UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

DIPLOMSKO DELO

SEGMENTACIJA UPORABNIKOV S PODATKOVNIM RUDARJENJEM: PRIMER TELEKOMUNIKACIJSKEGA PODJETJA

Ljubljana, september 2016 MARKO JORDAN

IZJAVA O AVTORSTVU

Podpisani Marko Jordan, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Segmentacija uporabnikov s podatkovnim rudarjenjem: Primer telekomunikacijskega podjetja, pripravljenega v sodelovanju s svetovalcem prof. dr. Jurijem Jakličem.

IZJAVLJAM

1. da sem predloženo delo pripravil samostojno;

2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;

3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbel, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;

4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;

5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;

6. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;

7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;

8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;

9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;

10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi.

V Ljubljani, dne 23.09.2016 Podpis študenta:_____

KAZALO

KAZALO TABEL

KAZALO SLIK

UVOD

Informacije so najbolj »vroče blago« v današnjem poslovnem svetu, saj je uspešnost poslovanja odvisna od tega kako dobro organizacije poznajo svoje stranke in kako dobro razumejo ter kako učinkovito upravljajo svoje poslovne procese, to pa je odvisno od informacij (Pareek, 2007, str. 1-2). Informacije se ustvarjajo, zbirajo, integrirajo, distribuirajo na različnih nivojih v poslovnih procesih in v današnjem svetu niso več redke, znanje na osnovi teh informacij pa je, saj so mnogi posamezniki že dosegli stanje »informacijske preobremenjenosti«, ker enostavno nimajo več dovolj časa za analiziranje ogromne količine informacij in njihove pretvorbe v koristno zanje (Pareek, 2007, str. 1). MacLennan, Tang in Crivat (2009, str. 2) opisujejo, da procesorska moč glede na Moorov zakon narašča eksponentno, kapacitete medijev za shranjevanje podatkov pa naraščajo po še hitrejši stopnji. Posledično je zmožnost shranjevanja podatkov presegla zmožnost procesiranja le teh. Velik delež podatkov, ki nastaja in se hrani v različnih poslovnih aplikacijah, celovitih programskih rešitvah (v nadaljevanju ERP), sistemih za upravljanje odnosov z odjemalci (v nadaljevanju CRM), spletnih strežnikih, podatkovnih strežniki itd., tako ostaja neobdelanih, organizacije pa tako postajajo bogate s podatki in hkrati revne z znanjem. Podatkovno rudarjenje pomaga zapolniti to informacijsko vrzel, saj pomaga pretvarjati podatke v informacije, te pa v uporabno znanje (Tsiptsis & Chorianopoulos, 2009, str. 2). S tega vidika je podatkovno rudarjenje mogoče razumeti tudi kot rezultat naravnega poteka razvoja informacijske tehnologije (Han, Kamber & Pei, 2012, str. 2). Pri procesu podatkovnega rudarjenja pa se je potrebno držati ustrezne metodologije, saj sicer obstaja nevarnost, da proces pripelje do napačnih ugotovitev, na podlagi katerih se lahko kasneje sklepajo napačne poslovne odločitve (Berry & Linoff, 2004, str. 44).

Namen diplomskega dela je na praktičnem primeru podjetja s področja telekomunikacij ter na dejanskih podatkih izvesti vedenjsko segmentacijo uporabnikov. Cilj je enotno bazo uporabnikov na osnovi različnih vzorcev uporabe storitev razdeliti na posamezne segmente in jih tudi opisati.

V prvem poglavju bom opredelil pojem podatkovnega rudarjenja, pojasnil njegovo vlogo v moderni informacijski družbi ter opisal posamezne tehnike podatkovnega rudarjenja, ki se v praksi uporabljajo.

V drugem poglavju bom pojasnil pomen uporabe ustrezne metodologije pri izvedbi procesa podatkovnega rudarjenja ter opisal različne metodologije, ki se pri podatkovnem rudarjenju najpogosteje uporabljajo. Natančneje bom opisal CRISP-DM metodologijo, ker jo bom kasneje tudi uporabil v praktičnem delu diplomske naloge.

Tretje poglavje je namenjeno opisu pomena podatkovnega rudarjenja pri učinkovitem upravljanju odnosov z odjemalcu oz. izvajanju učinkovite CRM strategije.

Četrto poglavje bom namenil opisu Microsoftovega algoritma za segmentacijo, ki ga bom tudi uporabil v praktičnem delu diplomske naloge. Pojasnil bom tudi različne metode razvrščanja, ter različne možnosti nastavitev, ki jih ta algoritem ponuja.

Peto poglavje je namenjeno opisu praktičnega primera izvedbe procesa podatkovnega rudarjenja, v okviru katerega bom na konkretnem primeru podjetja opisal glavne faze procesa in podal ugotovitve glede dobljenih rezultatov.

Na koncu bom podal sklepne misli.

1 OPREDELITEV PODATKOVNEGA RUDARJENJA

Han et al. (2012, str 1-6) opisujejo, da splošno znani izrek, da živimo v informacijski dobi, ne drži, saj dejansko živimo v podatkovni dobi. Eksplozivna rast razpoložljivih podatkov je po njihovem mnenju rezultat vse splošne informatizacije družbe na vseh področjih. Razne organizacije tako dnevno ustvarjajo in shranjujejo ogromne količine podatkov, te razmere pa ti avtorji opisujejo kot podatkovno bogato, a informacijsko revno stanje, zato so nujno potrebna zmogljiva in vsestranska orodja, ki omogočajo samodejno odkrivanje dragocenih informacij v velikih količinah podatkov in njihovo preoblikovanje v organizirano znanje. Potreba po taki obdelavi podatkov je vodila do pojava podatkovnega rudarjenja. Podatkovno rudarjenje tako Han et al. (2012, str. 2) opredeljujejo kot mlado, dinamično ter obetajoče področje, ki bo končno pripeljalo družbo iz podatkovne v informacijsko dobo, zato ga je s tega vidika mogoče razumeti kot rezultat naravnega poteka razvoja informacijske tehnologije.

V kontekstu podatkovnega rudarjenja se pogosto pojavlja tudi pojem poslovna inteligenca. Pareek (2007, str. 7-33) opisuje, da se je izraz poslovna inteligenca pojavil v sredini devetdesetih let dvajsetega stoletja in v vsebinskem smislu pomeni pretvorbo surovih podatkov v uporabno obliko, na podlagi katere se lahko sprejema poslovne odločitve. Poslovno inteligenco ta avtor vidi kot okvir, ki povezuje različne discipline, kot so podatkovno rudarjenje, statistične analize, napovedovanje in podporo odločanju in jo razume kot tehnologijo, ki omogoča pretvorbo podatkov v uporabne informacije, ki omogočajo organizacijam sprejemanje boljših, bolj utemeljenih odločitev. Podatkovno rudarjenje pa isti avtor označuje kot proces analiziranja podatkov in odkrivanja skritih vzorcev s pomočjo avtomatiziranih metodologij. Podobno tudi Han et al. (2012, str 27) vidijo podatkovno rudarjenje v samem jedru poslovne inteligence, kot njen poglaviten del, brez katerega mnoga podjetja ne morejo učinkovito analizirati podatkov in sprejemati dobrih poslovnih odločitev. Poleg podatkovnega rudarjenja pa se uporabljajo še izrazi, kot so strojno učenje, odkrivanje znanja v podatkovnih bazah ali napovedna analitika, ki imajo sicer nekoliko drugačen prizvok, pa se vendar toliko prekrivajo, da se jih v vsebinskem smislu lahko razume povsem enakovredno podatkovnemu rudarjenju (MacLennan et al., 2009, str. 1).

Tsiptsis in Chorianopoulos (2009, str. 2-3) opisujeta, da je cilj procesa podatkovnega rudarjenja pridobivanje znanja s pomočjo analiziranja velikih količin podatkov in uporabo različnih metod modeliranja, pri procesu pa gre za pretvorbo podatkov v koristne informacije oziroma znanje, na podlagi katerega je mogoče ukrepati.

Tudi D. T. Larose in C. D. Larose (2014, str. 2) opredeljujeta podatkovno rudarjenje kot proces odkrivanja uporabnih vzorcev in trendov v velikih zbirkah podatkov. Skoraj identično tudi tudi Berry in Lynoff (2004, str. 7) opredeljujeta podatkovno rudarjenje kot dejavnost raziskovanja in analiziranja velikih količin podatkov z namenom odkrivanja smiselnih vzorcev in pravil. Nadaljujeta pa, da podatkovno rudarjenje omogoča organizacijam izboljšavo marketinških, prodajnih in podpornih procesov saj jim omogoča boljši vpogled in s tem razumevanje uporabnikov.

1.1 Področja uporabe podatkovnega rudarjenja

MacLennan et al. (2009, str. 4-5) opisujejo, da je možno s podatkovnim rudarjenjem reševati širok spekter poslovnih problemov in navajajo naslednje najbolj pogoste scenarije:

- **Generiranje priporočil**. Ponujanje ustreznih priporočil kupcem so pomemben poslovni izziv za trgovce in ponudnike storitev. Kupci, ki so jim na voljo primerna in pravočasna priporočila, so navadno bolj lojalni in tudi kupujejo več. Primeri iz prakse so razne spletne trgovine, kjer se ob izbiri določenih artiklov generirajo priporočila za druge sorodne ali komplementarne artikle, ki bi tudi lahko zanimali kupca. Priporočila so izpeljana na podlagi analize nakupnega obnašanja vseh dosedanjih kupcev, ta pravila pa so nato uporabljena pri novem nakupu posameznega kupca.
- **Zaznavanje anomalij**. Pri iskanju anomalij gre za analizo podatkov s pomočjo podatkovnega rudarjenja in določitev tistih primerov, ki niso skladni z ostalimi. Na primer, pri kreditnih karticah lahko na ta način sistem označi posamezne transakcije kot sporne, ponudnik kartice pa prek klica pri uporabniku potem preveri, če je res on uporabil kartico oziroma gre za zlorabo. Zavarovalnice na ta način določajo potencialno sporne zahtevke, pri katerih gre lahko za prevare. Ker taka podjetja dnevno procesirajo ogromne količine zahtevkov, je nemogoče preverjati vsako posamezno zadevo. Podatkovno rudarjenje pa olajša delo v tej meri, da pomaga določiti tiste primere, ki so potencialno sporni. Zaznavanje anomalij pa se lahko uporablja tudi za preverjanje veljavnosti vnosnih podatkov in s tem pomaga preprečevati napake pri vnosih podatkov v informacijski sistem.
- **Analiza odhodov**. Podjetja v telekomunikacijski, bančni, zavarovalniški panogi so pogosto soočena s hudo konkurenco. Pridobivanje novih uporabnikov je drago, precej dražje kot pa zadržanje obstoječih uporabnikov. Analiza odhodov lahko pomaga odkriti uporabnike z veliko verjetnostjo prehoda h konkurenci in tudi določiti razloge za to. Te informacije pa podjetjem lahko zelo pomagajo pri poskusih zadržanja teh uporabnikov.

- **Upravljanje s tveganji**. Primer je uporabna pri odobravanju posojil. Bančnikom pomaga pri določitvi stopnje tveganja posojila in posledično določiti obrestno mero in ostale stroške posojila.
- **Segmentacija uporabnikov**. Segmentacija uporabnikov pomaga podjetjem pridobiti predstavo o njihovih uporabnikih. Uporabniki niso nepopisne množice posameznikov z nekimi »povprečnimi« lastnosti, ampak so to množice posameznikov, kje ima vsak svoje individualne lastnosti in potrebe. Dobro poznavanje uporabnikov omogoča podjetjem bolj prilagojene, bolj osebne odnose z uporabniki. Segmentacija uporabnikov omogoča določitev različnih vedenjskih in opisnih profilov uporabnikov, ti pa so potem osnova za izvedbo trženjskih programov in strategij, ki so prilagojene različnim skupinam uporabnikov.
- **Ciljno oglaševanje**. Spletni portali in spletne trgovine s pomočjo te tehnike ponujajo posameznih uporabnikom prilagojeno spletno vsebino. S pomočjo podatkov o preteklih nakupih ali pa s pomočjo podatkov o zgodovini iskanja ali obiska strani, lahko na primer posameznim uporabnikom prikazujejo oglase, ki so v skladu z njihovim vzorci obnašanja.
- **Napovedovanje**. Tehnika, ki je na primer uporabna za napovedovanje prihodnje prodaje, napovedovanje potrebne zaloge artiklov in druga podobna s časovno dimenzijo povezana vprašanja.

Poslovno področje pa ni edino področje, kjer se uporabljajo tehnike podatkovnega rudarjenja. Berry in Lynoff (2004, str. 7) poudarjata, da so orodja in tehnike podatkovnega rudarjenja enako uporabne tudi na drugih področjih, kot so astronomija, medicina, nadzor v industrijski proizvodnji in celo več, da algoritmi podatkovnega rudarjenja prvotno sploh niso bili zamišljeni z izrecnim namenom uporabe na poslovnem področju.

D. T. Larose in C. D. Larose (2014, str. 1) tako navajata celo primer uporabe podatkovnega rudarjenja v politične namene, kjer je to odigralo pomembno vlogo pri izvolitvi predsednika Baraka Obame na ameriških predsedniških volitvah leta 2012. Szkolar (2013) podrobneje opisuje ta primer predsedniških volitev in navaja, da so v Obamini administraciji zbrali ogromno podatkov s strani terenskih delavcev, zbiralcev prispevkov, socialnih omrežij in jih s pomočjo modela podatkovnega rudarjenja obdelali, ter tako identificirali profile tipičnih predstavnikov svojih volilcev. Potencialne volilce so nato ovrednotili glede na verjetnost njihove podpore ter glede na verjetnost udeležbe volitev, nakar so jih s pomočjo prilagojene in usmerjene kampanje nagovarjali, da so se tudi dejansko udeležili volitev. Taka usmerjena aktivnost jim je omogočila, da so svoja omejena sredstva kar najbolj učinkovito izkoristili. Medtem ko na drugi strani v taboru nasprotnika, Mitta Romneya, niso tako učinkovito izkoristili te tehnologije, na dan volitev pa so se jim celo sesuli analitični strežniki. Tudi ekipa njihovih podatkovnih analitikov je obsegala manj kot desetino velikosti ekipe analitikov Baraka Obame (Issenberg, 2012).

1.2 Algoritmi podatkovnega rudarjenja

V literaturi je mogoče zaslediti delitev algoritmov podatkovnega rudarjenja v dve skupini, najpogosteje se navaja delitev na nadzorovane (angl. *supervised*) in nenadzorovane (angl. *unsupervised*) algoritme (MacLennan et al., 2009, str.6; Larose , D. T., & Larose, C. D., 2014, str. 138). Nekateri avtorji pa navajajo tudi delitev na usmerjene (angl. *directed*) in neusmerjene (angl. *undirected*) algoritme (Berry & Linoff, 2004, str. 8), ali pa na napovedovalne in opisovalne algoritme (Tsiptsis & Chorianopoulos, 2009, str. 3). V vseh teh primerih pa gre za ločevanje na osnovi enakega kriterija, to je na osnovi prisotnosti oz. odsotnosti spremenljivke posebnega tipa, tako imenovane ciljne oz. odvisne spremenljivke. Usmerjeni algoritmi tako poskušajo določiti vrednost te ciljne spremenljivke, na podlagi analize ostalih neodvisnih spremenljivk. V skupino usmerjenih algoritmov spadajo klasifikacija, ocenjevanje in napovedovanje. Neusmerjeni algoritmi pa poskušajo odkriti vzorce ali podobnosti med skupinami zapisov brez uporabe neke ciljne spremenljivke. V to skupino spadajo npr. asociacije, razvrščanje v skupine, opisovanje in profiliranje.

V nadaljevanju bom navedel in na kratko opisal posamezne skupine algoritmov podatkovnega rudarjenja. Berry in Linoff (2004, str. 8-12) združujeta različne algoritme podatkovnega rudarjenja v skupine glede na naloge, ki se jih z njihovo pomočjo rešuje.

1.2.1 **Klasifikacija**

Klasifikacija (angl. *Classification*) je ena najbolj pogostih nalog podatkovnega rudarjenja, saj je razvrščanje stvari ena od prirojenih človeških lastnosti. Zaradi razumevanja sveta stalno razvrščamo snovi v elemente, živali v živalske vrste, pse v različne pasme, ljudi v različne rase. Klasifikacija pomeni analizo lastnosti objekta in določitev enega od v naprej definiranih razredov. V tehničnem smislu pa gre za določanje diskretne (kategorične) vrednosti ene (ciljne) spremenljivke na podlagi analize preostalih spremenljivk oz. preostalih lastnosti objekta, ki je predmet analize. S to nalogo rešujemo poslovne probleme, kot so napovedovanje odhodov, ocene tveganja, ciljno trženje, zaznavanje prevar, itd. Algoritmi, ki jih za klasifikacijo uporabimo, so odločitvena drevesa, naivni Bayes, razvrščanje v razrede in nevronske mreže.

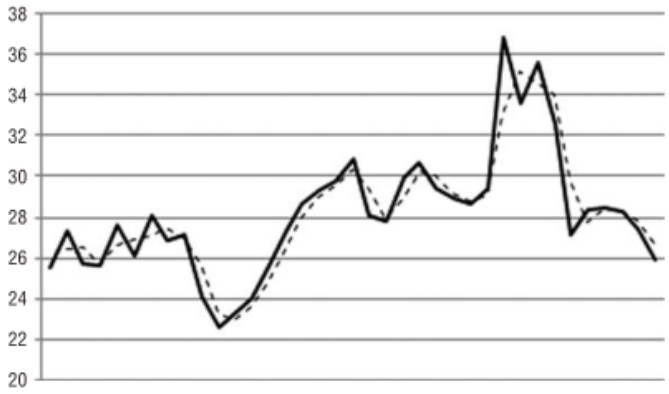
1.2.2 **Ocenjevanje**

Ocenjevanje (angl. *Estimation*) je podobna naloga klasifikaciji, le da namesto iskanja vzorcev, ki določajo kategorično (diskretno) vrednost ciljne spremenljivke, iščemo njeno numerično vrednost. Torej je tu ciljna spremenljivka zvezna, na primer prihodek, saldo, višina, hitrost, in ne kategorična. Primer uporabe je napovedovanje skupnega prihodka gospodinjstva ali pa na primer napovedovanje hitrosti vetra na osnovi temperature, zračnega tlaka in vlažnosti. Primer algoritmov pa je linearna regresija in časovne vrste.

