VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ

FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY

DEPARTMENT OF INFORMATICS

BUSINESS INTELLIGENCE - VYUŽITIE DATA MININGU VO FIREMNÝCH PROCESOCH

BUSINESS INTELLIGENCE - USE OF DATA MINING IN BUSINESS PROCESSES

BAKALÁRSKA PRÁCA

BACHELOR'S THESIS

AUTOR PRÁCE — TOMÁŠ SKALICKÝ

AUTHOR

VEDÚCI PRÁCE — Ing. JIŘÍ KŘÍŽ, Ph.D.

SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání bakalářské práce

Ústav:	Ústav informatiky
Student:	**Tomáš Skalický**
Studijní program:	Systémové inženýrství a informatika
Studijní obor:	Manažerská informatika
Vedoucí práce:	**Ing. Jiří Kříž, Ph.D.**
Akademický rok:	2019/20

Ředitel ústavu Vám v souladu se zákonem č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a se Studijním a zkušebním řádem VUT v Brně zadává bakalářskou práci s názvem:

Business Intelligence – využití data miningu ve firemních procesech

Charakteristika problematiky úkolu:

Úvod
Cíle práce, metody a postupy zpracování
Teoretická východiska práce
Analýza současného stavu
Vlastní návrhy řešení
Závěr
Seznam použité literatury
Přílohy

Cíle, kterých má být dosaženo:

Cílem práce je využití metod a nástrojů Business Intelligence pro podporu rozhodování.

Základní literární prameny:

FOTR, Jiří. Tvorba strategie a strategické plánování: teorie a praxe. Praha: Grada, 2012. Expert (Grada). ISBN 978-80-247-3985-4.

LABERGE, Robert. Datové sklady: agilní metody a business intelligence. Brno: Computer Press, 2012. ISBN 978-802-5137-291.

NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ. Business intelligence: jak využít bohatství ve vašich datech. Praha: Grada, 2005. Management v informační společnosti. ISBN 80-247-1094-3.

RUD, Olivia Parr. Data Mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001. Databáze. ISBN 80-722-6577-6.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2019/20

V Brně dne 29.2.2020

L. S.

doc. RNDr. Bedřich Půža, CSc. ředitel	doc. Ing. et Ing. Stanislav Škapa, Ph.D. děkan

Abstrakt

Cieľom danej bakalárskej práce je zoznámiť sa s pojmom Business Intelligence, rovnako ako aj s pojmom datamining a jeho využitím vo firemnej sfére. V úvodnej teoretickej časti priblížim nástroje Business Intelligence a datamining algoritmy. V nasledujúcej praktickej časti dané algoritmy využijem pre analýzu poskytnutých firemných dát. Následne získané analýzy môžu byť použité ako podpora pre firemné rozhodovanie.

Abstract

The aim of this bachelor thesis is to get acquainted with the concept of Business Intelligence as well as with the concept of data mining and its use in the company sphere. In the introductory theoretical part I will introduce the tools of Business Intelligence and data mining methods. In the following practical part I will use the methods for analysis of provided company data. The analysis obtained can be used as a support for company decision making.

Kľúčové slová

Business Intelligence, datamining, databáza, rozhodovací strom, lineárna regresia, zhluková analýza

Keywords

Business Intelligence, datamining, database, decision tree, linear regression, cluster analysis

Citácia

SKALICKÝ, Tomáš. *Business Intelligence - využitie data miningu vo firemných procesoch.* Brno, 2020. Bakalárska práca. Vysoké učení technické v Brně, Fakulta podnikatelská. Vedúci práce Ing. Jiří Kříž, Ph.D.

Business Intelligence - využitie data miningu vo firemných procesoch

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Jiřího Kříže, Ph.D.. Uviedol som všetky literárne pramene, publikaćie a ďalšie zdroje, z ktorých som čerpal.

.....

Tomáš Skalický

17. mája 2020

Poďakovanie

Týmto by som rád poďakoval svojmu vedúcemu bakalárskej práce pánovi Ing. Jiřímu Křížovi za konzultácie a pomoc pri tvorbe danej práce. Zároveň touto cestou ďakujem aj pánovi Ing. Martinovi Veselému za ochotu a prístup pri spolupráci, poskytnutí firemných dát a informácií, ktoré som použil ako podklad tejto práce.

Obsah

Úvod

Tému svojej bakalárskej práce som si vybral za základe neustále sa zvyšujúceho tlaku na úložné kapacity pre každoročne prudko rastúce množstvá zaznamenaných dát. Takýmto enormným nárastom sa dáta stávajú čím ďalej tým viac neprehľadnými a preto sa zvyšuje aj potreba práce s nimi, ich úprava a hlavne analýza pre ich možné budúce využitie.

Práve pre takúto analýzu je vhodné využiť postupy a nástroje Business Intelligence (BI), ktoré surové dáta premenia na informácie a poznatky, podporujúce firemné rozhodovanie. Vďaka správnemu a presnému využitiu týchto nástrojov sa hodnota firiem na trhu z hľadiska konkurencieschopnosti neustále zvyšuje.

Z celej oblasti BI sa budem zaoberať hlavne procesom datamining. Osobne ma zaujímajú hlavne jeho prediktívne a štatistické možnosti a metódy, ktoré si priblížime v nasledujúcich častiach.

V prvej časti práce sa budem zaoberať teoretickými pojmami a ich vysvetlením pre lepšie pochopenie ďalej rozoberanej problematiky. Bude sa jednať najmä o pojmy spojené s Business Intelligence a datamining.

Popíšem BI ako koncept, ktorý zastrešuje analytické a transakčné procesy (OLAP a OTLP), zber dát, ukladanie dát a riadenie vedomostí. Ďalej bude nasledovať charakteristika pojmu datamining ako celku. Po ňom popíšem jednotlivé datamining algoritmy.

V druhej časti priblížim spoločnosť, s ktorou pri písaní práce spolupracujem. Popíšem základné informácie o nej, jednotlivé pracovné pozície a čím sú špecifické. Väčšiu časť pozornosti budem venovať analýze pracovného procesu. Pomocou neho ukážem ako sa prijatá klientska požiadavka mení na hotový produkt/službu.

Nutnosťou je aj poukázať na zápis a ukladanie všetkých dát spojených s pracovným procesom do databázy, s ktorou budem pracovať.

V poslednej tretej návrhovej časti budem prezentovať dáta poskytnutej databázy, ich úpravu a následné použitie. Na základe výberu troch hlavných datamining algoritmov vytvorím návrhy predikčných alebo štatistických modelov pre každý z nich. Pomocou týchto modelov budem schopný dolovať informácie, ktoré budú môcť byť použité v ďalších analýzach alebo firemnom rozhodovaní.

Pri tvorbe budem dbať na čo najvhodnejší výber dát a následne čo najpresnejší výpočetný model. Následne vyhodnotím vytvorený návrh riešenia s možnosťami využitia vytvorených modelov.

V úplnom závere zhodnotím celkový prínos tejto práce pre firmu spolu so znalosťami, ktoré som počas tvorby modelov, návrhov a analýz nadobudol.

Kapitola 1

Cieľ práce, metódy a postupy spracovania

Cieľom práce je využitie metód a nástrojov Business Intelligence pre podporu rozhodovania.

V rámci procesu získavania, úpravy a používania dát som absolvoval niekoľko stretnutí, hovorov a konzultácií s viceprezidentom riadenia zdrojov spoločnosti. Po dohodnutí a schválení podmienok práce mi bola poskytnutá časť firemnej databázy dostačujúcich rozmerov a kvality dát pre plnohodnotnú analýzu a dolovanie informácií.

Pre úpravu samotných dát v databáze som zvolil metódu OLAP, jednu z hlavných funkcionalít BI, a pomocou kontingenčných tabuliek som upravil kvalitatívne dáta do potrebnej kvantitatívnej formy.

Následne som si zvolil program RapidMiner Studio spomedzi viacerých vo výbere, pre jeho jednoduchý prístup a interaktívny tutoriál. Po prejdení celého tutoriálu som bol schopný vytvoriť jednoduchý model so znalosťami väčšiny potrebných operátorov. Pre presnejšiu tvorbu kvalitnejších modelov som sa nechal inšpirovať internetovými článkami s danou témou.

Po zadaní operátorov, vybraní atribútov a celkovom prepojení sa vytvoril proces pripravený na testovanie. Po spustení sa zobrazili výstupy, na základe ktorých som mohol usúdiť, či je miera presnosti testovania dostatočne vysoká. Postupnými úpravami dát a procesu som získal presnosť vhodnú pre použitie a prezentáciu predikčného modelu s konkrétnymi, reálne interpretovateľnými výsledkami, ktoré firma môže ďalej využiť v procesoch rozhodovania.

Kapitola 2

Teoretické východiská

2.1 Business Intelligence

Business intelligence (BI) obsahuje stratégie a technológie používané podnikmi na analýzu údajov o obchodných informáciách.

Technológie BI poskytujú historické, súčasné a prediktívne pohľady na obchodné operácie. Medzi bežné funkcie technológií podnikovej inteligencie patrí podávanie správ, online analytické spracovanie, analytika, dolovanie údajov, dolovanie procesov, komplexné spracovanie udalostí, riadenie výkonnosti podniku, porovnávanie, dolovanie textu a prediktívna analýza. Technológie BI dokážu spracovať veľké množstvo štruktúrovaných a niekedy neštruktúrovaných údajov, aby pomohli identifikovať, rozvíjať a inak vytvárať nové strategické obchodné príležitosti. Ich cieľom je umožniť ľahkú interpretáciu týchto údajov. Identifikácia nových príležitostí a implementácia efektívnej stratégie založenej na poznatkoch môžu podnikom poskytnúť konkurenčnú výhodu na trhu a dlhodobú stabilitu.

Podnikateľské informácie môžu podniky využívať na podporu širokého spektra obchodných rozhodnutí od operatívnych po strategické. Medzi základné prevádzkové rozhodnutia patrí umiestnenie produktu alebo stanovenie ceny. Strategické obchodné rozhodnutia zahŕňajú priority, ciele a smery na najširšej úrovni. Vo všetkých prípadoch je BI najúčinnejšie, keď kombinuje údaje získané z trhu, na ktorom spoločnosť pôsobí (externé údaje) s údajmi, ako sú finančné a prevádzkové (interné údaje).

Ak sa kombinujú, externé a interné údaje môžu poskytnúť úplný obraz, ktorý nemožno odvodiť z nijakého jedinečného súboru údajov. Medzi nespočetné spôsoby použitia umož-

ňujú nástroje BI získavať informácie o nových trhoch, posudzovať dopyt a vhodnosť produktov a služieb pre rôzne segmenty trhu.

Aplikácie BI používajú údaje zhromaždené z dátového skladu (date warehouse) alebo z dátového servera. Dátový sklad obsahuje kópiu analytických údajov, ktoré uľahčujú podporu rozhodovania.

Podnikové informácie definujeme ako systémy, ktoré kombinujú zbieranie dát (data gathering), úložisko dát (data storage) a riadenie vedomostí (knowledge management) s analýzou na vyhodnotenie komplexných podnikových a konkurenčných informácií, ktoré sa majú predložiť orgánom s rozhodovacou právomocou, s cieľom zlepšiť kvalitu vstupov do rozhodovacieho procesu (5).

2.1.1 Online Analytical Processing (OLAP)

OLAP je technológia k rýchlemu odpovedaniu na viacrozmerné analytické dotazy. Nástroje OLAP umožňujú interaktívnu analýzu viacrozmerných údajov z mnohých uhlov pohľadu.

Skladá sa z troch základných analytických operácií:

1. **Roll-up:** jedná sa o zoskupenie údajov, ktoré je možné zhromaždiť a vypočítať v jedenej alebo viacerých dimenziách

   Príklad: všetky pobočky sú zoskupené pod názov firmy pre určenie celkových výnosov

2. **Drill-down:** jedná sa o techniku umožňujúcu používateľom detailnú navigáciu

   Príklad: predaj podľa jednotlivých produktov a služieb určitej pobočky

3. **Slice and dice:** jedná sa o funkciu výberu špecifickej množiny údajov kocky OLAP s možnosťou prezerať (Slice) časti z rôznych uhlov pohľadu.

   Príklad: pohľad na predaj produktu na základe dátumu predaja alebo zákazníka, alebo miesta predaja, atď.

Databázy OLAP využívajú viacrozmerný dátový model, ktorý umožňuje analytické dotazy s rýchlou časovou odozvou (5).

