POTENCIJAL ANALIZE KORISNIČKIH PODATAKA WEB SERVISA ZA RECENZIRANJE USLUGA

Birgmajer, Petar

SVEUČILIŠTE U SPLITU
EKONOMSKI FAKULTET

DIPLOMSKI RAD

POTENCIJAL ANALIZE KORISNIČKIH PODATAKA WEB SERVISA ZA RECENZIRANJE USLUGA

Mentor:
izv.prof.dr.sc. Maja Ćukušić

Student:
Petar Birgmajer 2152738

Split, kolovoz, 2018.

SADRŽAJ:

1. UVOD

1.1 Problem istraživanja

Proces „kopanja“ po podacima u svrhu pronalaženja skrivenih veza između podataka, donošenja zaključaka te predviđanja budućih trendova ne predstavlja nov koncept, a sam proces ima dugu povijest, iako se naziv „rudarenje podataka“ po prvi put spominje početkom 90-tih godina prošloga stoljeća.[1] Rudarenje podataka objedinjuje tri znanstvene discipline: statistiku (brojčana analiza veza između podataka), umjetnu inteligenciju (računalnu imitaciju ljudske inteligencije), te strojno učenje (algoritmi sa sposobnošću učenja na temelju podataka u svrhu predikcije ili donošenja odluka). [2]

No same tehnike analize podataka, koliko god efikasne bile, za korisnika ne mogu stvoriti vrijednost ukoliko nisu primijenjene na odgovarajućim podacima. Stoga možemo reći kako efektivno prikupljanje podataka preduvjet za kvalitetnu provedbu rudarenja podataka, odnosno izvlačenja korisnih zaključaka provedbom istih. Organizacije centraliziraju podatke prikupljene iz raznih, često geografski i odjelno diferenciranih izvora u jednu bazu podataka, stvarajući time skladište podataka (eng. *data warehouse).*[3] Na ovaj način, s podacima poslovanja okrupnjenim na jednom mjestu, moguće je odabrati one segmente podataka nad kojima će se vršiti analiza u svrhu odgovaranja na postavljena poslovna pitanja.

Prikupljanje podataka o poslovanju posebno je lako onim poduzećima s elektroničkim, odnosno online modelom poslovanja. Jedno od takvih poduzeća je i Yelp. Yelp je online servis osnovan 2004. godine kao platforma za pronalazak i recenziranje lokalnih pružatelja usluga kao sto su zubari, frizeri, mehaničari i ostali. Datas je Yelp sa 28 milijuna jedinstvenih posjetitelja mjesečno, te s 135 milijuna napisanih recenzija od strane korisnika do kraja drugog kvartala 2017. izrastao je u jednu od vodećih svjetskih platformi za pronalazak i recenziranje restorana i barova.[4]

[1] SAS (2017): Data Mining: What it is and why it matters, [Internet], raspoloživo na: https://www.sas.com/en_us/insights/analytics/data-mining.html, [19.9.2017].
[2] SAS (2016): Data Mining From A to Z: How to Discover Insights and Drive Better Opportunites, [Internet], raspoloživo na: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf, [19.9.2017]
[3]Investopedia (2017): Data Mining, [Internet], raspoloživo na: http://www.investopedia.com/terms/d/datamining.asp [19.9.2017]
[4] Yelp (2017): About us, [Internet], raspoloživo na: https://www.yelp.com/about, [19.9.2017]

Prema istraživanju[5] o utjecaju online recenzija na ponašanja potrošača, 84% potrošača vjeruje online recenzijama jednako koliko i preporukama poznanika, ali i 90% potrošača formira mišljenje o poslovnom subjektu čitanjem manje od 10 recenzija. O neupitnom utjecaju recenzija na ponašanje potrošača govori i podatak kako se čak 72% potrošača odlučuje na konzumaciju usluge nakon čitanja pozitivnih recenzija. Goleme količine podataka koje servisi ovog tipa posjeduju predstavljaju idealnu podlogu za analizu, ali i razvoj algoritama i mehanizama od koje koristi imaju potrošači, poslovni subjekti, oglašivači te mnogi drugi sudionici procesa.

