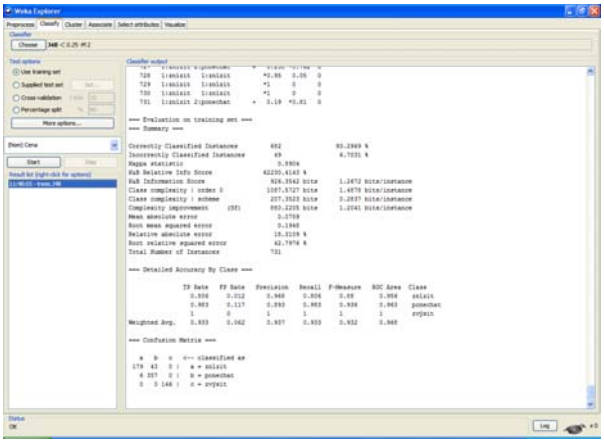
Každý krok je možné vrátiť späť kliknutím na tlačidlo Undo, vedľa tlačidla Edit v pravom hornom rohu panela predspracovania. Tlačidlom Edit je možné upravovať dáta.

Naše údaje chceme preskúmať pomocou rozhodovacích stromov. V hornej časti okna zvolíme box **Classify**. Tento box má textové pole, v ktorom je názov aktuálne vybraného triedenia, a jeho možnosti. Tlačidlom **Choose** si vyberieme z klasifikátor. V našom prípade **Choose → trees → J48.** V **Test Options** vyberieme **Use tranig set** a v ďalších možnostiach doplníme okrem predvolených výstupov aj možnosť **Output entropy** a **Output predictions**.

Nastavíme klasifikátor, možnosti skúšok a triedu. Proces učenia spustíme kliknutím na tlačidlo Štart. Keď školenie je kompletné, výstupná oblasť popisuje výsledky školenia a testovania. Preskúmame výstup.

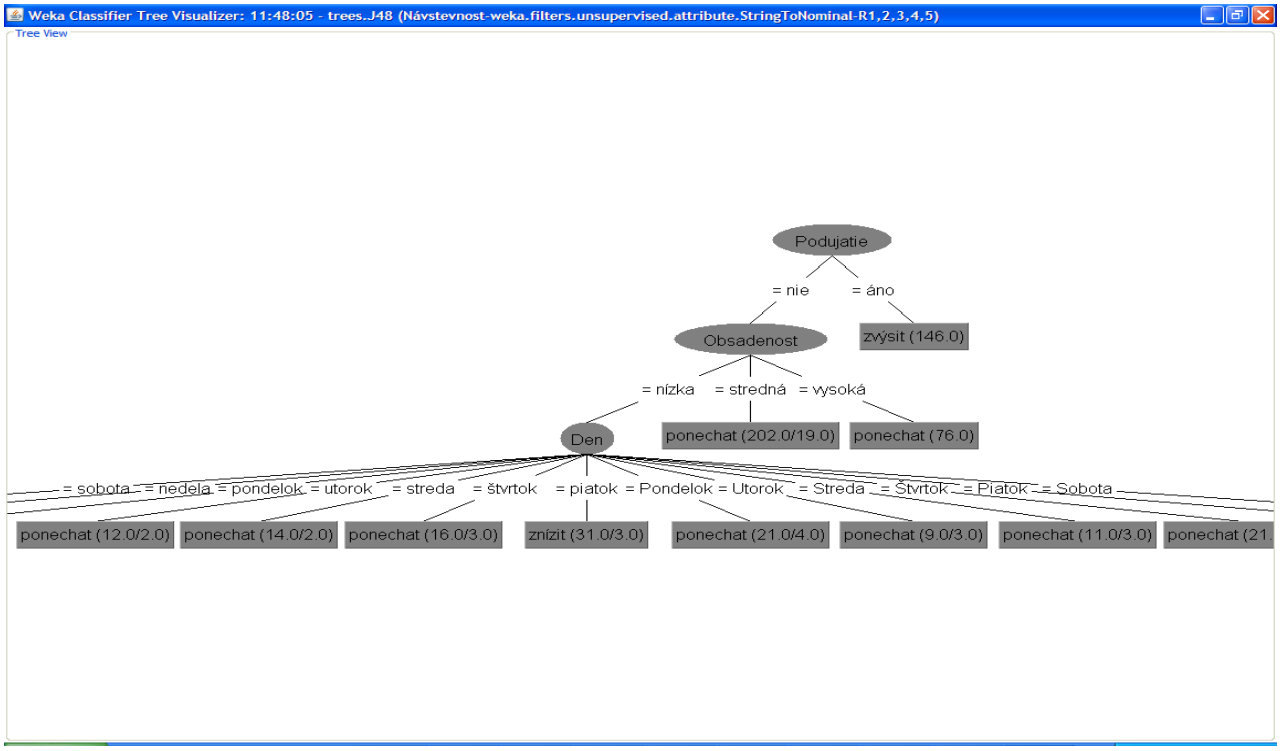
## 6.4 Ohodnotenie modelu





**Obrázok 4** Algoritmus v textovej podobe. Zdroj“ Vlastné spracovanie

V pravej časti obrazovky sa nám objaví rozhodovací strom vypracovaný algoritmom J48 v textovej podobe. Celý originál textový výstup je v prílohe C.