1.2.3 **Napovedovanje**

Napovedovanje (angl. *Prediction*) je zelo podobno klasificiranju in ocenjevanju, le da so zapisi klasificirani na podlagi ocenjene prihodnje vrednosti. Vhodni podatki so časovna serija podatkov, na podlagi katerih se s pomočjo tehnik strojnega učenja in statističnih tehnik, z upoštevanjem sezonskih vplivov, trenda in šuma v podatkih, izračunajo prihodnje vrednosti. Primer uporabe je napovedovanje prodaje za naslednji mesec ali pa npr. napovedovanje prihodnje vrednosti delnice. Slika 1 tako prikazuje dve krivulji. Polna krivulja prikazuje dejanske vrednosti časovne serije, črtkana krivulja pa prikazuje napovedane vrednosti, ki jih izračuna model na osnovi preteklih vrednosti.

Slika 1: Napovedovanje vrednosti delnice

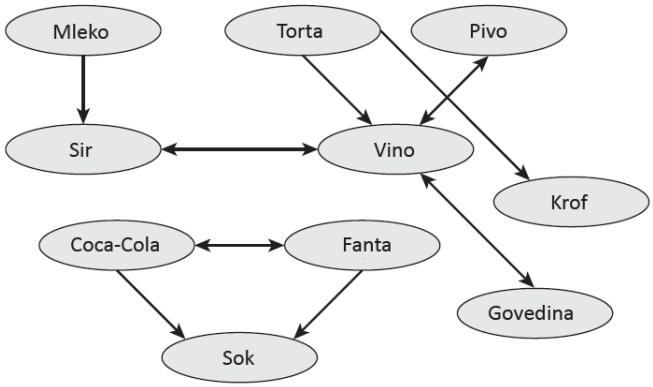


Vir: J. MacLennan, Z. Tang, & B. Crivat, Data Minig with Microsoft SQL Server 2008, 2009, str. 8.

1.2.4 **Asociacije**

Asociacijam pogosto pravimo tudi analiza nakupovalne košarice. Primer asociacije artiklov iz skupine prehrambenih proizvodov prikazuje Slika 2.

Slika 2: Asociacije artiklov



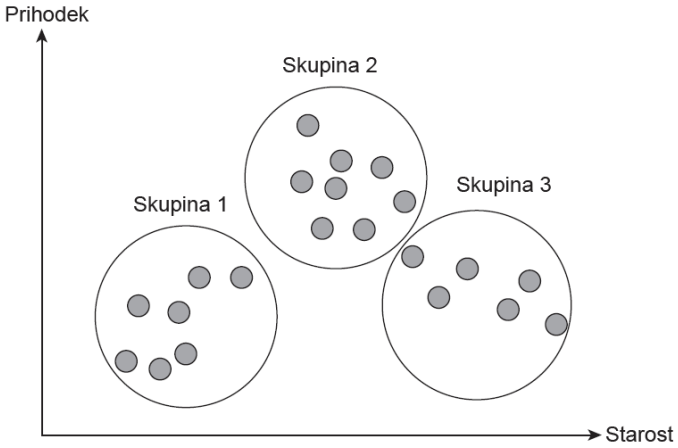
Vir: J. MacLennan, Z. Tang, & B. Crivat, Data Minig with Microsoft SQL Server 2008, 2009, str. 7.

Tipičen primer uporabe je določiti kateri artikli se pogosto prodajajo skupaj, torej v okviru iste transakcije (nakupovalne košarice). Trgovci pogosto te informacije uporabijo za razporeditev artiklov po prodajnih policah. Artikli, ki se pogosto prodajajo skupaj, so tako tudi razporejeni bližje drug drugemu. Lahko pa se uporabi tudi za odkrivanje priložnosti za navzkrižno prodajo (angl. *Cross-selling*) ali pa za sestavljanje za stranke atraktivnih paketov artiklov oziroma storitev.

1.2.5 **Segmentacija**

Pri segmentaciji oz. razvrščanju v skupine, segmente oz. grozde (angl. *Clusters*) se posamezne primere iz heterogene populacije razvršča v večje število bolj homogenih podmnožic, ki jim pravimo segmenti. Segmentacijo od klasifikacije ločuje to, da tu ne določamo v naprej definiranega razreda, posamezni primeri pa so razporejeni v skupine samo glede na medsebojno primerljivost svojih lastnosti. Cilj je doseči čim večjo homogenost znotraj skupine in veliko heterogenost med skupinami. Slika 3 tako prikazuje razvrstitev posameznih primerov v tri skupine. Razvrstitev je opravljena glede na njihove lastnosti, ki jih predstavljata spremenljivki prihodek in starost.

Slika 3: Razvrščanje v skupine



Vir: J. MacLennan, Z. Tang, & B. Crivat, Data Minig with Microsoft SQL Server 2008, 2009, str. 7.

1.2.6 **Opisovanje**

Včasih je cilj podatkovnega rudarjenja samo opisati trende in vzorce v podatkih, na način, da to poveča naše razumevanje vzrokov. Dober opis običajno že sam poda razlago, oziroma vsaj poda namig, kje začeti iskati razlago za različne vzorce in trende. Larose, D. T. in Larose, C. D. (2014, str. 8) opisujeta, da bi morali bili modeli podatkovnega rudarjenja čim bolj pregledni, rezultati modelov pa jasno opisovati vzorce in omogočiti čim bolj intuitivno razlago. Nekateri metode so bolj primerne za razumljivo razlago kot druge. Na primer,

odločitvena drevesa so močno orodje za opisovanje (angl. *Profiling*) in omogočajo človeku prijazno razlago rezultatov, medtem ko so nevronske mreže precej bolj težavne za interpretacijo.

2 METODOLOGIJE PODATKOVNEGA RUDARJENJA

Berry in Linoff (2004, str. 43-50) opisujeta pomen uporabe ustrezne metodologije pri izvedbi podatkovnega rudarjenja. Izhajata iz dejstva, da je zavedanje potencialnih napak in sprejetje ustreznih preventivnih ukrepov najboljši način za uspešno izvedbo podatkovnega rudarjenja. Skozi prakso so se namreč odkrili različni razlogi, zaradi katerih gre pri projektih podatkovnega rudarjenja lahko kaj narobe, v izogib temu pa so se razvili tudi primeri dobre prakse, ki so lahko uporabnikom v pomoč pri uspešni izvedbi projekta podatkovnega rudarjenja. Iz teh primerov dobre prakse se je sčasoma razvila metodologija. Potreba po uporabi metodologije pa se še povečuje s kompleksnostjo poslovnega problema. Ista avtorja navajata, da sta dve pogosti napaki oz. dva nezaželena izida podatkovnega rudarjenja:

- odkrivanje stvari, ki niso resnične, ter
- odkrivanje stvari, ki so sicer resnične, vendar niso koristne.

Pri tem opozarjata, da je odkrivanje stvari, ki niso resnične, precej bolj nevarno, kot odkrivanja stvari, ki so nekoristne, saj se na podlagi napačnih informacij lahko kasneje sklepajo pomembne poslovne odločitve. Precej bolj pogosta napaka je sicer odkrivanje stvari, ki so nekoristne, pri tem pa gre večinoma za odkrivanje stvari, ki so že poznane ali pa za odkrivanje stvari, ki so še nepoznane, vendar pa se jih morda zaradi zakonskih ali regulatornih razlogov sploh ne da uporabiti ali pa so dejavniki izven nadzora organizacije. Tako lahko model odkrije, da se določeni artikli, kot npr. sladoled ali pa klimatske naprave, bolje prodajajo v regijah z vročim podnebjem, vendar pa to spoznanje ne koristi, ker je težko vplivati na vreme v lokalnem okolju. Avtorja zaključujeta, da je metodologija namenjena temu, da usmerja izvedbo podatkovnega rudarjenja k uspešnemu izidu in olajša pot od poslovnega problema do stabilnega modela, ki ponuja koristne in merljive rezultate, na podlagi katerih je mogoče ukrepati.

V nadaljevanju bom opisal tri standarizirane metodologije, KDD (angl. *Knowledge Discovery in Databases*), CRISP-DM (angl. *Cross-Industry Standard Process for Data Mining*) in SEMMA (angl. *Sample-Explore-Modify-Model-Assess*).

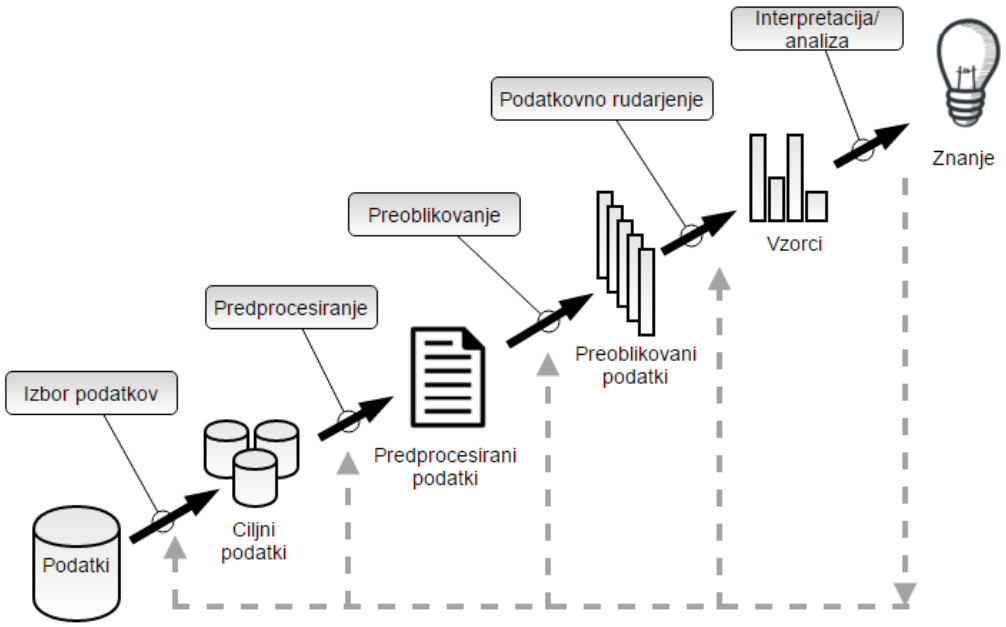
2.1 KDD metodologija

Fayyad, Piatetsky-Shapiro in Smyth (1996, str. 37) opisujejo podatkovno rudarjenje kot del procesa odkrivanja znanja v podatkovnih bazah. Pojav tega izraza sega v leto 1989, ko so na prvi KDD delavnici želeli s tem poudariti, da je znanje končni cilj raziskovanja podatkov, ta

izraz pa je bil hitro sprejet v krogih umetne inteligence in strojnega učenja (Fayyad et al., 1996, str. 37).

KDD metodologija se osredotoča na problem preslikave velikih količin podatkov, ki se zbirajo na raznih nivojih in so običajno zelo podrobni in preobsežni za razumevanje, v bolj kompaktno, združeno in bolj razumljivo obliko. Zgodovinsko gledano so za pojem iskanja uporabnih vzorcev v podatkih obstajala različna imena, vključno z podatkovno rudarjenje, pridobivanje znanja, odkrivanje informacij, podatkovna arheologija itd., termin podatkovno rudarjenje pa je bil večinoma sprejet pri statistikih in podatkovnih analitikih (Fayyad et al., 1996, str. 39). V pogledu Fayyad et al. (1996, str. 41) se KDD nanaša na celoten proces odkrivanja znanja v podatkih, podatkovno rudarjenje pa se nanaša samo na en korak v tem procesu, in sicer na algoritemski del, v katerem se iz podatkov pridobivajo vzorci (Slika 4). Ker se s podatkovnim rudarjenjem lahko misli na celoten KDD proces, ali pa samo na eno fazo v tem procesu, lahko to pri uporabnikih povzroča kar nekaj zmede.

Slika 4: Faze v KDD metodologiji



Vir: U. Fayyad et al., From Data Mining to Knowledge Discovery in Databases, 1996, str. 41

2.2 CRISP-DM metodologija

CRISP-DM je procesni model, ki je bil zasnovan leta 1996 s strani vodilnih podjetij na takrat še relativno novem področju podatkovnega rudarjenja (Chapman et al., 2000, str. 1). Ta podjetja so bila Daimler-Benz, Integral Solutions Ltd. (ISL), NCR in OHRA. V letu 1996 je že obstajalo veliko zanimanje za podatkovno rudarjenje, ni pa še bilo splošno sprejetega pristopa, ki bi organizacijam olajšalo izvedbo lastnih projektov podatkovnega rudarjenja.

Ker bom v praktičnem delu naloge uporabil to metodologijo, jo bom v nadaljevanju tudi podrobneje opisal.

2.2.1 Zgodovinsko ozadje nastanka CRISP-DM metodologije

DaimlerChrysler, takrat še Daimler-Benz, je že takrat močno prednjačil pred ostalimi organizacijami glede implementacije podatkovnega rudarjenja v svojih poslovnih procesih. SPSS (takrat še ISL) je ponujal storitve s področja podatkovnega rudarjenja in je v letu 1994 na trgu ponudil prvo komercialno orodje za podatkovno rudarjenje – Clementine. NCR pa je že takrat zaposloval večje število svetovalcev in specialistov s področja podatkovnega rudarjenja, saj je želel ponuditi dodano vrednost k svoji rešitvi za podatkovna skladišča - Teradata. OHRA pa je bila ena od večjih zavarovalnic z velikimi količinami podatkov, kar je takrat predstavljalo zelo zanimivo okolje za razne testne projekte podatkovnega rudarjenja. Vsem organizacijam je bilo skupno to, da bi s standariziranim procesom potencialnim uporabnikom lažje demonstrirali zrelost podatkovnega rudarjenja za integracijo v poslovne procese (Chapman et al., 2000, str. 2).

Leta 1997 so oblikovali konzorcij in pridobili finančna sredstva s strani Evropske komisije (Shearer, 2000, str. 13). V to leto spada tudi nastanek kratice CRISP-DM. CRISP-DM je bil zasnovan nevtralno tako z vidika uporabne na različnih industrijskih področjih, kot tudi glede uporabe različnih orodji za podatkovno rudarjenje. CRISP-DM procesni model se je v naslednjih letih intenzivno razvijal in v letu 2000 se je pojavila nova različica standarda CRISP-DM v. 1.0. Danes pa se CRISP-DM standard ne razvija več (Piatetsky, 2016). Brown (2015) navaja, da po koncu financiranja s strani Evropske komisije ni bilo več tako velikega komercialnega interesa za nadaljnji obstoj organizacije, zato se je ta sčasoma razpustila. Tudi spletna stran organizacije crisp-dm.org ni več dosegljiva, dokumentacija pa je na voljo preko spletne strani korporacije IBM.

2.2.2 Faze CRISP-DM metodologije

CRISP-DM je ciklični procesni model, sestavljen je iz šestih faz, hierarhično pa je razdeljen na štiri nivoje, saj se faze delijo na posamezne generične aktivnosti, ki se nato delijo na specializirane aktivnosti, te pa še na posamezne instance procesov.

Tabela 1 prikazuje razdelitev faz na generične aktivnosti, ki so v poudarjenem besedilu, v alinejah pa so navedeni dokumenti oz. izhodi (angl. *Outputs*), ki so rezultati posameznih aktivnosti.

Slika 5 predstavlja diagram procesa in prikazuje zaporedje faz v običajnem poteku izvajanja projekta podatkovnega rudarjenja. Zunanji krog predstavlja ciklično naravo samega procesa podatkovnega rudarjenja. Notranje povezave pa predstavljajo najpogostejše prehode med

fazami, s tem pa tudi najpomembnejše odvisnosti med posameznimi fazami. Zaporedje posameznih faz ni striktno določeno, pogosto je namreč potrebno tudi vračanje na predhodne faze.

Tabela 1: CRISP-DM Faze, aktivnosti in izhodni dokumenti

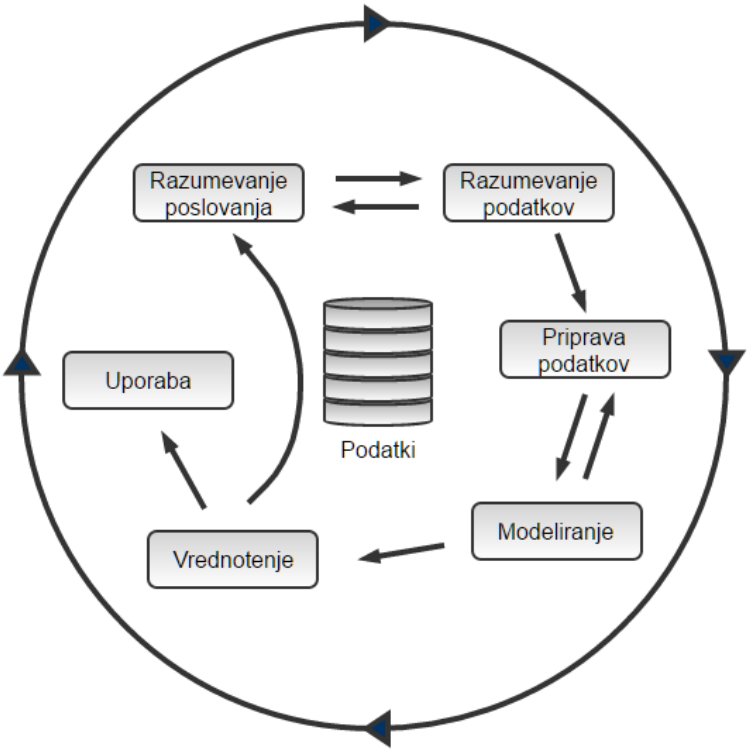
Razumevanje poslovanja	Razumevanje podatkov	Priprava podatkov	Modeliranje	Vrednotenje	Uporaba
Določitev poslovnih ciljev • Poslovno ozadje • Poslovni cilji • Kriteriji poslovne uspešnosti **Ocena stanja** • Popis virov • Zahteve, predpostavke in omenjitve • Tveganja in nepredvideni dogodki • Stroški in koristi **Določitev ciljev podatkovnega rudarjenja** • Cilji • Kriteriji uspešnosti **Izdelava projektnega načrta** • Projektni načrt • Ocena potrebnih orodij in tehnik	**Začetni izbor podatkov** • Poročilo o izboru podatkov **Opis podatkov** • Poročilo o značilnostih podatkov **Pregled podatkov** • Poročilo o pregledu podatkov **Ocena kakovosti podatkov** • Poročilo o kakovosti	**Izbor podatkov** • Opredelitev razlogov za izbor oz. za izključitev **Čiščenje podatkov** • Poročilo o čiščenju podatkov **Izpeljava novih atributov** • Poročilo o izpeljanih atributih **Integracija podatkov** • Poročilo o združenih podatkih **Formatiranje podatkov** • Poročilo o preoblikovanih podatkih • Nabor podatkov • Opisi	**Izbira tehnik modeliranja** • Tehnika modeliranja • Predpostavke **Načrt testiranja** • Priprava načrta za testiranje rezultatov **Izgradnja modelov** • Nastavitev parametrov • Modeli • Opisi modelov **Ocena modelov** • Ugotovitve • Revizija parametrov nastavitev	**Vrednotenje rezultatov** • Ocena rezultatov v odnosu do kriterijev uspešnosti • Odobreni modeli **Revizija procesa** • Pregled poteka procesa **Določitev naslednjih korakov** • Seznam možnih ukrepov • Odločitve	**Načrtovanje implementacije** • Načrt prenosa v produkcijsko okolje **Spremljanje in vzdrževanje** • Načrt nadzorovanja in vzrževanja **Priprava končnega poročila** • Končno poročilo • Končna predstavitev **Revizija projekta** • Dokument o izkušnjah

Vir: P. Chapman et al., CRISP-DM 1.0: Step-by-step data mining guide, 2000, str. 12.

