EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA PODNIKOVÉHO MANAŽMENTU

Evidenčné číslo: 104006/B/2019/36114651036800516

APLIKÁCIA DATA MININGU V PODNIKOVEJ PRAXI

Bakalárska práca

2019 **Andrej Macko**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA PODNIKOVÉHO MANAŽMENTU

APLIKÁCIA DATA MININGU V PODNIKOVEJ PRAXI

Bakalárska práca

Študijný program: ekonomika a manažment podniku

Študijný odbor: ekonomika a manažment podniku

Školiace pracovisko: Katedra informačného manažmentu

Vedúci záverečnej práce: doc. Ing. Anna Hamranová, PhD.

Bratislava 2019 **Andrej Macko**

Poďakovanie

Chcel by som sa poďakovať pani doc. Ing. Anne Hamranovej, PhD. za odbornú pomoc a usmernenie pri písaní bakalárskej práce, za cenné rady a informácie a v neposlednom rade za ochotu.

Abstrakt

MACKO, Andrej: *Aplikácia data miningu v podnikovej praxi.* - Ekonomická univerzita v Bratislave, Fakulta podnikového manažmentu, Katedra informačného manažmentu. - Vedúci záverečnej práce: doc. Ing. Anna Hamranová PhD. - Bratislava: FPM EU, 2019, 46 s.

Cieľom záverečnej práce je vysvetlenie základných pojmov, postupov a metód v oblasti hĺbkovej analýzy dát a následná analýza údajov z prostredia verejnej autobusovej dopravy prostredníctvom vybraných metód. Práca je rozdelená do troch kapitol. V prvej kapitole sú objasnené teoretické východiská potrebné na správnu aplikáciu hĺbkovej analýzy dát. Druhá kapitola je venovaná cieľu, metódam spracovania práce a podrobne konkrétnym dvom metódam analýzy dát a to klasifikácii pomocou rozhodovacieho stromu a zhlukovej analýze, pomocou ktorých je vybraný dátový súbor analyzovaný. Záverečná kapitola pozostáva z využitia metód popísaných v druhej kapitole. Výsledkom práce sú odporúčania pre podniky pôsobiace na Slovensku v odvetví dopravy získané na základe analýzy nami vybraného dátového súboru, Pri vizualizácii a aplikácii vybraných metód na dátový súbor bol použitý verejne dostupný softvér ako Power BI a WEKA.

Kľúčové slová: data mining, verejná autobusová doprava, rozhodovací strom, zhluková analýza

Abstract

MACKO, Andrej: *Application of Data mining in business practise.* - Economic university in Bratislava, Faculty of Business Management, Department of Information Management, Supervisor: doc. Ing. Anna Hamranová PhD. - Bratislava: FPM EU, 2019, 46 p.

The aim of this thesis is an explanation of basic concepts, procedures and methods of the data mining and an analysis of the public bus transport data through the selected methods. The thesis is divided to the three main chapters. In the first chapter, theoretical backround which is needed for right application of data mining is explained. Second chapter is dedicated to the description of the main aim of the thesis and the two methods which are later used for the data analysis of the dataset: Decision Tree and Cluster Analysis. Last chapter consists of using the methods described in the second chapter. The result of the thesis is to point out through analysis of our dataset on the possible recommendations for slovak companies in the transport segment. For the vizualization and applications of methods on the dataset two open-source software have been used including Power BI and WEKA.

Key words: Data Mining, Public Bus Transport, Decision Tree, Cluser Analysis

1 Obsah

Úvod

V súčasnej modernej spoločnosti predstavujú informácie a predovšetkým poznatky z nich plynúce často hranicu medzi úspechom a neúspechom. Využívaním technológií umožňujúcich uskladnenie veľkých objemov dát a rozsiahle možnosti internetového spojenia sa tieto kvantá údajov stali postupom času dostupnými.

Podniky, ktoré sú schopné premeniť tieto dáta na informácie a znalosti, ich môžu využiť k rýchlejšiemu a efektívnejšiemu rozhodovaniu a tým dosiahnuť konkurenčnú výhodu. Analýza verejne dostupných informácií umožňujúcich vývoj lepších a inovačných služieb pre obyvateľov sa dá taktiež vykonať aj vo verejnej doprave. Názov metódy, ktorá nám umožňuje realizovať danú analýzu je dolovanie dát alebo tiež data mining.

Cieľom práce je sprostredkovať stručný a ucelený pohľad na problematiku hĺbkovej analýzy údajov prostredníctvom riešenia problémov týkajúcich sa verejnej autobusovej dopravy pomocou praktického nasadenia príslušného softvéru na relevantné dáta z daného odvetvia. Ďalším cieľom je poukázať na dôležitosť zberu, uchovávania a následnej analýzy dát v konkurenčnom prostredí.

Práca je rozdelená do troch kapitol. V prvej kapitole sa venujeme teoretickému vymedzeniu problému spojeného s meškaním autobusov. Druhá kapitola obsahuje metódy, ktoré budú slúžiť ako základ pre analýzu samotných dát a v tretej kapitole si priblížime spôsob získania a upravenia konkrétneho dátového súboru a následne ho využijeme na jeho vizualizáciu. Ďalej sa oboznámime s pracovným prostredím aplikácie Weka, čo následne vyústi do samotnej analýzy nami získaných dát prostredníctvom využitia metód popísaných v druhej kapitole. V závere poukážeme prostredníctvom jednotlivých výsledkov práce na užitočnosť analýzy dát pre slovenské podniky v odvetví dopravy a poskytneme odporúčania na jednotlivé typy dát, ktoré by mohli byť zdrojom cenných informácii pre spoločnosti v tejto oblasti.

1 Súčasný stav riešenej problematiky doma a v zahraničí

V súčasnosti je využívanie informačných systémov samozrejmosťou v takmer každom odvetví, čo má za následok aj vznik a uskladňovanie čoraz väčšieho množstva dát a taktiež vznik mnohých nových oblastí súvisiacich s prácou s dátami a ich analýzou. Jedna z najrozšírenejších a najviac využívaných oblastí je hĺbková analýza dát - data mining (DM).

Témou bakalárskej práce je problematika využitia hĺbkovej analýzy dát, ktorej význam a uplatnenie sa v súčasnom modernom svete každým rokom zväčšuje. V tejto kapitole sa zameriame na priblíženie a spresnenie pojmov týkajúcich sa DM, ktorých znalosť je nutnou podmienkou správnej a úspešnej aplikácií samotnej ťažby dát.

1.1 Business Intelligence

Pojem „Business Intelligence“ (BI) je definovaný ako poskytovanie informácii a znalostí užívateľom pre rozhodovanie využitím rôznych zdrojov dát medzi ktoré patria štruktúrované ale aj neštruktúrované dáta. Informácie a dáta môžu byť uložené vo vnútri podniku ale aj mimo neho, môžu byť získané z rôznych zdrojov a štruktúrované viacerými spôsobmi a taktiež môžu byť buď kvantitatívne alebo kvalitatívne.[1]

BI je súbor procesov, aplikácii a technológií, ktorých cieľom je účinne a účelne podporovať rozhodovacie procesy vo firme. Podporujú analytické a plánovacie činnosti podnikov a organizácii a sú postavené na princípoch multidimenzionálnych pohľadov na dáta podniku. Aplikácie BI pokrývajú analytické a plánovacie funkcie väčšiny oblastí podnikového riadenia. To znamená predaja, nákupu, marketingu, finančného riadenia, kontroly, majetku, riadenia ľudských zdrojov a výroby.[2]

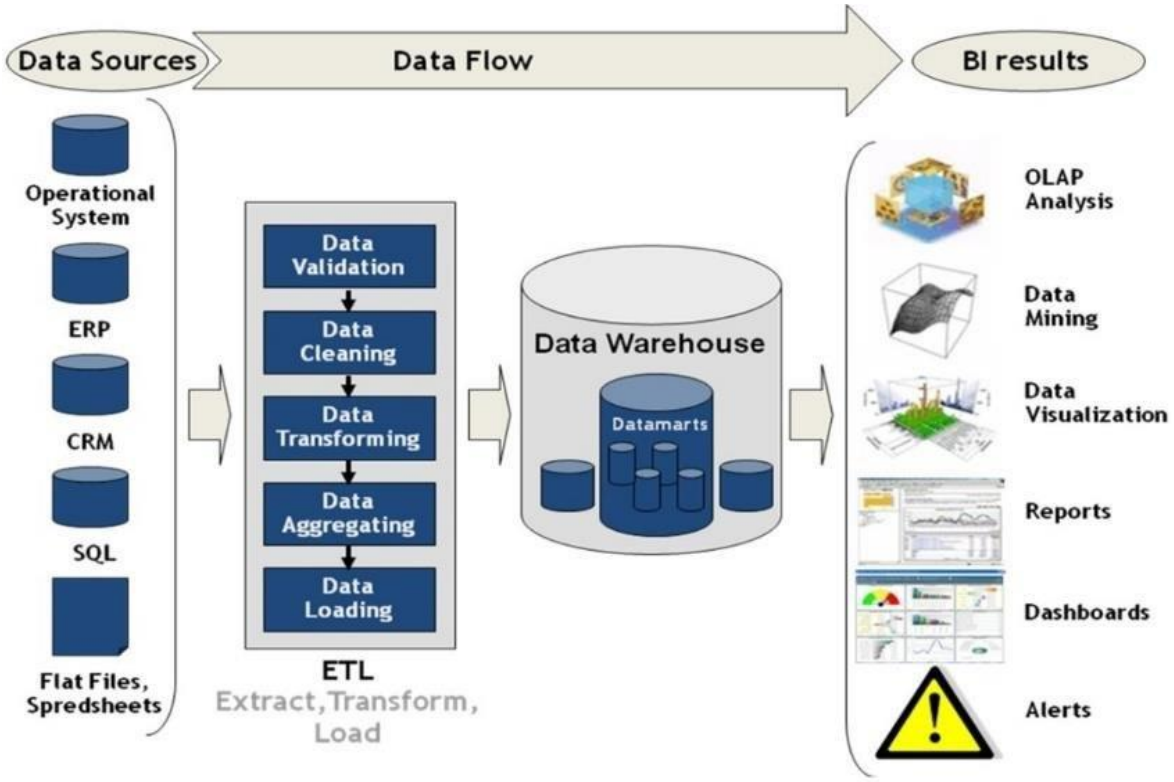
Hlavný výsledný výstup je znalosť, ktorá umožňuje samotné rozhodovanie, pričom vstupmi sú informácie a dáta. BI teda využíva dáta, ktoré môžu byť interné alebo externé a získané z rozličných zdrojov vrátane dátového skladu, a informácie, ktoré sú vytvárané pomocou vhodných analýz a následne prezentované v jednoduchej podobe vo forme zápisov a zoznamov.

[1] SABHERWAL, Rajiv - BECERRA-FERNANDEZ, Irma. *Business Inteligence: Practices, Technologies, and Management*. 2010. s. 6. ISBN 978-0470461709.
[2] SLÁNSKÝ, Dávid - POUR, Jan - NOVOTNÝ, Ota. *Business Intelligence: Jak využít bohatství ve vašich datech.* 2004. s. 19. ISBN 9788024710945.

Pojem BI môže byť využívaný dvoma rôznymi spôsobmi. V určitých prípadoch je používaný na označenie samotného výsledku tohto procesu, respektíve informácie alebo znalosti ktoré sú pre podniky užitočné pre ich podnikové aktivity a rozhodovanie. V iných prípadoch je tento pojem využívaný ako odkaz na proces, ktorým organizácia získava, analyzuje distribuuje dané informácie a znalosti.[3]

Pre podnikanie je kriticky dôležité lepšie pochopiť komerčnú stránku organizácie, ako sú zákazníci, trh, obstarávanie a zdroje a taktiež konkurenciu. BI sprostredkováva historický, súčasný ale taktiež predpovedá budúci pohľad na činnosti podniku.

Techniky klasifikácie a predikcie sú základom prediktívnej analýzy v BI., pre ktorú existuje mnoho využití v analýze trhov, obstarávania a predaja. Taktiež zhlukovanie má výraznú úlohu v riadení vzťahov so zákazníkmi, keďže vytvára skupiny zákazníkov založené na ich podobnostiach. Využitím techník charakterizácie môžeme lepšie pochopiť vlastnosti jednotlivých skupín a vyvíjať zákaznícke programy odmeňovania.[4]



Obrázok 1: Schéma BI
Zdroj: https://www.researchgate.net/figure/Scheme-BI-3-Business-Intelligence-Business-Intelligence-Business-Intelligence-is_fig9_284353121

[3] SABHERWAL, Rajiv - BECERRA-FERNANDEZ, Irma. *Business Inteligence: Practices, Technologies, and Management*. 2010. s. 6. ISBN 978-0470461709.
[4] HAN, Jiawei - KAMBER, Micheline – PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 27. ISBN 9780123814807.

Súčasti BI

Hlavnými súčasťami BI sú zdrojové dáta, dátové sklady pomocou ktorých sú uskladňované a konkrétne nástroje na prácu s nimi vrátane DM (Obrázok 1). .

OLTP (On Line Transaction Processing) systémy

OLTP systémy slúžia na realizáciu podnikateľských a ostatných transakcii v podniku. Sú uložené najmä v relačných databázach, zobrazujú aktuálny stav podniku a v priebehu jedného dňa sa môžu niekoľkokrát meniť. Príkladom môžu byť napríklad dáta v účtovníctve alebo v dokumentoch obchodných prípadov. Transakčné systémy realizujú ich spracovanie v reálnom čase a označujú sa ako OLTP (On Line Transaction Processing) systémy. Vzhľadom k analytickým aplikáciám sa dáta OLTP systémov chápu ako primárne, zdrojové alebo produkčné.[5]

Dátové sklady

„Dátový sklad je kolekciou integrovaných, predmetovo orientovaných databáz dizajnovaných na zlepšenie funkcií na podporu rozhodovania, kde každá zložka dát je relevantná pre nejaký moment v čase“.