**Obrázok 4.** Vizualizácia stromového grafu. **Zdroj**: Vlastné spracovanie

Tieto informácie nám poskytne textový výstup z testovania. Testovací režim je weka. **classifiers. trees.J48.** Testovanie prebieha na relácii „návstevnost“ **weka.filters. unsupervised. attribute**. Dáta sú typu **StringToNominal**. V súbore máme 731 inštancií a 5 atribútov: Deň, Dátum, Obsadenosť, Podujatie a Cena.

```
Podujatie = nie
|   Obsadenost = nízka
|   |   Den = sobota: znízit (38.0/2.0)

...
...
...

|   |   Den = Sobota: znízit (25.0)
|   Obsadenost = stredná: ponechat (202.0/19.0)
|   Obsadenost = vysoká: ponechat (76.0)
Podujatie = áno: zvýsit (146.0)
```

Testovací režim prebieha na trénovacích dátach v algoritme J48, kde je možne prerezávanie. Prvé číslo v zátvorke hovorí o počte prípadov, ktoré dosiahne list stromu. Druhé číslo hovorí o počte chybných prípadov. Počet listov vo výsledku je 16.

Predikácia na trénovacej množine má v každom riadku najprv číslo riadku. V prvom stĺpci je aktuálny stav a v druhom predikácia. Nasledujú stĺpce, kde je vyčíslená chyba a rozdelenie pravdepodobnosti pre jednotlivá inštancie.

```
=== Predictions on training set ===

inst. actual. predicted.      error. probability distribution

1   1:znízit  1:znízit        *0.947  0.053           0
2   1:znízit  1:znízit        *1      0               0

...

728 1:znízit  1:znízit        *0.95   0.05            0
729 1:znízit  1:znízit        *1      0               0
730 1:znízit  1:znízit        *1      0               0
731 1:znízit  2:ponechat      +0.19  *0.81            0
```

Zhrnutie vyhodnotenia trénovacej množiny nám hovorí že správne klasifikovaných prípadov v množine je v absolútnom hodnotení 682, čo predstavuje 93,3%. Nesprávne klasifikovaných

prípadov je v absolútnom hodnotení 49, čo predstavuje 6,7%. Percento správne klasifikovaných prípadov nazývame presnosť, alebo správnosť vzorky.

Takéto hodnotenie má aj určité nevýhody, ktoré sa ukážu napríklad pri odhade výkonnosti (nie je citlivý na rozdelenie triedy), takže je pravdepodobné, že bude potrebné vidieť aj niektoré ďalšie štatistiky.

Kappa štatistika je korigovaná miera zhody medzi klasifikáciou a správnou triedou. Hodnota väčšia ako 0 znamená, že triedenie je na tom lepšie, než náhodný výber. Ďalšie charakteristiky, ktoré hovoria o sile modelu, sú:

♦ **Mean absolute error** – Priemerná absolútna chyba 0,0759

♦ **Root mean squared error** – stredá kvadratická chyba predikácie 0,1948. Hodnota RMSE (Root Mean Square Error) vyjadruje mieru, o koľko sa skutočné hodnoty odlišujú od hodnôt predikovaných modelom.

♦ **Relative absolute error** – Relatívna Absolútna chyba 18,3109%

♦ **Root relative squared error**- Stredná relatívna kvadratická chyba 42,7976%

Celkový počet inštancií na ktorých bolo urobené štatistické skúmanie je 731. Podrobný prehľad, presnosť podľa tried vidíme v nasledujúcom výstupe. Hodnoty Precisions, Recall, F-measure a ROC Area (senzitivita, špecificita a ROC krivka) hovoria o meraní presnosti diagnostického testu. Porovnáva výsledky testu a skutočnosť. Výsledky sú zobrazené v kontingenčnej, alebo v matrix tabuľke.

ROC krivka sa používa na určenie optimálnej kritickej hodnoty, ktorá delí číselné hodnoty na pozitívne a negatívne hodnoty testu, a na porovnanie alternatívnych testov.

**=== Detailed Accuracy By Class ===**

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.806 | 0.012 | 0.968 | 0.806 | 0.88 | 0.956 | znízit |
| 0.983 | 0.117 | 0.893 | 0.983 | 0.936 | 0.963 | ponechat |
| 1 | 0 | 1 | 1 | 1 | 1 | zvýsit |
| Weighted Avg. | 0.933 | 0.062 | 0.937 | 0.933 | 0.932 | 0.968 |

Plocha pod ROC krivkou (AUC) vyjadruje pravdepodobnosť správneho zaradenia prvku, ktorý bol náhodne vybratý z jednej z tried. Odhad kvality modelu podľa ROC indexu môžeme rozdeliť do nasledovných skupín:

- 0,50 – 0,75 – použiteľný
- 0,75 – 0,92 – dobrý
- 0,92 – 0,97 – veľmi dobrý
- 0,97 – 1,00 – výborný[32]

V prípade kde ROC index dosahuje hodnotu pod 0,75 je nevyhnutné prehodnotiť aplikáciu modelu. Tieto modely patria do skupiny použiteľných modelov, avšak čím nižšia je hodnota ROC indexu, tým viac sa model dostáva ku spodnej hranici kategórie použiteľnosti a bude mať pravdepodobne skresľujúce výsledky.