2.1.2 Online Transaction Processing (OLTP)

OLTP je technológia na spracovanie a uloženie dát v databáze, kde systém čo najľahšie, najrýchlejšie a najbezpečnejšie reaguje na požiadavky používateľov. Systém OLTP sa v súčasnosti používa najmä v databázových aplikáciách.

Základným rozdielom OLTP oproti OLAP je, že u OLTP sú dáta priebežne a často modifikované obvykle viacerými užívateľmi, zatiaľ čo u OLAP sú dáta jednorazovo nahrávané a sú nad nimi realizované dotazy (5).

2.1.3 Zber dát (Data gathering)

Zber údajov je proces zhromažďovania a merania informácií o cieľových premenných v zavedenom systéme, ktorý potom umožňuje odpovedať na príslušné otázky a hodnotiť výsledky. Zber údajov je súčasťou výskumu vo všetkých študijných odboroch vrátane fyzikálnych a spoločenských vied, humanitných vied a podnikania. Zatiaľ čo metódy sa líšia podľa disciplíny, dôraz na zabezpečenie presného a čestného zberu zostáva rovnaký. Cieľom zhromažďovania všetkých údajov je zachytiť kvalitatívne dôkazy, ktoré umožnia, aby analýza viedla k formulácii presvedčivých a dôveryhodných odpovedí na položené otázky (5).

2.1.4 Ukladanie dát (Data storage)

Ukladanie dát je zaznamenávanie informácií na pamäťové médium. Ako príklady pamäťových médií môžeme uviesť:

- DNA a RNA
- Fonografický záznam
- Rukopis
- Magnetická páska
- Optické disky

Ukladanie počítačových údajov je jednou z hlavných funkcií počítača na všeobecné použitie a elektronické dokumenty sa môžu ukladať na oveľa menšom priestore ako papierové dokumenty. Čiarové kódy a rozpoznávanie znakov magnetického atramentu (MICR) sú dva spôsoby zaznamenávania strojom čitateľných údajov na papier (5).

2.1.5 Riadenie vedomostí (Knowledge management)

Riadenie vedomostí je proces vytvárania, zdieľania, využívania a riadenia znalostí a informácií organizácie. Poukazuje na multidisciplinárny prístup na dosiahnutie organizačných cieľov čo najlepším využitím znalostí.

Knowledge management je zavedenou disciplínou od roku 1991 a zahŕňa kurzy vyučované v oblasti podnikovej správy, informačných systémov, managementu, knižnice a informačných vied.

Úsilie v oblasti riadenia vedomostí sa zvyčajne zameriava na organizačné ciele, ako je lepší výkon, konkurenčná výhoda, inovácia, zdieľanie získaných skúseností, integrácia a neustále zlepšovanie organizácie (5).

2.2 Data mining

Dolovanie údajov je proces zisťovania vzorcov vo veľkých súboroch údajov zahŕňajúcich metódy strojového učenia, štatistík a databázových systémov. Jedná sa o oblasť informatiky a štatistiky s celkovým cieľom extrahovať informácie zo súboru údajov a transformovať ich do zrozumiteľnej štruktúry na ďalšie použitie.

Dolovanie údajov je krokom analýzy procesu zisťovania znalostí v databázach (Knowledge discovery in databases). Okrem prvotnej analýzy zahŕňa aj aspekty správy databáz a údajov, predbežné spracovanie údajov, úvahy o modeloch, následné spracovanie objavených štruktúr, vizualizáciu a online aktualizáciu.

Termín „dolovanie údajov“ nie je úplne správne pomenovanie, pretože cieľom je extrakcia vzorov a poznatkov z veľkého množstva údajov, ťažba samotných údajov. Často sú vhodnejšie všeobecnejšie pojmy ako analýza a analytika údajov alebo umelá inteligencia a strojové učenie.

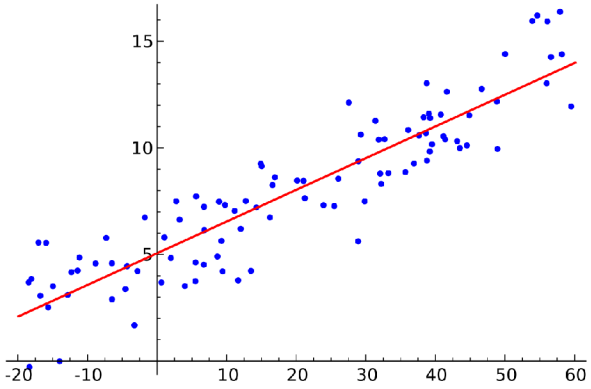
Rozdiel medzi analýzou údajov a ťažbou údajov spočíva v tom, že analýza údajov sa používa na testovanie modelov a hypotéz v súbore údajov. Ťažba (dolovanie) údajov využíva strojové učenie a štatistické modely na odhaľovanie skrytých vzorcov vo veľkom množstve údajov (13).

2.3 Data mining algoritmy

2.3.1 Lineárna regresia (Linear regression)

Lineárna regresia je najjednoduchší algoritmus regresnej analýzy, ktorý prebieha metódou najmenších štvorcov. Regresia vie určiť závislosť medzi vstupom a výstupom z tréningových dát a následne vie vypočítať závislú premennú, výstup. Rovnako ako sa klasifikácia používa na predpovedanie kategorických prvkov, na regresiu sa používa predpovedanie spojitej hodnoty.

Jedná sa teda o vyjadrenie korelácie pomocou korelačného diagramu. Jej výsledkom je spojitý výstup (2).



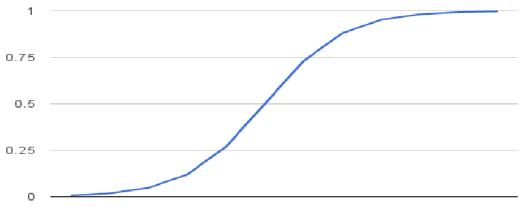
Obr. 2.1: Príklad grafu výstupu lineárnej regresie

2.3.2 Logistická regresia (Logistic regression)

Logistická regresia je štatistický model, ktorý využíva logistickú funkciu na modelovanie binárne závislej premennej.

Zaoberá sa problematikou odhadu pravdepodobnosti určitého javu na základe známych skutočností a odhaduje parametre logistického modelu.

Binárny logistický model má závislú premennú s dvoma možnými hodnotami, ako sú napríklad hodnoty *prešiel/neprešiel, výhra/prehra, živý/mŕtvy* alebo *zdravý/chorý*. Tieto dve hodnoty sú označené ako **0** a **1**.



Obr. 2.2: Príklad grafu logistickej regresie

Zdroj obr. 2.1: https://en.wikipedia.org/wiki/Linear_regression
Zdroj obr. 2.2: https://machinelearningmastery.com/logistic-regression

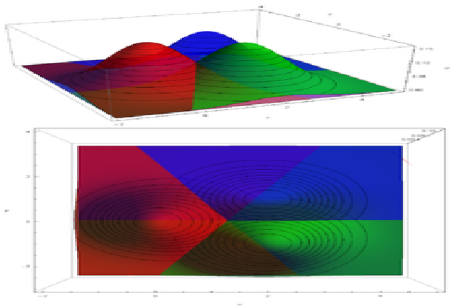
2.3.3 Diskriminačná analýza (Linear discriminant analysis)

Diskriminačná analýza je metóda viacrozmernej štatistickej analýzy, ktorá charakterizuje a oddeľuje dva alebo viac tried objektov, alebo udalostí, vytvorením rozhodovacieho pravidla.

Na rozdiel od zhlukovej analýzy sa používa v prípade, keď sú skupiny vopred známe, musia mať jednu alebo viacero kvantitatívnych premenných a kvalitatívnu premennú.

Jednoducho povedané, diskriminačná analýza je klasifikácia, a teda rozdelenie objektov do skupín, tried alebo kategórií rovnakého typu.

Všetky objekty sú charakterizované znakmi, ktoré môžeme pozorovať (4).



Obr. 2.3: Príklad viacrozmerného grafu diskriminačnej analýzy

2.3.4 Faktorová analýza (Factor analysis)

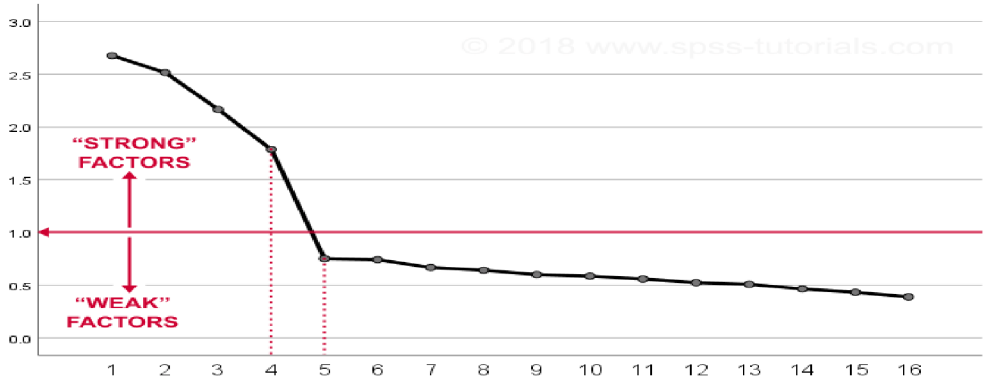
Faktorová analýza je viacrozmerná štatistická metóda, ktorá sa používa na redukciu veľkého počtu premenných na niekoľko interpretovateľných základných faktorov. Pomáha pri riešení súborov údajov, kde existuje veľké množstvo pozorovaných premenných, o ktorých sa predpokladá, že odrážajú menší počet základných (skrytých) premenných (6).

Sú definované dve hlavné metódy faktorovej analýzy:

1. **Analýza hlavných komponentov**- extrahuje faktory založené na celkovom rozptyle faktorov pre nájdenie najmenšieho počtu premenných, ktoré vysvetľujú najväčšie rozptyly

2. **Spoločná analýza faktorov**- extrahuje faktory založené na rozptyle zdieľanom týmito faktormi pre nájdenie skrytých základných faktorov

Zdroj obr. 2.3: https://cs.wikipedia.org/wiki/Diskrimina%C4%8Dn%C3%AD_anal%C3%BDza



Obr. 2.4: Príklad grafu faktorovej analýzy

2.3.5 Zhluková analýza (Cluster analysis)

Zhluková analýza alebo zhlukovanie je úlohou zoskupovania množiny objektov takým spôsobom, že objekty v rovnakej skupine (zhluky) sú si navzájom viac podobné ako v iných skupinách (zhlukoch).

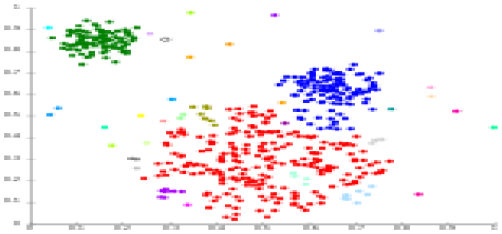
Zhluk je určený polohou stredu v „n-rozmernom“ priestore „n“ prvkov. Táto pozícia sa nazýva ťažisko (14). Je to bežná technika štatistickej analýzy údajov, ktorá sa používa v mnohých oblastiach vrátane strojového učenia, analýzy obrázkov, získavania informácií a počítačovej grafiky.

Samotná zhluková analýza nie je jeden špecifický algoritmus, ale viacero metód založených na miere vzdialenosti alebo podobnosti medzi prvkami. Zhlukovací algoritmus a nastavenia parametrov závisia od individuálneho súboru údajov a zamýšľaného použitia výsledkov (13).

Rozlišujeme tri úlohy zhlukovej analýzy:

1. Nájdenie vopred definovaného počtu zhlukov
2. Nájdenie nešpecifikovaného počtu množín zhlukov
3. Vytvorenie hierarchického stromu

Zdroj obr. 2.4: https://www.spss-tutorials.com/spss-factor-analysis-tutorial/

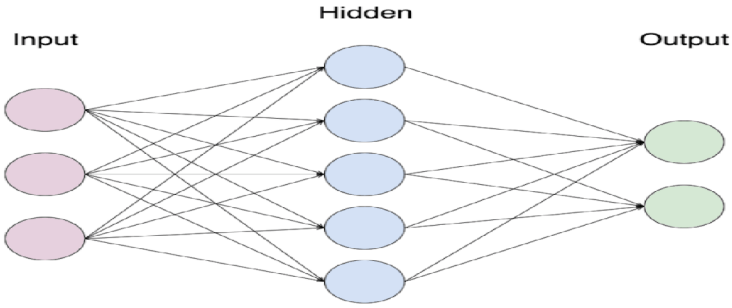


Obr. 2.5: Príklad grafu zhlukovej analýzy

2.3.6 Neurónové siete (Artificial neural network)

Umelá neurónová sieť (ANN), obvykle nazývaná neurónová sieť (NN), je matematický model alebo výpočtový model, ktorý je inšpirovaný štruktúrou a funkčnými aspektmi biologických neurónových sietí. Pozostáva z prepojenej skupiny umelých neurónov a spracováva informácie pomocou spojovacieho prístupu k výpočtu. Vo väčšine prípadov je neurálna sieť adaptívnym systémom, ktorý mení svoju štruktúru počas fázy učenia. Používa sa na modelovanie zložitých vzťahov medzi vstupmi a výstupmi alebo na zisťovanie vzorcov v údajoch.