Na základe tejto definície môžeme dátový sklad považovať za podnikové úložisko dát, ktoré je založené na podporovaní strategických rozhodnutí podniku. Funkciou dátového skladu je teda uchovávať historické dáta podniku integrovaným spôsobom, ktorý reprezentuje viaceré stránky podniku a jeho podnikateľskej činnosti. Dáta v dátových skladoch sa neaktualizujú, ale používajú sa na zodpovedanie dotazov konečných používateľov, ktorí sa považujú za tvorcov rozhodnutí. Dátové sklady sú väčšinou veľké, pričom môžu uschovávať miliardy záznamov.[6]

Dva najpodstatnejšie aspekty pre lepšie pochopenie dizajnového procesu dátového skladu sú špecifické typy dát uložených v dátovom sklade a taktiež súbor transformácií používaných na spracovanie dát do finálnej podoby tak, aby boli užitočné pre podporu rozhodovania.

[5] SLÁNSKÝ, Dávid - POUR, Jan - NOVOTNÝ, Ota. *Business Intelligence: Jak využít bohatství ve vašich datech.* 2004. s. 20, 21. ISBN 9788024710945.
[6] KANTARDZIC, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. 2011. s. 14. ISBN 978-0-470-89045-5.

Dátový sklad obsahuje nasledujúce kategórie dát, pričom ich klasifikácia je pričlenená k časovo závislým zdrojom dát:

1. Staršie podrobné dáta
2. Momentálne podrobné dáta
3. Mierne sumarizované dáta
4. Vysoko sumarizované dáta
5. Meta dáta

Na prípravu zdrojových elementárnych alebo derivovaných dát do dátového skladu používame štyri základné druhy transformácie dát, pričom každá z nich má svoje vlastné charakteristiky:

1. Jednoduché transformácie - tieto premeny dát sú základným prvkom pre ostatné, viac komplexné transformácie. Táto kategória zahŕňa manipuláciu s dátami, ktoré sú zamerané na jednu oblasť v čase, bez toho aby brali do úvahy ich hodnoty v iných oblastiach.
2. Čistenie dát - premeny zaručujúce konzistentné formátovanie a využitie určitej oblasti alebo príbuzných skupín oblastí. Môže zahŕňať napríklad formátovanie informácie o adrese.
3. Integrácia - je to proces mapovania operačných dát z jedného alebo viacerých zdrojov na novú dátovú štruktúru v dátovom sklade.

V mnohých prípadoch môžu mať podniky niekoľko lokálnych dátových skladov budovaných pre jednotlivé odbory, ktoré sa nazývajú aj „data marts“ (dátové trhoviská). Dátové trhovisko je dátový sklad, ktorý bol navrhnutý s cieľom uspokojiť potreby konkrétnej skupiny používateľov a jeho veľkosť závisí od rozsahu danej skupiny.[7]

Podstatou dátových trhovísk sú decentralizované dátové sklady, ktoré sa postupne integrujú do celopodnikového riadenia. V niektorých prípadoch dátové trhoviská slúžia aj po vytvorení celopodnikového dátového skladu ako určitý medzistupeň pri transformácii dát z produkčných databáz.

Dátové trhoviská boli pôvodne orientované najmä na oblasť marketingu pre podnikový archív, ktorý obsahuje všetky informácie spojené s novými, respektíve potenciálnymi zákazníkmi vo forme databázy, ktorá bola kompletne zameraná na

[7] KANTARDZIC, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. 2011. s. 14, 15. ISBN 978-0-470-89045-5.

spravovanie vzťahov so zákazníkmi. Napriek tomu, že dátové trhoviská sú vo všeobecnosti považované za podmnožinu dátového skladu, môžu byť vytvorené aj v prípade, že v podniku nie je žiadny integrovaný dátový systém. Zriadenie tematických štruktúr ako sú dátové trhoviská reprezentuje prvý a základný krok pre dosiahnutie informatívneho prostredia potrebného pre proces data miningu.[8]

OLAP systémy

Na druhej strane systémy pracujúce s analytickými informáciami využívajú primárne dáta vytvorené v OLTP systémoch. Pre svoje uloženie a operáciu s dátami sa pre tieto systémy zaviedol v osemdesiatych rokoch minulého storočia názov OLAP (On Line Analytical Processing). So zavedením pojmu „Business Inteligence“ (ktorý vo svojej podstate kopíruje vyššie zmienený význam výrazu OLAP) a súčasne s rozvojom nástrojov a technológii pre podporu analytických činností v organizácii sa však význam OLAP v určitej miere zúžil.

Užší význam definuje OLAP čisto technologicky, teda ako „informačnú technológiu založenú predovšetkým na koncepcii multidimenzionálnych databáz“. Ich hlavným princípom multidimenzionálna tabuľka umožňujúca rýchlo a pružne meniť jednotlivé dimenzie, a meniť tak pohľady užívateľa na modelovanú ekonomickú realitu.[9]

Porovnanie systémov OLTP a OLAP

Požiadavka pohľadu na dáta z viacerých hľadísk so sebou prináša aj požiadavku na optimalizované fyzické ukladanie dát, pričom väčšinou ide o historické, agregované, priebežne rozširované a ukladané dáta v jednoduchej štruktúre vhodnej pre analýzu a prispôsobené potrebám manažmentu.

Slánský, Pour a Novotný uvádzajú tieto základné rozdielové charakteristiky medzi systémami OLTP a OLAP:

- OLTP systémy sú primárne určené na získavanie dát. Tomu zodpovedá ich celková architektúra, a hlavne následne databázový model. Ten je charakteristický tabuľkami v tretej normálnej forme, veľkým počtom tabuliek a ich spojenie, snahou o nulovú redundanciu dát. Súčasne sú tabuľky indexované len v najnutnejších prípadoch.

[8] WANG, John. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. 2008. s. 11. ISBN 978-1599049526.
[9] SLÁNSKÝ, Dávid - POUR, Jan - NOVOTNÝ, Ota. *Business Intelligence: Jak využít bohatství ve vašich datech.* 2004. s. 21, 33. ISBN 9788024710945.

- Dáta sú do OLTP získavané v reálnom čase, v bežnej prevádzke prebiehajú desiatky až státisíce transakcii za minútu. Systémy sú tak zaťažované kontinuálne.
- Analytické systémy sú oproti OLTP určené primárne pre podporu dotazov. Z toho vyplýva ich architektúra a databázové modely, vyznačujúce sa zmenšeným počtom takzvaných denormalizovaných tabuliek, vyššou frekvenciou indexov, duplicitou uloženia dát a podobne.
- Drvivá väčšina analytických systémov aktualizuje svoje dáta periodicky (najčastejšie v denných a mesačných intervaloch). Najnovšie trendy síce umožňujú aj aktualizáciu dát v reálnom čase, avšak v dnešnej dobe sa zatiaľ jedná o výnimočné prípady. Zaťažovanie analytických systémov je z tohto dôvodu nárazové - veľké zaťaženie je typické pre obdobie nahrávania dát, a následne je možné pozorovať nepravidelnú záťaž podľa frekvencie a zložitosti jednotlivých analytických úloh, prebiehajúcimi nad analytickými systémami.

Kým OLTP systémy udržujú svoje dáta pri maximálnej úrovni detailov (teda na úrovni jednej transakcie so všetkými jej detailnými atribútmi), analytické systémy ukladajú len dáta relevantné pre analýzy, teda buď agregované na vyššej úrovni než jednotlivá transakcia, alebo zahrňujúce len niektoré jej atribúty.[10]

1.2 Data Mining

Data Mining (DM) je proces objavovania nových, užitočných korelácii, vzorov a trendov pomocou výberu z veľkého množstva dát nachádzajúcich sa v úložiskách využitím technológii na rozpoznávanie vzorov ako aj matematických a štatistických metód. Zdroje dát môžu obsahovať databázy, dátové sklady, prípadne aj dáta ktoré sú dynamicky vysielané do systému.[11]

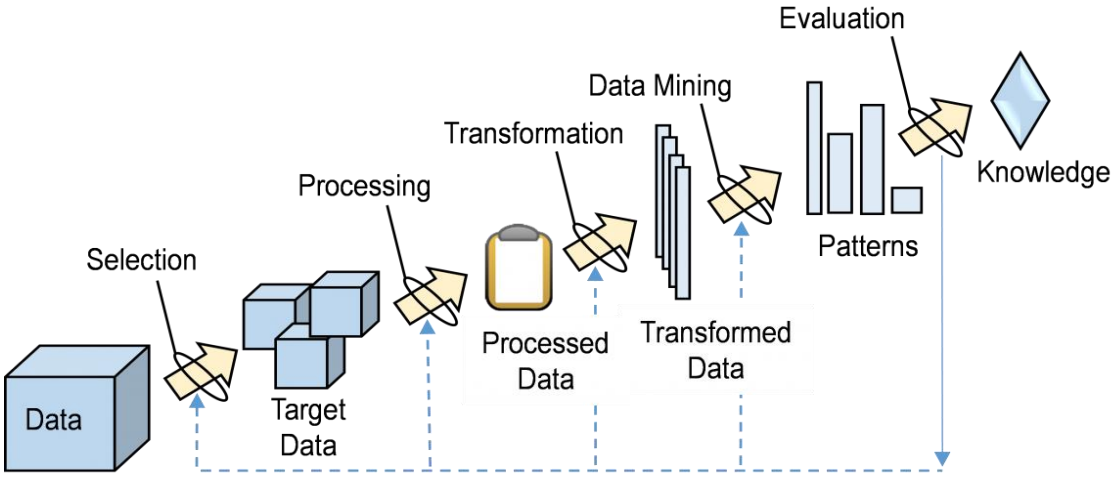
Napriek veľkému množstvu dát ktoré je v súčasnosti generované prostredníctvom informačných systémov a technológii, veľká časť z nich ešte nebola zozbieraná alebo analyzovaná. Avšak prostredníctvom DM môžu byť existujúce dáta roztriedené a informácie plynúce z nich môžu byť využité na ich plný potenciál.

[10] SLÁNSKÝ, Dávid - POUR, Jan - NOVOTNÝ, Ota. *Business Intelligence: Jak využít bohatství ve vašich datech.* 2004. s. 25, 26. ISBN 9788024710945.
[11] T.LAROSE, Daniel - D.LAROSE, Chantal. *Discovering Knowledge in Data: An Introduction to Data Mining*. 2. vyd. 2014. s. 1. ISBN 978-0-470-90874-7.

1.2.1 *Knowledge Discovery from Data (KDD)*

DM je častokrát vnímaný ako synonymum pre ďalší veľmi často používaný pojem a to „knowledge discovery from data“. KDD je proces hľadania znalostí v dátach, ktorý sa uskutočňuje pomocou využívania metód a algoritmov DM s cieľom extrahovať požadované poznatky z veľkého množstva dát.

Typický proces modelovania KDD zahŕňa metodológie na vyťahovanie a prípravu dát ako aj na tvorbu rozhodnutí po uskutočnení samotného procesu DM.



Obrázok 2: Proces hĺbkovej analýzy dát
Zdroj: https://behavior.lbl.gov/?q=node/11.

Tento proces pozostáva z nasledujúcich krokov:

1. Selekcia dát - výber dát podstatných pre cieľ analýzy z databázy.
2. Spracovanie dát - odstraňovanie nejasných a nekonzistentných dát a kombinovanie viacerých zdrojov dát.
3. Transformácia dát - dáta sú konsolidované do podoby vyhovujúcej ťažbe dát.
4. Data mining - samotná ťažba dát prostredníctvom metód a techník DM.
5. Hodnotenie dát - ohodnotenie vzorov dát.
6. Prezentácia dát - vizualizácia a využitie techník reprezentácie dát na prezentovanie získaných dát používateľom.[12]

1.2.2 *Základné ciele a úlohy data miningu*

V praxi sa medzi základné 2 ciele DM považuje predikcia a deskriptíva, na základe čoho rozlišujeme prediktívnu ťažbu dát, ktorá produkuje model systému opísaný dodanými

[12] HAN, Jiawei - KAMBER, Micheline – PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 7. ISBN 9780123814807.

dátami, a deskriptívnu ťažbu dát, ktorá produkuje nové a netriviálne informácie založené na súbore údajov ktorý máme k dispozícii.

Ciele deskripcie a predikcie sú dosahované pre nasledujúce úlohy DM:

1. Klasifikácia - začlenenie dát do niekoľkých preddefinovaných tried.
2. Regresia - získanie funkcie, ktorá mapuje dátovú položku na predikčnú premennú so skutočnou hodnotou.
3. Zhlukovanie - identifikovanie konečných kategórii a skupín na opis dát.
4. Sumarizácia - nájdenie jednotného opisu pre skupiny dát.
5. Modelovanie závislostí - nájdenie lokálneho modelu, ktorý opisuje významné závislosti medzi premennými prípadne medzi hodnotami vlastností dát v dátových súboroch alebo časti dátových súborov.
6. Detekcia zmeny - objavovanie najväčších a najvýznamnejších zmien v dátových súboroch.[13]

Typy dát vhodné na data mining

DM môže byť aplikovaný na ľubovoľný typ dát pokiaľ dané dáta majú význam pre cieľovú aplikáciu. Najčastejšie formy dát pre DM sú databázové dáta, dáta v dátových skladoch a transakčné dáta. Samotný DM však môže byť aplikovaný aj na iné typy dát ako napríklad dátové toky, grafické a sieťové dáta, priestorové dáta, ale aj textové a multimediálne dáta.

1.2.3 Vývoj data miningu

Od začiatku využívania prvkov data miningu v 60-tych rokoch sa informačné technológie systematicky vyvinuli z primitívnych systémov na spracovanie dát na sofistikované a výkonné databázové systémy. Výskum a vývoj v databázových systémoch od 70-tych rokov pokročil z hierarchických a sieťových databázových systémov na relačné databázové systémy (kde sú dáta skladovane v relačných tabuľkových štruktúrach), nástroje na modelovanie dát a metódy indexovania a prístupu. Používatelia navyše získali pohodlný a flexibilný prístup k dátam prostredníctvom dopytovacích jazykov, užívateľských rozhraní a manažmentu transakcií.