Ko pride do uporabe rezultatov projekt še ni zaključen. Novo pridobljeno znanje in razumevanje problema, ter izkušnje, ki so bile pridobljene med procesom, pogosto sprožijo nova, še bolj podrobna in še bolj osredotočena poslovna vprašanja in tako se celoten cikel

ponovi. V naslednjih ciklih se v procesu uporabi vse novo pridobljeno znanje, razumevanje problema, izkušnje, in to sproži spet nova vprašanja.

Slika 5: Faze CRISP-DM modela



Vir: P. Chapman et al., CRISP-DM 1.0: Step-by-step data mining guide, 2000, str. 10.

Shearer (2000, str. 14-18) podrobneje opisuje posamezne faze CRISP-DM procesa, čigar opis bom v strnjeni obliki povzel v nadaljevanju. Shearer tako glavne faze procesa, njihov pomen, ter njihove cilje opisuje kot:

- **Razumevanje poslovanja**. Začetna faza se osredotoča na razumevanje ciljev projekta in zahtev s poslovne perspektive, ki se kasneje transformira v definicijo ciljev podatkovnega rudarjenja in načina kako te cilje doseči. Shearer (2000, str. 14) zato to fazo navaja kot potencialno najbolj pomembno fazo procesa, saj je bistvena za kasnejšo pravilno izbiro relevantnih podatkov za analizo konkretnega poslovnega problema. Ključne aktivnosti v tej fazi so določitev poslovnih ciljev, ocena stanja, določitev ciljev podatkovnega rudarjenja in izdelava projektnega načrta.

- **Razumevanje podatkov**. Faza razumevanja podatkov se prične z začetnim izborom podatkov in nadaljuje z aktivnostmi, ki omogočajo spoznavanje s podatki in odkrivanje morebitnih težav s kvaliteto podatkov. V tej fazi pride do prvega spoznavanja s podatki, kar zajema začetni izbor podatkov, opis podatkov, pregled podatkov in oceno kakovosti podatkov.

- **Priprava podatkov**. Ta faza zajema aktivnosti povezane s pretvorbo surovih izvornih podatkov v končno podatkovno strukturo, ki se bo uporabila za uvoz v orodje za podatkovno rudarjenje. Te aktivnosti se navadno izvajajo večkrat, zajemajo pa izbor tabel, zapisov in atributov in tudi druge aktivnosti kot so čiščenje in pretvorbe podatkov. Podrobneje pa so to izbor podatkov, čiščenje podatkov, izpeljava novih atributov, integracija podatkov in formatiranje podatkov.

- **Modeliranje**. V tej fazi se uporabi različne tehnike modeliranja in poigrava z različnimi nastavitvami parametrov z namenom doseči optimalne nastavitve. Določene vrste poslovnih problemov lahko rešujemo z različnimi tehnikami , od katerih imajo določene posebne zahteve glede priprave podatkov, zato je tukaj pogosto potrebno vračanje na predhodno fazo priprave podatkov. Modeliranje zajema aktivnosti, kot so izbira tehnike modeliranja, priprava načrta testiranja, izgradnja modelov in ocena modelov.

- **Vrednotenje**. V tej fazi projekta že imamo zgrajene modele. Pred končno uporabo v produkcijskem okolju pa je potrebno temeljito ovrednotiti in revidirati predhodne korake in se prepričati, da izbrani modeli pravilno obravnavajo poslovne cilje. Ta faza zajema vrednotenje rezultatov, revidiranje procesa in določitev naslednjih korakov.

- **Uporaba**. Izgradnja modelov navadno še ne pomeni konca projekta. Iz podatkov pridobljeno znanje je potrebno organizirati in predstaviti na način, da je uporabno za odločitve v poslovnih procesih. To pogosto pomeni uporabo »živih« modelov v odločitvenih procesih, na primer prilagoditev spletnih strani posameznikom v realnem času. Aktivnosti v tej fazi so načrtovanje implementacije, spremljanje in vzdrževanje modelov, priprava končnega poročila in revizija projekta.

2.3 SEMMA metodologija

SEMMA je metodologija, ki jo je SAS namensko razvil za uporabo z svojim lastnim orodjem za podatkovno rudarjenje (SAS Enterprise Miner). Kratica SEMMA se nanaša na posamezne korake, ki si logično sledijo v procesu izvedbe projekta podatkovnega rudarjenja, in sicer:

- **Vzorči** (angl. *Sample*): iz velike količine podatkov se izlušči manjše relevantne vzorce, ki se jih lažje obdela.

- **Razišči** (angl. *Explore*): raziskava podatkov, kjer se išče nepričakovane trende in nepravilnosti, da se pridobi boljše razumevanje podatkov.
- **Spremeni** (angl. *Modify*): prilagoditev podatkov, kjer se ustvari, izbira in preoblikuje spremenljivke za potrebe modeliranja.
- **Modeliraj** (angl. *Model*): modeliranje podatkov s pomočjo programske opreme, ki samodejno išče kombinacije podatkov, ki napovedujejo željeni rezultat.
- **Oceni** (angl. *Assess*): ocena uporabnosti in zanesljivosti rezultatov.

Po drugi strani pa Dean (2014, str. 61) opozarja, da je napačno interpretirati SEMMA kot posebno metodologijo, ker se nanaša samo na logično organiziranost posameznih funkcionalnih sklopov v orodju SAS Enterpise Miner, ki omogočajo izvedbo osrednih aktivnosti pri podatkovnem rudarjenju. To orodje pa je mogoče seveda mogoče uporabiti tudi v sklopu katerega koli drugega procesnega modela.

2.4 Primerjalna analiza metodologij

Azevedo in Santos (2008, str. 182-185) ugotavljata, da obstajajo določene očitne vzporednice med KDD in SEMMA fazami. Vzorčenje je primerljivo z izborom podatkov, raziskovanje z predprocesiranjem, spreminjanje z preoblikovanjem itd. (glej Tabela 2) zato je SEMMA metodologijo mogoče videti kot praktično implementacijo KDD procesa za potrebe SAS Enterprise Miner programskega orodja. Primerjava KDD in CRISP-DM pa po njunem mnenju ni tako direktna in enostavna kot pri SEMMA. CRISP-DM je širša metodologija, saj vsebuje tudi korake, ki se morajo zgoditi pred KDD procesom in korake, ki mu morajo slediti. Faza »Razumevanje poslovanja« se lahko razume kot razumevanje poslovne domene in ciljev raziskovanja, faza »Uporaba« pa kot konsolidacija odkritega znanja in njegova uporaba v realnem okolju, ki sledi KDD.

Tabela 2: Primerjava faz med KDD, SEMMA in CRISP-DM

KDD	**SEMMA**	**CRISP-DM**
Pred-faza KDD	…….	Razumevanje poslovanja
Izbor podatkov	Vzorči	Razumevanje podatkov
Predprocesiranje	Razišči	
Preoblikovanje	Spremeni	Priprava podatkov
Podatkovno rudarjenje	Modeliraj	Modeliranje
Interpretacija/Analiza	Oceni	Vrednotenje modelov
Post-faza KDD	…….	Uporaba

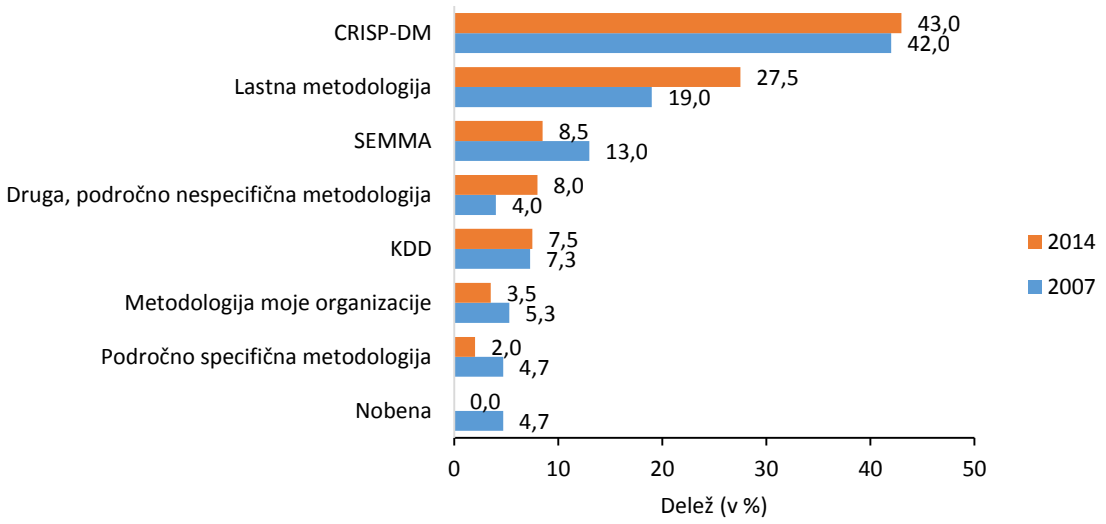
Vir: A. Azevedo & M. F. Santos. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW, 2008, str. 185.

Avtorja zaključujeta, da je tako SEMMA kot CRISP-DM mogoče razumeti kot implementacijo KDD metodologije, s tem, da je na prvi pogled CRISP-DM metodologija videti širša in celovitejša. Pri globlji analizi pa je jasno, da npr. faze vzorčenja pri SEMMA metodi sploh ne more biti brez predhodnega razumevanja vseh relevantnih vidikov, po drugi strani pa je tudi razumljivo, da uporaba znanja mora biti prisotna, saj je to osnovi razlog izvajanja podatkovnega rudarjenja.

2.5 Uporaba metodologij v praksi

Združenje KDnuggets je leta 2007 med uporabniki izvedlo anketo o uporabi metodologije pri projektih podatkovnega rudarjenja. V letu 2014 so ponovno izvedli enako anketo. Slika 6 prikazuje rezultate ankete, ki so prikazani primerjalno za obe leti.

Slika 6: Uporaba različnih metodologij pri podatkovnem rudarjenju



Vir: Piatetsky, G., CRISP-DM, still the top methodology for analytics, data mining, or data science projects, 2016

Piatetsky (2016) analizira rezultate ankete in ugotavlja, da je CRISP-DM še vedno vodilna metodologija pri projektih podatkovnega rudarjenja, vendar pa nujno potrebuje posodobitve in prilagoditve zaradi izzivov, ki jih prinašajo masovni podatki (angl. *big data*) in moderna podatkovna znanost. Piatetsky je tudi mnenja, da je CRISP-DM še vedno dober referenčni model, je pa dejstvo, da CRISP-DM nima zagotovljenega nadaljnjega razvoja, saj za njim ne stoji več organizacija, ki bi za to skrbela. Rezultat tega dejstva, ter tudi pomanjkanja nove, moderne metodologije, je po njegovem mnenju tudi občutno povečanje deleža uporabnikov, ki uporabljajo svojo lastno oz. druge, področno nespecifične metodologije.

3 STRATEGIJA UPRAVLJANJA ODNOSOV Z ODJEMALCI IN PODATKOVNO RUDARJENJE

Tsiptsis in Chorianopoulos (2009, str. 1-3) opredeljujeta pomen obstoja strategije upravljanja odnosov z odjemalci. Navajata, da so odjemalci najpomembnejši kapital organizacije in da se poslovnega uspeha ne da zagotoviti brez zadovoljnih odjemalcev, ki ostanejo zvesti organizaciji in z njo razvijejo dolgoročne odnose. Po njunem mnenju mora zato vsaka organizacija načrtovati in vpeljati jasno strategijo upravljanja odnosov z odjemalci. Upravljanje odnosov z odjemalci (angl. *Customer Relationship Management*, v nadaljevanju CRM) je tako strategija za gradnjo, upravljanje in krepitev zvestobe in posledično dolgo trajajočih odnosov z odjemalci. CRM pa mora biti v odjemalce osredotočen pristop, ki temelji na njihovem vpogledu in razumevanju. Ista avtorja poudarjata, da mora CRM strategija obsegati upravljanje odnosov z individualnimi odjemalci na osebnem nivoju, saj so tudi posamezni odjemalci različne individualne entitete, to pa je mogoče samo preko zaznavanja njihovega individualnega vedenja ter razumevanja njihovih individualnih potreb in želja.

V telekomunikacijah se sicer za odjemalce večinoma uporablja termin uporabniki. Tako tudi Agencija za komunikacijska omrežja in storitve Republike Slovenije (v nadaljevanju AKOS) v poročilu (AKOS, 2016, str. 8-42) govori o uporabnikih širokopasovnega dostopa do interneta, uporabnikih fiksne telefonije, uporabnikih mobilne telefonije itd. in ne o kupcih oz. odjemalcih, zato bom tudi sam v nadaljevanju večinoma uporabljal ta termin.

Tsiptsis in Chorianopoulos (2009, str. 1) za CRM strategijo navajata dva glavna cilja:

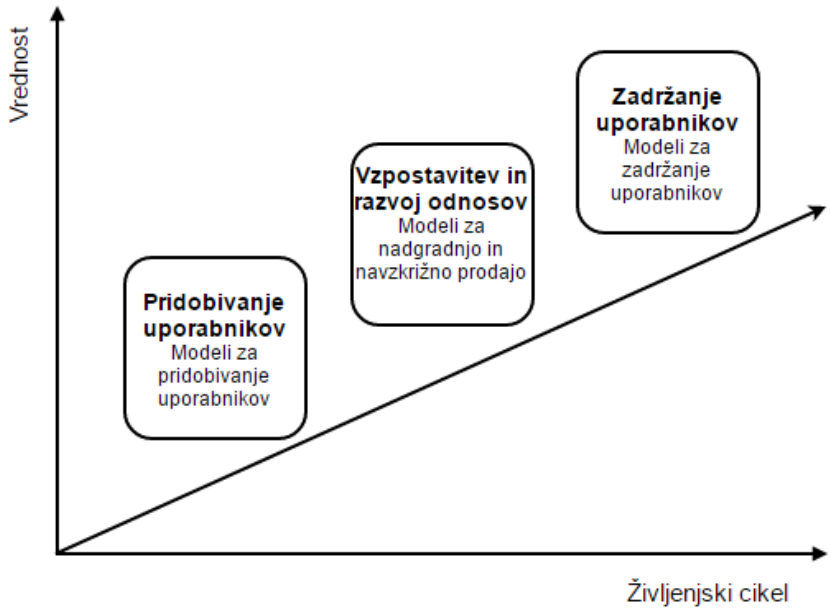
1. zadržanje odjemalcev preko njihovega zadovoljstva, ter
2. razvoj odjemalcev preko vpogleda v njihove potrebe, želje in obnašanje.

Pomembnost prvega cilja izhaja iz dejstva, da je pridobivanje novih uporabnikov težavno, kar še toliko bolj velja na zrelih trgih, kje ni novih uporabnikov. Nove uporabnike je takrat potrebno pridobivati s prevzemi od konkurenčnih podjetij, kar pa je lahko tudi do petkrat dražje, kot pa zadržanje obstoječih uporabnikov (Kotler & Keller, 2012, str. 140). Prepričevanje uporabnikov za prehod od konkurence je namreč zelo težavno in zahteva velike napore, še posebej, če so uporabniki z obstoječim ponudnikom zadovoljni.

Glede drugega cilja pa je glavno sporočilo, da povprečnega uporabnika v praksi ni. Baza uporabnikov namreč obsega osebe z različnimi individualnimi potrebami, obnašanjem in različnim potencialom, in jih je zato potrebno temu primerno individualno obravnavati. Vpogled v uporabnike in s tem poglobljeno razumevanje njihovih potreb pa omogoča ravno podatkovno rudarjenje. Interakcije z uporabniki na individualno prilagojenem nivoju povečuje njihovo zadovoljstvo in posledično dobičkonosnost.

Učinkovita CRM strategija torej pomeni optimalno prilagojeno upravljanje uporabnikov čez vse faze življenjskega cikla, od pridobivanja in vzpostavitve odnosov, njihovega razvoja in nadgradnje, do preprečevanja odhodov ali celo ponovnega pridobivanja izgubljenih uporabnikov (Slika 7).

Slika 7: Podatkovno rudarjenje in upravljanje življenjskega cikla uporabnikov



Vir: K. Tsiptsis & A. Chorianopoulos, Data Mining Techniques in CRM: Inside Customer Segmentation, 2009, str.5.