**=== Confusion Matrix ===**

```
 a    b     c   <-- classified as
179  43    0   |   a = znízit
 6   357   0   |   b = ponechat
 0   0     146 |   c = zvýsit
```

Na správne a nesprávne klasifikovaných prípadoch vieme ukázať percento testov, ktoré boli správne a nesprávne klasifikované. Čísla sú uvedené v matrix matrici, v riadkoch a stĺpcoch **a b c** zastupujúcich triedy.[33]

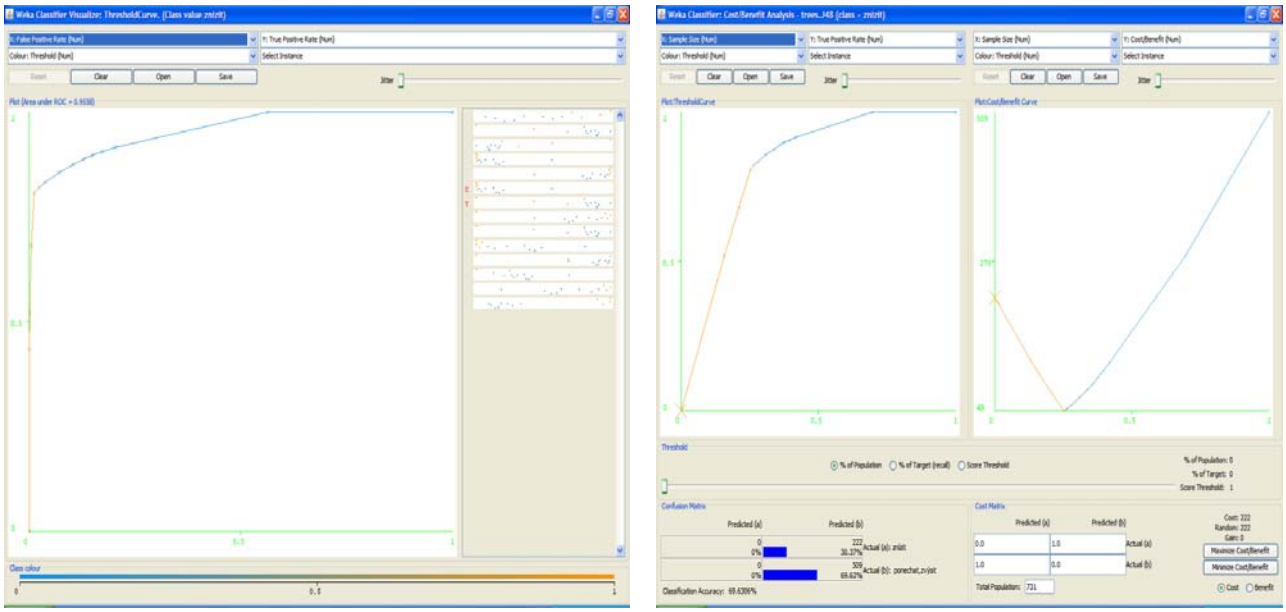
Platí že: aa + bb + cc = 179 + 357 + 146 = 682,
ba + ab + ac= 6 + 43+0= 49.

Weka v časti vizualizácia podporuje grafické znázornenie všetkých štatistických charakteristík, vrátane ROC krivky. Ak vyberiete panel Vizualizácia, nájdeme tu bodový graf matice pre všetky atribúty. Farebné značenie sa mení podľa aktuálne vybranej triedy.

[32] Rimančík M., **Štatistika pre prax**, Vydané nákladom vlastným 2007 ISBN 978-80-969813-1-1

[33] Ian H. Witten, Eibe F., **Data Mining: Practical Machine Learning Tools and Techniques**, Elsevier 2005 ISBA 0-12-088407-0

Je tu možné zmeniť veľkosť jednotlivých výkresov, alebo veľkosť bodu, prípadne odhaliť zakryté body. Je tiež možné vybrať iba podmnožinu atribútov pre zahrnutie do scatter plot matice a čiastková vzorka dát.



**Obrázok 5.** Vizualizácia štatistických charakteristík. **Zdroj**: Vlastné spracovanie

## 6.5 Sumarizácia výsledkov pri tvorbe prediktívneho modelu

Po teoretickom definovaní problému, analýze a vytvorení dataminingového modelu, interpretujeme dosiahnuté výsledky, a zhodnotíme open-source nástroje Weka, výhody a nevýhody tohto analytického nástroja.

Pri výbere správnej metódy skúmania je nevyhnutné poznať Weka prostredie. Na rozbor vstupov a grafické spracovanie údajov sme použili tabuľkový kalkulátor Excel a pri tvorbe modelov sme využívali pracovné prostredie Weka Explorer.

Ak hodnotíme prediktívny model, kde klasifikátorom je logistická regresia, môžeme vychádzať z hodnoty odds ratio. Výstupná zostava z Weky poskytujú odhad odds ratio pre všetky vstupné premenné a hodnoty triedy. Hodnota odds ratio určuje vplyv triedy, ak všetky ostatné premenné (atribúty) sú nezmenné.

Weka pre atribúty s nominálnou hodnotou zobrazuje početnosť. Pri numerickej hodnote nám dáva informácie o maximálnej a minimálnej hodnote, priemere a smerodajnej odchýlke. Tiež ku každej premennej udávaná informáciu o chýbajúcich hodnotách, unikátnych hodnotách a odlišných hodnotách.

Weka poskytuje  prívetivé grafické spracovanie údajov pomocou:

- histogramov
- maticových grafov

Nevýhody Weky:

- neposkytuje štatistické ukazovatele ako sú šikmosť a špicatosť
- maticové grafy nie je možné ich uložiť ako výstupnú zostavu.
- Weka poskytuje menej ukazovateľov na hodnotenie sily modelu ako iné nástroje.