Feed-forward sieť je umelá neurónová sieť, kde spojenia medzi jednotkami netvoria riadený cyklus. Informácie sa pohybujú iba jedným smerom, zo vstupných uzlov, cez skryté uzly k výstupným uzlom. V sieti tak nie sú žiadne cykly ani slučky (14).



Obr. 2.6: Architektúra neurónovej siete

Zdroj obr. 2.5: https://en.wikipedia.org/wiki/Cluster_analysis

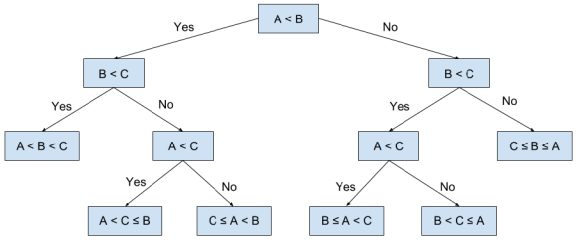
Zdroj obr. 2.6: https://laptrinhx.com/titanic-prediction-with-artificial-neural-network-in-r\-3087367370/

2.3.7 Rozhodovacie stromy (Decision tree)

Rozhodovacie stromy sú nástroj na podporu rozhodovania, ktorý používa stromový model rozhodnutí a ich možné dôsledky, vrátane výsledkov náhodných udalostí. Jedná sa o algoritmus založený na podmienených riadiacich príkazoch.

Rozhodovací strom tvoria uzly a vetvy. Základný uzol (koreňový uzol) sa ďalej rozvetvuje do ďalších uzlov, kde každý z nich predstavuje určitú vlastnosť objektu.

Rozvetvením jednotlivých uzlov sa vytvára štruktúra stromu (9).



Obr. 2.7: Príklad grafu vetvenia rozhodovacieho stromu

2.3.8 Naivný Bayesov klasifikátor (Naive Bayes classifier)

Naivný Bayesov klasifikátor je jednoduchý pravdepodobnostný klasifikátor založený na Bayesovej vete so silnými (naivnými) predpokladmi nezávislosti medzi vlastnosťami.

Klasifikátori sú trénovaní pomocou viacerých algoritmov založených na rovnakom princípe. Predpokladajú, že hodnota určitého znaku je nezávislá od hodnoty iného znaku vzhľadom na premennú triedy. To znamená, že každý zo znakov prispieva nezávisle k pravdepodobnosti, bez ohľadu na možné korelácie medzi nimi.

Je obľúbenou metódou s frekvenciami slov ako znakmi na kategorizáciu textu alebo hodnotenie dokumentov (spam, šport, politika) (8).

Zdroj obr. 2.7: https://elf11.github.io/2018/07/01/python-decision-trees-acm.html

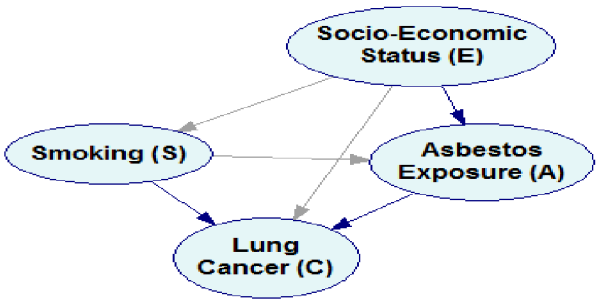
2.3.9 Sekvenčná analýza (Sequential pattern mining)

Analýza sekvencií je štatistická analýza, kde sa údaje získavajú nájdením vzorov a hodnoty sú dodávané v postupnosti. Údaje sa vyhodnotia hneď, ako sa zozbierajú, a ďalšie vzorkovanie sa zastaví na základe vopred definovaného pravidla zastavenia, len čo sa zaznamenajú významné výsledky. Preto je niekedy možné dospieť k záveru oveľa skôr, ako by bolo možné pri klasickom testovaní. Jedná sa o špeciálny prípad dolovania štruktúrovaných údajov, čo je proces zisťovania a získavania užitočných informácií z pološtrukturovaných súborov údajov.

2.3.10 Bayesove siete (Bayesian network)

Bayesova sieť (Bayesovská sieť, sieť viery, rozhodovacia sieť, Bayesov model, pravdepodobnostne riadený acyklický grafický model) je pravdepodobnostný grafický model, ktorý predstavuje skupinu premenných a ich podmienené závislosti prostredníctvom riadeného acyklického režimu.

Bayesovské siete sú ideálne na uskutočnenie udalosti, ktorá sa vyskytla a na predpovedanie pravdepodobnosti, že k tomu prispela niektorá z niekoľkých možných známych príčin (8).



Obr. 2.8: Príklad grafického zobrazenia Bayesovej siete

2.3.11 Algoritmus najbližších susedov (k-nearest neighbors algorithm)

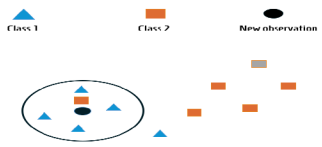
Algoritmus najbližších susedov je neparametrická metóda používaná na klasifikáciu a regresiu. Susedia sú prevzatí zo súboru objektov, pre ktoré je známa trieda (klasifikácia) alebo

Zdroj obr. 2.8: https://www.bayesfusion.com/bayesian-networks/

hodnota vlastnosti objektu (regresia). V oboch prípadoch vstup pozostáva z „k“ najbližších cvičných príkladov v priestore prvkov.

Výstup závisí od toho, či sa jedná o klasifikáciu alebo regresiu (10).

- **Klasifikácia**: objekt je klasifikovaný väčšinovým hlasom svojich susedov, pričom objekt je zaradený do triedy najbežnejšej spomedzi svojich najbližších susedov, výstupom je teda členstvo v triede
- **Regresia**: výstupom je hodnota vlastnosti pre objekt, ktorá predstavuje priemer hodnôt „k“ najbližších susedov



Obr. 2.9: Príklad zobrazenia k-NN

2.3.12 Samoorganizujúca sa mapa (Self-organizing map)

Samoorganizujúca sa mapa je typ umelej neurónovej siete, ktorá je trénovaná pomocou učenia bez dozoru, za účelom vytvorenia nízkorozmernej, diskriminačnej reprezentácie vstupného priestoru tréningových vzoriek (mapa). Uplatnením konkurenčného učenia sa na rozdiel od učenia korekcie chýb líšia od ostatných umelých neurónových sietí.

Majú dva režimy:

1. Tréning - vytvorenie mapy pomocou vstupných vzoriek
2. Mapovanie - automatická klasifikácia nového vstupného vektoru

Samoorganizujúce sa mapy sú užitočné pre vizualizáciu tým, že vytvoria nízkorozmerné zobrazenie vysokorozmerných dát (11).

Zdroj obr. 2.9: https://www.unemyr.com/k-nearest-neighbour-ai/

Kapitola 3

Analýza súčasného stavu

V danej kapitole sa budem zaoberať analýzou súčasného stavu firmy, v ktorej zadanú prácu robím. Vzhľadom na dodržiavanie firemného GDPR budem ďalej používať „XYZ“, ako označenie/pracovný názov firmy. Uvediem základné informácie o firme, čím sa zaoberá a v akej oblasti pôsobí. Ďalej popíšem kľúčové pracovné pozície, organizačnú štruktúru a približný model informačného toku.

3.1 Základné informácie o spoločnosti

Spoločnosť XYZ je vedúcou silou v oblasti profesionálnych prekladateľských služieb a lokalizačných technológií.

Jedná sa o nadnárodnú spoločnosť s medzinárodným pôsobením v Ázii, Európe a Severnej Amerike, ktorá ma širokosiahle skúsenosti v oblasti lokalizácie softvéru, technickej dokumentácie, e-learningu a multimédií pre popredné svetové spoločnosti. Počet zamestnancov tejto spoločnosti celkovo nepresahuje 500 osôb.

Pomáha iným spoločnostiam zvyšovať ich výnosy uvoľňovaním produktov a služieb na medzinárodné trhy prostredníctvom škálovateľného a modulárneho balíka služieb, ktorý prispôsobuje jazykové, kultúrne a technické aspekty výrobkov, služieb, dokumentácie a komunikácií v krajinách takmer celého sveta.

Spoločnosť XYZ poskytuje svoje služby v širokej škále odvetví.

Hlavné poskytované služby:

- **Preklad** - úpravy strojového prekladu, rýchly preklad, marketingový preklad, tlmočenie, zabezpečenie jazykovej kvality
- **Testovanie** - jazykové testovanie, funkčné testovanie, testovanie lokalizácie, správa chýb
- **Lokalizácia** - lokalizácia softvéru, lokalizácia užívateľskej podpory, lokalizácia webových stránok, lokalizácia e-learningu
- **Obchodné služby** - personálne služby, získavanie jazykových talentov
- **Multimédia** - prepisovanie, Voice Over, titulky, publikovanie na počítači
- **Technológia a umelá inteligencia** - NMT Custom Engine, WordsOnline TMS, analýza textu, naratívna generácia
- **Hry** - preklad a lokalizácia, zvuková produkcia, titulky, dabing a Voice Over, funkčné a jazykové testovanie, marketing a transcreation
- **Software** - preklad a lokalizácia, funkčné a jazykové testovanie, správa chýb, marketing a transcreation, agilné spôsoby dodania, UI a preklad e-learningu
- **Predaj a reklama** - preklad a lokalizácia, strojový preklad, popis produktov, marketing a transcreation, Global SEO
- **Marketing** - preklad a lokalizácia, transcreation, Global SEO a ASO, lokalizácia videa, Brand Research
- **TV a Film** - preklad a lokalizácia, transcreation, preklad skriptov, prepisovanie, Voice Over, titulky, video postprodukcia
- **E-learning** - preklad a lokalizácia, lokalizácia videa, prepisovanie, Voice Over, titulky, publikovanie na počítači, video postprodukcia
- **Cestovanie a turizmus** - preklad a lokalizácia, marketing a transcreation, lokalizácia webových stránok, online recenzie

3.2 WordsOnline

WordsOnline je lokalizačná platforma typu end-to-end[1]. Riadi všetky preklady akéhokoľvek typu obsahu a jazyka na všetkých cloudových médiách. Jedná sa o kombináciu neurálneho strojového prekladu a prekladovej pamäte s prekladateľskou komunitou s posilneným AI. WordsOnline poskytuje kontrolu, transparentnosť a škálovateľnosť.

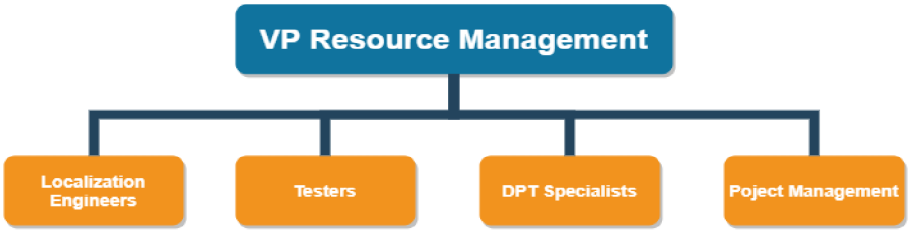
Na rozdiel od väčšiny prekladov, ktoré používajú zdĺhavý, manuálny, drahý a časovo náročný step-by-step postup, WordsOnline poskytuje nepretržitý prístup k publikovaniu a lokalizácii, ktorý je plne automatizovaný, založený na údajoch a súvislý s plne integrovanou jazykovou komunitou.

WordsOnline zjednodušuje pracovný postup prekladu prostredníctvom okamžitého pridelenia obsahu založeného na nárokoch globálnej sieti lingvistov. Platforma AI tiež sľubuje nepretržité doručovanie, s automatickým výberom obsahu pre pridanú QA a prioritnou kontrolou založenou na KPI s cieľom zabezpečiť vynikajúcu kvalitu prekladov.

3.3 Pracovné pozície

V tejto podkapitole budem špecifikovať štyri hlavné pracovné pozície vo firme. Jedná sa o *Local Engineers*, *Testers*, *DTP/Multimedia*a *Project Management*.