S obzirom na broj podataka i prirodu poslovnoga modela Yelp-a, rudarenje podataka, odnosno algoritmi generirani tehnikama istoga čine jedan od glavnih konkurentskih prednosti poduzeća. Yelp od 2013. godine periodično organizira Yelp Dataset Challenge, natjecanje u provođenju rudarenja podataka nad Yelp Open Datasetom, na koji se mogu prijaviti studenti i osobe iz cijelog svijeta, a za najuspješnije radove predviđene su novčane i druge nagrade. [6]

Yelp Open Dataset podset je podataka o oglašenim poslovnim subjektima, recenzijama te podacima korisnika aplikacije dostupan za korištenje u osobne, edukativne i akademske svrhe.[7] Upravo ovaj set podataka biti će podloga za provedbu tehnika rudarenja podataka u empirijskom dijelu ovoga rada.

1.2 Predmet istraživanja

U ovom istraživanju detaljno će se definirati pojam i obuhvat poslovne inteligencije, te navesti poslovni i tehnološki preduvjeti za implementaciju sustava iste u organizaciju. Koncept rudarenja podataka teoretski će se obraditi, a razjasniti će se sličnosti, razlike te preklapanja s povezanim područjima. Tehnike rudarenja podataka biti će teoretski obrađene, uz navođenje primjera primjene svake od navedenih tehnika u poslovnom okruženju.

U nastavku rada biti će analizirana svojstva seta podataka, te će biti poduzete korekcije i adaptacije potrebne za daljnje provođenje analize, a kao podloga za provođenje ovakvog

[5] Brightlocal (2016): Local Consumer Survey, [Internet], raspoloživo na: https://www.brightlocal.com/learn/local-consumer-review-survey/ [23.9.2017]
[6] Yelp (2017): Yelp Dataset Challenge, [Internet], raspoloživo na: https://www.yelp.com/dataset/challenge, [19.9.2017]
[7] Yelp (2017): Yelp Open Dataset, [Internet], raspoloživo na: https://www.yelp.com/dataset, [19.9.2017]

procesa poslužiti će definirani poslovni problem – predmet provođenja projekta. Ukupan set podataka sastoji se od 6 podsetova, od kojih svaki sadrži određenu jedinicu promatranu po različitim varijablama. U ovom radu biti će korišteni setovi podataka o poslovnim subjektima, te recenzijama korisnika. Set podatka o poslovnim subjektima poslužiti će nam u fazi razumijevanja poslovnog problema, kako bi se isti kvalitetnije definirao, te jasnije definirali rezultati. Nad setom podataka o recenzijama korisnika biti će provedena tehnika klasifikacija, odnosno pokušati će se razviti model sposoban za predviđanje ocjene na temelju teksta recenzija, te model za klasifikaciju recenzije kao one pozitivnog odnosno negativnog karaktera.

1.3 Istraživačka pitanja

U ovom radu pokušati će se dati odgovor na sljedeća istraživačka pitanja:

- Što je poslovna inteligencija i koji su preduvjeti za njenu uspješnu implementaciju u organizaciju?
- Što je rudarenje podataka i koja su povezana područja?
- Koje su najpopularnije metodologije rudarenja podataka i po čemu se razlikuju?
- Koje metode i tehnike rudarenja podataka nam stoje na raspolaganju? Kada i u koje svrhe ih je moguće koristiti?
- Može li se izraditi kvalitetan model u svrhu klasificiranja recenzije po ocjenama?
- Može li se izraditi kvalitetan model u svrhu klasificiranja recenzije kao one pozitivnog ili negativnog karaktera?
- Koje su karakteristike odabranoga dataseta?
- Na koji način pripremiti podatke za ulazak u model?
- Koji je od implementiranih algoritama točnije klasificira recenzije?
- Koje su najprigodnije tehnike vizualizacije podataka za prikaz dobivenih rezultata?

1.4 Ciljevi istraživanja

Unutar rada nastojati će se definirati poslovna inteligencija kao i njezina uloga i mogućnost primjene unutar organizacije. Teoretski će se obraditi pojam rudarenja podataka, objasniti će se i neke od pripadajućih tehnika koje korisnicima stoje na raspolaganju te će se detaljnije obraditi najpopularniji pristup planiranju i provođenju projekata rudarenja podataka.

Nakon uvodnih teoretskih razmatranja, cilj ovog rada jest primijeniti tehnike rudarenja podataka na odabranom setu podataka kako bi se odgovorilo na postavljena istraživačka pitanja.
Cilj je pružiti modele koji će zadovoljavati potrebe poslovnih subjekata kao i one korisnika same aplikacije.