Aj napriek týmto skutočnostiam je Weka klasifikátor dostatočnou alternatívou k iným nástrojom.

V prípade skúmania predikcie ceny v čase dní s malou návštevnosťou hotela je najlepším modelom rozhodovací strom. Pre logistickú regresiu je nevyhnutné aby cieľová premenná bola nominálna.

Využitím klasifikátora J48 sme dosiahli veľmi dobré výsledky. V rozhodovacom strome J48 je implementovaný algoritmus C4.5, ktorý je open-source možnosťou komerčného algoritmu C5.0. Základný editor Weka umožňuje nastaviť rôzne parametre pre skúmanie a vyhodnotenie modelu, ako je napríklad hladinu spoľahlivosti. Má však celkovo menší počet charakteristík určujúcich silu modelu.

V dataminingovom modely sme nevyužili možnosť identifikácie vstupných premenných. Je dobré doplniť modely aj o túto časť procesu. Weka tiež ponúka niekoľko spôsobov hodnotenia metód a výberu atribútov.

Aj napriek odlišnostiam oproti iným nástrojom Weka poskytuje plnohodnotný analytický nástroj, ktorý by mohol nájsť uplatnenie pre malé a stredné podniky.

## 6.6 Integrácia výsledku do manažérskeho rozhodnutia, prínosy práce

Bratislava je moderné, dynamické mesto, ktoré láka tisíce turistov. Návštevnosť stúpla, vlani prišlo asi 950.000 turistov. Cestovný ruch aj tak ide len zhruba na 40%, v hlavnom meste je obrovský prebytok hotelovej kapacity, rezervy sú stále veľké.

V posledných rokoch prichádza množstvo mladých a menej solventných cudzincov ktorý nás objavili prostredníctvom facebooku či youtube. Nie je však isté, či je mesto pripravené na náročnejších a bohatších klientov. Títo okrem hotelových a reštauračných služieb a prehliadok našich pamiatok požadujú aj výber zaujímavých a špecifických aktivít.

*„Nie je v silách Bratislavy „zmeniť siločiary“ a očakávať, že solventní turisti z Japonska, Ruska či Škandinávie tu pri dlhšej návšteve budú aj prespávať. Skôr tu takíto hostia strávia iba 4, 5 hodín, z čoho hodinu sedia na obede. „Treba však ďakovať, že aj za ten krátky čas dokážu minúť na suvenírochči alkohole 500 a viac eur.“*

(Marián Bilačič, Slovenskej spoločnosti priateľov cestovného ruchu)

Spravodajská spoločnosť CNN zaradila Bratislavu do prvej šestice najzaujímavejších pohraničných miest v Európe. Zaradila sa k mestám ako je francúzske Lille, švédske Malmö, španielsky San Sebastian, taliansky Terst či nórsky Kirkenes. Takéto informácie vo vplyvných zahraničných médiách upevnia kredit ktorejkoľvek turistickej destinácie. Môžeme teda očakávať, že aj v budúcnosti bude záujem turistov o Bratislavu a o Slovensko rásť..

*„Na tomto úspechu sa podpísala najmä blízkosť Viedne. „Bratislava významne ťaží z toho, že Viedeň je neďaleko a obe mestá spája Dunaj. Viedeň nás však zároveň v mnohých aspektoch limituje a zrejme nám nikdy nedá šancu poraziť ju napríklad ponukou kongresovej turistiky či kvalitou a kvantitou kultúrnych a spoločenských podujatí,“ priznáva odborník.“*

(Marián Bilačič, Slovenskej spoločnosti priateľov cestovného ruchu)

Minulý mesiac označil Eurostat Bratislavský kraj za piaty najbohatší spomedzi 272 regiónov Európskej únie. Kúpna sily dosiahla z hľadiska výšky HDP na obyvateľa 186% priemeru EÚ.

Zaradila sa za  Londýn, Luxembursko, Brusel a Hamburg. Je pravda že spomedzi slovenských miest je to najbohatšie mesto a mzdy tu sú v priemere najvyššie, na druhej strane z pohľadu obyvateľov prieskum radí Bratislavu na miesto, ktoré jej neprislúcha.

Bratislava je síce na oko bohatá, keby sa neposudzovala zo zlého hľadiska. A to preto že pri malom počte obyvateľov HDP na jedného vystúpi veľmi vysoko a veľká časť ziskov, ktorú v Bratislave vyprodukujú nadnárodné firmy a banky, sa nezdanená a nespotrebovaná vyvezie do zahraničia.[34]

Za týchto okolností je potrebné využiť všetky naše schopnosti a vedomosti ako danú situáciu zmeniť a využiť v prospech rozšírenia cestovného ruchu a pritiahnuť ďalších turistov.

Strojové učenie a algoritmy rozhodovacích stromov čoraz častejšie slúžia k získavaniu znalosti, ktoré môžeme ďalej využiť v rôznych oblastiach. Hlavným prínosom je ich schopnosť pracovať s dátami, ktoré nie sú úplné, alebo sa v nich nachádzajú chyby.

Cieľom nebolo popísať podrobný postup pri tvorbe rozhodovacieho stromu a jeho modelu v programe Weka, ale poukázať na výhody a nevýhody rozhodovacích stromov a informácií získaných pomocou nich.