Všetky štyri tvoria základný kameň firmy, pretože sú na seba priamo naviazané a jedna bez druhej nemôžu fungovať. Každá z pozícií má vlastné „podpozície“, kde jednotliví vedúci tímov rozdeľujú úlohy svojim pracovníkom. Z hľadiska čo najväčšej presnosti a prehľadnosti preto môžeme zobraziť organizačnú štruktúru tak ako je na nasledujúcom obrázku 3.1.



Obr. 3.1: Organizačná štruktúra

Zdroj obr. 3.1: interný

[1]End-to-end popisuje proces, berúci systém alebo službu od začiatku do konca a poskytuje kompletné funkčné riešenie, bez potreby získania čohokoľvek od tretej strany.

3.3.1 Local Engineers (ENG)

Local Engineers patria do produkčného personálu a podávajú správy vedúcemu lokalizačného inžinierstva.

Primárnou úlohou sekcie ENG je podpora projektovým manažérom a zákazníkom v technických oblastiach vrátane prípravy projektu, zabezpečenia kvality, software testovania, technického riešenia problémov a finalizácie projektu. Skúsení inžinieri sa podieľajú na validácii, optimalizácii a automatizácii existujúcich inžinierskych procesov, ako aj na navrhovaní a implementácii nových procesov.

Náplň práce:

- Localization Engineering (analýzy, príprava súborov a následné spracovanie)
- Lokalizácia SW/UA
- Online help lokalizácia
- Screenshotting
- Testovanie a oprava chýb

3.3.2 Testers (TEST)

Testers podávajú správy vedúcemu inžinierskej skupiny.

Primárnou úlohou sekcie TEST je vykonávanie manuálneho aj automatického software testovania, sledovanie chýb a opravy chýb v súlade s priemyselnými normami a potrebami spoločnosti. Skúsení testeri sa podieľajú na validácii a optimalizácii existujúcich testovacích procesov, ako aj na tvorbe testovacích scenárov a testovacích prípadov.

Náplň práce:

- Funkčné, kozmetické a iné špecifické typy testovania na identifikáciu chýb
- Zaznamenávanie chýb, nové testovanie a overenie opravy a zmeny, až kým sa nedosiahne nulová chyba
- Výskum vrátane preskúmania iných východiskových materiálov pred začatím a počas projektu, s cieľom rozšíriť vedomosti o testovaní

- Spolupráca s členmi tímu, s cieľom zabezpečiť dokončenie práce v súlade s termínmi projektu.
- Zabezpečenie kvality DTP/doc
- Zabezpečenie kvality dodávok súborov
- Vykonávanie pridelenej QA aktivity vedúcim testovacieho tímu

3.3.3 DTP/Multimedia specialists (DTP)

DTP/Multimedia specialists sú členmi produkčného personálu a zodpovedajú vedúcemu DTP oddelenia.

Primárnou úlohou je realizácia širokej škály počítačových, grafických a dokumentačných úloh v závislosti od potrieb klienta. Vo väčšine prípadov pracujú priamo so súbormi klienta. Rovnako tak sú schopní vytvárať vlastné textové alebo grafické návrhy.

Náplň práce:

- Práca s dokumentáciou a grafickými materiálmi v rôznych formátoch
- Tlač a tvorba online PDF
- Multi-tasking prostredníctvom mnohých projektov
- Práca s nástrojmi spoločnosti Adobe
- Odhady práce, koordinácia

3.3.4 Project management (PM)

Projektoví manažéri sú členmi produkčného personálu a podávajú správy vedúcemu tímu projektu alebo skupiny.

PM riadia projekt a projektový tím v súlade so stanovenými požiadavkami a potrebami od začiatku až do konca. Zodpovedajú za plánovanie a chod jednotlivých projektov. Poskytujú vysoko kvalitné služby na dosiahnutie cieľov v primeranom časovom rámci a cene.

Náplň práce:

- Ziskovosť projektov
- Vyrovnávanie pracovnej záťaže členov projektových tímov
- Plánovanie lokalizačných projektov a alokácia zdrojov, vypracovanie rozpočtu, harmonogramu a parametrov kvality.
- Poskytovanie podpory podnikovým manažérom, architektom riešení a programovým manažérom pri výpočte nákladov a plánovaní projektov, a pri prezentáciách zákazníkom
- Aktívna komunikácia s členmi tímu s cieľom identifikovať potenciálne organizačné a prevádzkové problémy.
- Zabezpečenie dokončenia všetkých administratívnych projektových procesov

3.4 Hardware a Software

Základom technického vybavenia firmy XYZ je predovšetkým výpočtová technika od spoločnosti Dell (stolové počítače a monitory), kladne hodnotená hlavne pre ich kvalitu a vysoký výkon, ktorý je pre výkonovo aj výpočetne (časovo) náročné softwarové aplikácie na poskytované služby prekladov, lokalizácií a grafických úkonov priam nevyhnutný.

Z hľadiska softwaru sú využívané produkty od spoločností Microsoft (Excel) pre tvorbu záznamov, reportov a Adobe (Photoshop, Illustrator) predovšetkým na zložitejšie grafické úpravy. Rovnako sa používajú programy pre výkonný management kvality a terminológie (Xbench) alebo pre uľahčenie prekladu a korekciu textu (SDL Trados Studio).

3.5 Analýza pracovného procesu

Pracovný proces v danej spoločnosti začína analýzou, prerokovaním a schválením požiadaviek a podmienok stanovených klientom.

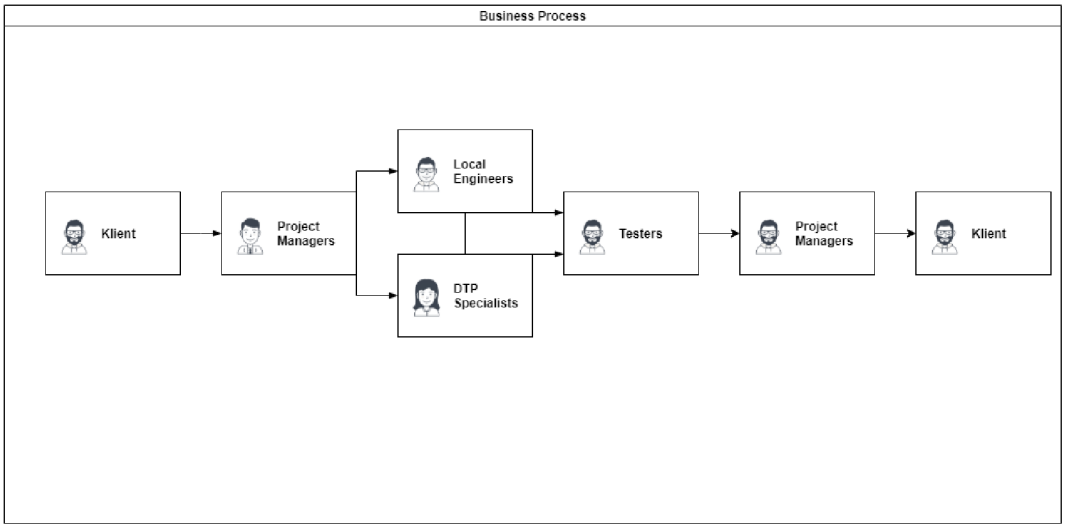
Projekt ako celok je následne rozdelený na jednotlivé časti v závislosti od typu a charakteru pracovných úkonov. Takto rozdelené časti projektu sú priradené jednotlivým projektovým manažérom. Tí následne stanovia a prerozdelia úlohy špecialistom (Local Engineers a DTP Specialists), ktorí zadané úlohy spracujú.

Local Engineers a DTP Specialists najskôr skontrolujú klientsky zdrojový súbor a jeho správnosť. V prípade nájdenia chýb ich opravia pomocou softwarových aplikácií.

Local Engineers sa zameriavajú na kontrolu štruktúry a textového obsahu súboru.

Úlohou DTP Specialists je nájdenie problémov v grafickom rozhraní.

Vzájomne spolupracujú pri odstraňovaní všetkých nedostatkov a chýb a následne upravia súbor na požadovaný formát, ktorý je určený k prekladu.



Obr. 3.2: Pracovný proces

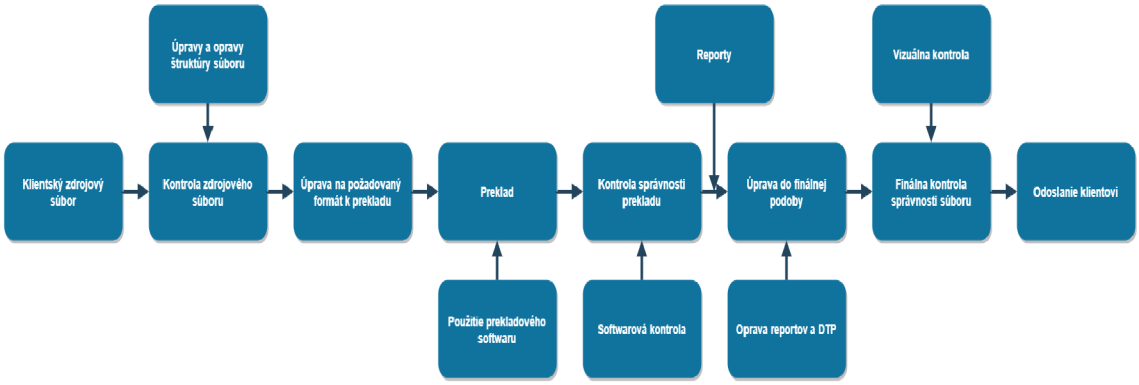
Ďalším krokom je samotný preklad textu, ktorý realizujú Local Engineers. Ten si vyžaduje nielen ich primerané jazykové kompetencie, ale hlavne využitie vhodných programov a aplikácií určených na prekladateľskú činnosť.

Zdroj obr. 3.2: interný

Po realizácii procesu prekladu je ďalším nevyhnutným krokom validácia výstupu pomocou softwaru, ktorý realizujú testeri. Tí podrobne skontrolujú textovú i grafickú časť. V prípade zistenia nedostatkov ich zdokumentujú a vytvoria report. Následne report pošlú príslušným špecialistom (ENG, DTP), ktorí ho využijú na odstránenie zistených chýb.

Local Engineers upravia súbor do finálnej podoby. Nasleduje posledná fáza kontroly „vizuálna kontrola“, ktorú vykonávajú testeri. V prípade nájdenia chýb sa proces opravy a následnej kontroly opakuje.

Po odstránení všetkých nedostatkov sa finalizuje výsledný produkt. Ten sa odosiela ako celok späť projektovým manažérom. Tí zaevidujú do firemnej databázy všetky údaje o projekte a výslednom produkte a odošlú ho klientovi.



Obr. 3.3: Informačný tok

Zdroj obr. 3.3: interný

3.6 Zhodnotenie súčasného stavu

XYZ je lokalizačno-prekladateľskou firmou, ktorá má k dispozícii profesionálny software na úpravu, editáciu, kontrolu a opravu štruktúry do klientom vyžiadaného výstupného formátu. O každom projekte si jednotliví pracovníci vedú záznamy a reporty, ktoré posielajú vyššie do firemného managementu (VP Resource Management).

VP Resource Management zaznamenáva dôležité, podstatné dáta pomocou tabuliek od Microsoft Excel. Zaznamenané dáta spoločne vytvárajú databázu, v ktorej budem v ďalšej kapitole vykonávať dolovanie dát (datamining). Takáto databáza je pomerne neprehľadná a ťažko sa s ňou pracuje z dôvodu veľkého množstva dát. Jednou z možností pre prácu a získanie podstatných údajov je využitie nástrojov Business Itelligence (BI) - Excelovských filtrov a kontingenčných tabuliek (Excel Pivot Tables), umožňujúcich vytvárať zostavy (reporty). Pivot Tables rovnako umožňujú prepájanie na databázové údaje v iných súboroch a formátoch ako je Excel. Všetky dané procesy spadajú pod koncepciu analytického spracovania OLAP (Online Analytical Processing).

Problém nastáva v situácii rozhodovania s výhľadom do budúcna. Nakoľko Excel Pivot Tables ponúkajú kvalitný prehľad len aktuálnych dát, nie je pomocou nich možné vytvoriť predikčný model, ktorý by vedel predpovedať napríklad možné budúce výnosy/náklady alebo zisk/stratu. To je moment, kedy do rozhodovacieho procesu vstupujú datamining nástroje a algoritmy.