1.5 Metode istraživanja

- **Deduktivna metoda** kao sustavna primjena deduktivnog načina zaključivanja u kojemu se iz općih sudova izvode posebni i pojedinačni zaključci.
- **Induktivna metoda** kao sustavna primjena induktivnog načina zaključivanja kojim se na temelju analize pojedinačnih činjenica dolazi do zaključka o općem sudu, od zapažanja konkretnih pojedinačnih slučajeva dolazi do općih zaključaka.
- **Metoda analize** kao postupak znanstvenog istraživanja raščlanjivanjem složenih pojmova, sudova i zaključaka na njihove jednostavnije sastavne dijelove i elemente.
- **Metoda sinteze** kao postupak znanstvenog istraživanja i objašnjavanja stvarnosti putem sinteze jednostavnih sudova u složenije
- **Metoda klasifikacije** kao sistemska i potpuna podjela općeg pojma na posebne, u okviru opsega pojma.
- **Metoda kompilacije** kao postupak preuzimanja tuđih rezultata znanstvenoistraživačkog rada, odnosno tuđih opažanja, stavova, zaključaka i spoznaja.
- **Metoda rudarenja podataka** kao postupak analiziranja velike količine podataka u svrhu pronalaženja uzoraka i izvlačenja korisnih informacija.

1.6 Doprinos istraživanja

Ovim istraživanjem pokušati će se doprinijeti znanju o iskorištavanju sirovih podataka u svrhu izvlačenja korisnih informacija, te formirati model sposoban za klasifikaciju recenzija po dodijeljenim ocjenama, odnosno po negativnom ili pozitivnom karakteru recenzije.

1.7 Struktura diplomskog rada

U uvodnom dijelu rada bit će definirani problem i predmet istraživanja, a definirati će se pitanja na koje ovo istraživanje treba dati odgovor. Odrediti će se ciljevi istraživanja te navesti metode koje će se koristiti u daljnjoj izradi rada, a uvodna razmatranja zaključiti će se očekivanim doprinosom istraživanja te strukturom samog diplomskog rada.

Drugo poglavlje rada biti će posvećen teoretskoj razradi koncepta poslovne inteligencije kao krovnog pojma koji obuhvaća sve metode potrebne za provedbu iste u poduzeću. Nadalje, teoretski će se obraditi proces prikupljanja i skladištenja podataka unutar poduzeća. Zaključno, analizirati će se preduvjeti koji moraju biti zadovoljeni za uspješnu implementaciju sustava poslovne inteligencije u organizaciju.

U trećem poglavlje pažnja će se posvetiti rudarenju podataka, pojam će se teoretski obraditi te će se obraditi odnos istoga s povezanim područjima. Tehnike rudarenja podataka biti će teoretski obrađene, uz navođenje primjera primjene istih u poslovnom okruženju. Zaključno, obraditi će se problem izostanka standardizirane metodologije za provođenje projekata rudarenja podataka, a najpopularnija metodologija te pripadajuće faze pobliže će se teoretski razraditi.

Četvrto poglavlje rada predstavlja praktičnu primjenu prethodno obrađene metodologije i tehnika na datasetu web servisa za recenziranje Yelp. Nakon upoznavanja s alatom, praćenjem odabrane metodologije provesti će se primjena odabranih tehnika rudarenja podatka na odabranom setu podataka

U zaključnom dijelu rada prodiskutirati će se uspješnost provedenih tehnika rudarenja nad podacima, te će se dobiveni rezultati usporediti s empirijskim dokazima sličnih istraživanja.

2. POSLOVNA INTELIGENCIJA

Svaka poslovna organizacija, ukoliko želi ostati konkurentna na suvremenim tržištima, mora kontinuirano pratiti svoje poslovno okruženje, te svoj vlastiti učinak u istom, kako bih pravovremeno mogla preoblikovati poslovne poteze i planove za budućnost.[8] Ono što uvjetuje navedenu sposobnost organizacije svakako jest uspješno implementiran sustav Poslovne inteligencije.

2.1 Definicija i obuhvat

Pod pojmom Poslovna inteligencija podrazumijevamo tehnologije, primjene, te skup najboljih praksi za prikupljanje, integraciju, analizu i prezentaciju informacija o poslovanju.[9] Sustavi poslovne inteligencije pružaju povijesne, trenutne i prediktivne poglede na poslovne operacije, najčešće koristeći podatke prikupljene u DW (*Data warehouse)* ili DM (*Data mart)*, te rjeđe iz operativnih podataka.