Využitie rozhodovacích stromov je široké, od ekonomických oblastí kde potrebujeme prehľadávať veľké množstvá dát, až po oblasti technické ako rozpoznávanie obrazov, navigácia alebo riadenie. To všetko môže byť na prospech cestovného ruchu a služieb.

V našom prípade sme zostavením modelu a zhodnotením výsledkov prišli k záveru že v mesiacoch január, apríl, júl, august a december návštevnosť klesá. Na základe toho môžeme urobiť opatrenia ako pritiahnuť zákazníka

Tiež sme zistili že v pravidelných intervaloch klesá návštevnosť koncom týždňa (piatok, sobota, nedeľa). Vzhľadom k tomu nám náš model hovorí, že v dňoch piatok až nedeľa je potrebé ceny nastaviť na nižšie.

[34] Soňa Pacherová, Pravda, 1.4.2014, **Milionárka na Dunaji? O Bratislave to neplatí**

Návštevnosť bude nižšia a teda nie je potrebné zabezpečiť externú firmu na upratovanie, a druhú obsluhu do reštaurácie, tiež je možné posúdiť či stav zásob je dostatočný, a aký počet raňajok je nutné naplánovať.

Výnimku tvoria víkendové dni počas ktorých prebieha v okolí hotela kultúrne, alebo športové podujatie. V prípade pravidelných podujatí ktorých je niekoľko, alebo pri dostatočnej informovanosti o podujatiach, je možné dopredu stanoviť ceny.

Pre rok 2013 sme zostavili plán známych podujatí a následne sme predikovali ceny. Tieto ceny, ich zvýšenie, alebo zníženie sme dopredu stanovili pre predaj priamo v hoteli, ale aj ceny na webportáloch cez ktoré sa predáva ubytovanie. Takto zamedzíme tomu, že ceny sú stanovené neskoro, zákazník má informačný náskok a objednáva si ubytovanie za nízku cenu na dátum v ktorom bude vysoká obsadenosť.

Tak isto ako sme stanovili dni keď je potrebné ceny upraviť vieme model jednoducho prestaviť a získať informácie o počte obsadených izieb v určité dni, a zariadiť potrebné množstvo personálu alebo prostriedkov a tovaru.

Bolo zaujímavé, pomerne jednoduché a lacné zostaviť takýto model voľne šíriteľným programom. Tento model je ľahko modifikovateľný na rovnakom datasete. Podľa nášho názoru aj napriek niektorým nevýhodám a menším problémom, s ktorými sme sa stretli, poskytuje dostatočné pracovné prostredie pre tvorbu prediktívnych analytických modelov využiteľné najmä pre malé a stredné podniky.

Využitím tohto nástroja má hotel možnosť znížiť náklady nie len zredukovaním výdavkov na personál a tovar, ale aj úplne eliminovať náklady na zavedenie prediktívnych modelov do praxe pomocou finančne nenáročného softvéru. Weka je distribuovaný pod GNU GPL licenciou, teda jeho nadobudnutie je bezplatné. Analytický nástroj Weka je neustále vyvíjaný a aktualizovaný. Pri aktualizáciách nevznikajú ďalšie dodatočné náklady.

# Záver

Obe úlohy ktoré sme si na začiatku dali, sa nám podarilo splniť. Náš prvý cieľ – realizácia vybraných postupov spracovania údajov hĺbkovej analýzy s využitím voľne dostupného nástroja pre datamining – sme úspešne splnili.

Datamining nám môže pomôcť pri riešení dôležitých úloh v procese rozhodovania. Existuje mnoho nástrojov na dolovanie dát. Pre úspešnú tvorbu dataminingových modelov je nevyhnutné nájsť vhodný softvérový nástroj, ktorý by zabezpečil tvorbu modelov, ale na druhej strane je potrebné vybrať nástroj, ktorý je finančne nenáročný pre malý podnik. Aj toto bolo dôvodom výberu témy diplomovej práce, pretože je nevyhnutné aby náklady na takýto nástroj boli pre hotel akceptovateľné  a výsledky spoľahlivé. Poukázali sme na možnosti znížiť náklady v prevádzke.

Druhým hlavným cieľom práce bolo vytvorenie prediktívneho modelu vo voľne šíriteľnom programe ako je napríklad Weka, a preverenie analytických nástrojov ktoré poskytuje ich stabilitu a spoľahlivosť. V analytickej časti práce sme skúmali štruktúru poskytnutých údajov a zostavili  sme prediktívny model. Realizovali sme vybrané postupy spracovania údajov a hĺbkovej analýzy s nástrojmi dostupnými vo Weke, na vstupných údajoch z hotela. Postupy a metódy sme zhodnotili.

Samotné vytvorenie analytického prediktívneho modelu bolo úspešné. Zostavili sme funkčný model predikácie ceny. V prípade ak sa v okolí hotela organizuje kultúrne podujatie, sme vopred pripravený, a máme vhodne nastavené ceny.

Metóda rozhodovacích stromov, ktorú sme v analýze použili je najvhodnejšia pre realizáciu nášho modelu. Aj napriek nevýhodám a problémom, s ktorými sme sa stretli považujeme dataminingový nástroj Weka za vhodnú alternatívu komerčného dataminingového nástroja. Odporúčame ju využívať malým podnikom, ktoré majú na jednej strane menšiu databázu údajov a na druhej strane nedisponujú dostatočnými finančnými prostriedkami.