Ťažba dát je v súčasnosti považovaná za jeden z kritických krokov k podnikateľskému úspechu. Jej využívanie bolo dokonca prítomné aj v druhej svetovej vojne, kde metódy dátovej analýzy boli používané na vojenské záležitosti a demografické

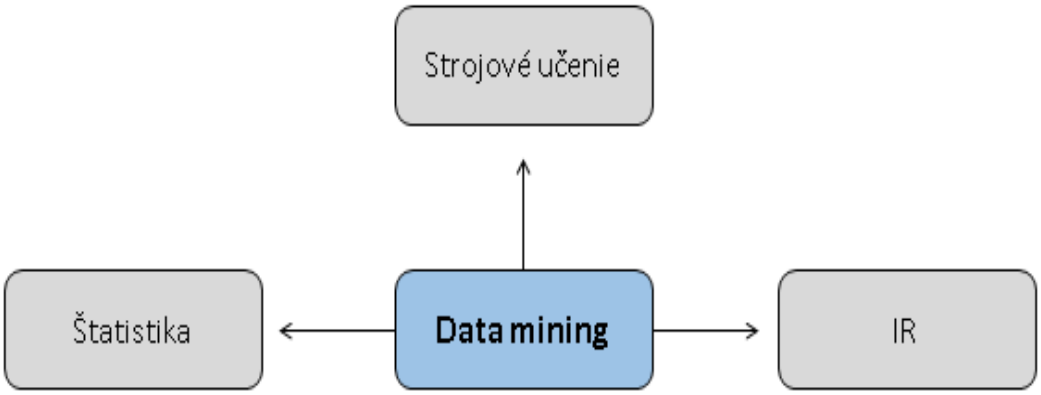
[13] KANTARDZIC, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. 2011. s. 3. ISBN 978-0-470-89045-5.

účely. Napriek tomu formálne techniky DM neboli propagované až do 90-tych rokov. Napriek tomu že dolovanie dát bolo v histórii známe pod viacerými označeniami ako napríklad „knowledge extraction“, „information discovery“, „data archalogy“ a „data pattern processing“, označenie „data mining“ je v súčasnosti najčastejšie používaný termín štatistami a dátovými analytikmi.[14]

Po vzniku systémov na riadenie databáz sa databázové technológie posunuli smerom k vývoju pokročilejších databázových systémov, skladovaniu dát ale taktiež k data miningu pre pokročilé dátové analýzy a webovým dátam. Tieto systémy zahŕňali nové a výkonnejšie dátové modely ako rozšírené relačné, objektovo orientované, objektovo relačné a deduktívne modely.[15]

1.3 Techniky využívané v DM

Ako sféra riadená hlavne aplikáciami, DM zahrňuje množstvo techník z rôznych oblastí ako štatistika, strojové učenie, rozpoznávanie vzorov, techniky získavania informácii a množstvo iných aplikácii..



Obrázok 3: Techniky využívané v DM
Zdroj: Vlastné spracovanie podľa https://www.oreilly.com/library/view/data-mining-concepts/9780123814791/xhtml/ST0090.html

Charakter výskumu a vývoja DM sa významne podieľa na úspechu samotného DM a jeho rozsiahleho využívania. V tejto časti popíšeme najviac využívané techniky v DM uvedené vyššie.

[14] WANG, John. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. 2008. s. 7. ISBN 978-1599049526.
[15] HAN, Jiawei - KAMBER, Micheline – PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 20. ISBN 9780123814807.

1.3.1 *Štatistika*

Štatistický model je množina matematických funkcií, ktorá opisuje správanie objektov v cieľovej triede týkajúce sa náhodných premenných a rozdelenia pravdepodobnosti spojené s nimi. Štatistické modely sú vo všeobecnosti využívané na modelovanie dát a dátových tried.[16]

Môžeme ho využiť napríklad na modelovanie nejasných a chýbajúcich dát a následne pri ťažbe dát vo veľkých súboroch môže proces ťažby dát využiť tento model na pomoc pri identifikácii a spracovaní nejasných prípadne chýbajúcich hodnôt v dátach.

Štatistický prieskum rozvíja nástroje na predikciu a predpoveď použitím dátových a štatistických modelov. Štatistika je užitočná pre ťažbu rozličných vzorov z dát ako aj pre pochopenie základných mechanizmov vytvárajúcich a ovplyvňujúcich vzory dát. Prediktívna štatistika modeluje dáta spôsobom, ktorý zodpovedá za náhodnosť a nejasnosť v pozorovaniach a je používaný na vyvodzovanie záverov o skúmanom procese.

Štatistické metódy môžu byť taktiež využité na overenie výsledkov ťažby dát. Napríklad, po „vydolovaní“ klasifikačného alebo predikčného modelu by mal byť tento model overený štatistickou hypotézou. Test štatistickej hypotézy tvorí štatistické rozhodnutia použitím experimentálnych dát.

1.3.2 *Strojové učenie*

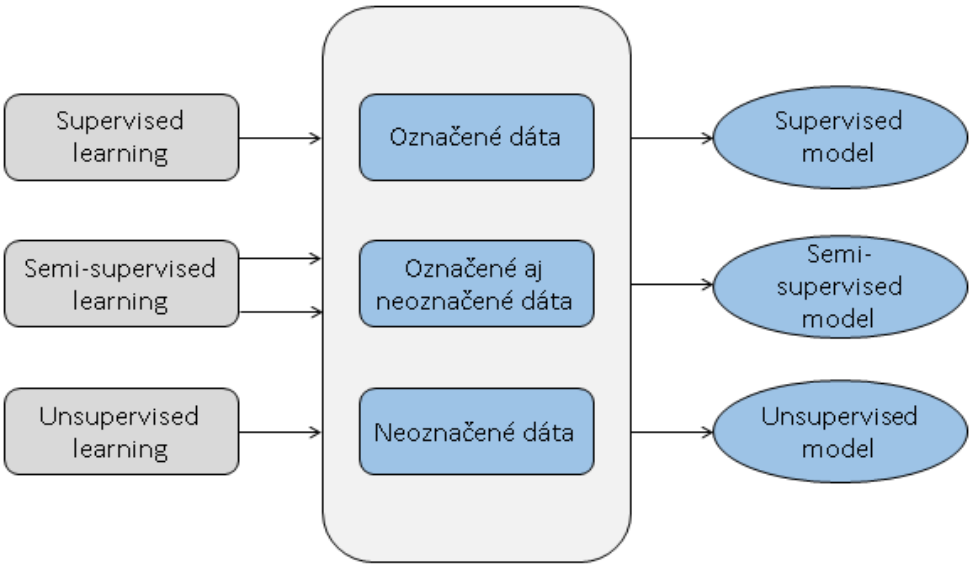
Strojové učenie skúma ako sa môžu počítače učiť (alebo zlepšiť svoju výkonnosť) využitím dát. Hlavná oblasť výskumu umožňuje počítačovým programom automatické učenie na rozpoznávanie komplexných vzorov a vytváranie inteligentných rozhodnutí na základe dát. Napríklad, typický problém strojového učenia je naprogramovanie počítača na automatické rozpoznávanie ručne písaných poštových kódov po učení sa z niekoľkých súborov príkladov.

Medzi základne typy strojového učenia patria:

- „Supervised learning“, respektíve učenie s dohľadom, je v podstate synonymum pre klasifikáciu. Dohľad v učení súvisí s označenými príkladmi v cvičných dátových súboroch. Napríklad pri probléme s rozpoznávaním poštových kódov, súbor obrázkov ručne písaných poštových kódov a ich korešpondujúce strojovo čitateľné preklady sú používané v testovacích príkladoch, ktoré dohliadajú na učenie klasifikačného modelu.

[16] HAN, Jiawei -KAMBER, Micheline - PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 23. ISBN 9780123814807.

- „Unsupervised learning“, učenie bez dohľadu, je naopak synonymum pre zhlukovanie. Proces učenia je bez dohľadu preto, že vstupné príklady nie sú triedovo označené. Typicky ho využívame na objavovanie tried v rámci údajov. Túto metódu môžeme využiť napríklad ak ako vstupné dáta použijeme súbor obrázkov ručne písaných číslic. Predpokladajme, že bude nájdených desať zhlukov dát. Tieto zhluky môžu zodpovedať desiatim číslicam od nula po desať. Avšak, keďže testovacie dáta nie sú označené, naučený model nám nedokáže objasniť význam týchto nájdených zhlukov.
- „Semi-supervised learning“ je skupina techník strojového učenia, ktorá využíva pri učení modelu označené, ale aj neoznačené vzorové príklady. Označené testovacie príklady môžu byť využité na učenie sa modelov tried a neoznačené príklady môžu byť využité na zmierňovanie hraníc medzi triedami.
- „Active learning“ je prístup strojového učenia, ktorý necháva samotných užívateľov zohrávať aktívnu úlohu v procese učenia. Tento prístup môže od používateľa vyžadovať označenie určitého príkladu, ktorý pochádza zo súboru neoznačených príkladov alebo bol syntetizovaný programom. Cieľom je zlepšiť kvalitu modelu aktívnym vyžadovaním vedomostí od používateľov, pričom počet vyžiadaní na označenie príkladov je obmedzený.[17]



Obrázok 4: Typy strojového učenia
Zdroj: Vlastné spracovanie podľa https://viblo.asia/p/an-introduction-to-generative-adversarial-networks-gans-a-semi-supervised-learning-3P0lPmqq5ox

[17] HAN, Jiawei - KAMBER, Micheline – PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 24, 25. ISBN 9780123814807.

1.3.3 Získavanie informácii (Information retrieval)

„Information retrieval“ (IR) je veda týkajúca sa hľadania dokumentov a informácii v dokumentoch a je taktiež jednou z techník, ktoré DM využíva (Obrázok 4). Dokumenty môžu byť textové, ale aj multimediálne a môžu sa vyskytovať na sieti.

Typické prístupy IR využívajú pravdepodobnostné modely. Napríklad, textový dokument môže byť považovaný balík slov, respektíve súbor rôznych slov vyskytujúcich sa v dokumente. Jazykový model dokumentu je funkcia hustoty pravdepodobnosti, ktorá generuje súbor slov v dokumente. Podobnosti dvoch dokumentov môžu byť merané podobnosťou medzi ich jazykovými modelmi.[18]

1.4 Aplikácia dátovej analýzy v doprave

Manažment dopravy v súčasnosti vyžaduje analýzu veľkého množstva dát v reálnom čase za účelom poskytovať aktuálne informácie o stave premávky všetkým skupinám používateľov. Moderné autá sú vybavené niekoľkými senzormi, ktoré dokážu produkovať užitočné dáta na analýzu situácie v premávke.

Využívaním mobilných komunikačných technológii môžu byť takéto dáta integrované a agregované z veľkého množstva áut, čo umožňuje inteligentným dopravným systémom monitorovanie stavu premávky vo veľkej oblasti za relatívne nízku cenu.[19]

1.4.1 GPS dáta

Existujú mnohé štúdie, ktorých cieľom je riešenie problémov spojených s verejnou dopravou pomocou GPS dát z taxi služieb. Niektoré z nich na podporu a zvýšenie ziskovosti taxi služieb navrhli modely a implementovali algoritmy pomocou ktorých vedia určiť a nájsť najlepšie možné parkovacie miesto na ktorom čakať na potenciálnych zákazníkov, optimalizovať svoju cestu a tak zabezpečiť kvalitu poskytovaných služieb. Keďže GPS dáta taxi služieb reflektuje cestovné vzory ľudí, analýza reálnych GPS dát taxi služieb môže zlepšiť kvalitu dopravných služieb.[20]

[18] HAN, Jiawei - KAMBER, Micheline - PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. s. 26. ISBN 9780123814807.
[19] CERVONE, Guido - LIN, Jessica - WATERS, Nigel. *Data Mining for Geoinformatics*: Methods and Applications. 2014. s. 83. ISBN 978-1461476689.
[20] LUO, Xudong - XU YU, Jeffrey - LI, Zhi. *Advanced Data Mining Applications*: 10th International Conference. 2014. s. 295. ISBN 978-3-319-14717-8.

1.4.2 Oblasť autobusovej dopravy

V husto obývaných oblastiach je určitá miera časovej nestability hlavne na vysoko frekventovaných trasách. Na takýchto trasách je dôležitejšia pravidelnosť postupu ako plnenie času príchodu autobusu na zastávku. Malé meškanie autobusu má za následok väčší počet cestujúcich na ďalšej zastávke. Tento zvýšený počet nastupujúcich cestujúcich spôsobí predĺženie času nástupu a tým pádom aj ďalšie navýšenie meškania autobusu. Na druhej strane, ďalší autobus bude mať menej cestujúcich, kratší nástupný čas a žiadne meškanie. Toto spôsobí „efekt snehovej gule" a v určitom čase sa dané autobusy stretnú na jednej zo zastávok na danej trase. Tento fenomén má viacero pomenovaní: „Bangkok effect", „Bus Platooning", „Bus Clumping" alebo „Bus Bunching" (BB). V práci budeme používať posledné z nich (BB).

Výskyt BB núti kontrolórov podnikať kroky na zabránenie nestability postupu dopravnej prevádzky a dodržiavanie časového plánu. BB môže spôsobovať viacero problémov, ako napríklad ešte dlhšie meškanie, plné autobusy, znížený komfort v autobusoch, dlhšie čakanie na zastávkach a nižšej miery dôveryhodnosti časového rozvrhu.. Pomocou identifikácie vzorov túkajúcich sa BB síce nemôžeme jeho výskyt úplne vylúčiť, ale môžeme tieto informácie použiť ako podklad na zlepšenie časového rozvrhu. Cieľom teda nie je podobné udalosti eliminovať, ale zmierniť ich dopad.[21]

Aplikácia dátovej analýzy vo verejnej doprave je predmetom ďalších kapitol bakalárskej práce.

[21] PERNER, Petra. *Applications and Theoretical Aspects: 12th Industrial Conference*. 2012. s. 77, 78. ISBN 978-3642314872.

2 Cieľ práce, metodika práce a metódy skúmania

V druhej kapitole je špecifikovaný hlavný cieľ práce, čiastkové ciele potrebné na jeho dosiahnutie a metódy, ktoré boli použité na splnenie jednotlivých cieľov.