Glavna svrha Poslovne inteligencije u organizaciji jest pružiti izvršnim direktorima, menadžerima te operativnim radnicima podlogu za donošenje informiranijih i kvalitetnijih poslovnih odluka. Tako poslovni subjekti između ostalog koriste BI (*Business Intelligence – Poslovna Inteligencija)* kako bih smanjili troškove, identificirali nove poslovne prilike, te uočili i unaprijedili neefikasne poslovne procese.

Poslovna inteligencija kao disciplina i kao tehnološki zasnovan proces čini nekoliko međusobno povezanih aktivnosti, koje uključuju:[10]

1. Rudarenje podataka (*eng. Data mining)*
2. Online analytical processing - OLAP
3. Izrada „Querya“ – Upita u bazu podataka
4. Izvještavanje

U nastavku rada kratko će se definirati svaka od navedenih aktivnosti, te pojasniti uloga i povezanost svake u kontekstu primjene u Poslovnoj inteligenciji.

[8] Maheshwari, A. (2014): Business Inteligence and Data Mining, Business Expert Press, New York; str. 3.
[9] OLAP.com (2017): What is Business Intelligence (BI)? [Internet], raspoloživo na: http://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/ [31.07.2018]
[10] Finannces Online (2017): What Is the Purpose of Business Intelligence in a Business? [Internet], raspoloživo na: https://financesonline.com/purpose-business-intelligence-business/ [01.08.2018]

2.1.1 Rudarenje podataka

Rudarenje podatak proces je analiziranja skrivenih uzoraka u podacima iz različitih perspektiva u svrhu pretvaranja istih u korisne informacije. Podaci na kojima se ovakve analize temelje nalaze se u za poslovni subjekt zajedničkim područjima, kao što su skladišta podataka (*eng. Data Warehouse)* ili Data martovi. S obzirom da će u nastavku ovog rada detaljnijoj obradi rudarenja podataka biti posvećeno čitavo poglavlje, za sada ćemo ovu definiciju smatrati dostatnom kako bi se bolje razumjela poveznica rudarenja podataka s ostalim aktivnostima u kontekstu Poslovne inteligencije.

2.1.2. OLAP

OLAP (Online Analytical Processing) alat je koji omogućava korisnicima analiziranje podatka iz perspektive više dimenzija, ali i različitih baza podataka. Za ilustraciju koncepta poslužiti ćemo se primjerom, pa tako je korisniku OLAP sustava omogućeno uspoređivanje prodaje određenog artikla u određenim periodima na određenoj lokaciji, s prodajom istog artikla u istom vremenskom razdoblju ali na različitoj lokaciji.[11] Ovakva integraciju omogućuje OLAP server, koristeći kojega serverski klijent organizira i uspoređuje podatke.

Višedimenzionalnost kao svojstvo OLAP sustava podrazumijeva činjenicu da se svaki atribut podataka promatra kao zasebna dimenzija. Za razliku od relacijskih baza podataka gdje su podaci pohranjeni u tradicionalnom dvodimenzionalnom redak-stupac formatu, podaci u OLAP sustavima pohranjeni su po multidimenzionalnoj strukturi baze podataka – OLAP kocki, u kojoj vrijednosti na „križanju“ različitih dimenzija predstavljaju konsolidirane vrijednosti po atributima, odnosno dimenzijama. [12]

Princip OLAP kocke omogućava korisniku interaktivnu analizu podataka. Korisniku su na raspolaganju brojne operacije nad podacima u kocki, kako bi se omogućili različiti pogledi na podatke koje ona sadržava, te na taj način omogućujući podlogu za interaktivno postavljanje upita nad podacima i analizu podataka.

[11] TechTerms (2016): OLAP Definition [Internet], raspoloživo na: https://techterms.com/definition/olap [01.08.2018]
[12] OLAP.com (2016): OLAP for Multidimensional Analysis [Internet], raspoloživo na : http://olap.com/olap-definition/ [01.08.2018]

Neke od najčešće korištenih operacija nad podacima u OLAP kocki uključuju:[13]

1. Roll up – označava operaciju agregiranja nad podacima unutar OLAP kocke bilo podizanjem pogleda podataka na viši hijerarhijski stupanj dimenzije (npr. sa pogleda prodaje po poslovnicama na pogled prodaje po gradovima), ili potpunim uklanjanjem jedne dimenzije (npr. uklanjanjem prostorne dimenzije iz analize prodaje).
2. Roll down – Označava postupak suprotan prethodno definiranoj Roll up operaciji, te nam omogućava detaljniji pregled podataka. Moguće ga je ostvariti spuštanjem pogleda na niži hijerarhijski stupanj dimenzije (npr. sa pogleda prodaje po gradovima na pogled prodaje po regijama), ili dodavanjem dodatne dimenzije (npr. dodavanjem vremenske dimenzije u analizu prodaje).
3. Slicing – Tehnika „vađenja" jedne dimenzije iz kocke, koja za rezultat ima novu kocku, odnosno „podkocku".
4. Dicing – Tehnika koja nam omogućuje „vađenje" dvije ili više dimenzija iz postojeće kocke, koja za rezultat ima novu kocku, odnosno „podkocku".
5. Scoping – Ova tehnika omogućuje korisniku odabiranje podseta podataka koje želi pregledavati, te za koje želi da se događaju automatska osvježavanja te punjenja novim podacima.