Tsiptsis in Chorianopoulos (2009, str. 2-3) CRM delita na operativni CRM in analitični CRM. Operativni CRM avtorja definirata kot različne CRM programske rešitve, ki omogočajo beleženje vhodnih in izhodnih interakcij s odjemalci, upravljanje marketinških kampanj in klicnih centrov. Teke programske rešitve tipično podpirajo prodajne procese, marketing kampanje, podporne procese, klicne centre ter omogočajo avtomatizirane komunikacije in interakcije z odjemalci. Beležijo vso relevantno zgodovino in tudi druge uporabne informacije o odjemalcih. Zagotavljajo tudi konsistenten pregled odnosov med odjemalci in podjetjem, ki je kot tak na voljo na vseh kontaktnih mestih. Vendar pa avtorja opozarjata, da so to le orodja, ki so v pomoč pri učinkovitem upravljanjem odnosov z odjemalci. Za to, da se lahko pripelje prave informacije do pravih uporabnikov, pa je potreben analitični CRM. Analitični CRM pa po opisu teh dveh avtorjev obsega analiziranje podatkov o odjemalcih, uporabo modelov podatkovnega rudarjenja in analiziranje vzorcev, vse z namenom zagotavljanja boljšega upravljanja odnosov z odjemalci. Rezultate analitičnega CRM je zato potrebno povezati z operativnim CRM, kjer so ti podatki v pomoč pri operativnem upravljanju odnosov z odjemalci. Takšna integracija je nujno potrebna, da so operativne interakcije z odjemalci sploh mogoče na osebno prilagojenem nivoju.

Tsiptsis in Chorianopoulos (2009, str. 5) navajata, da organizacije stalno stremijo k večjemu tržnemu deležu oz. k večjemu deležu porabe svojih uporabnikov (angl. *Share of Wallet*) in opisujeta naslednja področja v okviru CRM, kjer je podatkovno rudarjenje v veliko pomoč.

3.1 Segmentacija uporabnikov

Segmentacija uporabnikov je postopek razdelitve baze strank v različne, vendar na znotraj homogene skupine, ki se jih potrebuje zaradi razvoja prilagojenih trženjskih strategij glede na karakteristike posameznih skupin. Tsiptsis in Chorianopoulos (2009, str. 4) navajata več vrst segmentacij.

Vedenjska segmentacija (angl. *Behavioral Segmentation*) je tako združevanje uporabnikov glede na vzorce uporabe storitev oziroma proizvodov, čemur pravimo tudi vzorci vedenja. Avtorja pa omenjata, da se segmentacija lahko pripravi tudi s pomočjo poslovnih pravil, vendar ima ta pristop po njunem mnenju pomembne pomanjkljivosti. Segmentacija na osnovi poslovnih pravil namreč temelji na osebnem dojemanju poslovnega analitika, kar pomeni, da vprašljiva objektivnost take analize, poleg tega pa lahko posameznik učinkovito obravnava samo nekaj segmentacijskih atributov. Na drugi strani pa je mogoče s pomočjo podatkovnega rudarjenja analizirati veliko atributov in tako ustvariti s strani podatkov izpeljane vedenjske segmente oz. naravne skupine uporabnikov, ki imajo tudi globlji poslovni pomen in poslovno vrednost.

Sheme razdelitve se lahko pripravijo tudi glede na sedanjo, prihodnjo ali ocenjeno vrednost posameznih uporabnikov. Segmentacija glede na vrednost (angl. *Customer Value Segmentation*) pa tako lahko pripravi osnovo za prioritetno obravnavno tistih uporabnikov, ki so za organizacijo pomembnejši.

3.2 Kampanje neposrednega trženja

Kampanje neposrednega trženja avtorja opredeljujeta kot namensko kontaktiranje uporabnikov preko pošte, interneta, elektronske pošte, telefona ali drugih komunikacijskih kanalov, z namenom preprečevanje odhodov oz. preprečevanja prekinitve naročniškega razmerja (angl. *Customer Retention*), pridobivanja novih uporabnikov (angl. *Customer Acquisition*), ali z namenom prodaje dodatnih, več obstoječih ali bolj donosnih proizvodov oz. storitev svojim obstoječim uporabnikom (angl. *Cross-Deep-Up-selling Campaigns*).

Pri tem pa opozarjata, da v kolikor kampanje niso dobro premišljene, lahko vodijo k nepotrebnem zapravljanju virov, pri uporabnikoh pa lahko zaradi večkratnega nepotrebnega, nezaželenega kontaktiranja, dosežejo celo negativen učinek. V tej luči podatkovno rudarjenje predvsem s pomočjo klasifikacijskih modelov lahko vodi k pripravi namenskih in usmerjenih kampanj. Z analiziranjem podatkov uporabnikov se lahko odkrije različne profile

uporabnikov, ki se jih kasneje usmerjeno kontaktira. Avtorja navajata naslednje scenarije, kjer so klasifikacijski modeli uporabni za optimizacijo trženjskih kampanj:

- **Modeli pridobivanja uporabnikov**: uporabni za prepoznavanje potencialno dobičkonosnih strank na raznih seznamih potencialnih uporabnikov, tako da se glede na primerljive lastnosti išče »klone« obstoječih dobrih uporabnikov.
- **Modeli navzkrižne, poglobljene prodaje ali nadgradnje:** uporabni za odkrivanje nakupnega potenciala v obstoječi bazi uporabnikov.
- **Modeli prostovoljnih odhodov:** uporabni za odkrivanje zgodnjih znakov, ki že kažejo namen prenehanja uporabe storitev oz. nakazujejo namen prehoda h konkurenci.

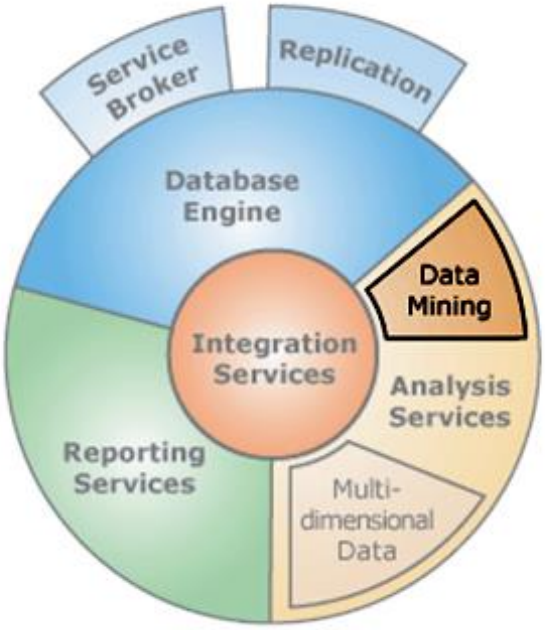
Ob pravilni uporabi taki modeli omogočajo identifikacijo primernih uporabnikov za posamezne kampanje, tako da kampanje zajemajo samo posameznike s povečano verjetnostjo določenega izida. Uspešnost tako pripravljenih kampanj je precej boljša, kot pa je uspešnost kampanj, ki so pripravljene samo na osnovi naključnega izbora uporabnikov ali pa na podlagi poslovnih pravil ali celo osebne intuicije analitikov. Meri za uspešnost modelov podatkovnega rudarjenja pravimo dvig (angl. *Lift*) in prikazuje koliko učinkovitejša je tako pripravljena kampanja v primerjavi z naključnim izborom uporabnikov.

4 MICROSOFTOV ALGORITEM ZA SEGMENTACIJO

Microsoftov algoritem za razvrščanje v skupine oz. segmentacijo (angl. *Clustering Algorithm*) je na voljo v okviru komponente za podatkovno rudarjenje, ki je ena ključnih funkcionalnosti v okviru analitičnega strežnika (angl. *Analysis* Services). Analitični strežnik pa je tudi polno integriran z integraciskimi storitvami (angl. *Integration Services*), poročilnimi storitvami (angl. *Reporting Services*) in podatkovnim strežnikom (angl. *Database Engine*). Je torej del precej širšega produkta z imenom Microsoft SQL Server 2014. Slika 8 prikazuje komponente tega produkta.

Algoritem uporablja iterativne tehnike za razvrščanje primerov (angl. *cases*) v množici podatkov v različne skupine oz. segmente (angl. *clusters*), v katerih imajo posamezni primeri medsebojno podobne lastnosti, ki pa se razlikujejo od lastnosti primerov v drugih segmentih. Tako pridobljene skupine so koristne za analiziranje podatkov, iskanje anomalij v podatkih in ustvarjanje napovedi. Algoritem odkriva relacije med podatki, na katere ni mogoče logično sklepati samo s priložnostnim opazovanjem. Algoritem spada v skupino neusmerjenih algoritmov in zato nima definiranega napovednega atributa. Je torej brez ciljne spremenljivke, ki jo napovedujemo. Model se nauči izključno iz medsebojnih razmerij, ki obstajajo v podatkih posameznih primerov in med skupinami, ki jih identificira.

Slika 8: Komponente Microsoft SQL strežnika

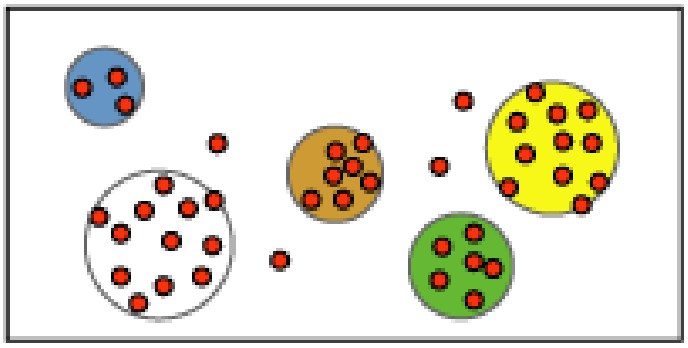


Vir: SQL Server Analysis Services - Data Mining, 2016

4.1 Opis delovanja algoritma za segmentacijo

Microsoftov algoritem za razvrščanje v skupine oz. segmentacijo najprej identificira povezave med zapisi v množici zapisov na osnovi katerih potem kreira vrsto skupin oz. segmentov. Vizualno si to lahko predstavljamo z diagramom raztrosa (Slika 9). Posamezne točke predstavljajo posamezne primere, krogi, ki obkrožajo posamezne skupine primerov, pa predstavljajo identificirane skupine, ki jih je na osnovi povezav med primeri algoritem odkril.

Slika 9: Najdene skupine primerov



Vir: Microsoft Clustering Algorithm, 2016.

Ko so posamezne skupine oz. segmenti ustvarjeni, v naslednjem koraku algoritem preračuna kako dobro ti segmenti predstavljajo tako združene primere in poskuša s ponovnim formiranjem definirati nove, boljše segmente, ki bi še bolje predstavljali lastnosti tako združenih primerov. Algoritem iterativno izvaja ta proces vse dokler s ponovnim formiranjem ne more več izboljšati dobljenih rezultatov.

Microsoftov algoritem za razvrščanje v skupine ponuja dve različni metodi razvrščanja:

- **Metoda EM** (angl. *EM Clustering*), je mehka metoda razvrščanja v skupine. Mehka metoda pomeni, da posamezen zapis oziroma primer vedno pripada večim skupinam, za vsako skupino pa je izračunana verjetnostjo članstva v tej skupini.

- **Metoda voditeljev** (angl. *K-Means Clustering*), je tako imenovana trda metoda razvrščanja v skupine. Posamezen primer pripada točno eni skupini, algoritem torej vsak primer dodeli samo eni skupini, kjer je verjetnost članstva v tej skupini točno 1, medtem ko je verjetnost članstva tega primera v drugih skupinah točno 0. Zato tej metodi tudi pravimo trda metoda razvrščanja.

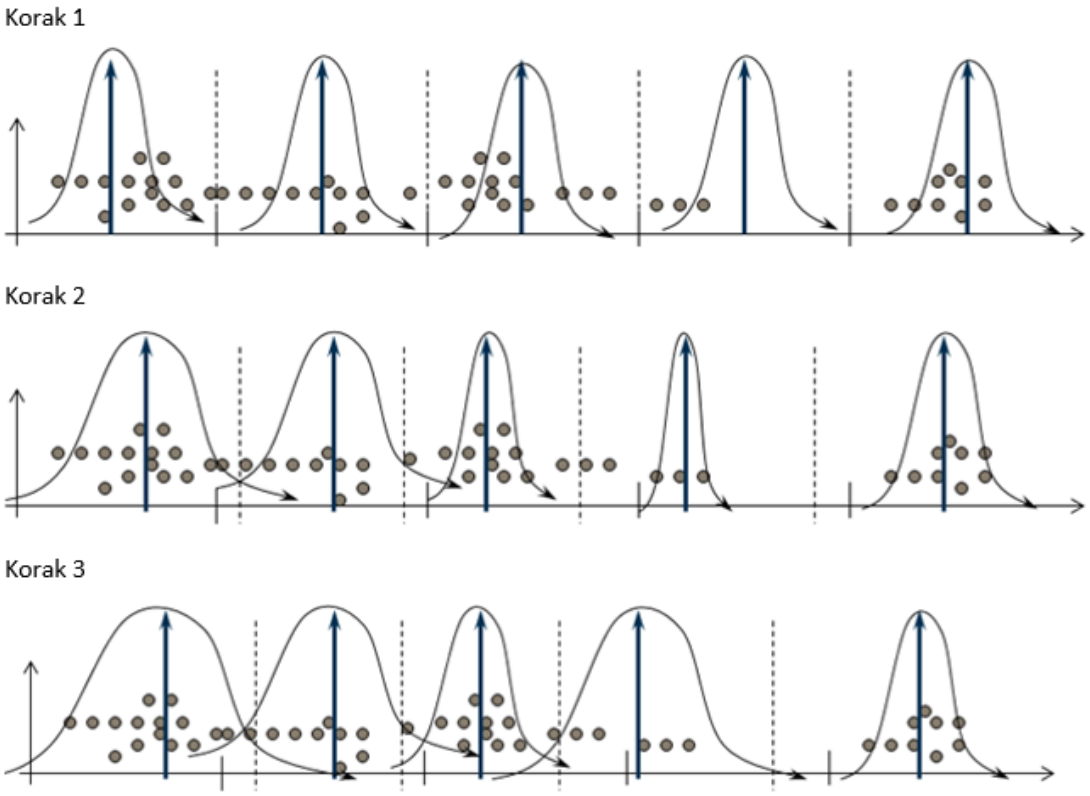
4.1.1 Metoda EM

Metoda EM namesto razdalje uporablja verjetnostno mero za določanje članstva v skupinah. Skupinam se za vsako dimenzijo določi zvonasta krivulja z aritmetično sredino in standardnim odklonom (Slika 10). Vsak primer tako pade znotraj določene krivulje in se tako določi v skupino z določeno verjetnostjo. Ker se krivulje prekrivajo, primeri padejo v različne skupine, zato tej metodi tudi pravimo mehka metoda razvrščanja.

V procesu razporejanja primerov EM algoritem iterativno ponavlja dva koraka. Postopek razvrščanja grafično prikazuje Slika 10. V prvem koraku pričakovanj (angl. *expectation step*) algoritem za vsak posamezen primer izračuna verjetnost članstva v vsaki od začetno določenih skupin. V drugem koraku maksimizacije (angl. *maximization step*) pa algoritem uporabi dejanske podatke o članstvih, da ponovno prilagodi krivulje oz. parametre porazdelitve. Proces se zaključi, ko med iteracijami ne pride več do napredka v pokritosti primerov in se tako vzpostavi končno stanje, kar prikazuje Slika 10 v koraku 3.

Če se med procesom pojavijo prazne skupine, ali če je zasedenost skupin pod določenim pragom, se za take skupine določi nove začetne točke oz. novo seme (angl. *seed*) in proces grajenja modela se na novo zažene.

Slika 10: Razvrščanje v skupine pri EM metodi



Vir: D. Sarka. Data Mining Algorithms – EM Clustering, 2015b

Rezultati EM metode so torej izračunane verjetnosti. To pomeni, da vsak primer pripada vsaki skupini, vendar ima članstvo v posamezni skupini različno verjetnost.

EM metoda je privzeta metoda v Microsoftovem algoritmu za razvrščanje v skupine, predvsem zaradi določenih prednosti v primerjavi z metodo voditeljev. Med bolj pomembnimi prednostmi je to, da zahteva samo en pregled (angl. *Scan*) podatkov in da bo ta metoda delovala tudi v primeru omejene količine sistemskega pomnilnika RAM (angl. *Random Access Memory*).

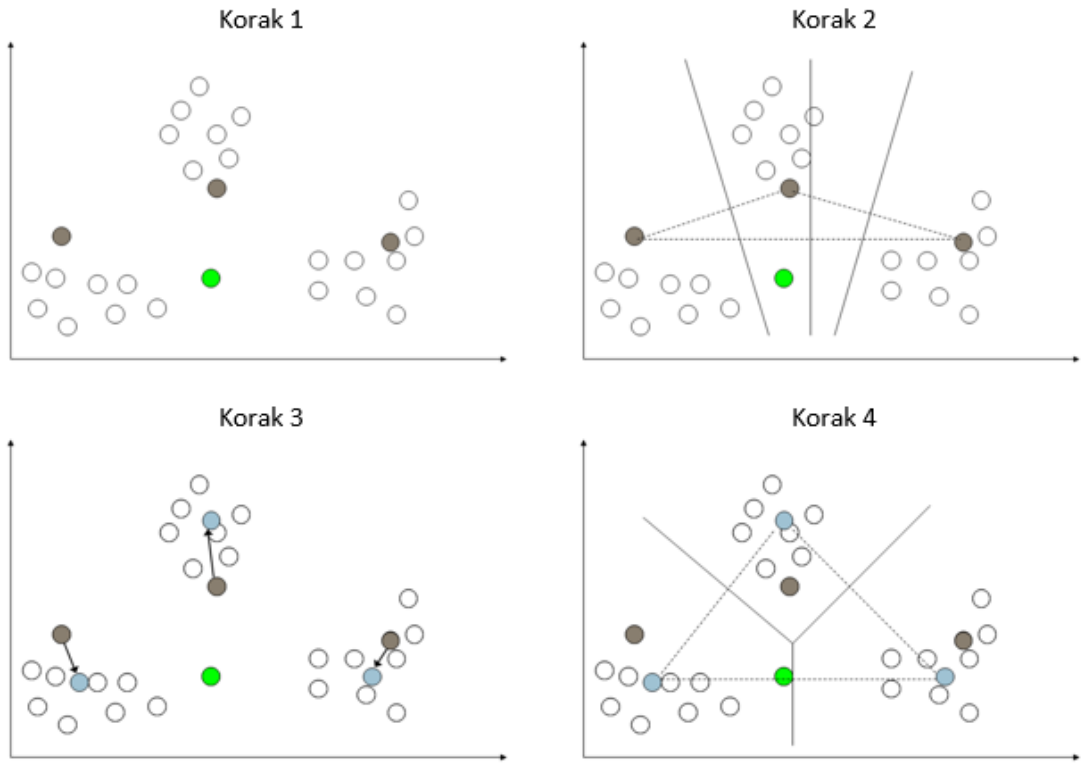
Microsoftov algoritem ponuja še dve dodatni možnosti načina graditve modela: prilagodljiv (angl. *scalable*) in neprilagodljiv (angl. *non-scalable*) EM. Privzeto uporablja prilagodljiv način. Prilagodljiv način pomeni, da algoritem vzame prvih 50.000 zapisov in poskusi zgraditi model. V kolikor je uspešen, uporabi samo te podatke. V nasprotnem primeru vzame dodatnih 50.000 zapisov. V neprilagodljivem načini pa algoritem vzame celoten set zapisov, in to ne glede na njegovo velikost. V tem načinu algoritem lahko zgradi točnejše skupine, vendar pa so zahteve po sistemskem pomnilniku lahko zelo obsežne. Iterativna obdelava podatkov je v primeru prilagodljivega načina precej hitrejša, algoritem pa tudi precej bolje izkoristi procesor, kot pa v primeru neprilagodljivega načina. Tako je ta način tudi do trikrat

hitrejši kot pa v primeru neprilagodljivega načina in to celo takrat, ko sistemski pomnilnik zadostuje za celoten set podatkov. V večini primerov pa izbira prilagodljivega načina zaradi vseh svojih performančnih prednosti, ki jih ponuja, ne vodi v slabšo točnost modela.