Nakoľko poskytnutá databáza nie je vo vhodnom stave pre dolovanie dát, je nutné aby som si ju upravil za pomoci vyššie uvedených nástrojov BI. Jednotlivé postupy úpravy, výber vhodných dát ako aj dolovanie popíšem v nasledujúcej kapitole.

Kapitola 4

Vlastný návrh riešenia

V tejto kapitole sa budem zaoberať úpravou poskytnutej firemnej databázy a následne dolovaniu podstatných a dôležitých dát pre čo najvhodnejší tréning dát a následnú predikciu. Pomocou rôznych dolovacích metód budem schopný podať vedúcim pracovníkom managementu reálne a čo najpresnejšie výsledky. Tie budú môcť ďalej využiť v procese firemného rozhodovania a budúceho vývoja firmy.

Všetky obrázky v tejto kapitole budú vlastne vytvorené v prostredí RapidMiner Studio a Microsoft Excel.

4.1 Úprava databázy

Ako som už spomínal, jednotlivé zaznamenané dáta sa uschovávajú v rámci firemnej databázy pomocou tabuliek programu Microsoft Excel. Nakoľko databáza nie je vo vhodnom stave a forme pre priame využitie dolovacích metód a algoritmov je nutné dáta zobraziť vo vyhovujúcej podobe.

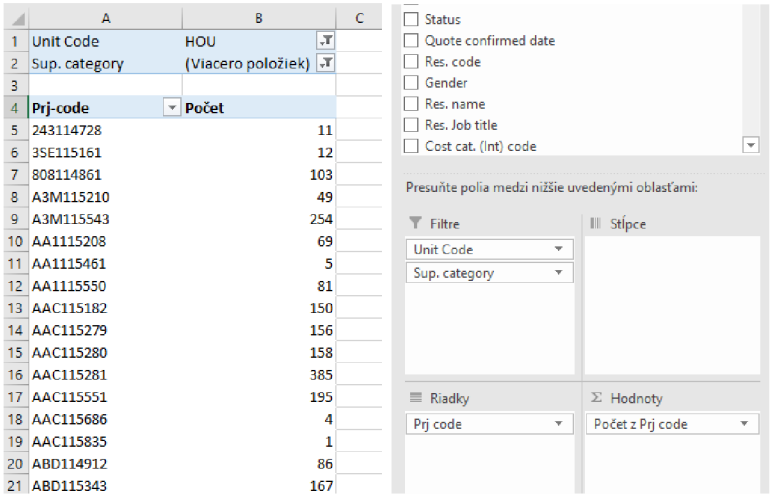
Type	Document code	Prj code	Prj status	Customer	WO	Service
PO	3172419	WAR114351	Completed	23133	WAR269698	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269698	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269699	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269699	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269697	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269697	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269697	Crowdsourcing
PO	3172419	WAR114351	Completed	23133	WAR269696	Crowdsourcing
PO	3171801	ADB114230	Completed	23579	ADB267635	Engineering
PO	3171794	ADB114230	Completed	23579	ADB267629	Engineering
PO	3174303	HB1114385	Completed	22930	HB1270039	PM
PO	3174311	HB1114385	Completed	22930	HB1270039	PM
PO	3174311	HB1114385	Completed	22930	HB1270040	Engineering
PO	3174771	HB1114515	Completed	22930	HB1271752	Engineering
PO	3174197	AZL114287	Completed	22773	AZL268616	Resourcing

Obr. 4.1: Ukážka z pracovnej databázy

Pre tento účel som si zvolil kontingenčné tabuľky (Pivot Tables) spadajúce pod technológiu OLAP. Pivot Tables sú nástrojom spracovania a usporiadania dát s cieľom upozorniť na užitočné informácie. Tento nástroj je vbudovaný a poskytovaný samotným programom Excel. Výsledkom úprav v kontingenčnej tabuľke teda bude sumarizácia údajov, ktorá môže obsahovať sumy, priemery alebo iné štatistiky vzhľadom na potreby užívateľa.

Poskytnutá databáza obsahovala niekoľko tisíc riadkov údajov o projektoch, zákazníkoch, zamestnancoch atď. Pomocou funkcie filtru som odstránil prázdne alebo nežiadúce prvky, ktoré by prípadne mohli znehodnotiť dolovací algoritmus aj s jeho výsledkami. V tomto konkrétnom prípade bolo nutnosťou filtrovať položku *Sup. category* z dôvodu nejasností a chybovosti určitých položiek (nedefinované položky, nezmyselné názvy). Oproti pôvodnému počtu všetkých riadkov tabuľky sa po filtrovaní počet znížil len o pár riadkov, čo znamená že odfiltrovaním som výrazne nezasiahol do štruktúry, práve naopak, zvýšil som šance na presnejšie modelovanie.

Ná základe obrázku 4.2 vidieť, že do funkcie filtru je priradená ešte jedna položka *Unit Code*. Samotná položka obsahuje údaje o jednotke hodnotenia pracovného výkonu. To znamená že obsahuje hodnoty ako: *HOU* (hodinová sadzba), *CHR* (sadzba za slovo), *DAY* (denná sadzba) a *PGS* (sadzba za stranu). V tomto konkrétnom prípade som sa zameral hlavne na hodinovú sadzbu (HOU) a preto všetky modely a analýzy budú pracovať len v „hodinovom režime“.



	A	B	C
1	Unit Code	HOU	
2	Sup. category	(Viacero položiek)	
3			
4	**Prj-code**	**Počet**	
5	243114728	11	
6	3SE115161	12	
7	808114861	103	
8	A3M115210	49	
9	A3M115543	254	
10	AA1115208	69	
11	AA1115461	5	
12	AA1115550	81	
13	AAC115182	150	
14	AAC115279	156	
15	AAC115280	158	
16	AAC115281	385	
17	AAC115551	195	
18	AAC115686	4	
19	AAC115835	1	
20	ABD114912	86	
21	ABD115343	167	

Obr. 4.2: Kontingenčná tabuľka

Ďalším a finálnym krokom som výsledné analýzy získané z kontingenčných tabuliek zjednotil a spojil do jednej spoločnej tabuľky určenej pre konkrétny typ modelu dátového dolovania. Pre každý jeden model bolo nutné úpravami vytvoriť samostatnú tabuľku (databázu) v závislosti na požadovaný výsledný stav a predikciu budúcich stavov.

Takto upravené dáta boli pripravené na import do zvoleného softwaru pre dolovanie dát (datamining). V tomto prípade som si zvolil program **RapidMiner Studio** pre jeho jednoduchosť používania, interaktívny tutoriál, licenčné možnosti pre edukačné účely a hlavne množstvo možností pre tvorbu analytických a prediktívnych modelov.

RapidMiner je softwarová platforma, ktorá poskytuje integrované prostredie pre prípravu údajov, ťažbu textu a prediktívnu analýzu pomocou schém, modelov a algoritmov. Poskytuje GUI (Graphical User Interface) na navrhovanie a vykonávanie analytických pracovných tokov „procesov“. Procesy sa skladajú z viacerých „operátorov“. Operátor vykonáva jednu úlohu v rámci procesu a výstup každého operátora tvorí možnosť vstupu pre ďalší. Celkový výsledok analýzy alebo predikcie vzniká prepojením operátorov a výstupného bodu s následným spustením procesu (14).

4.2 Návrh lineárnej regresie (LIN_REG)

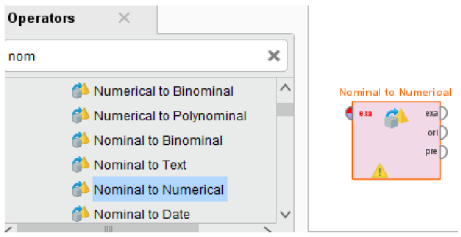
Model lineárnej regresie som vytvoril za účelom zistenia miery závislosti zvolených atribútov *Cost Cat. Code* (ID nákladovej kategórie do ktorej projekt spadá) a *Supplier type* (typ dodávateľa) na *New HOU Value* (celkové náklady za projekt).

Cost Cat. Code a *Supplier type* obsahujú číselné hodnoty, ktoré vyjadrujú celkový počet hodín strávených na projektoch a *New HOU Value* obsahuje číselné hodnoty vyjadrujúce celkovú sumu nákladov v eurách.

Výstupom tohto návrhu by mal byť predikčný model, schopný predpovedať možné budúce hodnoty celkových nákladov na projekt s konkrétnou (čo najviac možnou) presnosťou.

4.2.1 Tvorba modelu LIN_REG

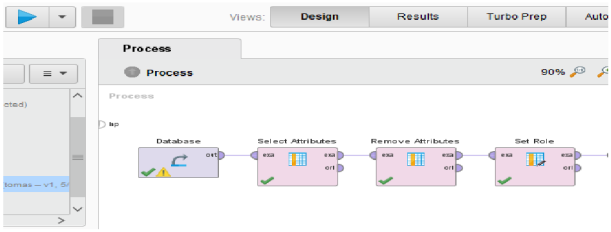
Návrh začínal importom vhodne upravenej databázy. Vzhľadom na to, že výstupom lineárnej regresie bol spojitý výstup, museli byť všetky atribúty číselného dátového typu. V prípade iného než číselného dátového typu bolo možné upraviť ich priamo v programe a to použitím operátorov pre zmenu typu atribútu.



Obr. 4.3: Operátory pre zmenu dátového typu

Najčastejšie mnou využívaným operátorom pre zmenu dátového typu bol *Nominal to Numerical*, ktorý z každej jedinečnej textovej hodnoty vytvoril binárny atribút. Takto vytvorené binárne atribúty boli naplnené hodnotami „0“ alebo „1“.

Po importe databázy bolo nutné zvoliť atribúty závislé na sledovanom pomocou operátora *Select Attributes*. V mojom prípade som kvôli vysokému počtu atribútov zvolil metódu selekcie všetkých a následne obrátenej selekcie (invert selection), ktorá mi po zvolení atribútov vykonala opačnú činnosť, zvolené atribúty odstránila z databázy. Tým som zabezpečil, aby mi do procesu vstupovali len tie významovo potrebné.



Obr. 4.4: Proces úpravy databázy pre lineárnu regresiu v programe RapidMiner Studio

Ďalším krokom bolo pridanie operátoru *Set Role*. Pomocou neho som si vybral atribút, ktorý som chcel sledovať a predikovať. Jednalo sa o celkové náklady *New HOU Value*. Po realizácii tohto kroku som mohol konštatovať, že databáza je upravená a pripravená pre tvorbu modelu lineárnej regresie.

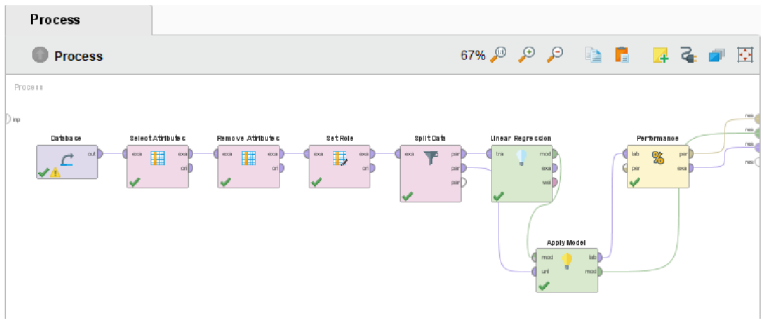
Pokračoval som pridaním operátora *Split Data*, ktorý rozdelil databázu na menšie partície v požadovanom pomere. V prípade tvorby len tréningových dát by som doň zadal jedinú hodnotu „1.0“, ktorá vyjadruje percentuálne (100%) použitie celkového obsahu na trénovanie modelu.

V prípade tvorby samotnej predikcie bolo nutné rozdeliť databázu na tréningovú partíciu a na partíciu predikčnú (zvolený atribút na predikciu bol odstránený a nahradený prázdnym pre zápis vypočítaných hodnôt). Všeobecne najideálnejší pomer rozdelenia je „0.7“ (70%) na tréning a „0.3“ (30%) na predikciu.

S takto rozdelenou databázou bolo možné pridať ďalší operátor *Linear Regression*. Tento operátor vypočítal model lineárnej regresie na základe dát na vstupe. V danom prípade využíval ku kalkulácii dataset tvorený 70% partíciou.

Následne bolo nutné pridať operátor *Apply Model*, do ktorého vstupovala predikčná partícia *Split Data* (30%) a výstup z *Linear Regression*. *Apply Model* aplikoval vypočítaný model na predikčnú partíciu.