2.1 Hlavný a čiastkové ciele práce

Hlavným cieľom bakalárskej práce je využiť výsledky analýzy pri riadení spoločnosti a tým poukázať na možnú užitočnosť zberu a analýzy takýchto dát pre slovenský podnik SAD Lučenec. Keďže v dnešnej dobe je analýza dát jedným z najdôležitejších faktorov úspechu, je nevyhnutné aby sa tomuto trendu prispôsobili aj podniky, ktoré ešte v súčasnosti nemajú vybudovanú dostatočnú základňu na to, aby ich získavanie a analýza dát mohla prebiehať efektívne a bez znevýhodňovania daného podniku v konkurenčnom prostredí.

Čiastkové ciele

Na dosiahnutie hlavného cieľa boli naformulované nasledovné čiastkové ciele:

- Analýza riešenej problematiky v literatúre z oblasti BI a DM.
- Získanie dátového súboru, ktorý bude predstavovať predmet našej analýzy.
- Úprava dátového súboru a jeho vizualizácia v programe Power BI
- Podrobenie dátového súboru klasifikácii a zhlukovej analýze pomocou programu Weka
- Vyhodnotenie výsledkov analýz

2.2 Metodika práce a použité metódy

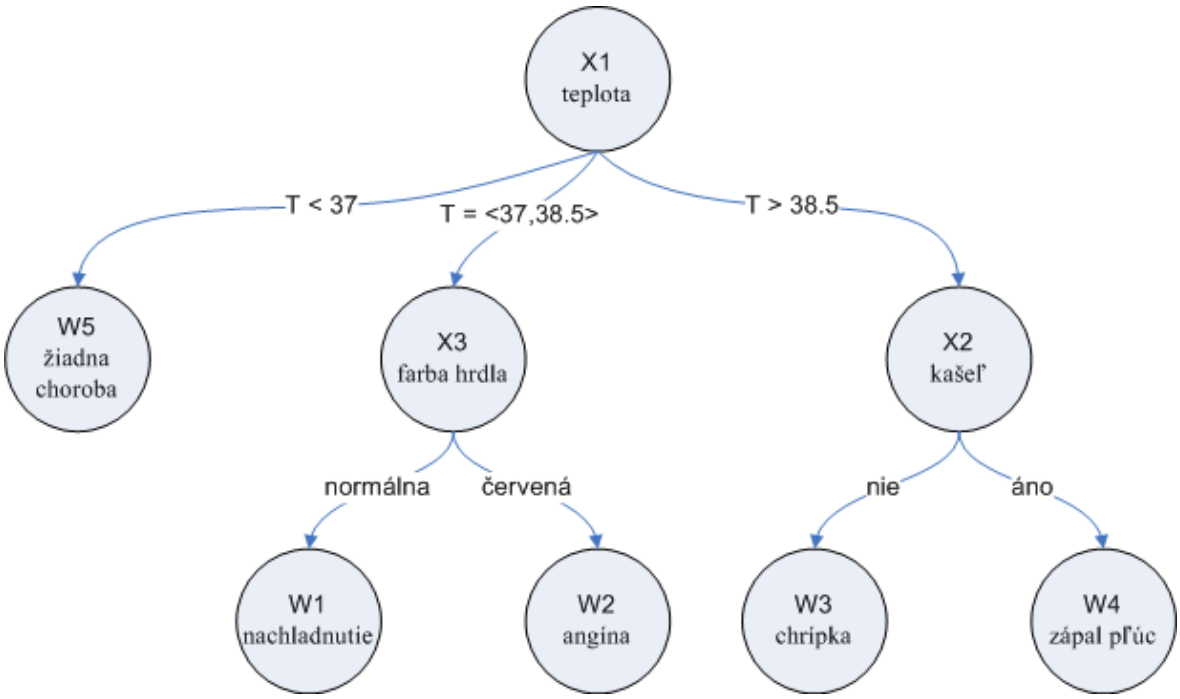
Pri analýze súčasného stavu riešenej problematiky boli použité štandardné metódy vedeckej práce: analýza, syntéza, indukcia a dedukcia.

V aplikačnej časti sme sa zamerali konkrétne na:

- hĺbkovú analýzu údajov a ich vizualizáciu pomocou programu Power BI
- dve dataminingové metódy: klasifikáciu prostredníctvom rozhodovacích stromov a zhlukovú analýzu.

2.2.1 *Rozhodovací strom*

Rozhodovacie stromy slúžia grafické vyobrazenie situácií, ktoré sú predmetom rozhodovacích analýz. Zobrazujú celý rozhodovací proces vrátane všetkých jeho etáp a kritérií na základe ktorých rozhodovanie prebieha.[22] V našom prípade bude takýto rozhodovací strom aj výsledkom klasifikačnej analýzy nášho dátového súboru. Jednoduchý vzor ako rozhodovací strom môže vyzerať tak, ako je znázornený na Obrázku 5, pričom kritériá sú znázornené písmenom X a výsledné hodnoty písmenom W.



Obrázok 5: Vzorový rozhodovací strom
Zdroj:< *http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0708/vano/index.html*>

2.2.2 *Zhluková analýza*

Zhluková analýza predstavuje rozklad súboru na niekoľko homogénnych podsúborov takým spôsobom, aby si boli štatistické jednotky v jednom zhluku čo najpodobnejšie a najbližšie a štatistické jednotky z rôznych zhlukov čo najodlišnejšie a najvzdialenejšie.

Tento rozklad súboru je uskutočňovaný využívaním zhlukovacích postupov a metód. Na základe výsledného systému postupov ich klasifikujeme na hierarchické a nehierarchické zhlukové postupy a metódy. Hierarchické postupy sa používajú ako začiatočné, respektíve prieskumné riešenia zhlukovania, ale výsledné riešenie je doplnené

[22] ROKACH, Lior - Z MAIMON, Oded. *Data Mining With Decision Trees: Theory and Applications*. 2014. s. 10. ISBN 978-9814590075.

nehierarchickým postupom. Na základe tohto môžeme hierarchické a nehierarchické postupy považovať za navzájom sa doplňujúce.[23]

Zhlukovú analýzu a jej algoritmy je možné v súčasnosti použiť vo viacerých oblastiach, ako napr.: marketing (hľadanie skupín zákazníkov s podobnými vlastnosťami (správaním) na základe vlastností zákazníkov obsiahnutých v databáze), knižnice (zoskupovanie kníh), poisťovanie (identifikácia poistných skupín, podvodov), web stránky (klasifikácia dokumentov) a mnohých ďalších.

[23] WU, Junjie. *Advances in K-Means Clustering.* 2014. s.3. ISBN 978-3642447570.

3 Výsledky práce a diskusia

Tretia časť bakalárskej práce pozostáva z aplikácie popísaných teoretických východísk na analýzu konkrétnych dát. Našim cieľom bude dané dáta vizualizovať, analyzovať a následne výsledok týchto analýz pochopiť a ohodnotiť ich využiteľnosť v podnikoch na Slovensku v oblasti verejnej dopravy Na vizualizáciu datasetu využijeme program Power BI a na analýzu nami získaných dát použijeme voľne dostupný analytický nástroj určený na hĺbkovú analýzu dát známy pod názvom Weka Open Source Machine Learning Software.[24]

Nami využívaný súbor dát bude pozostávať z dát týkajúcich sa zahraničnej verejnej mestskej autobusovej dopravy, konkrétnejšie dát získaných v meste New York City prostredníctvom systému na zaznamenávanie zlyhaní a omeškaní autobusov, ktorý je verejne prístupný a aktualizuje sa v reálnom čase. Hlavným dôvodom výberu práve tohto dátového súboru bolo to, že slovenskí dopravcovia sa na zber a uchovávanie dát sústreďujú len minimálne. Menšie dopravné spoločnosti nemajú dostatok zdrojov popri vykonávaní svojej činnosti na investície v tejto oblasti a väčšie buď preferujú iné možnosti využitia svojich zdrojov, prípadne uchovávané dáta nezverejňujú.

3.1 Charakteristika podniku SAD Lučenec

Inšpiráciou pre výber témy práce bola akciová spoločnosť SAD Lučenec, pôsobiaca práve v oblasti verejnej autobusovej dopravy. Akciová spoločnosť vznikla v roku 2002 transformáciou z pôvodného štátneho podniku SAD Lučenec.

V súčasnosti tento dopravca zabezpečuje výkon prímestskej dopravy v okresoch Lučenec, Rimavská Sobota, Revúca a Poltár, dve diaľkové linky a mestskú dopravu v Lučenci a Rimavskej Sobote. Spoločnosť zamestnáva viac ako 300 zamestnancov a prevádzkuje viac ako 200 autobusov. Tie v priebehu dňa zastavia v 243 obciach v okresoch Lučenec, Rimavská Sobota, Revúca, Poltár, Poprad, Rožňava, Veľký Krtíš, Detva, Zvolen a Banská Bystrica, najazdia priemerne 38855 kilometrov, zastavia na 887 zastávkach 38812 krát a odvezú 39 100 cestujúcich, z toho 14 500 tvoria deti.

Tento slovenský dopravca bohužiaľ taktiež nerealizuje zber dát v dostatočnom množstve a kvalite, preto ich nepodrobujú žiadnej hlbšej dátovej analýze. Tým pádom aj ich informácie a podklady, na základe ktorých realizujú strategické rozhodnutia vrátane

[24] FRANK, Eibe - HALL, Mark - WITTEN, Ian. *Data Mining: Practical Machine Learning Tools and Techniques*. 2016. [online].[cit. 8.4.2019]. Dostupné na internete: <https://www.cs.waikato.ac.nz/ml/weka/>

tvorby časového rozvrhu alebo analýzy stavu jednotlivých vozidiel nie sú dostatočne detailné a obsiahle. Práve preto by investícia v tejto oblasti a z nej plynúca analýza dát prostredníctvom autobusov by mohla predstavovať ďalší krok pri optimalizácii a koordinácii podnikových zdrojov spoločnosti.

3.2 Dátový súbor

Ako predmet na aplikáciu teoretických poznatkov sme si vybrali dátový súbor týkajúci sa verejnej mestskej autobusovej dopravy v meste New York City získaný systémom na zaznamenávanie zlyhaní a omeškaní autobusov, ktorý poskytuje bližšie informácie o jednotlivých meškaniach.

Použitím tohto dátového súboru sa pokúsime poukázať na výhody, ktoré môže jednotlivým spoločnostiam hĺbková analýza dát v konkurenčnom prostredí priniesť a ktoré by mohli slúžiť ako motivačný faktor pri rozhodovaní podniku o investícii v tejto oblasti.

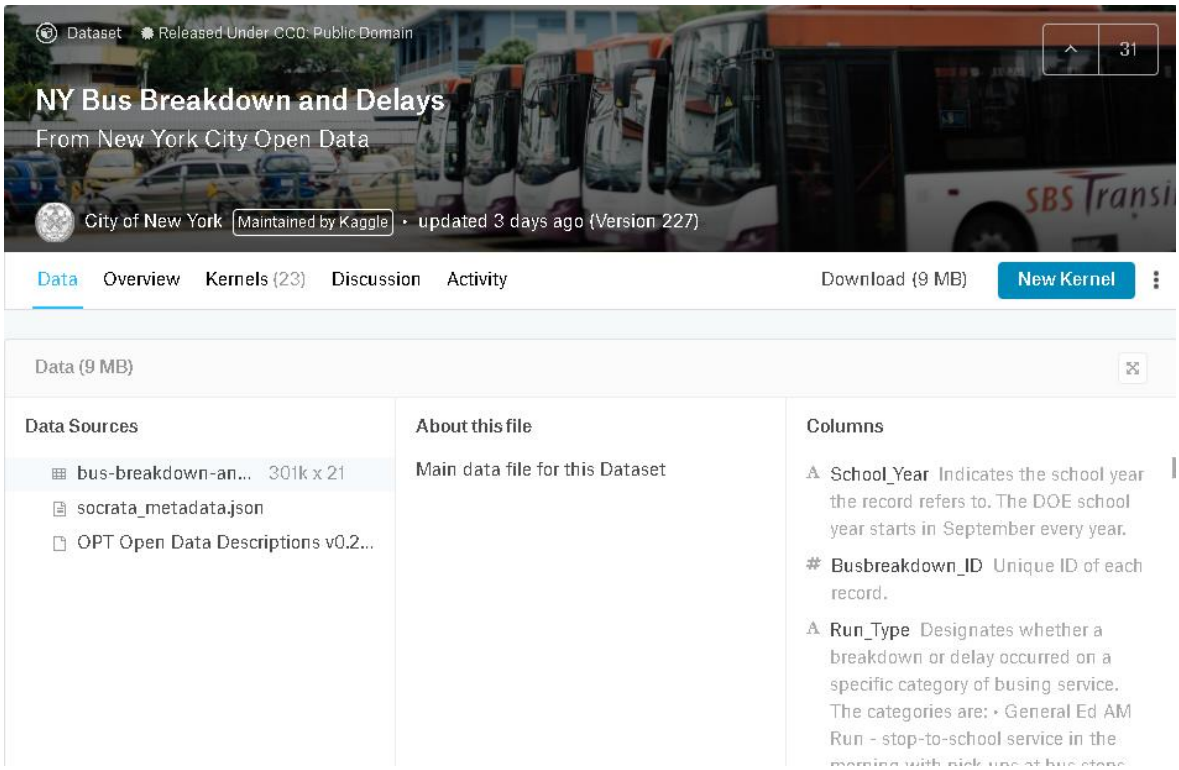
V práci sa konkrétne zameriame na šesť atribútov opisujúcich jednotlivé meškania, a to identifikačné číslo meškania, číslo autobusu, číslo trasy, štvrť v ktorej sa meškanie vyskytlo a dĺžku daného meškania a na základe analýzy ich hodnôt sa pokúsime poukázať na možnú mieru užitočnosti dát v prípade, že ich podnik má k dispozícii v dostatočnom množstve a kvalite. Tá by sa následne v prípade investície v tejto oblasti a následnému zberu a analýze dát mohla prejaviť najmä pri vytváraní a optimalizácii časového rozvrhu, v lepšej evidencii technického stavu vozidiel a v lepšej pripravenosti na výskyt kritických situácií.