4.1.2 **Metoda voditeljev**

Metoda voditeljev za razvrščanje v skupine je dobro poznana metoda razvrščanja na osnovi minimizacije razlik med primeri znotraj skupine in obenem maksimizacije razlik oz. razdalje med skupinami. Beseda »means« v angleškem imenu metode se nanaša na sredine, to so posebne točke skupin, ki jim pravimo centroidi oz. voditelji. »K« v angleškem imenu metode pa se nanaša na število začetnih točk oz. semen (angl. *seeds*), ki jih algoritem arbitrarno določi na začetku za sprožitev procesa razvrščanja. Proces razvrščanja prikazuje Slika 11.

Slika 11: Razvrščanje v skupine pri metodi voditeljev



Vir: D. Sarka. Data Mining Algorithms – K-Means Clustering, 2015a

Algoritem izhaja iz geometrije, zato si lahko vizualno predstavljamo razporeditev posameznih primerov v prostoru, kjer atributi predstavljajo posamezne dimenzije. Algoritem v prvem koraku inicialno določi »K« fiktivnih točk oz. semen in jih postavi v prostor kot centroide oz. voditelje. Vrednosti dimenzij oz. atributov teh začetnih voditeljev so lahko čisto naključno določene.

Slika 11 prikazuje začetne voditelje (semena) kot temne točke. Algoritem nato v drugem koraku dodeli vsak primer najbližjemu voditelju in na ta način določi začetne skupine. Ko so te začetne skupine določene, algoritem lahko izračuna dejanske voditelje skupin, kot je to prikazano v koraku 3. V koraku 4 se ponovno razporedi posamezne primere v skupine glede na izračunane voditelje. Posamezni primeri ob tem preskočijo v druge skupine, kar je prikazano s primerom točke zelene barve. Zaradi teh preskokov algoritem ponovno izračuna nove voditelje in ponovno razporedi vse primere v skupine. Pri razvrščanju se torej iterativno ponavlja koraka 3 in 4 vse dokler primeri nehajo prehajati med skupinami in se zato sestava skupin ne spreminja več.

Metoda voditeljev za razliko od metode EM dodeli vsak primer v točno eno skupino in ne dovoljuje nobenega dvoma o članstvu v tej skupini. Članstvo v skupini je izraženo kot oddaljenost posameznega primera od voditelja.

Tipično se metodo voditeljev uporablja za izdelavo modelov v primeru zveznih atributov, ker je pri njih relativno preprosto izračunati oddaljenost od voditelja. Vendar pa je Microsoftov algoritem prilagojen, da omogoča uporabo metode voditeljev tudi za diskretne atribute.

Tudi pri metodi voditeljev imamo tako kot pri EM metodi na voljo dva načina vzorčenja podatkov. Prvi je neprilagodljiv, ki prebere vse zapise v množici podatkov, drugi pa prilagodljiv, ki privzeto prebere samo prvih 50.000 zapisov. Samo v primeru, da prvih 50.000 zapisov ne omogoča izgradnje dobrega modela, algoritem uporabi še naslednjih 50.000 zapisov.

4.2 Prilagoditve Microsoftovega algoritma za razvrščanje v skupine

Microsoftov algoritem za razvrščanje skupine omogoča različne nastavitve, ki vplivajo na obnašanje, performance in natančnost podatkovnega modela. V nadaljevanju bom opisal parametre, ki so na voljo v verziji strežnika Microsoft SQL Server 2014 standardne izdaje, ki sem jo uporabljal in ki je glede nastavitvenih parametrov nekoliko omejena.

4.2.1 **Nastavitev parametrov algoritma**

CLUSTERING_METHOD

Ta parameter omogoča izbiro ene od štirih metod razvrščanje, ki so na voljo (Tabela 3). Metode sem podrobneje opisal v poglavju 4.1.

Tabela 3: Metode razvrščanja v skupin

ID	Metoda
1	prilagodljivi EM
2	neprilagodljivi EM
3	prilagodljivi K-Means
4	neprilagodljivi K-Means.

Vir: Microsoft Clustering Algorithm Technical Reference, 2016

Privzeta metoda je 1 (prilagodljivi EM).

CLUSTER_COUNT

S tem parametrom določimo število skupin, ki naj jih algoritem določi. Če tako določeno število skupin ne more biti ustvarjeno, algoritem naredi toliko skupin, kot je mogoče z danimi podatki. Če ta parameter nastavimo na vrednost 0, s tem algoritmu določimo, da uporabi hevristični način za določanje najprimernejšega števila skupin.
Privzeta nastavitev je 10.

MINIMUM_SUPPORT

S tem parametrom določimo minimalno število primerov, ki jih mora vsebovati posamezna skupina. Če je število primerov v neki skupini manjše od tukaj določene vrednosti, se ta skupina obravnava kot prazna in se zato zavrže. V kolikor tu določimo previsoko število, lahko s tem izgubimo povsem legitimne skupine.

Privzeta nastavitev je 1.

MODELLING_CARDINALITY

S tem parametrom se določi število vzorčnih modelov, ki so zgrajeni med procesom razvrščanja. Z manjšim številom se sicer lahko izboljša performance, vendar je to ob tveganju, da s tem zgrešimo kakšne dobre vzorčne modele.

Privzeta nastavitev je 10.

STOPPING_TOLERANCE

Parameter določa vrednost, ki se uporablja za ugotavljanje, kdaj je dosežena konvergenca in zato algoritem preneha z izgradnjo modela. Konvergenca je dosežena, ko je skupna sprememba verjetnosti segmentov manjša od razmerja tega parametra in velikosti modela.

Privzeta nastavitev je 10.

4.2.2 Nastavitev vrste atributov

Microsoftov algoritem za razvrščanje v skupine omogoča različne nastavitve vrste atributov. Obvezen atribut je atribut za ključ (angl. *Key Attribute*), ki je enolični identifikator posameznega primera. Vhodni atributi (angl. *Input Attribute*), so vsi atributi, ki se uporabljajo pri gradnji modela in so tudi obvezni. Atributi z oznako »Predict Only« (angl. *Predictable Attribute*) pa so opcijski, njihova posebnost pa je v tem, da se ne uporabljajo pri izgradnji modela ampak je distribucija teh vrednosti v skupine narejena šele po tem, ko so te že formirane.

5 IZVEDBA PODATKOVNEGA RUDARJENJA V PRAKSI

Praktični primer je razvrščanje uporabnikov v skupine je bil izveden na realnih podatkih podjetja, ki je eden od večjih operaterjev mobilne telefonije v Sloveniji.

V današnjem času imajo uporabniki mobilne telefonijo na voljo velik nabor storitev, saj je čas, ko je mobilna tehnologija omogočala le uporabo govornih klicev, že zdavnaj minil. AKOS tako v svojem kvartalnem poročilu (AKOS, 2016, str. 15-17) opisuje, da moderni način življenja vse bolj pogojuje odvisnost od mobilnih telefonov in da si mnogi uporabniki življenja brez mobilnih telefonov ne predstavljajo več. Ponudba storitev, namenjenih uporabnikom mobilne telefonije, je po navedbi AKOS zelo raznolika, saj uporabniki že dolgo niso več zadovoljni samo z govornimi in sporočilnimi storitvami. Njihove potrebe so vedno večje, zato povprašujejo tudi po različnih drugih storitvah, med katerimi je na prvem mestu dostop do interneta. AKOS tudi navaja, da je posledično mobilna telefonija prevzela vodilno vlogo pri uvajanju novih storitev, vse več uporabnikov pa poleg komunikacijskih storitev uporablja mobilni telefon tudi zaradi možnosti fotografiranja, predvajanja glasbe, dostopa do interneta, pošiljanje elektronske pošte, nalaganje in uporabo različnih mobilnih aplikacij itd.

Nekateri uporabniki mobilne telefone uporabljajo le občasno, drugi pa so praktično odvisni od stalne dosegljivosti ali stalnega dostopa do informacij in si zato normalnega življenja brez mobilnih telefonov ne morejo več predstavljati. Vse to namiguje na različne vzorce uporabe storitev pri različnih uporabnikih.

Tržno okolje mobilne telefonije je v Sloveniji zelo konkurenčno, saj je na trgu mobilne telefonije konec leta 2015 delovalo kar sedem operaterjev mobilne telefonije, stopnja penetracije aktivnih uporabnikov na prebivalstvo pa je dosegla 114 odstotkov (AKOS, 2016, str. 17). Stopnja penetracije se računa kot število aktivnih uporabnikov glede na število prebivalstva, AKOS za aktivne uporabnike šteje vse naročnike z veljavno pogodbo in vse

predplačnike, ki so v zadnjih treh mesecih opravili ali prejeli klic, poslali SMS, MMS sporočilo oz. uporabili podatkovne storitve. uporabili katero drugo storitev.

Tsiptsis in Chorianopoulos (2009, str. 291) navajata, da se v takem okolju močne konkurence in hitrih sprememb ni dovolj osredotočati le na pridobivanje novih strank (angl. *Customer Acquisition*), ker to postaja vedno težje. Organizacije se morajo takrat neizogibno osredotočati tudi na zadržanje svojih uporabnikov (angl. *Customer Retention*) in na pridobivanja večjega deleža porabe s strani svojih obstoječih uporabnikov, kot pa le na večji tržni delež. Taka notranja rast je po njunem mnenju lažje dosegljiva. Zadovoljni in zvesti uporabniki so ključ do poslovnega uspeha. Za doseganje tega cilja se morajo ponudniki osredotočati v svoje uporabnike in razumevanje njihovih potreb, njihovega obnašanja in njihovih želja, vedenjska segmentacija uporabnikov pa je eden ključnih pristopov, ki tako razumevanje omogoča.

Na srečo se v telekomunikacijah podatki o uporabi storitev zelo podrobno beležijo v obliki posebnih zapisov o uporabi storitev oz. CDR zapisov (angl. *Call Detail Record*), ki se obdelujejo v procesu zaračunavanja porabe. Ko so ti podatki obdelani in agregirani na primernem nivoju, pa predstavljajo tudi zelo dober vir za druge analize.

Ker so podatki o porabi storitev in proces njihove obdelave ključni za razumevanje izvornih podatkov za analizo, bom v nadaljevanju oboje nekoliko podrobneje opisal. Opis je povzetek s spletne strani Telecom Billing Guide (Telecom Billing Guide, 2015).

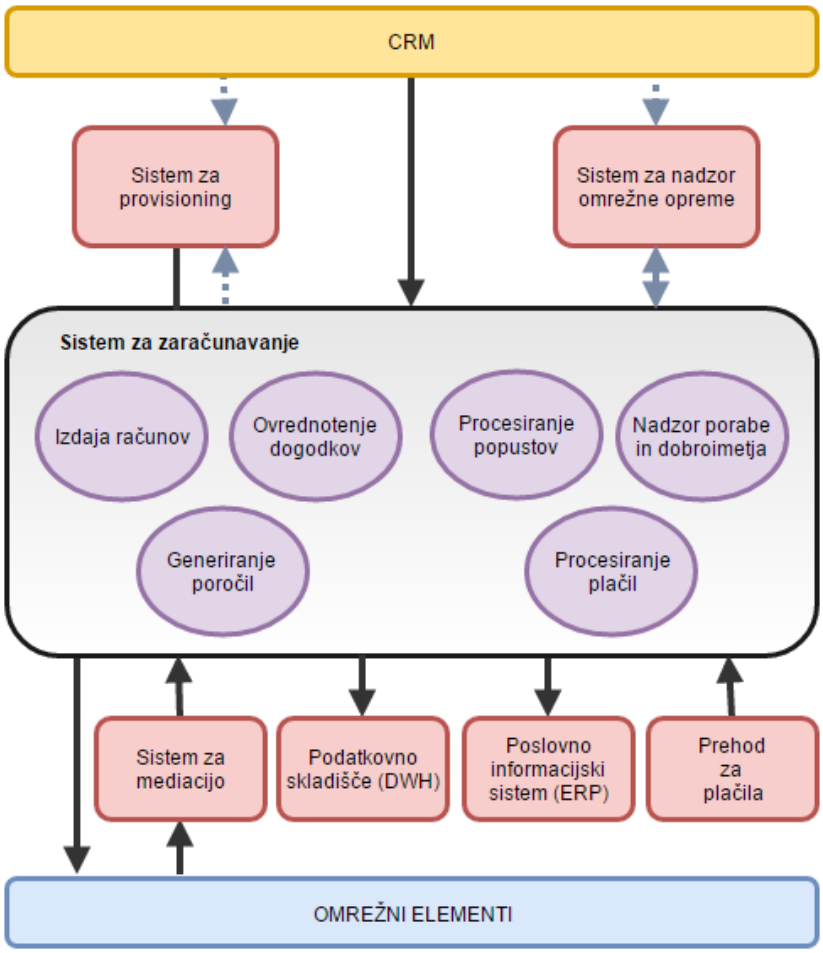
5.1 Zaračunavanje porabe

Zaračunavanje porabe (angl. *Billing*) je proces zbiranja podatkov o uporabi storitev, njihovega združevanja in obdelave z namenom izdaje računov. Sistemi za zaračunavanje (angl. *Billing Systems*) pa so kompleksne programske rešitve, ki učinkovito izvajajo različna z obračunom povezana, ponudnikom storitev pa omogočajo veliko mero prilagodljivosti, da lahko ponujajo svoje storitve v različnih prilagodljivih paketih. Slika 12 prikazuje arhitekturo takega sistema.

Ko uporabnik uporabi katero od storitev mobilne telefonije, se na omrežnih elementih generirajo podatki o uporabi storitve. Omrežni elementi (npr. telefonska centrala, SMS center, MMS center) so kombinacija programske in strojne opreme, njihov namen pa je nadzor in beleženje dogodkov za določen tip storitve. Dogodek je vsak posamezen primer uporabe storitve, ki se zabeleži na omrežju in je kasneje podvržen obračunu. Ko uporabnik opravi npr. telefonski klic, se zabeleži dogodek, ki vsebuje podatek o vrsti storitve (v tem primeru je to govorni klic), kot tudi parametre o dolžini klica, času klica, izvorni številki, klicani številki, itd. Mediacijski sistem od omrežnih elementov pridobi podatke o podrobnostih dogodkov in jih preoblikuje ter shrani v obliki CDR zapisov. Oblika CDR

zapisov je povsem prilagojena sistemu za zaračunavanje, kjer se ti zapisi v nadaljevanju ovrednotijo (angl. *Call Rating*). Pri vrednotenju CDR zapisov se npr. preveri, če gre za lokalni klic, mednarodni klic, klic na posebne komercialne številke, čas klica itd. in se določi vrednost tega dogodka.

Slika 12: Arhitektura sistema za zaračunavanje



Vir: Telecom Billing Guide, 2015.

Ovrednoteni dogodki se shranijo, saj se potrebujejo za izdajo računa, ki se predvidoma zgodi enkrat mesečno. Pri izdaji računov se dodajo še druge zaračunane postavke, ki niso povezane z uporabo storitev, kot npr. mesečna naročnina, popusti, obroki za nakup opreme itd.

5.2 Zapisi o uporabi storitev

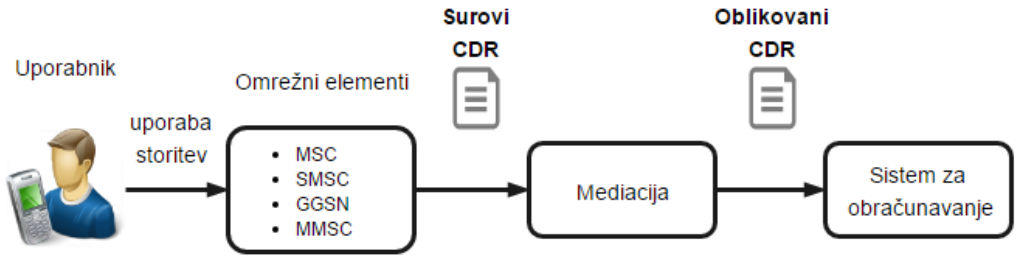
CDR zapis je opis zaračunljivega dogodka (klica) z vsemi atributi, ki so potrebni za obračun. Različni omrežni elementi zbirajo podatke o uporabi storitev in ob tem generirajo različne vrste zapisov CDR. Ker pa gre v novejšem obdobju za različne tipe storitev in ne samo za

klice, se lahko ekvivalentno uporablja tudi kratica UDR (angl. *Usage Detail Record*). V primeru mobilne telefonije so najpogostejši omrežni elementi:

- MSC sistem (angl. *Mobile Switching Centre*) za nadzor govornih klicev,
- SMSC sistem (angl. *SMS Centre*) za beleženje prometa besedilnih SMS sporočil (angl. *Short Message Service*),
- GGSN sistem (angl. *Gateway GPRS support node*) za nadzor podatkovnega prometa,
- MMSC sistem (angl. *MMS Centre*) za beleženje prometa multimedijskih sporočil (angl. *Multimedia Message Service*), v nadaljevanju MMS sporočil, ter
- podatki o gostovanju (angl. *Roaming*), ki se beležijo preko tujih operaterjev in njihovih omrežnih elementov.