Posledným krokom bolo pridanie operátoru *Performance (Regression)*. Ten slúži na vyhodnotenie výkonnosti regresných úloh podľa istých voliteľných kritérií. Prioritne sa používa kritérium RMSE (Root Mean Squared Error). Používa sa ako miera rozdielov medzi modelom predpovedanými hodnotami a pozorovanými hodnotami. Hodnota RMSE je vždy nezáporná alebo nulová. Nulová v praxi nie je takmer nikdy dosiahnutá, pretože by to znamenalo dokonalé prispôsobenie sa údajom. Zároveň ma zaujímala miera SC (Squared Correlation), ktorá vyjadruje korelačný koeficient medzi predikovaným atribútom a ostatnými.



Obr. 4.5: Celkový proces lineárnej regresie v programe RapidMiner Studio

4.2.2 Výsledok modelu LIN_REG

Na základe výsledku procesu môžem tvrdiť, že predikčný model vedel pomocou atribútov *Cost Cat. Code* a *Supplier type* predpovedať hodnoty celkových nákladov na projekt *New HOU Value* s presnosťou **90.7%** .

Obr. 4.6: Výsledné hodnoty RSME a SC

Na nasledujúcom obrázku 3.7 možno vidieť ako vyzerajú predikované hodnoty *prediction(New HOU Value)* oproti trénovaným hodnotám *New HOU Value* .
Z dôvodu zobrazenia citlivých položiek boli pôvodné hodnoty nákladov nahradené fiktívnymi.

Row No.	New_HO...	prediction(N...	BCIP	BGIP
334	21163.048	20153.584	112	0
259	21010.693	19131.032	0	0
44	20002.304	19921.311	38	0
260	17522.024	16348.988	0	0
365	16586.281	14736.046	0	129
51	16230.362	14774.102	0	0
56	16205.591	14951.248	0	0
153	16105.002	18425.240	0	0

Obr. 4.7: Predikované náklady

4.3 Návrh rozhodovacieho stromu(DEC_TREE)

Model rozhodovacieho stromu som vytvoril so zámerom zjednodušenia procesu rozhodovania v prípade výberu konkrétneho zákazníka pre ďalšiu budúcu spoluprácu. V mnou vytvorenom procese som skúmal závislosti atribútov spojených so zákazníkom, definovaným jeho jedinečným kódom (Customer ID).

Výstupom tohto návrhu by mal byť graf rozvetveného stromu s koncovými vetvami určujúcimi, či sa jedná o zákazníka, ktorý so spoločnosťou XYZ spolupracoval v rokoch 2019 a 2020 (potencionálne stály zákazník) alebo len v roku 2019.

4.3.1 Tvorba modelu DEC_TREE

Návrh začínal tak, ako v predošlom modeli lineárnej regresie, importom vhodne upravenej databázy. Vzhľadom k využitiu klasifikácie (triedenia) bolo v tomto prípade nutné, aby práve sledovaný (Label) atribút bol v nominálnej forme. To znamená, že som si ho mohol pomocou úprav vytvoriť/pridať do databázy ešte pred importom alebo použiť poskytované operátory na zmenu typu atribútu ako na obrázku 4.3. V konkrétnom prípade bolo možné použiť operátory *Nominal to Numerical* alebo *Nominal to Binominal*. Vzhľadom na rozsah a typ vybraných dát som zvolil úpravu a vloženie priamo do databázy pomocou znalostí so vstavanými Excel funkciami.

Po importe databázy nasledovalo zvolenie vhodných a potrebných atribútov pomocou operátora *Select Attributes*. Opäť som využil možnosť operátora pre inverziu vybraných atribútov, čo znamenalo odstránenie celkového výberu. Do procesu mi tým pádom vstupovalo niekoľko číselných a práve jeden nominálny atribút, pre splnenie klasifikačnej podmienky.

Atribúty číselného typu, ktoré vstupovali do procesu:

- *Cost*: vyjadruje celkové náklady vynaložené pre splnenie požiadaviek stanovených klientom od zadania až po odovzdanie hotového produktu/služby
- *Days*: vyjadruje celkový počet pracovných dní využitých pre splnenie požiadaviek stanovených klientom od zadania až po odovzdanie hotového produktu/služby
- *PrjNum*: vyjadruje celkový počet projektov zhotovených pre klienta

- *Responsibles*: vyjadruje celkový počet zamestnancov zodpovedných za jednotlivé úkony a projekty

Atribút nominálneho typu, ktorý vstupoval do procesu:

- *Again*: vyjadruje hodnoty textového formátu „Yes“ alebo „No“, reprezentujúce spoluprácu zákazníka v rokoch 2019 a 2020

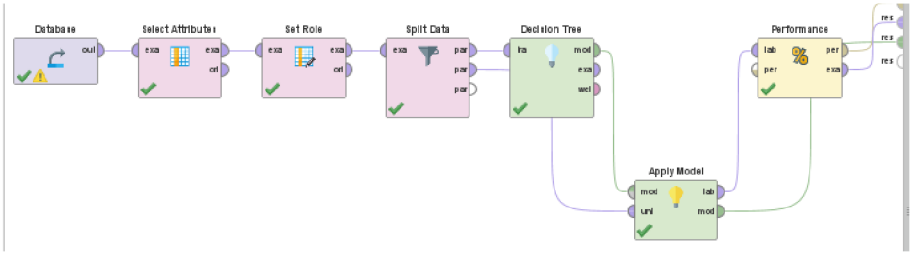
Ďalším krokom bolo pridanie operátoru *Set Role* . Pomocou neho som vyberal nominálny atribút (Label), od ktorého som chcel, aby bol konečným prvkom vetiev rozhodovacieho stromu. Jednalo sa o atribút *Again*, ktorý obsahoval hodnoty textového typu „Yes“ a „No“. Tieto hodnoty reprezentovali skutočnosť, či jednotlivý zákazník spolupracoval so spoločnosťou v roku 2019 a zároveň aj v roku 2020. Ukončením výberu som zároveň ukončil selekciu a úpravy prvkov databázy. Od toho momentu bola databáza pripravená na tvorbu požadovaného modelu rozhodovacieho stromu.

Nasledovalo pripojenie operátoru *Split Data* . V ňom som databázu rozdelil na menšie, mnou určené partície. Opäť som využil pomer 7:3. Tento pomer reprezentoval rozdelenie dát databázy na dve partície. Prvá obsahovala 70% celkových dát určených na trénovanie modelu a druhá obsahovala 30% zvyšných dát určených na testovanie modelu. Voľbu dát, ktoré následne idú do určených partícií som prenechal operátoru a to zvolením možnosti *shuffled sampling* (náhodný výber vzoriek) funkcie *sampling type* (typ vzorkovania). Pri väčšom objeme dát tým zaručoval vysokú kvalitu tréningu modelu vedúcu k čo najpresnejšej predikcii/testovaniu.

Po rozdelení databázy prišlo na rad pripojenie operátoru *Decision Tree*, ktorý z dát na vstupe (vybraných 70%) vypočítal model rozhodovacieho stromu. Odporúčaná funkcia pre prispôsobenie výsledného grafického zobrazenia je *maximal depth*. Túto funkciu som nastavil na hodnotu 7 kvôli prehľadnosti a vyjadruje maximálnu hĺbku stromu. Pre lepšie pochopenie danej štruktúry je možné predstaviť si mriežku, v ktorej riadky symbolizujú vetvy stromu a stĺpce predstavujú hĺbku stromu.

Ďalej som zapojil do procesu operátor *Apply Model*. Doň vstupovala partícia s 30% výberom dát z operátoru *Split Data* a výstup z operátoru *Decision Tree*. *Apply model* aplikoval vypočítaný model na testovaciu (30%) partíciu.

Posledným krokom bolo pridanie operátoru *Performance (Classification)*. Ten zobrazil presnosť (Accuracy[%]) modelu stromu a tabuľku vzťahu medzi testovacími a tréningovými hodnotami „Yes“ a „No“.



Obr. 4.8: Celkový proces rozhodovacieho stromu v programe RapidMiner Studio

4.3.2 Výsledok modelu DEC_TREE

Podľa výsledku procesu môžem tvrdiť, že rozhodovací strom mal na základe zvolených atribútov *Cost*, *Days*, *PrjNum*, *Responsibles* a *Again* presnosť (Accuracy) **81.97%**. Daná presnosť vyjadrovala s koľko percentnou úspešnosťou bolo možné predpovedať, aký zákazník na základe daných atribútov by dlhodobo (s výhľadom minimálne 1 rok) so spoločnosťou spolupracoval.

accuracy: 81.97%

	true Yes	true No	class precision
pred. Yes	11	9	55.00%
pred. No	2	39	95.12%
class recall	84.62%	81.25%	

Obr. 4.9: Presnosť predikcie s tabuľkou predikcii jednotlivých hodnôt

Tabuľka na obrázku 4.9 zobrazuje Accuracy a zároveň presnosť predikovaných hodnôt. Z tabuľky tak možno vyčítať, že najvyššia percentuálna presnosť (95.12%) bola pri predikcii hodnôt „No“ kde počet protichodných tvrdení medzi predikovanými a skutočnými hodnotami „No“ bol 2, zatiaľ čo zhoda bola u 39 prípadov. Naopak najmenšia percentuálna presnosť (55%) bola pri predikcii hodnôt „Yes“ kde počet protichodných tvrdení medzi predikovanými a skutočnými hodnotami „Yes“ bol 9, zatiaľ čo zhoda bola u 11 prípadov.

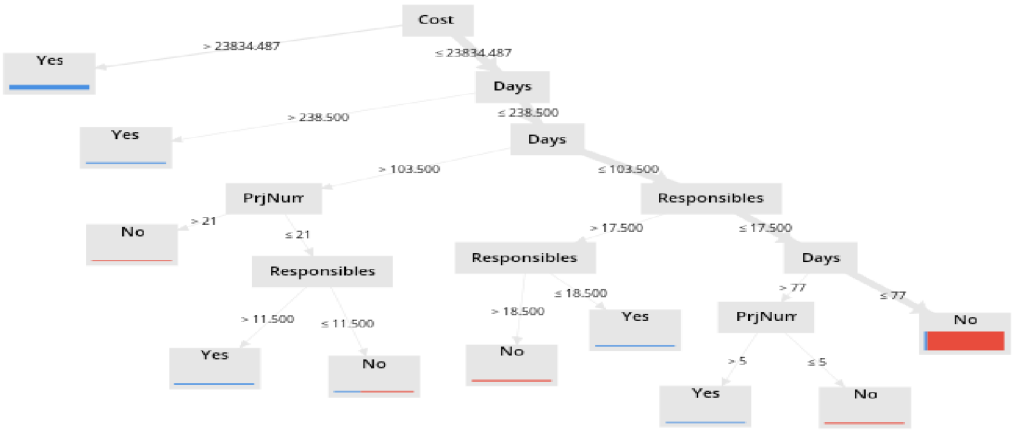
Vzhľadom na hodnoty tak bolo možné predpokladať, že celková presnosť by bola oveľa vyššia pri použití väčšieho množstva dát, ako pre tréning, tak aj pre testovanie (predikciu).

Z dôvodu zobrazenia citlivých položiek boli pôvodné hodnoty nákladov nahradené fiktívnymi.

Row No.	Again	prediction(A...	confidence(...	confidence(...	Days	Responsibles	PrjNum	Cost
1	Yes	Yes	1	0	245	58	17	105411.325
2	No	No	0.040	0.960	5	2	3	35.469
3	Yes	Yes	1	0	243	31	16	20870.853
4	No	No	0.040	0.960	3	4	1	32.417
5	No	No	0.040	0.960	27	6	2	1045.908
6	Yes	Yes	1	0	254	19	7	13071.148
7	Yes	Yes	1	0	255	18	23	63334.178
8	No	Yes	1	0	171	21	19	2845.762
9	No	No	0.040	0.960	37	7	5	642.181
10	No	Yes	1	0	102	11	18	3284.766
11	No	No	0.040	0.960	51	5	3	2440.752
12	No	No	0.040	0.960	26	6	3	812.895
13	No	No	0.040	0.960	30	2	1	395.098

Obr. 4.10: Graf rozhodovacieho stromu

Celý stromový graf určený k podpore rozhodovania ako aj jeho popis v podmienkovej forme možno vidieť na nasledujúcich obrázkoch 4.11 a 4.12.