Spôsob získania dátového súboru

Na získanie daného dátového súboru je potrebné najprv navštíviť webovú stránku platformy Kaggle <https://www.kaggle.com/new-york-city/ny-bus-breakdown-and-delays>, patriacu pod spoločnosť GoogleInc.. Kaggle predstavuje on-line komunitu dátových vedcov umožňujúcu používateľom vyhľadávať a publikovať dátové súbory.

Po načítaní hlavnej stránky je nevyhnutné sa na platformu zaregistrovať a následne prihlásiť pre možnosť stiahnutia dátových súborov. Po prihlásení si na hornej lište vyberieme zložku „datasets“. Následne pomocou vyhľadávača medzi dátovými súbormi nájdeme náš konkrétny dátový súbor týkajúci sa meškaní autobusov pod názvom „NY Bus Breakdown and Delays“ a otvoríme ho, čím sa dostaneme na vyššie uvedenú webovú stránku.

Prostredníctvom samotnej platformy Kaggle môžeme v prípade potreby do daného dátového súboru nahliadnuť a taktiež si pozrieť percentuálne zastúpenie hodnôt v rámci jednotlivých atribútov dátového súboru (Obrázok 6). Bližšie informácie slúžiace na vysvetlenie atribútov dátového súboru nájdeme pod zložkou „Columns“ a informácie o dátovom súbore ako celku po prepnutí na záložku „Overview“. Po oboznámení s dátovým súborom ho už len jednoducho skopírujeme pomocou ikony „Download“.



Obrázok 6: Nahliadnutie na dátový súbor prostredníctvom platformy Kaggle.
Zdroj: Vlastné spracovanie.

3.3 Vizualizácia dátového súboru v programe POWER BI

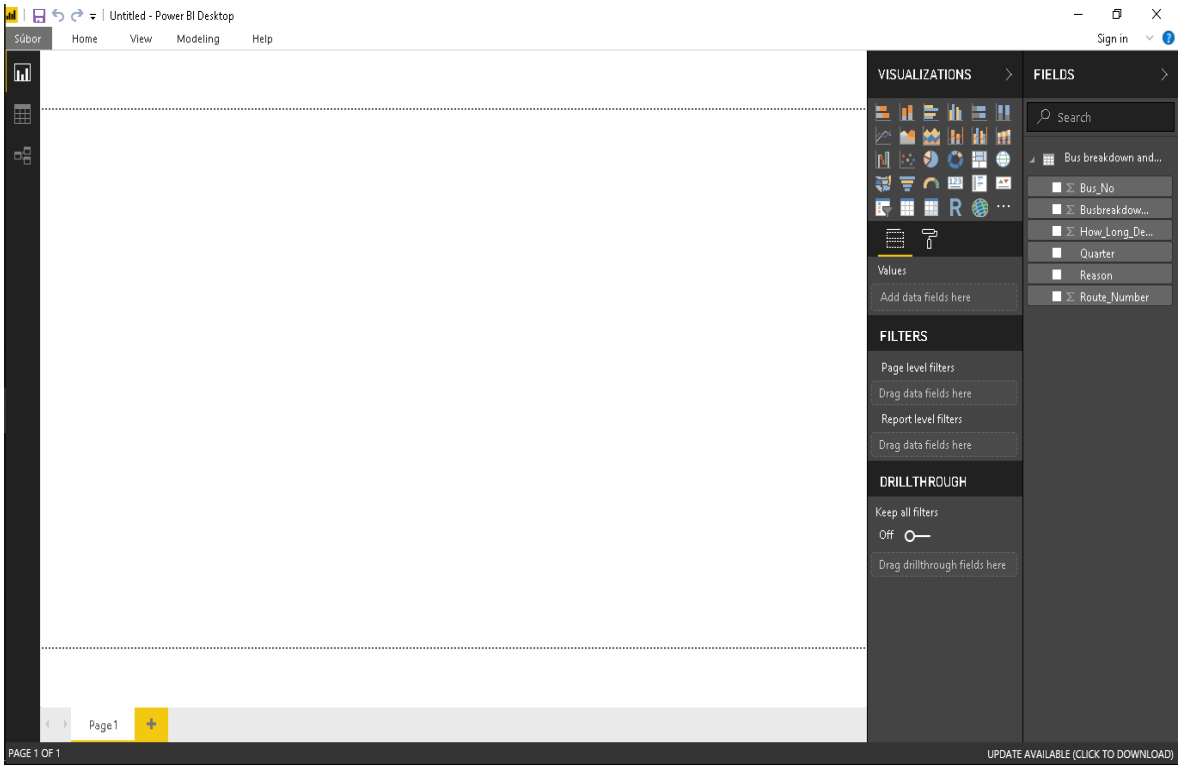
Po skopírovaní nášho dátového súboru si predtým než ho podrobíme analýze vytvoríme prostredníctvom programu Power BI interaktívny dashboard, ktorý nám poslúži na lepšie pochopenie samotných hodnôt atribútov a vzťahov medzi nimi.

Power BI je podnikový analytický nástroj, ktorý poskytuje základňu na rýchle a informované rozhodovanie.[25] Umožňuje prístup k dátam zo stoviek podporovaných zdrojov ako napríklad Dynamics 365, Salesforce, Azure SQL DB, Excel, alebo SharePoint. Jeho cieľom je poskytnúť používateľovi efektívne a podrobné spracovanie,

[25] *Power BI*. [online].[cit. 8.4.2019]. Dostupné na internete: <https://powerbi.microsoft.com/en-us/get-started/>

analýzu a vizualizáciu dát pomocou interaktívnych dashboardov pri zachovaní jednoduchého a zrozumiteľného prostredia pre používateľa.

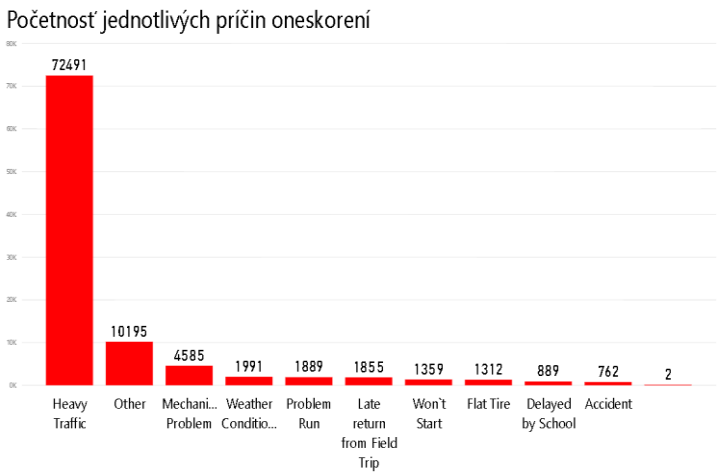
Keďže sú nami stiahnuté dáta vo formáte CSV (comma separated values - hodnoty oddelené čiarkami), ktorý program Power BI podporuje, môžeme ich do neho rovno načítať. Atribúty dátového súboru sa nám po jeho načítaní zobrazia v pravej časti používateľského prostredia pod zložkou „Fields“ (Obrázok 7).



Obrázok 7: Používateľské prostredie programu Power BI po načítaní dátového súboru.
Zdroj: Vlastné spracovanie

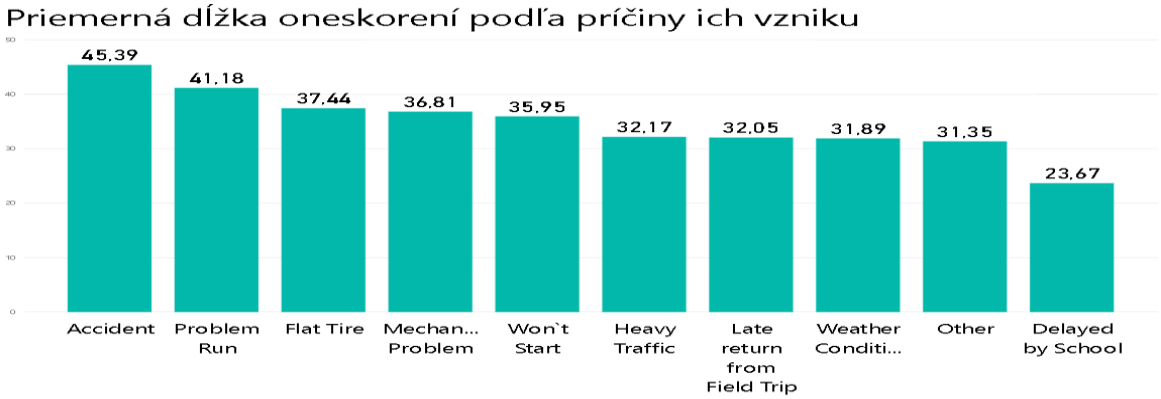
Po načítaní datasetu si zvolíme vhodnú formu vizualizácie v časti „Visualizations“ a taktiež vhodný atribút pre danú vizualizáciu. V našom konkrétnom prípade zvolíme atribúty, ktorých vizualizácie môžu priniesť podniku najväčší prínos a to dôvod oneskorenia, dĺžka oneskorenia a trasy na ktorých sa oneskorenie vyskytlo. Následne ich vizuálne upravíme do nami požadovanej podoby a optimalizujeme ich veľkosť.

Vytvorené vizualizácie obsahujú množstvo užitočných informácii pre podnik ako napríklad to, že najvýznamnejšia časť oneskorení bola spôsobená hustotou premávky a to, že štvrte Manhattan a Bronx sa podieľali na viac ako polovici oneskorení (Obrázok 8b). V priemere najdlhšie boli oneskorenia vyvolané dopravnou nehodou, ktorých dĺžka bola priemerne viac ako 45 minút (Obrázok 8c). Naopak najmenej času zabrali meškania vyvolané nástupom študentov, priemerne 23,67 minút (Obrázok 8c).

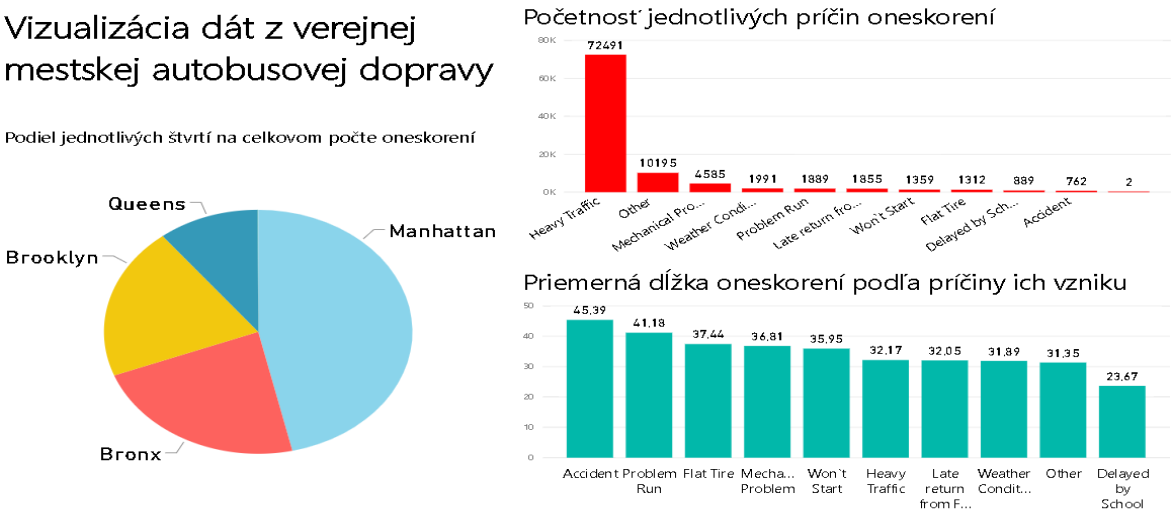


Obrázok 8a) Početnosť jednotlivých príčin oneskorení

Obrázok 8b) Podiel jednotlivých štvrtí na celkovom počte oneskorení



Obrázok 8c) Priemerná dĺžka oneskorení podľa príčiny ich vzniku



Obrázok 8d) Kompletná interaktívna vizualizácia (dashboard) vytvorená pomocou Power BI

Obrázok 8a) až 8d): Grafické znázornenie datasetu prostredníctvom Power BI
Zdroj: Vlastné spracovanie

Vizualizácie môže podnik využiť, tak ako v tomto prípade, pri prezentovaní štruktúry určitého problému vo svojom internom prostredí medzi manažérmi, ale aj ako prostriedok na prezentovanie výsledkov firmy v externom prostredí pred investormi alebo dôležitými partnermi. Prostredníctvom nich môže jednoducho a zrozumiteľne poukázať na zvyšovanie úrovne dosiahnutých výsledkov a na pozitívny vývoj kľúčových ukazovateľov. V tomto prípade vie aj pomocou samotných vizualizácii na ktorú oblasť sa pri riešení problému zamerať a na základe toho si naplánovať plán riešenia daného problému.

3.4 Analýza pomocou programu Weka

Weka je voľne dostupný analytický nástroj určený na hĺbkovú analýzu dát, vyvinutý na Waikatskej Univerzite na Novom Zélande, pričom celý systém je písaný v Jave a je dostupný pod verejnou licenciou. Systém je schopný správne fungovať takmer na všetkých platformách a využíva kolekciu učiacich sa algoritmov pre širokú škálu úloh na riešenie reálnych problémov spojených s ťažbou dát. Obsahuje nástroje na prípravu, klasifikáciu, regresiu, zhlukovanie a vizualizáciu dát ale taktiež grafické používateľské prostredie pre jednoduchý prístup k týmto nástrojom.[26]

Všetky techniky využívané analytickým nástrojom Weka sú založené na predpoklade, že dáta sú prístupné ako samostatné ploché súbory alebo spojenie, kde je každý bod dát popísaný pevným počtom vlastností. Nie je síce schopný viacnásobnej relačnej ťažby dát, avšak obsahuje nástroj na konvertovanie súboru prepojených databázových tabuliek na jednu tabuľku, ktorá je vyhovujúca na spracovanie prostredníctvom Weky. Ďalšia podstatná oblasť, ktorá v súčasnosti nie je obsiahnutá pomocou zahrnutých algoritmov je sekvenčné modelovanie.