Surovi CDR zapisi, ki jih generirajo omrežni elementi, se preko mediacijskega sistema pretvorijo v obliko, ki je prilagojena sistemu za obračunavanje storitev. Proces zajema CDR podatkov prikazuje Slika 13.

Slika 13: Zajem podatkov o porabi



Vir: Telecom Billing Guide, 2015.

CDR zapis vsebuje podatke uporabi zapisov in tudi druge uporabne informacije. Obvezni atributi CDR zapisa so:

- odhodna številka (A številka),
- dohodna številka (B številka),
- začetek klica (datum in čas),
- dolžina klica (trajanje),
- vrsta klica (glasovni klic, SMS, podatkovni klic), ter
- unikatni identifikator zapisa.

CDR zapise je potrebno po obdelavi za potrebe obračunavanja v skladu z določbami Zakona o elektronskih komunikacijah (Ur.l. RS, št. 109/2012), izbrisati oz. spremeniti tako, da se jih ne da več povezati z določeno ali določljivo osebo. V konkretnem primeru podjetja se CDR zapisi za potrebe obračuna hranijo tri mesece, nato pa se ti podatki brišejo. Podatki o porabi,

ki so bili za potrebe analize pridobljeni v podatkovnem skladišču, pa so bili za posamezni uporabniški račun že agregirani na mesečnem nivoju, kar pa je za potrebe vedenjske segmentacije uporabnikov povsem primeren nivo agregacije.

5.3 Razumevanje poslovanja

V praktičnem delu predstavljam primer segmentacije predplačniških uporabnikov mobilne telefonije. Za potrebe trženja želi podjetje na osnovi vedenja oz. vzorcev uporabe storitev te uporabnike razdeliti v različne segmente, saj bi to omogočalo:

- identifikacijo različnih potreb uporabnikov in nato razvoj njim prilagojenih produktov in storitev, kar bi vodilo v povečanje uporabe storitev s strani obstoječih uporabnikov, lahko bi pa tudi privabilo nove uporabnike s strani konkurence, ter
- pripravo prilagojenih trženjskih strategij za posamezne segmente.

Poslovni cilj je torej razvoj prilagojenih produktov ter njihovo prilagojeno trženje za različne segmente uporabnikov, ter s tem doseganje večje notranje rasti. Poslovni cilj pa je v nadaljevanju potrebno preslikati v cilj podatkovnega rudarjenja. Cilj podatkovnega rudarjenja je tako razdelitev enotne baze uporabnikov na različne segmente ter opis njihovih lastnosti.

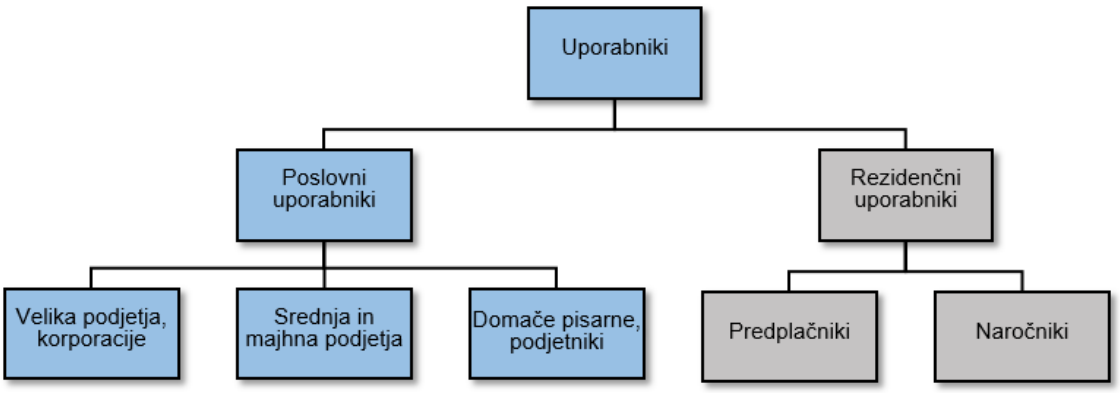
Tsiptsis in Chorianopoulos (2009, str. 292-294) opisujeta, da se v mobilni telefoniji tipično kategorizira uporabnike mobilne telefonije glede na vrsto naročniškega razmerja oz. vrsto poslovnega odnosa v tako imenovane jedrne segmente, ki so nato izhodišče za nadaljnjo segmentacijo. To razdelitev bom v nadaljevanju opisal, grafično pa jo prikazuje Slika 14.

Prvi nivo segmentacije predstavlja razdelitev na poslovne in rezidenčne uporabnike oz. fizične osebe. Rezidenčni uporabniki se nadalje delijo na predplačnike in naročnike. Naročniki so uporabniki s pogodbenim razmerjem, na podlagi katerega se jim storitve zaračunavajo na mesečnem nivoju glede na preteklo porabo in običajno tudi predstavljajo večino uporabnikov. Predplačniki pa nimajo pogodbenega razmerja, storitve pa zakupijo v naprej v obliki dobroimetja, ki se jim potem sproti obračunava glede na uporabo storitev. Poslovne uporabnike pa se v drugem nivoju segmentacije deli še glede na velikost podjetja na velika podjetja in korporacije, srednja in majhna podjetja, ter na domače pisarne in podjetnike.

Pri tem ta avtorja navajata to segmentacijo kot osnovo, na podlagi katere se lahko izvaja še podrobnejšo segmentacijo uporabnikov, se pa je pri tem potrebno osredotočiti na vsak posamezni jedrni segment posebej. Posamezni jedrni segmenti uporabnikov predstavljajo tako različne skupine uporabnikov, da zahtevajo prilagojen pristop. Pri predplačnikih je pri vedenjski analizi potrebno upoštevati tudi atribute kot so pogostost in vrednost polnitev

uporabniškega računa, ki pa jih pri drugih segmentih sploh ni na voljo. Po drugi strani pa so predplačniki anonimni uporabniki, saj so brez pogodbenega razmerja, niti se jim ni potrebno kako drugače registrirati, posledično pa za njih ne obstajajo atributi iz sklopa demografskih podatkov, kot so npr. spol in starost. Poslovni uporabniki tudi predstavljajo povsem drugačen tržni segment kot so recimo rezidenčni uporabniki.

Slika 14: Jedrni segmenti uporabnikov v mobilni telefoniji



Vir: K. Tsiptsis & A. Chorianopoulos, Data Mining Techniques in CRM: Inside Customer Segmentation, 2009, str. 293

Tako se tudi v praktičnem delu diplomske naloge pri segmentacija uporabnikov osredotočam samo na en jedrni segment uporabnikov, na segment predplačnikov.

5.4 Razumevanje podatkov

Za potrebe analize obnašanja uporabnikov sem iz podatkovnega skladišča potreboval podatke o uporabi različnih storitev in podatke o uporabniških računih. Podatki o uporabi storitev so bili za različne storitve shranjeni v različnih tabelah, in sicer v tabeli s podatki o odhodnih klicih (Tabela 4), tabeli s podatki o dohodnih klicih (Tabela 5) in tabeli s podatki o ostalih odhodnih storitvah (Tabela 6). Podatki so bili agregirani na mesečnem nivoju na osnovi porabe v določenem koledarskem mesecu.

Poleg podatkov o uporabi storitev sem potreboval tudi podatke o uporabniških računih, saj je bilo potrebno analizo omejiti le na segment predplačnikov. Podatki o uporabniških računih so bili na voljo v podatkovnem skladišču kot mesečni zajem stanja (angl. *Snapshot*) uporabniških računov. Tabela je vsebovala še ogromno drugih sistemskih atributov, ki pa za potrebe analize niso bili potrebni. Omejil sem se samo na tiste podatke, ki sem jih potreboval za določitev predplačniških računov, saj so bili le ti predmet analize. Uporabljene atribute o uporabniških računih, ki sem jih potreboval za pripravo podatkovne strukture, prikazuje Tabela 7.

Tabela 4: Podatki o odhodnih klicih

Atribut	Opis atributa
YEARMONTH	ID obdobja
ACCOUNT_ID	ID uporabniškega računa
MO_MINUTES_ONNET	Količina odhodnih minut v lastnem omrežju
MO_COUNT_ONNET	Število odhodnih klicev v omrežju
MO_MINUTES_MOBILE	Količina odhodnih minut v druga mobilna omrežja
MO_COUNT_MOBILE	Število odhodnih klicev v druga mobilna omrežja
MO_MINUTES_FIX	Količina odhodnih minut v fiksna omrežja
MO_COUNT_FIX	Število odhodnih klicev v fiksna omrežja
MO_MINUTES_INTL	Količina odhodnih minut v tujino
MO_COUNT_INTL	Število odhodnih klicev v tujino
MO_MINUTES	Skupna količina odhodnih minut
MO_COUNT	Število vseh odhodnih klicev
DATEINSERTED	Datum vnosa zapisa

Tabela 5: Podatki o dohodnih klicih

Atribut	Opis atributa
YEARMONTH	ID obdobja
ACCOUNT_ID	ID uporabniškega računa
MT_MINUTES_ONNET	Količina dohodnih minut iz lastnega omrežja
MT_COUNT_ONNET	Število dohodnih klicev iz lastnega omrežja
MT_MINUTES_MOBILE	Količina dohodnih minut iz drugih mobilnih omrežij
MT_COUNT_MOBILE	Število dohodnih klicev iz drugih mobilnih omrežij
MT_MINUTES_FIX	Količina dohodnih minut iz fiksnih omrežij
MT_COUNT_FIX	Število dohodnih klicev iz fiksnih omrežij
MT_MINUTES_INTL	Količina dohodnih minut iz tujine
MT_COUNT_INTL	Število dohodnih klicev iz tujino
MT_MINUTES	Skupna količina dohodnih minut
MT_COUNT	Število vseh dohodnih klicev
DATEINSERTED	Datum vnosa zapisa

Tabela 6: Podatki o ostalih odhodnih storitvah

Atribut	Opis atributa
YEARMONTH	ID obdobja
ACCOUNT_ID	ID uporabniškega računa
MO_SMS_COUNT_NAT	Število poslanih SMS sporočil v nacionalnem prometu
MO_SMS_COUNT_INTL	Število poslanih SMS sporočil v tujino
MO_SMS_COUNT	Skupno število poslanih SMS sporočil
MO_MMS_COUNT_NAT	Število poslanih MMS sporočil v nacionalnem prometu
MO_MMS_COUNT_INTL	Število poslanih MMS sporočil v tujino
MO_MMS_COUNT	Skupno število poslanih MMS sporočil
GPRS_COUNT	Število podatkovnih klicev
GPRS_MBYTES	Količina prenesenih podatkov (v MB)
DATEINSERTED	Datum vnosa zapisa

Tabela 7: Podatki o uporabniških računih

Atribut	Opis atributa
REFRESH_DATE	Datum stanja
YEARMONTH	ID obdobja
ACCOUNT_ID	ID uporabniškega računa
LAST_TP	Paket
ACC_STATUS	Status računa
PREPAID_STATUS	Predplačniški status računa

Ker pa je bilo podatkovno skladišče postavljeno na Oracle platformi, sem se že na začetku soočil s problemom prenosa podatkov v Microsoft okolje, saj sem za vse izvoze in obdelave podatkov uporabljal Microsoftova orodja. Microsoftovo orodje za integracijo podatkov SSIS (angl. *Sql Server Integration Services,* v nadaljevaju SSIS) pa v osnovi ne omogoča neposredne povezave na verzijo Oracle podatkovnega strežnika, ki je bila v uporabi. Možno bi bilo sicer podatke izvoziti, npr. v tekstovne datoteke in te nato uvoziti v Microsoft podatkovno bazo, vendar pa že zaradi količine podatkov to ni bila praktična rešitev. Poleg tega pa sem želel pripraviti direktno integracijo tudi zaradi morebitnih potreb po ponovnih izvozov podatkov v kasnejših mesecih. V ta namen sem za integracijo analitične baze in podatkovnega skladišča pripravil namensko proceduro prenosa podatkov v obliki SSIS paketa. Problem povezljivosti pa sem rešil z namestitvijo dodatnih Oracle ODAC komponent (angl. *Oracle Data Access Components*), v sklopu katerih so bili tudi gonilniki za Oracle podatkovni dostop iz Microsoft Visual Studio okolja.

5.5 Priprava podatkov

Podatki se v podatkovnih bazah večinoma nahajajo v surovi obliki in kot taki niso primerni za analize, ker so nepopolni in vsebujejo veliko šuma. Šum v podatkih predstavljajo naključne napake oz. variabilnost vrednosti opazovanega atributa (Han et. al., 2012, str. 89). Podatki lahko zajemajo atribute, ki so zastareli ali nepotrebni, manjkajoče vrednosti, odstopanja oz. ekstremne vrednosti ali pa so v obliki, ki ni primerna za podatkovno rudarjenje.

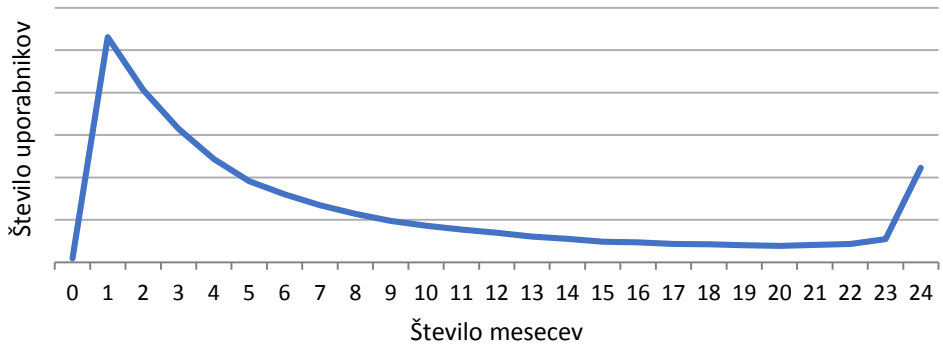
Pred izvedbo podatkovnega rudarjenja jih je torej potrebno pripraviti, kar zajema med drugim čiščenje in preoblikovanje podatkov, s čimer se poskuša zmanjšati GIGO učinek (angl. *Garbage In – Garbage Out*). GIGO učinek poenostavljeno pomeni, da se ne more pričakovati kvalitetnih rezultatov modela na osnovi nekvalitetnih podatkov na vhodu v model. Odvisno od izvornih podatkov pa lahko priprava podatkov zajema od 10 do 60 odstotkov celotnega časa v okviru procesa podatkovnega rudarjenja (Larose, D. T. & Larose, C. D., 2014, str. 17).

Izvorne podatke sem pridobil iz podatkovnega skladišča, kjer se že ob procesih polnjenja podatkovnega skladišča (angl. *Extract, Transform and Load*, v nadaljevanju ETL) izvaja tudi sprotno preverjanje prenešenih podatkov. Morebitne napake se tako že sproti popravljajo oz. se lahko v skrajnem primeru zaradi napak ETL proces tudi ponovi. Izvorni podatki so bili tudi že pretvorjeni v obliko mesečnih agregiranih vrednosti za posamezne uporabniške račune, kar je bil za potrebe te analize povsem primeren nivo agregacije. Oblika izvornih podatkov in njihova kvaliteta je tako pozitivno vplivala na čas, ki je bil potreben za pripravo podatkov.

Za izvedbo vedenjske analize sem prenesel podatke za dvoletno obdobje, saj sem na podlagi statistike predplačniških računov ocenil, da je to najbolj primerno obdobje. Pri tem je pomembno poudariti, da imajo predplačniški računi lahko zelo sporadične vzorce uporabe, saj je bil posamezen predplačniški račun v tem obdobju lahko v različnih statusih, ki so vplivali na možnost uporabe storitev. Status predplačniškega računa je odvisen od preteklega časa od zadnje polnitve računa. Samo v statusu »aktiven« ima uporabnik možnost nemotene uporabe tako odhodnih, kot tudi dohodnih storitev, vsi ostali statusi pa pomenijo vsaj blokado odhodnih storitev, ali pa celo blokado vseh storitev in tako pomenijo vsaj določeno omejitev uporabe storitev. Posamezni meseci z nizko porabo ali pa porabo celo brez porabe tako pri predplačniških uporabnikih niso nobena redkost.

Poleg tega so uporabniki v obravnavanem časovnem obdobju lahko ne glede na status aktivno uporabljali storitve različno število mesecev. Distribucijo števila mesecev aktivne uporabe storitev v obravnavanem obdobju prikazuje Slika 15. Da je uporabnik v določenem koledarskem mesecu aktivno uporabljal storitve, sem določil s kriterijem, da je moral v tem mesecu opraviti vsaj en odhodni ali dohodni klic, vsaj en odhodni SMS oz. MMS ali pa uporabiti prenos podatkov.

Slika 15: Porazdelitev mesecev aktivne uporabe storitev



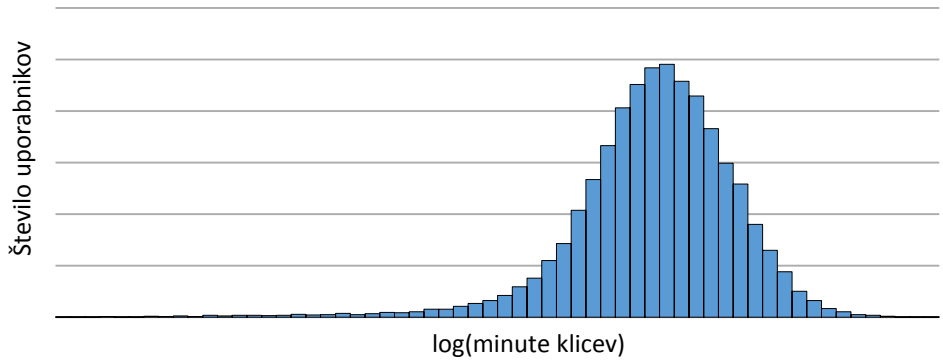
Da bi zagotovil medsebojno primerljivost uporabnikov z različnim obdobjem aktivnosti, sem skupno porabo posameznega računa delil s številom aktivnih mesecev in tako zagotovil

primerljivost na osnovi povprečne mesečne porabe. V analizo pa sem zajel samo uporabniške račune z minimalno tremi meseci aktivne uporabe storitev, saj povprečne mesečne vrednosti za krajše obdobje ne bi bile smiselne, poleg tega pa je potrebno vsaj neko minimalno obdobje, da se pri uporabniku izoblikuje reprezentativen vzorec uporabe storitev.