# Zoznam použitej literatúry


1. Witten I.H., Eibe F., Elsevier 2005 **Data Mining: Practical Machine Learning Tools and Techniques**, ISBA 0-12-088407-0

2. **Weka Tutoriál**, http://www.cs.waikato.ac.nz/ml/weka/index.html

3. Sarnovský J., Liguš J., Benko P., Košice 2001, **Kybernetika a manažment**, http://web.tuke.sk/ kybernetika/kam

4. Sarnovský J., Liguš J., Benko P., Košice 2001, **Kybernetika a manažment**, http://web.tuke.sk/ kybernetika/kam

5. Kostík Ľ., Saloky T., 2006, Riadenie a identifikácia systémov, **Niektoré z problémov pri získavaní dát pomocou rozhodovacích stromov**, AT&P journal PLUS2, Riadenie a identifikácia systémov

6. Novotný O., Pour J., Slánsky D., 2005, **Business Intelligence –** Jak využít bohatství ve vaších dátech, Grada Publishing, ISBN 80-247-1094-3

7. Berka P., Academia, Praha 2003, **Dobývaní znalostí z databází**, ISBN 80-200-1026–9

8. Kostík Ľ., Saloky T., 2006, Riadenie a identifikácia systémov, **Niektoré z problémov pri získavaní dát pomocou rozhodovacích stromov**, AT&P journal PLUS2

9. Hand D.,Mannila H.,Smyth P., **Principles of Data Mining,** MIT Press, 2001, IBAN 0-261-08290-X

10. Han J., Kamber M.; Diane C., **Data Mining Concepts and Technigues**; 2006; ISBN 978-1-55860-901

11. Dietterich T., **Ensemble Learning,** The MIT Press 2002, www.eecs.wsu.edu/~holder/ courses/ CptS570/fall07/ papers/Dietterich00.pdf

12. Rimančík M., **Štatistika pre prax**, Vydané nákladom vlastným 2007 ISBN 978-80-969813-1-1

## Stránky na internete


1. Radek Kafka, 2013, SystemOnline**, Implementace BI řešení ve výrobní firmě,** www.systemonline.cz/ business-intelligence/ implementace-bi-reseni-ve-vyrobni-firme-1.htm

2. Tomáš Třminek, 2012, SystemOnline, **Prínosy a náklady Business Intelligence,** www.systemonline.cz/ business-intelligence/prinosy-a-naklady-business-intelligence.htm

3. Dostál M. 2012/5, SystemOnline, **Věnujte pozornost BI governance**, www.systemonline.cz/business-intelligence/ venujte-pozornost-bi-governance.htm

4. **KDnuggets™**, www.kdnuggets.com/polls

5. Komora M., Hečková G., SystemOnline, **Rychlejší analýza dat s in-memory BI,** www. systemonline. cz/ business-intelligence/ rychlejsi-analyza-dat-s-in-memory-bi-1.htm

6. Lacko Ľ., Asseco, **BI: Trendy, prehľad riešení, poskytovateľov, možností a cien**, asseco.com /ce/assets/ Uploads/ attachments/news-items/NMinfoware2013060720.pd

7. **SAP**, www.sap.com/sk/index.html

8. **ORACLE**, www.oracle.com/sk/index.html?ssSourceSiteId=ocomen

9. **SAS,** www.sas.com/offices/europe/slovakia/

10. **SYBASE**, www.sybaseproducts.com

11. **IBM**, www-03.ibm.com/software/products/en/category/business-intelligence

12. **Microsoft**, www.microsoft.com/slovakia/sqlserver/default.mspx

13. **Cognos**, www-03.ibm.com/software/products/sk

14. **Rapidminer**, http://rapidminer.com/products/rapidminer-studio/

15. **Weka,** http://www.cs.waikato.ac.nz/ml/weka/downloading.html

16. Arnošt P., 22.8.2001, ROOT.cz, **Co je to „Open Source software“**, www.root.cz/clanky/co-je-to-open-source-software

17. Janů S., 18.4.2014, Zive.cz, **Open-source je poprvé kvalitnější než proprietární software**, www.zive.cz

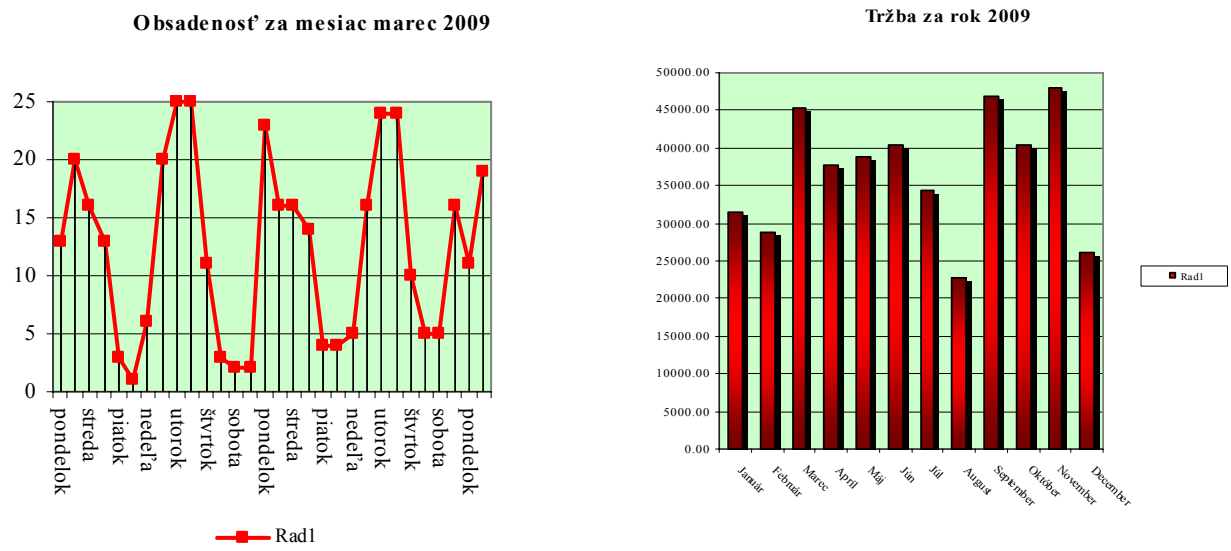
18**. Hotel SET** homepage, www.hotelset.sk