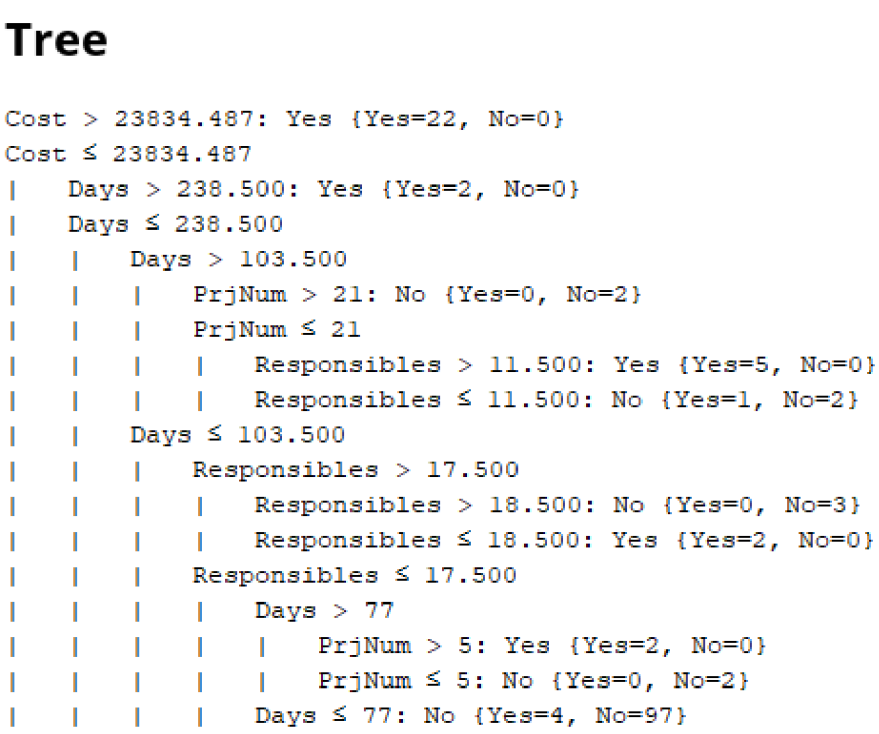


Obr. 4.11: Tabuľka testovacích a predikovaných dát

Podľa grafu a rovnako aj podľa popisu (obrázky 4.11 a 4.12) možno vidieť podmienky, ktoré je nutné splniť pre dosiahnutie výsledku na koncových listoch stromu.

Napríklad môžem vidieť, že v prípade celkových nákladov vyšších ako 23834.487 má spoločnosť 100% istotu budúcej spolupráce u 22 zákazníkov nachádzajúcich sa v danej vetve.

Naproti tomu, ak sa počet celkových nákladov nachádzal pod daný limit, celkový počet zamestnancov zodpovedných za jednotlivé úkony a projekty bol menší ako 17.5 a zároveň počet celkových dní strávených na všetkých projektoch bol pod hodnotou 77,viem povedať že 96% (Yes=4, No=97) všetkých zákazníkov nachádzajúcich sa v danej vetve nebude ďalej s firmou spolupracovať.



Obr. 4.12: Popis rozhodovacieho stromu v podmienkovej forme

4.4 Návrh zhlukovej analýzy (CLUSTER)

Model zhlukovej analýzy som vytvoril za účelom zobrazenia atribútov (prvkov) databázy, ktoré medzi sebou súvisia a sú si istými vlastnosťami podobné. Pre tento konkrétny návrh som si zvolil atribút *Country* ako sledovaný (Label), pre možnosť získania zaujímavých výsledkov čo sa riadenia projektov v rámci jednotlivých zákazníkov z rozličných krajín spolupracujúcich so spoločnosťou týka.

Takto vybrané a do skupín (zhlukov) zoradené dáta možno použiť v ďalších analýzach, prípadne priamo v rozhodovaní o budúcej spolupráci so zákazníkmi z daných krajín.

Výstupom tohto návrhu modelu by mali byť predovšetkým grafy, poukazujúce na rozdelenie krajín do určitého, vhodne zvoleného, počtu zhlukov na základe atribútov, v ktorých medzi nimi nastáva istá miera podobnosti. Grafy možno využiť najmä na prezentačné účely ako vizuálnu pomôcku pre lepšie vysvetlenie a pochopenie problematiky rozhodovania.

4.4.1 Tvorba modelu CLUSTER

Návrh začínal tak ako v predošlom modeli rozhodovacieho stromu. Bolo nutné zvoliť import vhodne upravených dát poskytnutej pracovnej databázy. Celý model zhlukovej analýzy sa zaoberal súvislosťami a podobnosťami len číselných atribútov, preto bolo možné využiť len numericky vyjadrené atribúty alebo použiť programom poskytovaný operátor na zmenu typu atribútu ako na obrázku 4.3. V danom prípade sa jednalo o možnosti *Numerical to Polynominal* alebo *Numerical to Binominal*. Vzhľadom na rozsah a typ vybraných dát som zvolil úpravu a vloženie priamo do databázy pomocou znalostí so vstavanými Excel funkciami.

Po importe databázy nasledovalo zvolenie vhodných a potrebných atribútov pomocou operátora *Select Attributes*. Možnosť vybraného operátoru mi poskytla inverznú funkciu. To znamená, že v prípade selekcie atribútov a zvolenia možnosti *invert selection* boli vybrané atribúty odstránené a nešli tak do ďalšieho procesu. Z hľadiska možnosti lepšej prehľadnosti výsledných dát som do procesu zapracoval aj atribút *Country* s nominálnymi hodnotami, ktorý ale nemal priamy vplyv na celkový výpočetný proces.

Atribút nominálneho typu, ktorý sa účastnil procesu:

- *Country*: zobrazuje názvy vybraných klientskych krajín

Atribúty číselného typu, ktoré vstupovali do procesu:

- *Customers*: vyjadruje celkový počet zákazníkov danej krajiny
- *lang. s.*: vyjadruje celkový počet jazykov vstupujúcich do procesu prekladu v rámci danej krajiny
- *lang. t.*: vyjadruje celkový počet jazykov vystupujúcich z procesu prekladu v rámci danej krajiny
- *Prj code*: vyjadruje celkový počet uskutočnených projektov
- *Prj type*: vyjadruje celkový počet typov projektov
- *Service*: vyjadruje celkový počet poskytnutých služieb zákazníkom
- *Unit Code*: vyjadruje celkový počet druhov jednotiek práce (hodina, deň, počet strán, počet písmen) použitých v rámci danej krajiny
- *Year*: vyjadruje celkový počet rokov spolupráce

Ďalším krokom bolo pridanie operátoru *Set Role* . Pomocou neho som vybral nominálny atribút (Label), ktorý slúžil len ako pomôcka pre lepšiu orientáciu vo výstupných dátach. Jednalo sa o atribút *Country*.

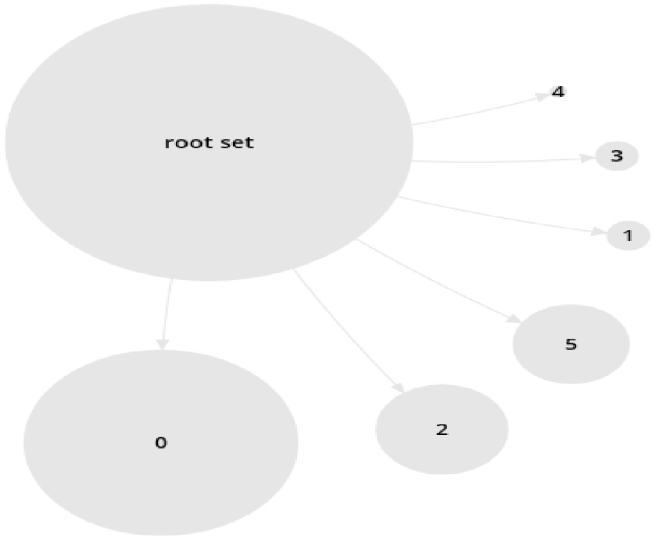
Posledným krokom bolo pripojenie operátoru *Clustering (k-Means)*. Tento operátor vykonal zhlukovanie pomocou algoritmu *k-means*. Daný algoritmus na základe určeného počtu „k“ vytvoril príslušný počet zhlukov, ktoré sú následne naplnené podobnými prvkami databázy. Podobnosť medzi prvkami je založená na mierke vzdialenosti medzi nimi.

Vzhľadom na veľkosť upravenej databázy som zvolil 6 ako hodnotu odpovedajúcu „k“. To znamená, že na výstupe bolo vytvorených práve 6 zhlukov, do ktorých boli rozdelené jednotlivé krajiny na základe podobnosti ich atribútov.

4.4.2 Výsledok modelu CLUSTER

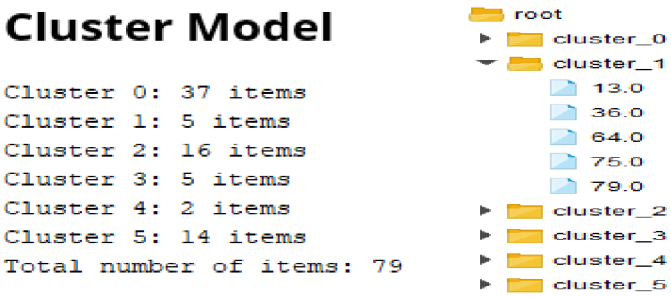
Výsledok modelu tvorilo 6 zhlukov. Na nasledujúcom obrázku 4.13 možno vidieť graf rozdelenia prvkov do zhlukov a veľkostné porovnanie vytvorených zhlukov oproti veľkosti všetkých dát v *root set*.

Môžem teda povedať, že zo všetkých zhlukov mal *cluster_0* najväčšie zastúpenie dát a *cluster_4* najmenšie zastúpenie.



Obr. 4.13: Graf rozdelenia prvkov do zhlukov

Podľa nasledujúceho obrázku 4.14 možno vidieť rozpis všetkých zhlukov s presným počtom prvkov, ktoré obsahujú. Teda z celkového počtu prvkov 79 práve 37 prvkov spadalo na základe podobných vlastností atribútov do *cluster_0*. Daný obrázok je teda číselným vyjadrením obrázku 4.13.



Obr. 4.14: Popis rozdelenia prvkov do zhlukov

Tak ako môžem vidieť počet zastúpených prvkov v jednotlivých zhlukoch, mám rovnako aj možnosť pohľadu do nich podľa obrázku vpravo. Pomocou daného pohľadu môžem vidieť všetky krajiny podľa ich ID hodnoty, ktoré mali v istej miere vzájomne podobné atribúty.

V prípade potreby zistenia polohy konkrétnej krajiny v istom zhluku bola súčasťou výstupu aj celková tabuľka hodnôt atribútov všetkých krajín tak ako išli do vstupu procesu. Táto tabuľka bola rozšírená o atribút *cluster* zobrazujúci konkrétny zhluk, do ktorého krajina na základe podobných atribútov patrila. Z hľadiska prehľadnosti som v obrázku 4.15 zobrazil len prvých 13 krajín (riadkov).

Row No.	id	Country	cluster	Customers	Prj code	Year	Prj type	lang. s.	lang. t.	Unit Code	Service
1	1	Algena	cluster_0	1	1	1	1	1	1	1	1
2	2	Argentina	cluster_5	49	171	2	5	7	13	3	14
3	3	Australia	cluster_0	16	47	1	4	2	6	2	8
4	4	Austria	cluster_2	26	66	1	4	4	6	3	6
5	5	Belarus	cluster_0	8	45	1	4	3	7	2	3
6	6	Belgium	cluster_5	66	260	3	6	8	14	3	13
7	7	Benin	cluster_0	1	1	1	1	1	1	1	1
8	8	Bosnia and H...	cluster_5	48	194	2	4	4	15	3	10
9	9	Brazil	cluster_3	75	430	2	6	7	17	3	13
10	10	Bulgaria	cluster_2	14	82	1	4	3	4	3	6
11	11	Cameroon	cluster_0	3	4	1	2	2	2	1	2
12	12	Canada	cluster_5	61	235	2	6	5	20	3	15
13	13	China	cluster_1	111	535	3	6	8	25	4	15

Obr. 4.15: Tabuľka rozdelenia do zhlukov

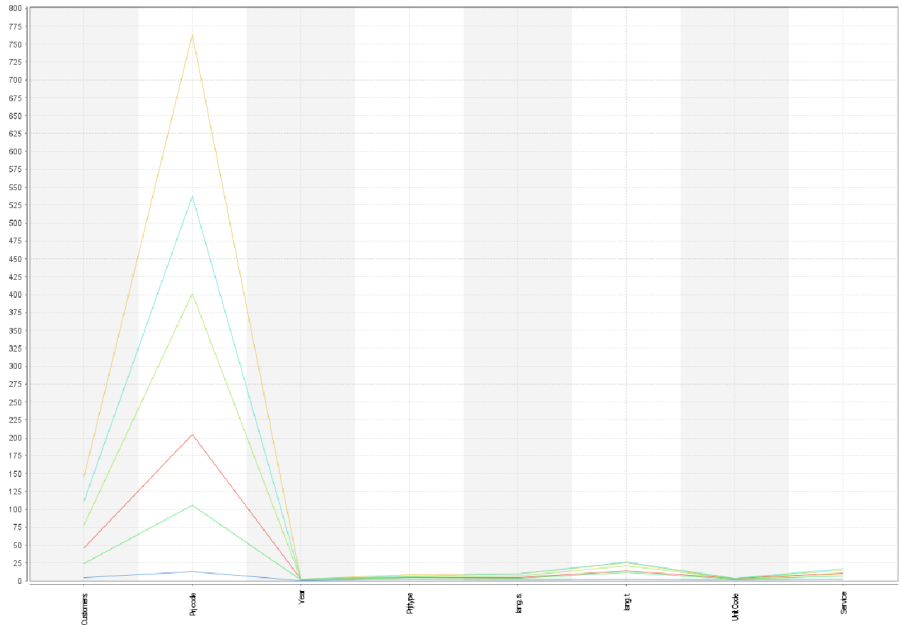
Ďalším podstatne potrebným výstupom pre ďalšie analýzy a rozhodovanie bola tabuľka priemerných hodnôt ťažiska jednotlivých atribútov, zobrazená na obrázku 4.16. Každý zhluk obsahoval atribúty s takto spriemerovanými hodnotami. Tie v podstate vyjadrovali akýsi stred (ťažisko), okolo ktorého sa pohybovali všetky hodnoty atribútu krajín patriacich do daného zhluku.