5.5.1 **Čiščenje podatkov**

Han et al. (2012, str. 88) opredeljuje čiščenje podatkov kot aktivnosti, s katerimi želimo odpraviti nepopolnosti izvornih podatkov, torej odpraviti manjkajoče vrednosti, zgladiti šum, odkriti odstopanja in odpraviti nedoslednosti v podatkih. Kljub temu, da so bili podatki pridobljeni iz podatkovnega skladišča in zato do določene more že preoblikovani in prečiščeni, sem še sam preveril porazdelitev vrednosti posameznih atributov. Določene vrednosti, med katerimi je bila npr. količina odhodnih klicev, so se izrazito logaritemsko normalno porazdeljevale (Slika 16).

Slika 16: Porazdelitev količine odhodnih klicev

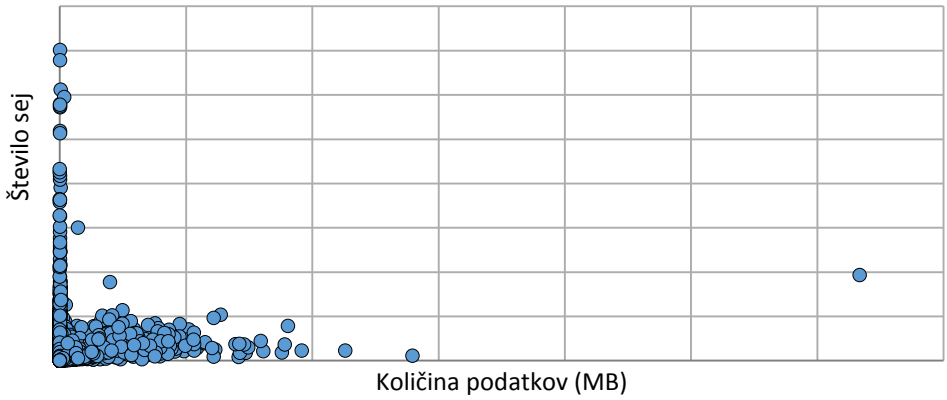


Pri porazdelitvi količine prenosa podatkov pa je bilo opaziti kar nekaj odstopanj oz. ekstremnih vrednosti (angl. *Outliers*), kar se dobro vidi v diagramu raztrosa, ki ga prikazuje Slika 17. Odstopanja so ekstremne vrednosti, ki so porazdeljene v nasprotju s trendom preostalih podatkov, njihova identifikacija pa je zelo pomembna, ker lahko predstavljajo napake v podatkih, pa tudi v primeru, ko gre za veljavne vrednosti, lahko te vodijo k nezanesljivim rezultatom, ker so določene statistične metode zelo občutljive na prisotnost ekstremnih vrednosti (Larose , D. T., & Larose, C. D., 2014, str. 22).

Slika 17 tako prikazuje kar nekaj odstopanj v količini prenesenih podatkov, kar predstavljajo primeri na desni strani grafikona. Še večjo pozornost pa so vzbudili primeri z velikim številom podatkovnih sej ob relativno majhni količini prenesenih podatkov. Porazdelitev teh primerov je bila na prvi pogled zelo nenavadna, saj je glede na vrednosti izgledalo, da je prišlo do napake pri ETL proceduri in da sta bili količini za število sej in količino prenesenih podatkov zamenjani. Se je pa kasneje izkazalo, da gre v teh primerih za regularne vrednosti,

veliko število vzpostavljenih sej ob hkratnem majhnem prenosu podatkov pa je bilo povezano z načinom delovanja določenih mobilnih aplikacij. Podatka za število sej sicer v modelu nisem uporabil, sem pa kljub temu te primere ekstremnih odstopanj odstranil, kar je predstavljalo 0,2 odstotka primerov celotne populacije.

Slika 17: Diagram raztrosa glede na število sej in prenesene količine podatkov

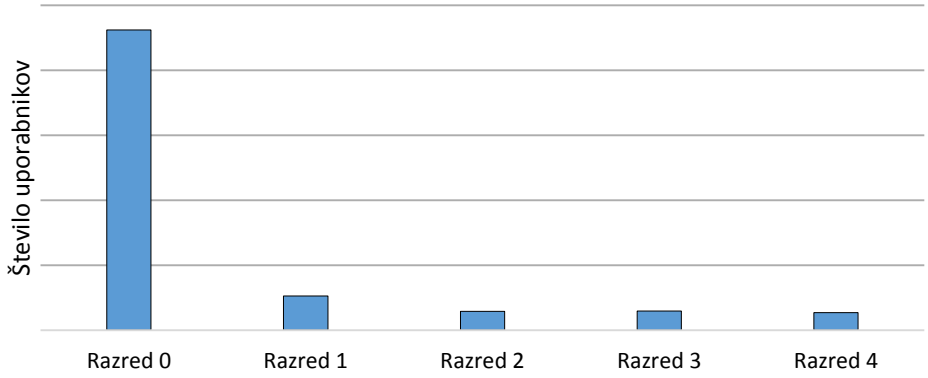


5.5.2 **Integracija in transformacija podatkov**

Pri transformaciji podatkov gre za preoblikovanje podatkov v obliko, ki je primerna za podatkovno rudarjenje. Podatke sem iz različnih izvornih tabel zapisal v enotno tabelo, kjer je posamezen zapis predstavljal posamezen uporabniški račun. Identifikator posameznega računa je predstavljal ključ, podatki o uporabi storitev pa so bili zapisani kot atributi tega zapisa. Atributi so bili v obliki povprečnih mesečnih vrednosti za posamezen račun.

Han et. al. (2012, str. 89) predlagajo odpravo šuma z različnimi tehnikami glajenja podatkov (angl. *Smooting*), med katerimi navajata tudi razdelitev v razrede (angl. *Binning*). Ta tehnika se uporablja tudi za diskretizacijo numeričnih vrednosti v razrede (Han et. al., 2012, str. 115). Pri količini prenesenih podatkov sem jo nameraval uporabiti že kot tehniko odpravljanja šuma. Microsoft orodje omogoča tudi avtomatično diskretizacijo posameznih atributov, vendar pa meje razredov, ki so bile na ta način določene, niso bile smiselne. Razred, ki je bil npr. določen v intervalu med 7,82 MB in 136,02 MB dejansko nima nekega globokega vsebinskega pomena. Chakrabarti et. al (2009, str. 103) zato predlagajo tehniko intuitivnega postavljanja mej, s katerimi postavimo razdelitev v bolj razumljive, naravne intervale. Za količino prenesenih podatkov sem tako postavil bolj naravne meje razdelitve, obenem pa sem pazil na čim bolj enakomerno zastopanost razredov. Razmejitve so bile tako postavljene pri vrednosti 0 MB, 5 MB, 100 MB, 2000 MB, zadnji razred pa je ostal odprt. Distribucijo v razrede na podlagi teh vrednosti pa prikazuje Slika 18. Razred 0 je največji, saj ta zajema vse uporabnike, ki sploh niso uporabljali prenosa podatkov.

Slika 18: Razdelitev v razrede



5.5.3 **Izpeljava novih atributov**

Izpeljani atributi so novi atributi, ki so izračunani iz originalnih atributov. Po drugi strani pa so bili tudi originalni atributi, ki sem jih uporabil za analizo, po svoje tudi že sami izpeljani atributi, saj so bili v obliki mesečno agregiranih prometnih podatkov. Razlog izpeljave novih atributov je v tem, da z njimi opišemo uporabnikovo vedenje na bolj razumljiv način, kar nam je potem lahko v pomoč pri opisovanju segmentov. Tak primer je bil npr. podatek o povprečni mesečni količini prenesenih podatkov, katerega diskretizacijo v razrede sem opisal v poglavju 5.5.2. V kasnejših iteracijah modeliranja sem nato izpeljal še večje število novih atributov, ki so se nato izkazali zelo uporabni pri opisovanju posameznih segmentov. To so bila npr. razmerja med uporabo določene vrste storitve v primerjavi z vsemi storitvami in pa indikatorji, ki so definirali, če je bil nek uporabnik sploh aktivni uporabnik določene vrste storitev. Da je bil nek uporabnik aktivni uporabnik neke storitve, sem upošteval v kolikor je povprečno mesečno porabil vsaj eno enoto te storitve. Primerljivost enot različnih storitev sem zagotovil na način, da je bila minuta klica primerljiva z enim poslanim SMS ali MMS sporočilom oz. s prenosom 1 MB podatkov.

5.5.4 **Redukcija podatkov**

Zaradi omejitev strojne opreme, saj sem podatke procesiral na virtualnem strežniku z dokaj omejenimi sistemskimi viri, ter tudi zaradi hitrejšega izvajanja posameznih iteracij modeliranja, sem z naključnim vzorčenjem pripravil vzorec 30.000 primerov. Posamezne iteracije podatkovnega rudarjenja so se tako izvedle relativno hitro. V končni verziji modela pa sem tudi precej omejil število uporabljenih atributov, saj so se določeni atributi v različnih iteracijah modeliranja izkazali za premalo pomembne, da bi jih bilo smiselno obdržati. Končni seznam uporabljenih atributov prikazuje Tabela 8.

Tabela 8: Uporabljeni atributi

Atribut	Vrsta atributa	Opis atributa
ACCOUNT ID	Ključ	ID uporabniškega računa
AVG TOPUP	Vhodni	Znesek povprečne polnitve računa
GPRS RATIO	Vhodni	Razmerje med porabljenimi GPRS enotami in vsemi enotami
MMS RATIO	Vhodni	Razmerje med porabljenimi MMS enotami in vsemi enotami
SMS RATIO	Vhodni	Razmerje med porabljenimi SMS enotami in vsemi enotami
VOICE RATIO	Vhodni	Razmerje med porabljenimi enotami klicev in vsemi enotami
GPRS MBYTES AVG	Vhodni	Povprečna mesečna količina prenesenih podatkov
MMS AVG	Vhodni	Povprečno mesečno število poslanih MMS sporočil
MO MINUTES AVG	Vhodni	Povprečna mesečna količina minut odhodnih klicev
MT MINUTES AVG	Vhodni	Povprečna mesečna količina minut dohodnih klicev
SMS AVG	Vhodni	Povprečno mesečno število poslanih SMS sporočil
TOPUPS AMOUNT AVG	Vhodni	Povprečni mesečni znesek polnitev računa
UNITS AVG	Vhodni	Povprečni mesečna poraba enot
USAGE GPRS	Vhodni	Aktivni uporabnik GPRS storitev (Da/Ne)
USAGE MMS	Vhodni	Aktivni uporabnik MMS storitev (Da/Ne)
USAGE MO CALLS	Vhodni	Aktivni uporabnik odhodnih govornih klicev (Da/Ne)
USAGE MT CALLS	Vhodni	Aktivni uporabnik dohodnih govornih klicev (Da/Ne)
USAGE SMS	Vhodni	Aktivni uporabnik SMS storitev (Da/Ne)

5.6 Modeliranje

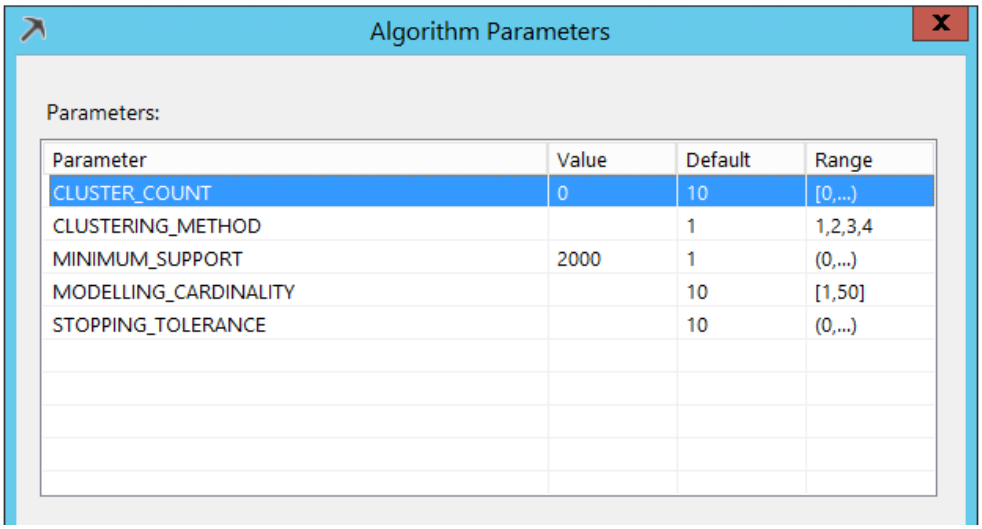
Za izvedbo modeliranja sem uporabil Microsoft segmentacijski algoritem. Prilagodil sem določene parametre nastavitev, za ostale parametre pa sem pustil privzete vrednosti.

Za parameter CLUSTERING_METHOD sem pustil privzeto vrednost, ki pomeni metodo 1 (Prilagodljivi EM). Vzrok za to izbiro je bil v tem, da je metoda EM v primerjavi z metodo voditeljev dala boljše rezultate, saj so bili segmenti precej bolj različni in jih je bilo zato tudi lažje opisati. Pri metodi voditeljev, ki sem jo tudi preizkusil, so si bili segmenti precej bolj podobni. V konkretnem primeru pa med metodo 1 in metodo 2, ki je neprilagodljivi EM, ne bi bilo nobene razlike, saj je bilo v vzorcu 30.000 zapisov, razlika med obema metodama pa se pojavi šele pri vzorcih z več kot 50.000 zapisov.

Za parameter MINIMUM_SUPPORT, ki predstavlja minimalno število primerov, ki jih mora vsebovati posamezen segment, sem nastavil vrednost 2.000. S tem sem določil dovolj visoko zastopanost segmentov, ki tako niso zajemali premalo uporabnikov, da bi jih bilo še smiselno ločeno marketinško obravnavati.

CLUSTER_COUNT parameter, s katerim se določi število ustvarjenih segmentov, sem nastavil na vrednost 0 in s tem določil modelu, da s pomočjo hevristike sam določi ustrezno število različnih segmentov, ki so ob tem dovolj različni in zato še smiselni. Slika 19 prikazuje zaslonsko sliko nastavitev parametrov algoritma. Pomen posameznih parametrov sem podrobneje opisal v poglavju 4.2.1.

Slika 19: Nastavitev parametrov algoritma



5.7 Vrednotenje

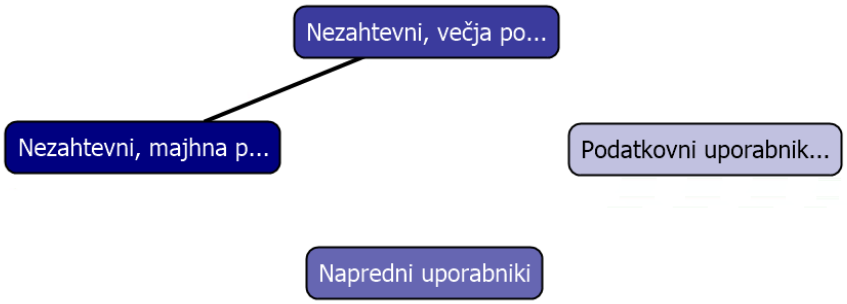
Microsoft segmentacijski algoritem najdene segmente privzeto poimenuje z besedo »Cluster« in z zaporedno številko segmenta (angl. *Cluster Index*). Pri prezentaciji rezultatov pa je dobro posamezne segmente kratko poimenovati z oznakami, ki jim dajo tudi nek vsebinski pomen, kar pa je pri modelih, ki jih lahko sestavlja tudi po več deset spremenljivk, navadno precej težko. MacLennan et al. (2009, str. 305) navajajo, da najbolj učinkovita poimenovanja izhajajo iz osebnega razumevanja poslovnega problema, pri opisovanju segmentov pa predlagajo naslednji postopek:

1. naredi površinski pregled najdenih segmentov,
2. izberi segment in opiši razlike med izbranim segmentom in splošno populacijo,
3. opiši kako je ta segment različen od ostalih segmentov,
4. preveri, da postavljene trditve o izbranem segmentu držijo,
5. označi oz. poimenuj izbrani segment, ter
6. ponovi predhodne korake še za preostale segmente.

Analysis Services ponuja različna orodja oz. poglede (angl. *View*), ki so v pomoč pri razumevanju in opisovanju segmentov. Posamezno orodje ponuja pogled iz določene perspektive, ki pa samostojno ne ponuja zadostnih informacij za celovito razumevanje posameznih segmentov, zato se pri opisovanju vedno uporablja kombinacijo različnih pregledov. Opisovanja segmentov sem se tako lotil z uporabo diagrama najdenih segmentov (angl. *Cluster Diagram*), pregleda profilov segmentov (angl. *Cluster Profiles*), pregleda lastnosti segmentov (angl. *Cluster Characteristics*) in pregleda razlikovanj segmentov (angl. *Cluster Discrimination*) ter s kombinacijo teh pogledov opisal posamezne segmente.

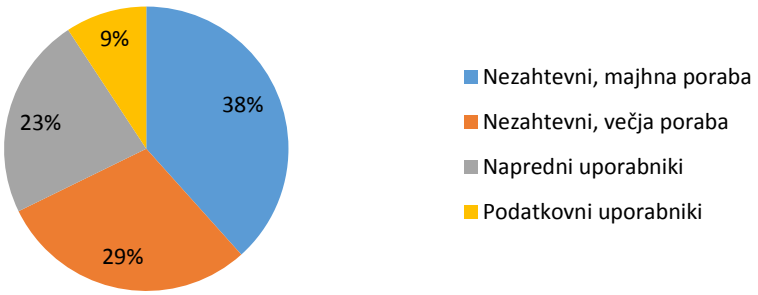
Pri opisu sem se opiral na najbolj izrazite lastnosti posameznih segmentov. Slika 20 prikazuje diagram najdenih segmentov. Povezave med segmenti nakazujejo medsebojno podobnost segmentov, intenziteta barve pa je sorazmerna z velikostjo posameznih segmentov. Iz diagrama najdenih segmentov uporabnikov tako izhaja, da je algoritem našel štiri segmente uporabnikov, kjer sta si dva segmenta uporabnikov medsebojno dokaj podobna in je zato med njima močna povezava, ostala dva segmenta pa nimata podobnih lastnosti, zato tudi nimata medsebojne povezave, niti povezav z drugimi segmenti.