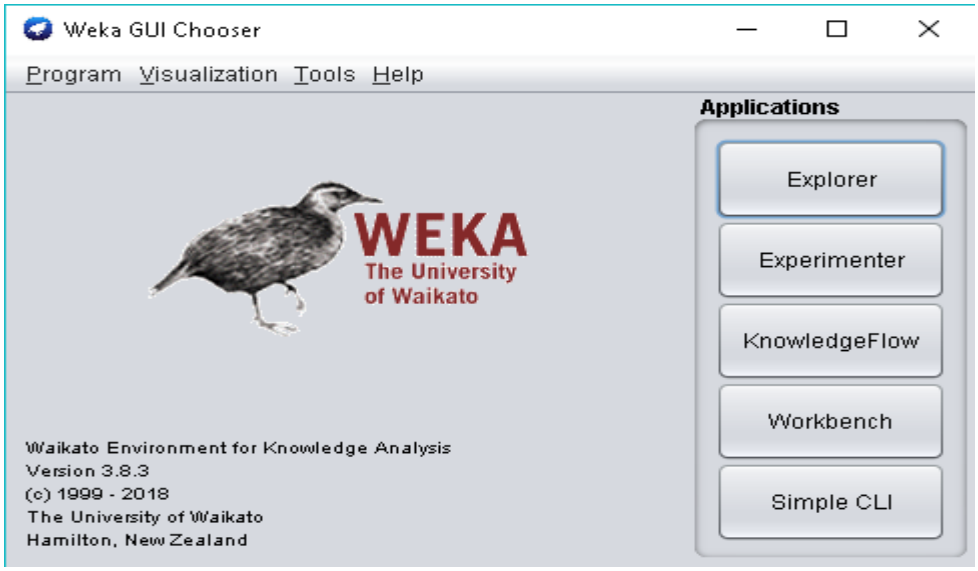
Analytický nástroj Weka je voľne dostupný na stiahnutie zo stránky <https://www.cs.waikato.ac.nz/ml/weka/>. Na výber je buď stiahnutie špecifického inštalačného programu podľa platformy alebo Java jar súbor, ktorý používateľovi umožní spustiť Weku v bežnej podobe, pokiaľ je na zariadení nainštalovaná Java.

Pracovné prostredie programu Weka

Pre správne pochopenie analytického softvéru Weka je potrebné si priblížiť jednotlivé nástroje, ktoré Weka ponúka, a prostredie, v ktorom budeme pracovať. Prvým

[26] FRANK, Eibe - HALL, Mark - WITTEN, Ian. *Data Mining: Practical Machine Learning Tools and Techniques*. 2016. [online].[cit. 8.4.2019]. Dostupné na internete: <https://www.cs.waikato.ac.nz/ml/weka/>

krokom je samotné spustenie Weky, ktoré docielime dvojitým kliknutím na weka.jar súbor. Po spustení programu sa nám na obrazovke zobrazí dialógové okno s názvom „Weka GUI Chooser“, ktoré slúži ako vstup do analýzy.



Obrázok 9: Úvodné dialógové okno v programe Weka
Zdroj: Vlastné spracovanie

Na výber nám poskytne nasledovné 5 základných rozhraní, ktoré slúžia na prácu s dátovými súbormi:

- Explorer - rozhranie Explorer je jednoduché používateľské prostredie, v ktorom sú obsiahnuté všetky hlavné „balíky“ Weky spolu s nástrojom na vizualizáciu, ktorý umožňuje vizualizáciu dátových súborov a predpovedí klasifikátorov a klastrov v dvoch rozmeroch.
- Experimenter - rozhranie Experimenter slúži ako pomoc pri riešení základných praktických otázok pri aplikovaní techník klasifikácie a regresie. Toto rozhranie používateľovi umožňuje automatizovať celý proces tým, že zjednoduší spustenie klasifikátorov a filtrov s rozličnými nastaveniami parametrov vo väčšom súbore dát, zhromažďovanie výkonnostných štatistických dát a uskutočňovanie testov významnosti.
- KnowledgeFlow - umožňuje navrhnúť konfigurácie pre spracovanie dátového toku. Základným nedostatkom rozhrania Explorer je, že uchováva všetko v hlavnej pamäti a pri otvorení dátového súboru ho automaticky celý načíta. To je dôvod prečo dané rozhranie môže byť použité len na úlohy malej, prípadne strednej veľkosti. Weka však obsahuje aj algoritmy na spracovanie veľkých súborov dát.

Výhodou rozhrania Knowledge Flow je to, že umožňuje spájať boxy predstavujúce učiace algoritmy a dátové zdroje do podoby, ktorá používateľovi vyhovuje. To nám dovoľuje špecifikovať dátový tok spojením komponentov reprezentujúcich dátové zdroje, nástroje na predbežné spracovanie, učiace algoritmy, vyhodnocovacie metódy a moduly na vizualizáciu. Pokiaľ sú filtre a učiace algoritmy schopné dodatočného učenia, dáta budú načítané a spracované dodatočne.

- Workbench - predstavuje zjednotené grafické používateľské prostredie, ktoré kombinuje predošlé tri do jednej aplikácie. Je ľahko konfigurovateľné, čo umožňuje používateľovi špecifikovať ktoré aplikácie sa objavia spolu s nastaveniami súvisiacimi s nimi.
- Simple CLI – nástroj na priame zadávanie príkazov z príkazového riadku.

3.5 Príprava dátového súboru

Po skopírovaní súboru je potrebné ho upraviť do podoby, ktorú Weka podporuje. Štandardným formátom programu Weka je formát ARFF (Attribute Relation File Format), avšak Weka podporuje, okrem iných, aj formát CSV (comma separated values – hodnoty oddelené čiarkami). Po skopírovaní bude dátový súbor vo formáte CSV, ale jeho úpravu zrealizujeme v programe excel a následne opäť uložíme vo formáte CSV. Úprava v exceli pozostáva z odstránenia nepotrebných atribútov a ponechania len tých, ktoré budú predmetom našej analýzy (Obrázok 10).

	A	B	C	D	E	F
1	Busbreakdown_ID	Bus_No	Route_Number	Reason	Quarter	How_Long_Delayed
2	1426983	889041	X2381	Heavy Traffic	Bronx	40
3	1394710	889041	K8468	Won`t Start	Brooklyn	40
4	1240620	889041	Q2624	Heavy Traffic	Queens	25
5	1480179	870537	M846	Mechanical Problem	Manhattan	22
6	1477852	774603	X056	Other	Bronx	10
7	1250187	746912	X128	Late return from Field	Bronx	20
8	1374536	746198	X187	Heavy Traffic	Bronx	22
9	1373710	718905	X686	Heavy Traffic	Bronx	22
10	1377944	718809	Q2957	Heavy Traffic	Queens	10

Obrázok 10: Upravený dátový súbor načítaný v Exceli
Zdroj: Vlastné spracovanie

Ďalším krokom je načítanie upraveného súboru vo formáte CSV do textového editora. Po jeho načítaní je nutné pridať názov datasetu, ku ktorému sa vzťahuje menovka

@relation, v našom prípade teda @relation Bus_breakdown_and_delays a taktiež pridať informáciu o jednotlivých atribútoch pomocou @attribute. Ku každej takto pridanej informácii je potrebné uviesť aj typ daného atribútu, pričom Weka rozlišuje štyri základné typy atribútov, ktorými sú nominal, numeric, string a date. V našom prípade sú v dátovom súbore obsiahnuté len atribúty typu nominal a numeric. Preto k atribútom týkajúcim sa identifikačného čísla meškania, čísla autobusu, čísla trasy a dĺžky meškania v minútach pridáme rozpoznávací typ numeric a k atribútom týkajúcich sa dôvodu a miesta vzniku meškania pridáme jednotlivé varianty, ktoré môžu hodnoty atribútov dosahovať. Takto upravené dáta uložíme pomocou textového editora vo formáte ARFF, podporovanom programom Weka (Obrázok 11a, 11b).

```
Súbor  Úpravy  Formát  Zobraziť  Pomocník
Busbreakdown_ID;Bus_No;Route_Number;Reason;Boro;How_Long_Delayed
1426983;889041;X2381;Heavy Traffic;Bronx;40
1394710;889041;K8468;Won`t Start;Brooklyn;40
1240620;889041;Q2624;Heavy Traffic;Queens;25
1480179;870537;M846;Mechanical Problem;Manhattan;22
1477852;774603;X056;Other;Bronx;10
1250187;746912;X128;Late return from Field Trip;Bronx;20
1374536;746198;X187;Heavy Traffic;Bronx;22
1373710;718905;X686;Heavy Traffic;Bronx;22
```

Obrázok 11a) Dátový súbor pred úpravou

```
Súbor  Úpravy  Formát  Zobraziť  Pomocník
@relation Bus_breakdown_and_delays

@attribute Busbreakdown_ID numeric
@attribute Bus_No numeric
@attribute Route_Number numeric
@attribute Reason {Accident,MechanicalProblem,Other,HeavyTraffic
@attribute Quarter {Bronx,Brooklyn,Manhattan,Westchester,Queens,
@attribute How_Long_Delayed_IN_MINS numeric


@data
1513187,2418,9967,LatereturnfromFieldTrip,Brooklyn,40
1504770,2418,9967,LatereturnfromFieldTrip,Brooklyn,22
```

Obrázok 11b) Dátový súbor po úprave

Obrázok 11a) a 11b): Príprava dátového súboru
Zdroj: Vlastné spracovanie

3.6 Analýza dát

Upravený stiahnutý dátový súbor podrobíme analýze pomocou analytického nástroja Weka, konkrétne využitím dataminingových techník klasifikácie a zhlukovej analýzy. Výsledné modely tejto analýzy následne vyhodnotíme a posúdime ich užitočnosť a možnosť využitia ich podstaty v podnikovej praxi.

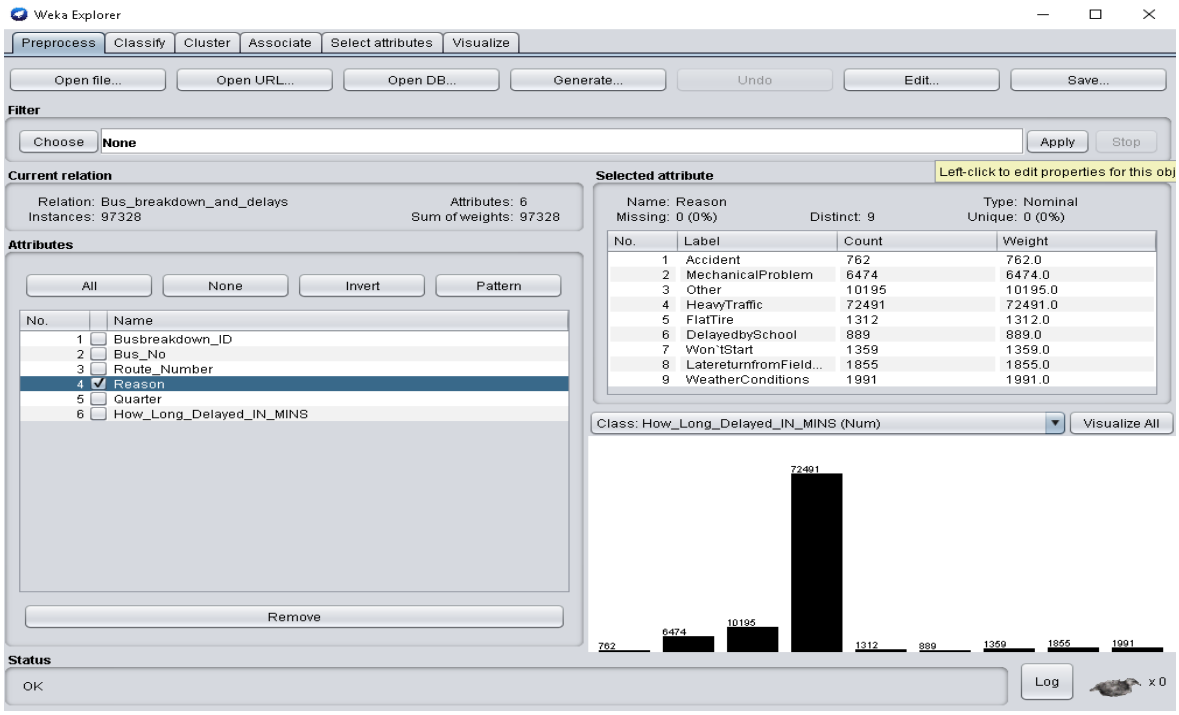
3.6.1 Príprava na analýzu

Po spustení analytického nástroja Weka dvojitým kliknutím na weka.jar súbor sa na obrazovke zobrazí úvodné dialógové okno s názvom „Weka GUI Chooser“, v ktorom následne zvolíme rozhranie Explorer, pomocou ktorého budeme samotnú analýzu dátového súboru realizovať. Po otvorení rozhrania Explorer si pomocou záložky „Open file“ v ľavom hornom rohu do Weky načítame náš konkrétny dátový súbor vo formáte ARFF (Obrázok 11b).

Už v tomto stave nám Weka poskytne množstvo užitočných informácii o našom datasete. Pod nápisom „Current relation“ môžeme vidieť, že nami načítaný súbor obsahuje 97328 prípadov, ktorých hodnota sa vzťahuje k jednému zo šiestich rôznych atribútov. Atribúty, ktorých hodnoty v našom datasete analyzujeme sa nazývajú Busbreakdown_ID (predstavuje identifikačné číslo jednotlivého prípadu omeškania), Bus_No (číslo autobusu, ktorého sa meškanie týkalo), Route_Number (číslo trasy, na ktorej sa meškanie vyskytlo), Reason (dôvod, ktorý spôsobil omeškanie autobusu), Quarter (štvrť, na ktorej sa meškanie vyskytlo) a nakoniec atribút How_Long_Delayed_IN_MINS (dĺžka omeškania autobusu v minútach).