19. Soňa Pacherová, Pravda, 1.4.2014, **Milionárka na Dunaji? O Bratislave to neplatí**

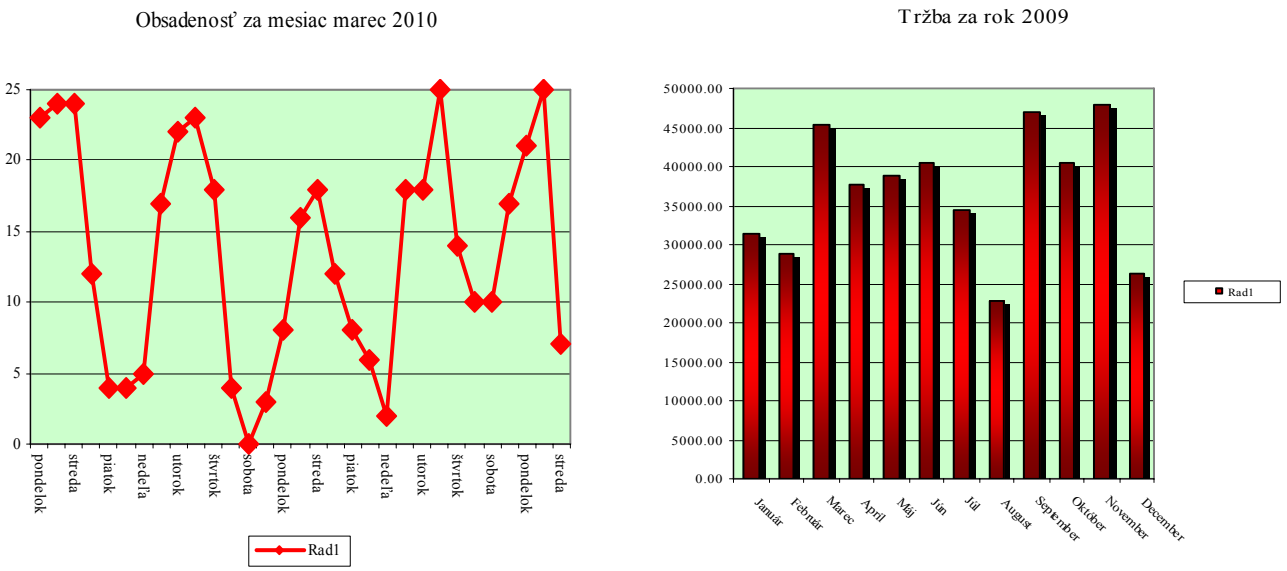
# Prílohy

## Príloha A

### Grafy agregovaných tržieb a obsadenosti



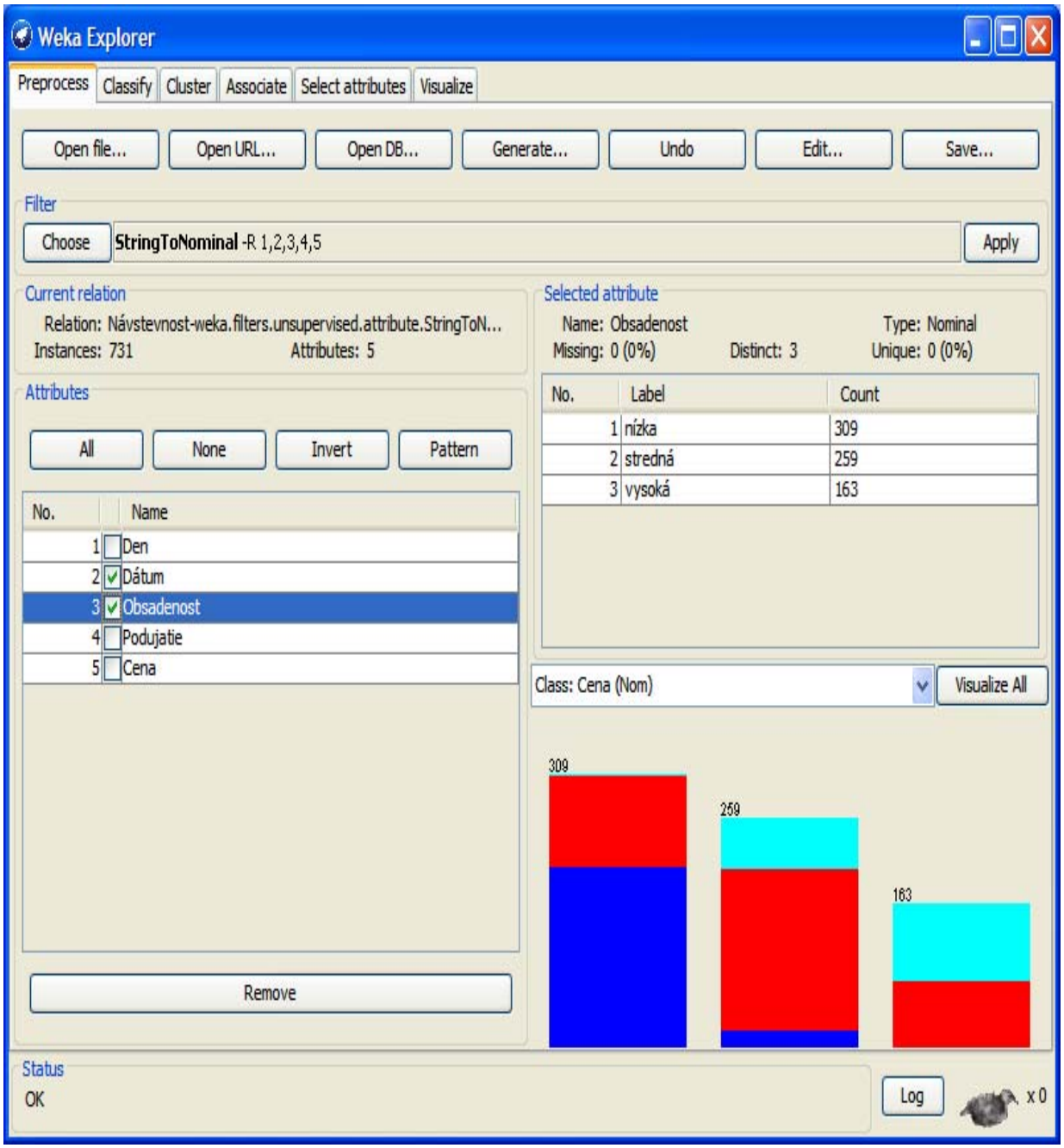

**Obrázok 1.** Tržba a návštevnosť za 2009. Zdroj: Vlastné spracovanie




**Graf 1.** Tržba a návštevnosť za 2010. Zdroj: Vlastné spracovanie

**Príloha B**

Okno predspracovania údajov v programe Weka



# Príloha C

### Textový výstup rozhodovacieho stromu

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Návstevnost weka.filters.unsupervised.attribute.
StringToNominal-R1,2,3,4,5
Instances:    731
Attributes:   5
              Den
              Dátum
              Obsadenost
              Podujatie
              Cena
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Podujatie = nie
|   Obsadenost = nízka
|   |   Den = sobota: znízit (38.0/2.0)
|   |   Den = nedela: znízit (71.0)
|   |   Den = pondelok: ponechat (18.0/2.0)
|   |   Den = utorok: ponechat (12.0/2.0)
|   |   Den = streda: ponechat (14.0/2.0)
|   |   Den = štvrtok: ponechat (16.0/3.0)
|   |   Den = piatok: znízit (31.0/3.0)
|   |   Den = Pondelok: ponechat (21.0/4.0)
|   |   Den = Utorok: ponechat (9.0/3.0)
|   |   Den = Streda: ponechat (11.0/3.0)
|   |   Den = Štvrtok: ponechat (21.0/5.0)
|   |   Den = Piatok: znízit (20.0/1.0)
|   |   Den = Sobota: znízit (25.0)
|   Obsadenost = stredná: ponechat (202.0/19.0)
|   Obsadenost = vysoká: ponechat (76.0)
Podujatie = áno: zvýsit (146.0)

Number of Leaves  :     16
Size of the tree :  19

* * *

...
```

```
...
...

* * *



Time taken to build model: 0.03 seconds