Slika 20: Diagram najdenih segmentov uporabnikov



Najdene segmente sem glede na njihove najbolj izrazite lastnosti preimenoval v nezahtevni – majhna poraba, nezahtevni – večja poraba, napredni uporabniki in podatkovni uporabniki. Razdelitev populacije v segmente prikazuje Slika 21.

Slika 21: Razdelitev populacije na segmente



Posamezni segmenti imajo naslednje karakteristike:

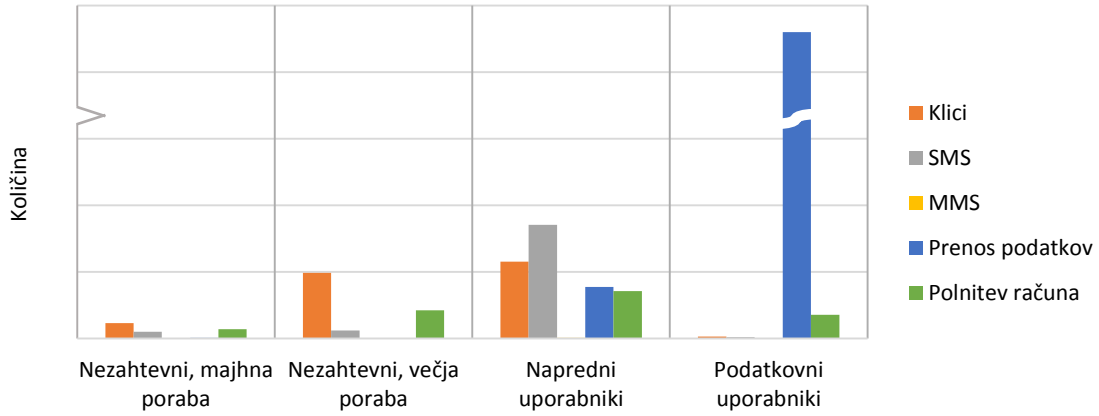
- **Nezahtevni, majhna poraba**: uporabniki pretežno uporabljajo samo govorne klice, uporabe ostalih storitev pa je zelo malo. Mesečno porabijo malo enot in od vseh segmentov uporabnikov tudi najmanj polnijo predplačniški račun.
- **Nezahtevni, večja poraba**: Ta segment je po strukturi porabe prvemu segmentu precej podoben. Tudi pri teh uporabnikih prevladuje uporaba govornih klicev, porabijo pa tudi

nekaj enot ostalih storitev, v količinskem smislu pa porabijo precej več enot od prvega segmenta in tudi precej več polnijo svoj predplačniški račun

- **Napredni uporabniki**: ti uporabniki uporabljajo vse tipe storitev. Ta segment zajema praktično vse aktivne uporabnike MMS storitev. Uporabniki v tem segmentu porabljajo veliko količino SMS sporočil, saj uporaba te vrste storitve tudi prevladuje pred uporabo govornih klicev. Tudi v količinskem smislu uporabljajo največ storitev, skupno mesečno število porabljenih enot je pri njih največje, posledično pa tudi najbolj polnijo svoj predplačniški račun.
- **Podatkovni uporabniki**: pri teh uporabnikih izrazito prevladuje uporaba prenosa podatkov. Pri veliki večini uporabnikov pa sploh ni zaslediti uporabe govornih ali drugih storitev, z izjemo prenosa podatkov. To nakazuje na možnost, da ti uporabniki svoje SIM kartice (angl. *Subscriber Identity Module*) uporabljajo namensko za podatkovni dostop, morda celo v napravah, kot so usmerjevalniki, tablični računalniki itd., ki niso prvenstveno namenjene uporabi govornih klicev ali pa jih celo sploh ne omogočajo.

Slika 22 prikazuje primerjavo segmentov glede na povprečne mesečne količine porabe. Opazno je izrazito odstopanje količine prenosa podatkov v segmentu podatkovnih uporabnikov. Po drugi strani pa uporaba SMS storitve izrazito odstopa v segmentu naprednih uporabnikov, kjer poraba te storitve presega porabo ostalih storitev, primerjalno gledano pa je poraba SMS storitve v ostalih segmentih skoraj zanemarljiva. Glede na povprečni mesečni znesek polnitev računa je najbolj zanimiv segment napredni uporabniki, kar pa je glede na strukturo in količino porabe tudi pričakovano.

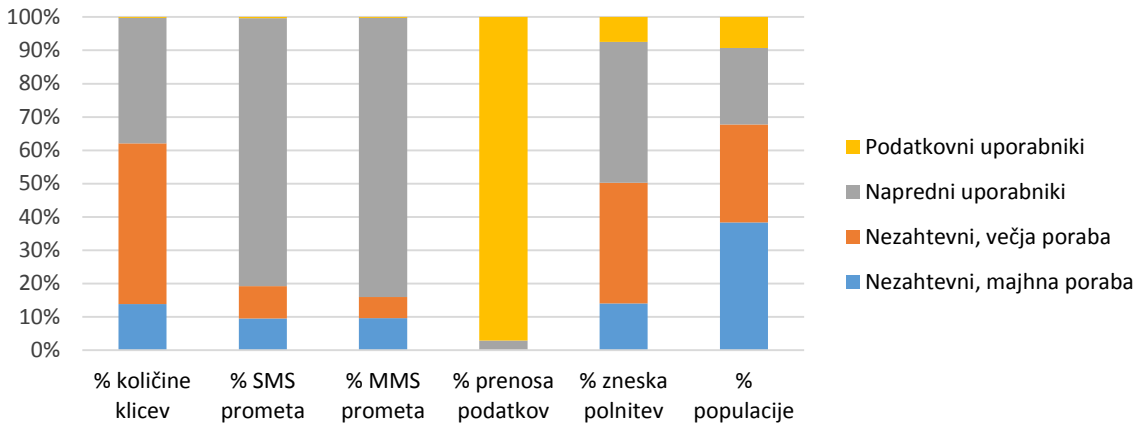
Slika 22: Primerjava povprečne mesečne porabe po segmentih



Opombi: Zaradi zaupnosti podatkov količine niso razkrite. Količina prenosa podatkov v segmentu podatkovni uporabniki izrazito presega ostale količine, zato je zaradi preglednosti grafikona skala ordinatne osi prelomljena oz. skrajšana.

Zanimal me je tudi kakšen je pomen najdenih segmentov z vidika strukture skupne porabe storitev v opazovanem obdobju, zato sem pripravil še porazdelitev skupne porabe po segmentih, kar prikazuje Slika 23. Ker pa ti podatki niso bili vključeni že v samem modelu, sem moral najprej rezultate segmentacije s pomočjo DMX (angl. *Data Mining Extensions*) poizvedbe izvoziti v ločeno SQL tabelo in jih naknadno povezati z dodatnimi atributi, ki sem jih potreboval za analizo.

Slika 23: Deleži porabe storitev po segmentih



Iz grafa se lahko razbere, da skoraj celotno skupno količino prenosa podatkov pokrije segment podatkovnih uporabnikov in da ti uporabniki praktično ne uporabljajo nobenih drugih storitev. Segment naprednih uporabnikov pokriva večino skupnega SMS in MMS prometa. Največji delež prihodka glede na skupen znesek polnitev računa prinašajo napredni uporabniki, po velikosti pa je ta segment šele na tretjem mestu, zato je z vidika dobičkonosnosti ta segment marketinško najbolj zanimiv. Po drugi strani ima segment podatkovnih uporabnikov relativno majhen delež skupnega zneska polnitev glede na skupno število porabljenih enot. Po dodatni analizi se je izkazalo, da so uporabniki ob polnitvi računa z določenim večjim zneskom polnitve pridobili še 30 dnevni neomejeni prenos podatkov. Po prenosu določene količine podatkov bi sicer morala nastopiti omejitev hitrosti prenosa, vendar pa se zaradi tehnične napake to ni zgodilo. Nekateri uporabniki so to napako s pridom izkoristili in prenesli nesorazmerno velike količine podatkov. Podatki o količinah porabe v okviru redne oz. posebne porabe pa za obravnavano obdobje v podatkovnem skladišču žal niso bili na voljo zato jih v tej analizi tudi nisem mogel upoštevati. V prihodnje pa bi bilo nujno upoštevati tudi takšne podatke, saj bi ti občutno povečali razumevanje posameznih segmentov.

Z identifikacijo in opisom različnih segmentov uporabnikov je bil sicer dosežen cilj podatkovnega rudarjenja v tehničnem smislu, vendar pa s tem poslovni cilj še ni dosežen. Poslovni cilj je na podlagi pregleda in razumevanja sestave baze uporabnikov, ki je

pridobljeno s podatkovnim rudarjenjem, pripraviti posameznim segmentom uporabnikov prilagojeno ponudbo storitev.

Med možnimi rešitvami bi lahko bila priprava različnih, prilagojenih predplačniških tarif ali pa možnost zakupa določene večje količine storitev v paketu (klicev, SMS sporočil ali prenosa podatkov). Za nekatere uporabnike bi bila smiselna tudi nadgradnja oz. prehod iz predplačniškega razmerja na naročniško razmerje. Podatkovnim uporabnikom, ki npr. potrebujejo samo prenos podatkov za tablične računalnike, bi lahko tako storitev ponudili tudi v obliki dodatne SIM kartice samo za podatkovni prenos, ki bi jo lahko pridobili v okviru svojega obstoječega naročniškega razmerja za govorne storitve. Taka ponudba bi lahko privabila tudi uporabnike s strani konkurence. Različne rešitve bi bilo seveda možno tudi selektivno tržiti posameznim segmentom uporabnikov, vendar pa to predstavlja izziv za tržnike v podjetju in že presega okvir tega diplomskega dela.

SKLEP

Rast razpoložljivih podatkov je rezultat splošne informatizacije družbe na vseh področjih. Organizacije tako dnevno ustvarjajo in shranjujejo velike količine podatkov, kar pa vodi v podatkovno bogato, a hkrati informacijsko revno stanje, saj ti podatki v podatkovnih bazah večinoma ostajajo neobdelani. Podatkovno rudarjenje je v samem jedru poslovne inteligence in predstavlja tisto področje, ki omogoča preoblikovanje surovih podatkov v uporabno obliko, na podlagi katerih se nato lahko sprejemajo dobre in utemeljene poslovne odločitve.

Zvesti in zadovoljni uporabniki, ki z organizacijo razvijejo dolgoročne odnose, so ključ do poslovnega uspeha. To je še posebej pomembno na zrelih trgih, kjer organizacije izgubljenih uporabnikov ne morejo enostavno nadomestiti z novimi. Organizacije morajo zato razviti jasno strategijo upravljanja odnosov z uporabniki skozi vse faze življenjskega cikla, ki mora temeljiti na vpogledu ter razumevanju njihovih individualnih potreb in želja.

V diplomski nalogi sem izvedel vedenjsko segmentacijo uporabnikov na podlagi analize uporabe storitev mobilne telefonije. Trg mobilne telefonije v Sloveniji velja za zrel trg, v takih razmerah pa se ni dovolj osredotočati samo na pridobivanje novih strank oz. na tržni delež, ampak se je enako pomembno osredotočati tudi na obstoječe uporabnike in s tem na notranjo rast, saj je ta takrat lažje dosegljiva. Vedenjska segmentacija je postopek razdelitve enotne baze uporabnikov v različne skupine, ki so določene glede na identificirane vzorce uporabe storitev. Taka razdelitev omogoča razvoj prilagojenih produktov oz. storitev, ki bolje zadovoljujejo potrebe posameznih skupin, omogoča pa tudi njihovo prilagojeno trženje, kar vodi v večjo porabo obstoječih uporabnikov, ki organizacijam pomeni večjo notranjo rast.

Pri procesu podatkovnega rudarjenja je zelo pomembno uporabiti ustrezno metodologijo, saj nas sicer proces lahko vodi v odkrivanje stvari, ki niso resnične ali pa stvari, ki so sicer resnične, a nam z vidika poslovnega problema niso koristne. Zato sem pri procesu podatkovnega rudarjenja sledil CRISP-DM standardu, ker je ta najbolj razširjen, je pa tudi prosto dosegljiv, dobro dokumentiran in nevtralen do različnih področij uporabe.

Izvorni podatki za analizo so izhajali iz podatkov o porabi, ki se beležijo v obliki CDR zapisov, ki so sicer v osnovi namenjeni obračunavanju, so pa tudi dober vir podatkov za druge analize. Podatke sem pridobil iz podatkovnega skladišča, kjer so bili že pretvorjeni in združeni v obliko mesečne porabe po posameznih storitvah. Kljub temu, da so bili podatki iz podatkovnega skladišča deloma že preverjeni in prečiščeni, je bilo potrebno ponovno preveriti porazdelitve posameznih atributov in odstraniti primere odstopanj, ki bi sicer lahko motili modele in vodili k nezanesljivim rezultatom. Izpeljal sem tudi kar nekaj novih atributov, ki so se pokazali zelo uporabni pri določanju in kasneje pri opisovanju lastnosti ustvarjenih segmentov.

V končni verziji je algoritem našel štiri različne segmente uporabnikov, ki so se zelo razlikovali glede vzorcev uporabe posameznih storitev in so bili po svoje presenetljivi, saj na osnovi dosedanjih analiz niso bili pričakovani. Celovitega pregleda in razumevanja sestave baze uporabnikov se namreč ne da pridobiti z enostavnimi poizvedbami oz. z ad hoc analizami, ki so običajno izvedene na manjšem obsegu atributov in zato pomanjkljive. Poleg tega pa so pogosto delane na osnovi osebne intuicije analitikov in je zato lahko vprašljiva njihova objektivnost. Z izvedbo segmentacije pa sem pridobil s strani podatkov izpeljane segmente, ki imajo zato precej večjo poslovno vrednost. Pokazala pa se je tudi izrazita iterativna narava procesa podatkovnega rudarjenja. Na osnovi rezultatov posameznih vmesnih iteracij sem se namreč pogosto vračal v fazo priprave podatkov, odstranjeval nepomembne atribute ter generiral nove, ki so nato bolje definirali posamezne segmente.

LITERATURA IN VIRI

1. Agencija za komunikacijska omrežja in storitve Republike Slovenije. (2016). *Poročilo o razvoju trga elektronskih komunikacij za četrto četrtletje 2015*. Ljubljana: Agencija za komunikacijska omrežja in storitve Republike Slovenije.
2. Azevedo, A. & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *Proceedings of the IADIS European Conference on Data Mining* (str. 182-185). Amsterdam: IADIS - International Association for Development of the Information Society.
3. Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (2nd ed.). Indianapolis: Wiley Publishing, Inc.
4. Brown, M. S. (2015, 6. oktober). *CRISP-DM: The dominant process for data mining* [videoposnetek]. Najdeno 19. januarja 2016 na spletnem naslovu https://www.youtube.com/watch?v=civLio11SjQ
5. Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., Kamber, M., Lightstone, S. S., Nadeau, T. P., Neapolitan, R. E., Pyle, D., Refaat, M., Schneider, M., Teorey, T. J., & Witten, I. H. (2009). *Data Mining: Know it all*. Burlington: Morgan Kaufmann Publishers.
6. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinrtz, T., Shearer, C., & Wirth, R. (2000), *CRISP-DM 1.0: Step-by-step data mining guide*. b.k.: CRISP-DM consortium
7. Dean, J. (2014). *Additional praise for Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Hoboken: John Wiley & Sons, Inc.
8. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37-54.
9. Han, J., Kamber, M. & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). Waltham: Elsevier Inc.
10. Hofmann, M., Klinkenberg, R. (2014). *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Boca Raton: Taylor & Francis Group.
11. Issenberg, S. (2012, 19. december). How Obama's Team Used Big Data to Rally Voters. *MIT Technology Review magazine*. Najdeno 10. februarja 2016 na spletnem naslovu https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters
12. Kotler, P., Keller, K. L. (2012). *Marketing management* (14th ed.). New Jersey: Pearson Education, Inc.
13. Larose , D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd ed.). Hoboken: John Wiley & Sons, Inc.
14. MacLennan, J., Tang Z., & Crivat B. (2009). *Data Mining with Microsoft SQL Server 2008*. Indianapolis: Wiley Publishing, Inc.

15. *Microsoft Clustering Algorithm*. Najdeno 17. januarja 2016 na spletnem naslovu https://msdn.microsoft.com/en-us/library/ms174879(v=sql.120).aspx
16. *Microsoft Clustering Algorithm Technical Reference*. Najdeno 17. januarja 2016 na spletnem naslovu https://msdn.microsoft.com/en-us/library/cc280445(v=sql.120).aspx
17. Pareek, D. (2007). *Bussiness Intelligence for Telecommunications*. New York: Auerbach Publications.
18. Piatetsky, G. (2014, 28. oktober), CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets.* Najdeno 22. maja 2016 na spletnem naslovu http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html
19. Sarka, D. (2015a, 17. april), Data Mining Algorithms – K-Means Clustering. *SQLblog*. Najdeno 7. januarja 2016 na spletnem naslovu http://sqlblog.com/blogs/dejan_sarka/archive/2015/04/17/data-mining-algorithms-k-means-clustering.aspx
20. Sarka, D. (2015b, 12. maj) Data Mining Algorithms – EM Clustering. *SQLblog*. Najdeno 7. januarja 2016 na spletnem naslovu http://sqlblog.com/blogs/dejan_sarka/archive/2015/05/12/data-mining-algorithms-em-clustering.aspx
21. Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5*(4), 13-22.
22. *SQL Server Analysis Services - Data Mining*. Najdeno 5. februarja 2016 na spletnem naslovu https://technet.microsoft.com/en-us/library/bb510517(v=sql.105).aspx
23. Szkolar, D. (2013, 24. januar), Data Mining in Obama's 2012 Victory. *iSchool*. Najdeno 14. februarja 2016 na naslovu http://infospace.ischool.syr.edu/2013/01/24/data-mining-in-obamas-2012-victory/
24. *Telecom Billing Guide*. Najdeno 15. novembra 2015 na spletnem naslovu http://www.tutorialspoint.com/telecom-billing/quick-billing-guide.htm
25. Tsiptsis, K., & Chorianopoulos, A. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation*. Chippenham: John Wiley and Sons.
26. Zakon o elektronskih komunikacijah. *Uradni list RS* št. 109/2012.