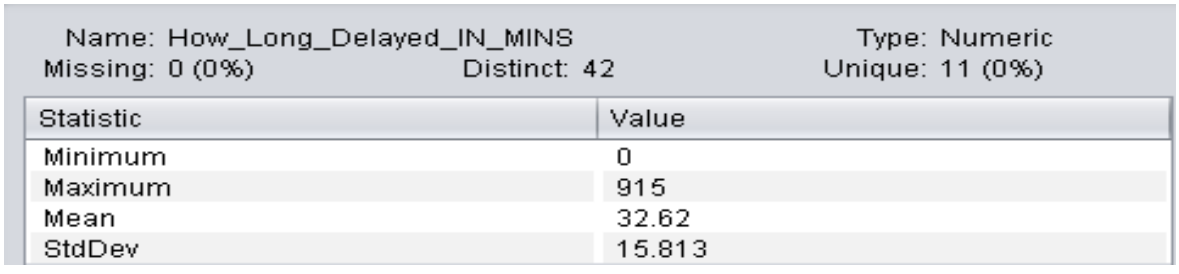
Po kliknutí na ľubovoľný atribút sa v pravej obrazovky zobrazia podrobnejšie informácie, ktoré ho opisujú. Napríklad po kliknutí na atribút „Reason“ vieme v pravej časti obrazovky zistiť typ daného atribútu, ktorý je v tomto prípade nominálny, počet chýbajúcich hodnôt, množstvo druhov hodnôt a počet špecifických prípadov pre daný atribút, ktoré predstavujú hodnoty pre daný atribút vyskytujúce sa len v jednom konkrétnom prípade. Pod týmito údajmi sa v okne „Preprocess“ nachádzajú typy hodnôt atribútov, ktoré v dátovom súbore dosahujú. V našom prípade je ich konkrétne deväť, pričom každý z nich predstavuje určitý dôvod, ktorý bol príčinou meškania autobusu. Pri každom z nich taktiež môžeme vidieť početné zastúpenie tejto hodnoty v celom dátovom súbore. Graf znázorňujúci veľkosť zastúpenia jednotlivých hodnôt v rámci daného atribútu sa nachádza v pravom dolnom rohu, pomocou ktorého vieme ľahko zistiť, že najčastejšia

príčina meškania autobusov v našom konkrétnom dátovom súbore bola zvýšená hustota premávky, ktorá meškanie zapríčinila presne v 72491 prípadoch (Obrázok 12).



Obrázok 12: Používateľské prostredie rozhrania Explorer po načítaní dátového súboru
Zdroj: Vlastné spracovanie

Pri kliknutí na numerický typ atribútu, ktorým je v našom prípade napríklad atribút How_Long_Delayed_IN_MINS predstavujúci dĺžku meškania v minútach sa nám v pravej časti obrazovky zobrazia jeho štatistické hodnoty, respektíve jeho minimálna (0) a maximálna hodnota (915), priemer (32,62) a veľkosť štandardnej odchýlky (15,813%) (Obrázok 13).



Obrázok 13: Štatistické hodnoty dátového súboru
Zdroj: Vlastné spracovanie

Atribút môžeme odstrániť jeho označením a následným kliknutím na „Remove“ v ľavom spodnom rohu. Uskutočnené zmeny sa dajú vrátiť pomocou tlačidla „Undo“.

Využitím tlačidla „edit“ môžeme hľadať špecifické hodnoty v dátovom súbore a v prípade potreby ich zmeniť alebo zmazať.

3.6.2 Klasifikácia

V ďalšom kroku náš dátový súbor podrobíme analýze prostredníctvom využitia jednej z techník data miningu, konkrétne klasifikácie. Na túto analýzu použijeme upravený dátový súbor o veľkosti 1000 prípadov, ktorý obsahuje prípady oneskorení autobusov na trase číslo 1 a k nim informácie o počte študentov, ktorí na autobus v priebehu jeho jazdy nastúpili. Pokúsime sa tak nájsť súvislosť a mieru ovplyvnenia meškania vyvolanú nastupovaním študentov.

Ako prvé je potrebné sa v používateľskom prostredí analytického nástroja Weka prepnúť do zložky „Preprocess“ v ľavom hornom rohu na záložku „Classify“. Táto záložka nám poskytne možnosť využitia množstva učiacich klasifikačných a regresných algoritmov ako napríklad BayesNet, MP5, SimpleLogistic, DecisionTable, JRip, J48 a mnohých ďalších. Tieto modely odhadnú presnosť výsledku predikčného modelu a ukážu nesprávne predikcie na modeli samotnom.

Ak na záložke „Classify“ vyberieme učiaci algoritmus pomocou tlačidla „Choose“, verzia klasifikátora sa nám objaví vedľa tlačidla, zahŕňajúca aj parametre špecifikované pomocou znaku mínus. Weka nám umožní použitie len tých klasifikátorov, ktoré sú pre daný dátový súbor z hľadiska typu jeho hodnôt použiteľné.

Na prácu s naším dátovým súborom použijeme algoritmus s názvom M5P, ktorý patrí do podskupiny prostriedkov označených ako rozhodovacie stromy, ktoré patria medzi najzrozumiteľnejšie a najčastejšie využívané prostriedky Weky. Pred spustením daného algoritmu však v časti „Test option“ otvoríme náš dátový súbor. Zložka „More options“ nám ponúka ďalšie možnosti na špecifikáciu výsledku modelu. Následne už len spustíme nami vybraný algoritmus prostredníctvom tlačidla „Start“ a počkáme, kým nám Weka model vypracuje.

Po vypracovaní modelu pomocou algoritmu M5P sa nám v pravej časti rozhrania zobrazí rozhodovací strom v textovej podobe, pričom v úvodnej časti výstupu sú pomenované základné parametre rozhodovacieho stromu a to názov učiacej schémy, názov dátového súboru, počet prípadov, počet atribútov a ich názvy, testovací mód a následne rozhodovací strom v textovej podobe. Pod ním nasledujú 3 zistené pravidlá (označené ako LM 1 až LM 3) pre daný dátový súbor, ktoré algoritmus našiel. Tieto pravidlá sú základom

rozhodovacieho stromu a predstavujú akýsi výsledkov jednotlivých variant, ktoré môžu v skutočnosti nastať (Obrázok 14).

```
=== Run information ===

Scheme:       weka.classifiers.trees.M5P -M 4.0
Relation:     Bus_breakdown_and_delays
Instances:    1000
Attributes:   2
              How_Long_Delayed_IN_MINS
              Number_Of_Students_On_The_Bus
Test mode:    10-fold cross-validation

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

M5 pruned model tree:
(using smoothed linear models)

Number_Of_Students_On_The_Bus <= 9.5 : LM1 (775/88.972%)
Number_Of_Students_On_The_Bus >  9.5 :
|   Number_Of_Students_On_The_Bus <= 13.5 : LM2 (178/119.832%)
|   Number_Of_Students_On_The_Bus >  13.5 : LM3 (47/138.486%)

LM num: 1
How_Long_Delayed_IN_MINS =
        + 17.5679

LM num: 2
How_Long_Delayed_IN_MINS =
        + 18.5717


LM num: 3
How_Long_Delayed_IN_MINS =
        + 20.9695

Number of Rules : 3

Time taken to build model: 0.23 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.1511
```

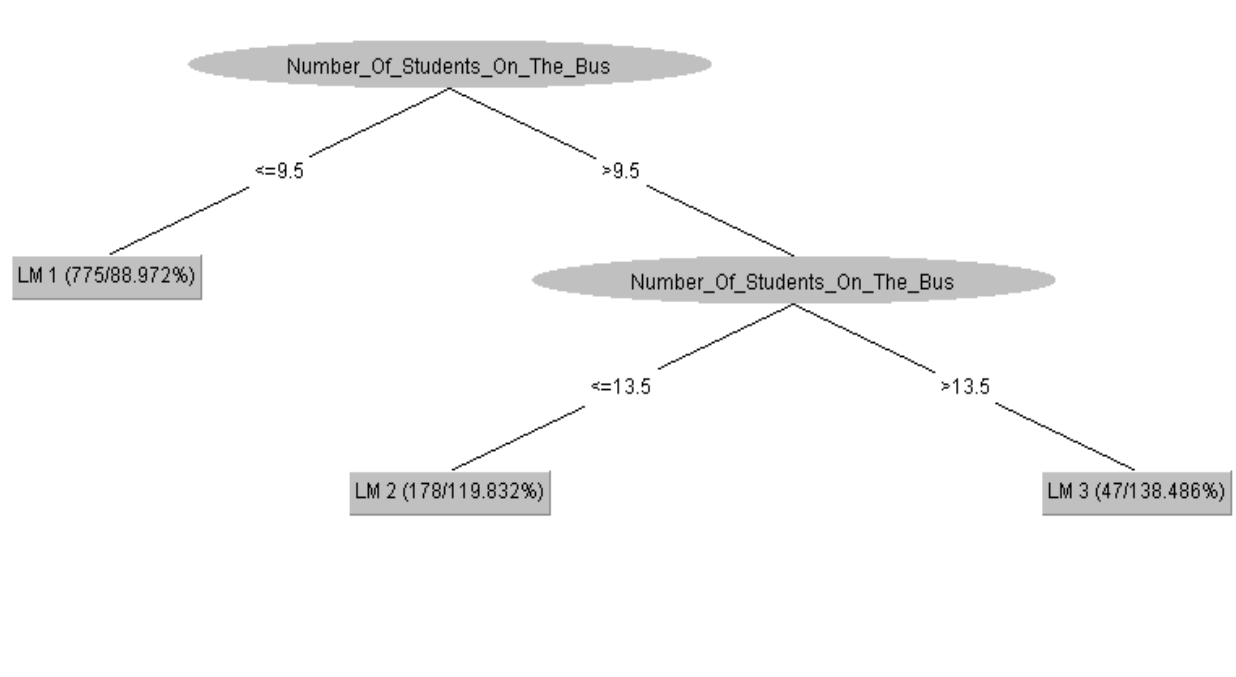
Obrázok 14: Rozhodovací strom v textovej podobe
Zdroj: Vlastné spracovanie

V závere rozhodovacieho stromu môžeme nájsť štatistické ohodnotenie nášho modelu. Nami sledované atribúty (dĺžka meškania a počet študentov nastupujúcich na daný autobus) dosiahli hodnotu korelačného koeficientu na úrovni 0,1511, čo predstavuje

relatívne slabú ale existujúcu závislosť medzi hodnotami nami sledovaných atribútov. To znamená, že počet študentov na danej trase mierne ovplyvňuje aj dĺžku meškania autobusu.

3.6.3 Možnosť využitia výsledkov získaných metódou klasifikácie

Kvantifikáciu miery vplyvu nastupujúcich študentov na celkovú dĺžku meškania vie následne podnik využiť ako významný zdroj informácii pri tvorbe časových rozvrhov. Keďže podnik disponuje dátami týkajúcimi sa počtu študentov nastupujúcich na konkrétny autobus a aj presnom čase danej jazdy, dokáže si dané informácie rozdeliť podľa času ich vzniku. To znamená, že ak si podnik kvantifikuje priemerný počet študentov na danom autobuse počas jednotlivých hodín v priebehu dňa, dokáže na základe toho zistiť v akej miere je na základe toho potrebné v jednotlivých hodinách dopravu posilniť prípadne oslabiť, čo prispieva k lepšiemu využívaniu jeho zdrojov.



Obrázok 15: Rozhodovací strom v grafickej podobe
Zdroj: Vlastné spracovanie

Na Obrázku 15 môžeme vidieť, že „cesta“ k výsledkom rozhodovacieho stromu je ovplyvnená počtom študentov na konkrétnom autobuse, pričom jeho výsledky predstavujú 3 zistené pravidlá (označené ako LM 1 až LM 3), ktoré algoritmus v dátovom súbore našiel. To znamená, že pre prípady s počtom študentov 9 a menej platí pravidlo č. 1 (s priemernou dĺžkou meškania 17,57 minúty), pre prípady s počtom študentov 10 až 13 pravidlo č. 2 (18,57 minúty) a pre prípady s väčším počtom študentov pravidlo č. 3 (20,97

minúty). Dané pravidlá sa od seba odlišujú priemernou dĺžkou meškania, pričom rozdiel medzi nimi môžeme vidieť na Obrázku 14.

3.6.4 Zhluková analýza

Pomocou zhlukovej analýzy môžeme prostredníctvom Weky podrobiť analýze dátové súbory, ktorých atribúty pozostávajú z číselných ale aj slovných hodnôt. To znamená, že nám môže priniesť užitočné informácie o datasete ako celku.

Dátový súbor budeme analyzovať využitím algoritmu SimpleKMeans, ktorý dokáže analyzovať číselné atribúty (v našom prípade identifikačné číslo oneskorenia, číslo autobusu, číslo trasy a dĺžku meškania) aj nominálne atribúty (dôvod meškania a štvrť na ktorej sa vyskytlo).

Na analýzu použijeme dva upravené dátové súbory opisujúce prípady oneskorení pri najviac frekventovanej trase označenej číslom 1. Oba budú predstavovať prípady oneskorení na tej istej trase, avšak u iného autobusu. Konkrétne sa bude jednať o autobusy s najvyšším počtom zapríčinených oneskorení na danej trase označené číslami 9302 a 362. Keďže autobus číslo 9302 má jazdné hodiny na danej trase len od 6:00 do 8:00 ráno a autobus číslo 362 jazdí na danej trase počas celého dňa, upravíme údaje získané autobusom číslo 362 pre zachovanie relevantnosti tak, aby sme analyzovali len údaje získané v danom čase. To nám prinesie užitočné informácie, ktoré budú môcť byť využité pri plánovaní a optimalizácii rozvrhu trás.

Pri analýze sa zameriame na rozdiely v štruktúre meškaní daných dvoch autobusov, pokúsime sa z danej analýzy získať informácie, ktoré by podnik mohol využiť pri následnom tvorení alebo upravovaní časového rozvrhu, prípadne hodnotení efektívnosti svojich zamestnancov a autobusov.