=== Predictions on training set ===

inst. actual.predicted,      error, probability distribution
1   1:znízit 1:znízit        *0.947  0.053       0
2   1:znízit 1:znízit        *1      0           0

...

728 1:znízit 1:znízit        *0.95   0.05        0
729 1:znízit 1:znízit        *1      0           0
730 1:znízit 1:znízit        *1      0           0
731 1:znízit 2:ponechat       +0.19  *0.81        0

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances     682            93.2969%
Incorrectly Classified Instances   49             6.7031%
Kappa statistic                    0.8904
K&B Relative Info Score            62230.4143%
K&B Information Score              926.3542 bits  1.2672
bits/instance
Class complexity | order 0         1087.5727 bits 1.4878
bits/instance
Class complexity | scheme          207.3523 bits  0.2837
bits/instance
Complexity improvement (Sf)        880.2205 bits  1.2041
bits/instance
Mean absolute error                0.0759
Root mean squared error            0.1948
Relative absolute error            18.3109%
Root relative squared error        42.7976%
Total Number of Instances          731

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area   Class
0.806   0.012    0.968     0.806  0.88      0.956      znízit
0.983   0.117    0.893     0.983  0.936     0.963      ponechat
1       0        1         1      1            1       zvýsit
```

```
Weighted Avg. 0.933  0.062  0.937   0.933     0.932   0.968

=== Confusion Matrix ===

 a    b     c   <-- classified as
 179  43    0   |   a = znízit
 6    357   0   |   b = ponechat
 0    0     146 |   c = zvýsit
```

**Príloha D**

Vizualizácia modelu rozhodovacieho stromu