Zjednodušene, na príklade povedané, hodnota atribútu *Prj code* zhluku *cluster_5* vyjadrovala priemernú hodnotu všetkých hodnôt *Prj code* daného zhluku. Jednalo sa teda o hodnotu, na základe ktorej si boli ostatné podobné.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
Customers	4.622	110.600	24.562	77.600	145	45.571
Prj code	12.892	537.400	105.312	401.800	764	204.857
Year	1.027	2.800	1.875	2.400	2.500	2.143
Prj type	2.216	6.200	4.125	5.200	8.500	5
lang. s.	1.784	9.600	3.812	7.200	9.500	5.143
lang. t.	3.027	26.200	12	21.400	25	13.429
Unit Code	1.730	3.200	2.875	3.200	2.500	3.286
Service	2.730	16.800	6.812	15	10	10.429

Obr. 4.16: Tabuľka priemerných hodnôt ťažiska jednotlivých atribútov

Na nasledujúcom obrázku 4.17 možno vidieť graf významnosti atribútov jednotlivých zhlukov. Podľa neho vidím a môžem jednoznačne identifikovať atribút, ktorý bol najpodstatnejší z hľadiska triedenia prvkov do určených zhlukov. Jedná sa o *Prj code*, celkový počet uskutočnených projektov.

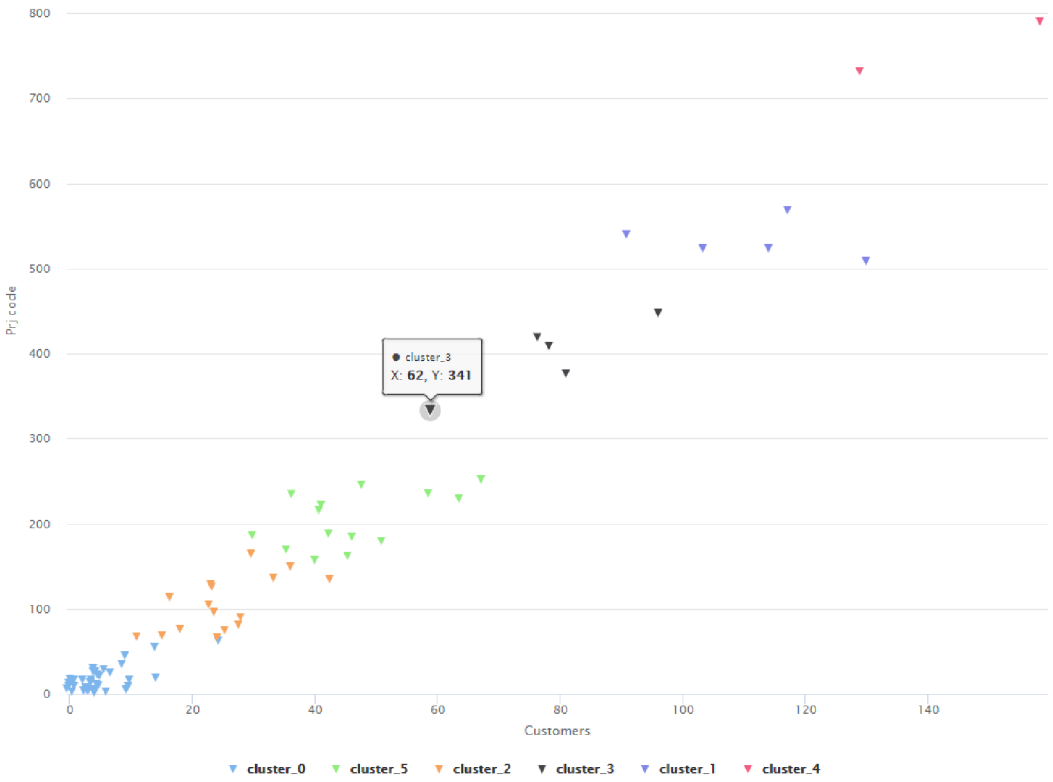


Obr. 4.17: Graf významnosti jednotlivých atribútov

Posledným výstupom zhlukovej analýzy bol graf zastúpenia krajín na obrázku 4.18. Graf mal štandardne dve súradnicové osi (X,Y), s možnosťou meniť a zadávať vybrané atribúty v rámci oboch osí. Výsledný graf bol bodový, kde každý bod prezentoval jednu krajinu. Body boli farebne rozdelené podľa typu zhluku, ktorý reprezentovali.

Práca s daným grafom bola zaujímavá hlavne z hľadiska možných výstupov, zobrazených vo zvolenej (v tomto prípade bodovej) grafickej podobe.

Na obrázku 4.18 tak možno vidieť jedno mnou vybrané bodové zastúpenie. Každý bod vyjadruje závislosť počtu zákazníkov jednej krajiny na počte celkových projektov vykonaných v rámci tej samej krajiny. Pomocou takto vytvoreného grafu je možné prezentovať prípadné budúce smerovanie spolupráce na projektoch a rozširovanie poľa pôsobnosti a služieb do ďalších krajín.



Obr. 4.18: Príklad grafu zastúpenia krajín v závislosti od počtu zákazníkov (os X) a počtu projektov (os Y)

4.5 Zhodnotenie návrhu riešenia

Návrh riešenia pozostával z návrhov a tvorby troch algoritmov dolovania dát (datamining). Jednalo sa o návrh *modelu lineárnej regresie*, návrh *modelu rozhodujúceho stromu* a návrh *modelu zhlukovej analýzy*. Dané tri modely som vybral so zámerom čo najvecnejšej prezentácie dolovaných a predikovaných dát spoločnosti XYZ. Každý z modelov pracoval s inou dátovou štruktúrou, kvôli prezentácii výsledných dát z rôznych uhlov pohľadu.

- **LIN_REG**: úlohou návrhu lineárnej regresie bolo najmä nájdenie silných vzťahov medzi skúmanými atribútmi za účelom získania modelu, ktorý by bol schopný pomocou daných atribútov predikovať možné budúce celkové náklady

  Využitie: optimalizácia celkových nákladov

- **DEC_TREE**: úlohou návrhu rozhodovacieho stromu bolo vytvorenie podmienkových vetiev, na základe zvolených atribútov, ktoré poukazovali na skúmaný výsledok (spolupráca klientov v rokoch 2019 a 2020) zobrazený v koncových vetvách pri splnení všetkých podmienok vedúcich od koreňa stromu, cez jednotlivé uzly, až po koncové listy („Yes“ a „NO“)

  Využitie_A: plánovanie budúcej spolupráce s vernými klientmi
  Využitie_B: odhalenie prvkov vplývajúcich na dlhodobú spoluprácu

- **CLUSTER**: úlohou návrhu zhlukovej analýzy bolo objavenie prvkov, ktoré mali klienti v rámci rozličných krajín podobné a následne roztriedenie daných krajín podľa daných nájdených podobných prvkov do zhlukov

  Využitie_A: expanzia služieb do nových krajín
  Využitie_B: rozšírenie služieb v stávajúcich krajinách

Z dôvodu zobrazenia citlivých položiek boli pôvodné hodnoty nákladov vo všetkých modeloch nahradené fiktívnymi. Niektoré hodnoty boli len zľahka pozmenené, no drvivá väčšina hodnôt bola ponechaná v pôvodnej forme. Aj napriek menším zásahom do štruktúry dát si dovoľujem poukázať na možnosť reálneho využitia všetkých modelov vo firemných procesoch, a to hlavne vďaka nadpriemerným výsledkom kvality a presnosti zobrazenia skúmaných a predikovaných dát.

Záver

Cieľom mojej bakalárskej práce bolo využitie metód a nástrojov Business Intelligence (BI) pre podporu rozhodovania vo firemných procesoch spoločnosti XYZ, so zameraním na dolovanie dát (datamining).

Spoločnosťou poskytnutú databázu som upravil pomocou nástrojov BI (kontingenčné tabuľky) do podoby vyhovujúcej pre ďalšie analýzy a tvorbu predikčných a štatistických modelov.

Na analýzu a tvorbu modelov som zvolil program RapidMiner Studio, pomocou ktorého som dáta ďalej upravoval a formoval do vyhovujúcej podoby. Spojením už upravených dát formou operátorov som vytvoril tri funkčné modely. Model lineárnej regresie, model rozhodovacieho stromu a model zhlukovej analýzy. Každý z nich plnil inú predikčnú alebo štatistickú úlohu a to vzhľadom na sledovanú oblasť dát. Tým som demonštroval spôsob, akým možno využiť dáta v rámci procesu dolovania dát.

Na úplný záver môžem napísať, že všetky vytvorené návrhy modelov majú reálne uplatnenie, nakoľko ich výstupné hodnoty a grafy predstavujú kvalitný podklad pre ďalšie rozšírenie analýz alebo pre firemný management ako podpora rozhodovania.

Zoznam použitej literatúry

(1) FOTR, Jiří. Tvorba strategie a strategické plánování: teorie a praxe. Praha: Grada, 2012. Expert (Grada). ISBN 978-80-247-3985-4.

(2) Lineárna regresia: Algoritmy - Učenie s učiteľom [online]. [cit. 2020-05-16]. Dostupné z: https://smnd.sk/mcibula/alg/linreg.html

(3) Linear regression analysis study. CURRICULUM IN CARDIOLOGY - STATISTICS [online]. 2018, , 33-36 [cit. 2020-05-16]. DOI: 10.4103/jpcs.jpcs_8_18. Dostupné z: http://www.j-pcs.org/text.asp?2018/4/1/33/231939

(4) Diskriminační analýza DA [online]. [cit. 2020-05-16]. Dostupné z: https://meloun.upce.cz/docs/research/chemometrics/methodology/4da.pdf

(5) NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ. Business intelligence: jak využít bohatství ve vašich datech. Praha: Grada, 2005. Management v informační společnosti. ISBN 80-247-1094-3.

(6) Exploračná faktorová analýza [online]. [cit. 2020-05-16]. Dostupné z: https://statistikapspp.sk/exploracna-faktorova-analyza/

(7) Introduction to Bayesian Networks [online]. [cit. 2020-05-16]. Dostupné z: https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e

(8) Bayesian network. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2020 [cit. 2020-05-16]. Dostupné z: https://en.wikipedia.org/wiki/Bayesian$_n etwork$

(9) LABERGE, Robert. Datové sklady: agilní metody a business intelligence. Brno: Computer Press, 2012. ISBN 978-802-5137-291.

(10) Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python R) [online]. [cit. 2020-05-16]. Dostupné z: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

(11) Self-organizing map. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2020 [cit. 2020-05-16]. Dostupné z: https://en.wikipedia.org/wiki/Self-organizing$_map$

(12) Self Organizing Maps [online]. [cit. 2020-05-16]. Dostupné z: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4

(13) RUD, Olivia Parr. Data Mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001. Databáze. ISBN 80-722-6577-6.

(14) RapidMiner Studio: Visual workflow designer for the entire analytics team [online]. [cit. 2020-05-16]. Dostupné z: https://rapidminer.com/products/studio/

Zoznam použitých obrázkov

Prílohy