Po zvolení záložky „Cluster“ tak isto ako pri klasifikácii si pomocou tlačidla „Choose“ otvoríme okno ponúkajúce nám vhodné algoritmy pre náš dátový súbor, v ktorom si zvolíme „SimpleKMeans“. Následne ešte upravíme hodnotu klastrov dvojkliknutím na názov algoritmu, ktorá je prednastavená na hodnotu 2. Nám však vyhovuje bližšie špecifikovanie oneskorenia podľa jeho veľkosti, preto zvolíme väčšiu hodnotu klastrov, konkrétne hodnotu 4. Po upravení tejto hodnoty už len spustíme zhlukovací algoritmus. Výsledky sú znázornené na Obrázkoch 16a, 16b.

Priemer	Cluster 1	Cluster 2	Cluster 3	Cluster 4

```
Reason                      HeavyTraffic    HeavyTraffic    HeavyTraffic    HeavyTraffic    HeavyTraffic
  Accident                     0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  MechanicalProblem            1.0 (  0%)      1.0 (  2%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  Other                        7.0 (  6%)      4.0 (  8%)      2.0 ( 15%)      0.0 (  0%)      1.0 ( 16%)
  HeavyTraffic                98.0 ( 89%)     42.0 ( 85%)     11.0 ( 84%)     40.0 ( 97%)      5.0 ( 83%)
  FlatTire                     1.0 (  0%)      1.0 (  2%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  DelayedbySchool              0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  Won`tStart                   0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  LatereturnfromFieldTrip      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  WeatherConditions            2.0 (  1%)      1.0 (  2%)      0.0 (  0%)      1.0 (  2%)      0.0 (  0%)

How_Long_Delayed_IN_MINS        33.7064         40.2041         53.7692         22.3659         14.6667
                               +/-11.9763       +/-0.9996       +/-6.3791       +/-1.8676       +/-3.6148



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       49 ( 45%)
1       13 ( 12%)
2       41 ( 38%)
3        6 (  6%)
```

Obrázok 16a) Výsledky zhlukovej analýzy autobusu číslo 362 v aplikácií Weka

Priemer	Cluster 1	Cluster 2	Cluster 3	Cluster 4

```
Reason                      HeavyTraffic    HeavyTraffic    HeavyTraffic    HeavyTraffic            Other
  Accident                     0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  MechanicalProblem            3.0 (  1%)      0.0 (  0%)      0.0 (  0%)      1.0 (  0%)      2.0 ( 10%)
  Other                       12.0 (  4%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)     12.0 ( 60%)
  HeavyTraffic               270.0 ( 90%)     25.0 ( 96%)     67.0 ( 97%)    178.0 ( 97%)      0.0 (  0%)
  FlatTire                     0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  DelayedbySchool              0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  Won`tStart                   0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  LatereturnfromFieldTrip      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)      0.0 (  0%)
  WeatherConditions           13.0 (  4%)      1.0 (  3%)      2.0 (  2%)      4.0 (  2%)      6.0 ( 30%)

How_Long_Delayed_IN_MINS        21.2584         41.3846         10.2899         21.8579           27.45
                               +/-8.4791        +/-3.9098        +/-1.177        +/-1.2586       +/-6.9697



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       26 (  9%)
1       69 ( 23%)
2      183 ( 61%)
3       20 (  7%)
```

Obrázok 16b) Výsledky zhlukovej analýzy autobusu číslo 9302 v aplikácií Weka

Obrázok 16a) a 16b): Výsledky zhlukových analýz
Zdroj: Vlastné spracovanie

Algoritmus nám podľa zadania v prvom dátovom súbore našiel 4 zhluky, ktoré obsahujú prípady zoskupené na základe ich vzájomnej podobnosti. To znamená, že každý z nich obsahuje prípady, ktoré sú si vzájomne podobné v rámci daného zhluku, ale popritom iné od prípadov v ostatných zhlukoch. V našom prípade teda rozdelil prípady oneskorení do 4 skupín podľa ich dĺžky a príčiny vzniku. Prvý stĺpec s priemernou dĺžkou meškania (33,71 minút pri prvom autobuse respektíve 21,26 minút pri druhom) obsahuje všetky hodnoty dátového súboru a zvyšné štyri predstavujú jednotlivé zhluky.

Na výsledkoch našej analýzy môžeme vidieť, že aj keď sa u oboch autobusoch jednalo o tú istú trasu, výsledky sú relatívne podstatne odlišné. Aj keď celková štruktúra príčin vzniku oneskorení je podobná, celková priemerná dĺžka oneskorenia a dĺžka oneskorenia v najviac zastúpených zhlukoch sa výrazne odlišuje. Kým pri autobuse číslo 362 je najviac zastúpený (45%) zhluk číslo 1 s priemernou dĺžkou oneskorenia vyššou ako 40 minút, u autobusu číslo 9302 je to zhluk číslo 3 (61%) s priemernou dĺžkou oneskorenia viac ako 21 minút.

Podobné výsledky môžeme vidieť aj na percentuálnom podiele zhluku s najkratšou mierou oneskorenia na celkovom zastúpení, ktorý u autobusu číslo 362 predstavuje až 14 minút a má zastúpenie len 6%, pričom pri autobuse číslo 9302 je to len 10 minút a má zastúpenie až 23%. Mechanické problémy sa pri oboch prípadoch vyskytli len výnimočne a preto výsledky ovplyvnili len minimálne. Príčinu teda môžeme skôr hľadať v horšom rozhodovaní šoféra v hustej premávke, prípadne pri pomalšom riešení špecificky vyskytujúcich sa problémov u nás klasifikovaných ako „Ostatné“.

V oboch prípadoch tvorili prvé tri zhluky najmä prípady vyvolané hustotou premávky, avšak v prípade autobusu číslo 9302 to bolo v každom z nich cez 90%, zatiaľ čo u autobusu číslo 362 to bolo v prvých dvoch zhlukoch 85%, respektíve 84%. To znamená, že v nich boli vo väčšej miere zastúpené aj iné prípady, v tomto prípade to boli dôvody klasifikované ako ostatné. Na základe toho vieme, že mali podobnú štruktúru oneskorenia ako ostatné prípade v danom zhluku vyvolané hustotou premávky a ich dĺžka oneskorenia ktorú priemerne vyvolali bola relatívne veľká. Pre porovnanie sa môžeme pozrieť na zhluk číslo 4 v prípade autobusu číslo 9302, kde väčšinu prípadov tvoria práve prípady vyvolané „ostatnými“ dôvodmi. V tomto je však ich dĺžka podstatne kratšia.

3.6.5 Možnosť využitia výsledkov získaných metódou zhlukovej analýzy

Porovnávanie je jedným z najefektívnejších spôsobov využitia výsledkov analýzy dát. Či už ide ako v tomto prípade o porovnávanie autobusov, prípadne o porovnávanie jednotlivých trás alebo zastávok. Každé z nich podniku môže priniesť cenné informácie na optimalizáciu jeho podnikových činností.

V tomto prípade môže podnik na základe porovnávania autobusov na konkrétnych trasách ohodnotiť ich efektívnosť, optimalizovať časový rozvrh na daných trasách a v prípade potreby podniknúť kroky k zlepšeniu danej situácie, ktoré sa môžu týkať ako aj personálnych otázok v podobe vodičov tak aj technických otázok v podobe servisných opatrení prípadne výmeny autobusov na trasách s rozličným profilom.

Záver

V súčasnej dobe je pri momentálnej rýchlosti vývoja technológii pre podniky pôsobiace v akejkoľvek oblasti životne dôležité aby tvrdo pracovali na prispôsobení sa danému trendu. S neustálym vývojom technológií je taktiež spojený aj nárast objemu dát, ktoré so sebou používanie týchto technológií prináša. A práve využívanie týchto dát predstavuje z roka na rok dôležitejší faktor pre dosiahnutie úspechu v konkurenčnom prostredí. Jedným zo spôsobov využívania takýchto dát je aj samotný data mining, pomocou ktorého sme si na danom datasete prostredníctvom relatívne ľahko zrealizovateľnej hĺbkovej analýzy poukázali na časť možností, ktoré nám analýza dát poskytuje.

Splnenie hlavného cieľa bolo dosiahnuté prostredníctvom naplnenia jednotlivých čiastkových cieľov. Ešte pred samotnou analýzou sme získali vhodný dátový súbor z verejnej autobusovej dopravy. Upravili sme ho pomocou textového editora do podoby, ktorá je potrebná na jeho analýzu v programe Weka.

Následne bol dátový súbor podrobený klasifikácii. Výsledkom klasifikácie bolo zistenie vplyvu počtu študentov nastupujúcich na konkrétny autobus počas jeho trasy na dĺžku meškania. Na základe toho vie podnik určiť v akej miere je nutné posilniť dopravu v hodinách, v ktorých je priemerný počet nastupujúcich študentov v priebehu jednotlivých trás najvyšší a v ktorých hodinách si môže dovoliť využiť autobusy na iných trasách.

Ďalej sme na dátovom súbore realizovali zhlukovú analýzu. Tú sme uplatnili pri porovnávaní jázd dvoch autobusov, pričom porovnávanie jednotlivých podnikových zdrojov ako také je základom ich efektívneho využívania. Na základe štruktúry podnikových dát v tejto oblasti podnikania, predmetom takého porovnávania môžu byť či už spomínané autobusy, ale aj jednotlivé trasy samotné prípadne šoféri autobusov.

Využiteľnosť zistení, ktoré hĺbková analýza prináša, je v podnikovej praxi takmer neobmedzená a jedinú hranicu predstavuje kvalita a množstvo dát. V súčasnosti bohužiaľ väčšina slovenských podnikov v odvetví dopravy nerealizuje zber dát v dostatočnom množstve a kvalite potrebnej pre analýzy. To isté platí zatiaľ aj pre dáta publikované Ministerstvom dopravy a výstavby SR a taktiež pre dáta akciovej spoločnosti SAD Lučenec, ktorej dáta slúžia momentálne skôr len na evidenciu jednotlivých zdrojov spoločnosti.

Z tohto dôvodu by som na základe mojej krátkej osobnej skúsenosti s data miningom dopravným spoločnostiam na Slovensku, vrátane akciovej spoločnosti SAD Lučenec, v budúcnosti navrhol realizovať v rámci ich možností zber a následnú analýzu dát spojených s nielen meškaním, ale aj počasím, presnejšou špecifikáciou vozidiel a aj detailnejšími informáciami o vodičoch. Pridanie podobných typov dát k tým, ktoré sme v práci analyzovali, by mohlo vytvoriť výbornú údajovú základňu pre ich spoločnosť v porovnaní s ostatnými spoločnosťami v danom odvetví, ktorú by mohli využívať aj prostredníctvom hĺbkovej analýzy na udržanie konkurencieschopnosti. Myslím si, že pokiaľ slovenské dopravné spoločnosti využijú svoje zdroje na zlepšenie práve v oblasti získavania a analýzy dát, môžu pri správnom využití získaných dát získať konkurenčnú výhodu a lepšie optimalizovať využívanie svojich ostatných zdrojov, z čoho môžu v konečnom dôsledku profitovať aj napriek potrebe počiatočnej investície.

Zoznam použitej literatúry

1. CERVONE, Guido - LIN, Jessica - WATERS, Nigel. *Data Mining for Geoinformatics: Methods and Applications*. 2014. 166 s. ISBN 978-1461476689.
2. *Data Mining*. [online].[cit. 8.4.2019]. Dostupné na internete: <https://behavior.lbl.gov/?q=node/11>
3. FRANK, Eibe - HALL, Mark - WITTEN, Ian. *Data Mining: Practical Machine Learning Tools and Techniques*. 2016. [online].[cit. 8.4.2019]. Dostupné na internete: <https://www.cs.waikato.ac.nz/ml/weka/>
4. HAN, Jiawei - KAMBER, Micheline - PEI, Jian. *Data Mining: Concepts and Techniques*. 2011. 744 s. ISBN 9780123814807.
5. KANTARDZIC, Mehmed. *Data Mining: Concepts, Models, Methods, and Algorithms*. 2011. 555 s. ISBN 978-0-470-89045-5.
6. LAROSE, T. Daniel - D.LAROSE, Chantal. *Discovering Knowledge in Data: An Introduction to Data Mining*. 2. vyd. 2014. 222 s. ISBN 978-0-470-90874-7.
7. LUO, Xudong - XU YU, Jeffrey - LI, Zhi. *Advanced Data Mining Applications: 10th International Conference*, 2014. 760 s. ISBN 978-3-319-14717-8.
8. *NY Bus Breakdown and Delays*. [online].[cit.8.4.2019]. Dostupné na internete: <https://www.kaggle.com/new-york-city/ny-bus-breakdown-and-delays>
9. OCHOA-ZEZZATTI, Carlos. *Transfer of living fish issues in different types of containers using a bin packing algorithm* [online].[cit.8.4.2019] Dostupné na internete: <https://www.researchgate.net/figure/Scheme-BI-3-Business-Intelligence-Business-Intelligence-Business-Intelligence-is_fig9_284353121>
10. POSPELOVA - Olga. *An introduction to Generative Adversarial Networks*. [online].[cit.8.4.2019]. Dostupné na internete: <https://viblo.asia/p/an-introduction-to-generative-adversarial-networks-gans-a-semi-supervised-learning-3P0lPmqg5ox>
11. *Power BI*. [online].[cit. 8.4.2019]. Dostupné na internete: <https://powerbi.microsoft.com/en-us/get-started/>
12. ROKACH, Lior - Z MAIMON, Oded. *Data Mining With Decision Trees: Theory and Applications*. 2014. 328 s. ISBN 978-9814590075.
13. SABHERWAL, Rajiv - BECERRA-FERNANDEZ, Irma. *Business Inteligence: Practices, Technologies, and Management*. 2010. 304 s.. ISBN 978-0470461709.
14. SLÁNSKÝ, Dávid - POUR, Jan - NOVOTNÝ, Ota. *Business Intelligence: Jak využít bohatství ve vašich datech*. 2004. 192 s. ISBN 9788024710945.

15. VAŇO, Jakub. Data mining a technika rozhodovacích stromov. [online].[cit.8.4.2019]. Dostupné na internete: <http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0708/vano/index.html>
16. WANG, John. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. 2008. 1803 s. ISBN 978-1599049526.
17. WU, Junjie. Advances in K-Means Clustering. 2014. 180 s. ISBN 978-3642447570.