2.1.3. Queryi – upiti u bazu

Query označava zahtjev za podacima iz baze podataka nad jednom ili kombinacijom više tablica. Standardan jezik za izvršavanje ovakvih upita jest Microsoftov Structured Query Language (SQL). Sa SQL-om kao podlogom razvijene su razna proširenja istog jezika, uključujući MySQL, Oracle SQL i NuoDB. [14]

Queryiji mogu izvršavati nekoliko različitih zadataka. Njihova primarna namjena jest povlačenje podataka iz baze prema specifičnim kriterijima koje u okviru programske sintakse postavlja korisnik. Također, queryiji omogućavaju korisniku provođenje brojnih računskih operacija nad podacima, umetanje novih, brisanje i zamjenu postojećih podataka, kao i automatizaciju operacija vezanih za Data management.[15]

[13] ECS (2015): OLAP operations [Internet], raspoloživo na: http://athena.ecs.csus.edu/~olap/olap/OLAPoperations.php [01.08.2018]
[14] Technopedia (2015): What does Query mean? [Internet], raspoloživo na https://www.techopedia.com/definition/5736/query [03.08.2018]
[15] TechTarget: SearchSQLServer (2014): What is Query? [Internet], raspoloživo na: https://searchsqlserver.techtarget.com/definition/query [03.08.2018]

2.1.4. Izvještavanje

Izvještavanje u okviru Poslovne Inteligencije podrazumijeva proces primanja odnosno isporuke informacija u strukturiranom obliku krajnjim korisnicima, organizacijama ili aplikacijama, putem programskog alata za podršku Poslovnoj inteligenciji. [16]

Iako se izvještavanje može kategorizirati na brojne načine, u ovom će radu izvještavanje biti definirano sa stajališta različitih nivoa u okviru organizacije, odnosno prema razlici u potrebama za informacijama svake od razina unutar organizacije.

Tablica 1 Zahtjevi za informacijama prema razinama u organizaciji

Kriteriji / Korisnici	Specifični ciljevi	Konkretne mjere i detaljnost podataka	Vremenski okvir
Top Menadžment	Dugoročni ciljevi	Visoko agregirani KPI-ovi.	Mjesečni, kvartalni te godišnji izvještaji.
Menadžment srednje razine	Kratkoročni ciljevi	Agregirani podaci s mogućnošću drill-down operacije.	Tjedni i mjesečni izvještaji.
Linijski menadžment, voditelji timova i operativno osoblje	Svakodnevni operativni ciljevi	Podaci s visokom razinom detalja.	Dnevni izvještaji i izvještaji po satu.

Izvor: SearchBusiness Analytics, https://searchbusinessanalytics.techtarget.com/

Iz tablice 1 vidljivo je kako top menadžment donosi odluke u svrhu ostvarenja dugoročnih ciljeva organizacije, stoga za potporu u procesu odlučivanja treba informacije koje pokrivaju širok spektar organizacije kako bi sagledali krupnu sliku stanja organizacije. Poslovna inteligencija na ovoj razini poduzeća mora odgovarati na ovakve zahtjeve, stoga izvještaji moraju sadržavati visoko agregirane podatke u obliku KPI-ova (Key Performance Indicator), odnosno ključnih pokazatelja performansi poduzeća. KPI-ovi se često ne prikazuju kao apsolutne vrijednosti, već kao indikatori pozitivnog ili negativnog odstupanja od očekivane ili prihvatljive vrijednosti.[17]

S obzirom da se radi o visoko agregiranom podacima možemo zaključiti kako izvještaji pripremljeni za ovu razinu dozvoljavaju više „kašnjenja“ informacija, odnosno nije potreban uvid u svakodnevne oscilacije u poslovanju.

[16] Passioned Group (2017): BI Reporting, [Internet] raspoloživo na: https://www.passionned.com/business-intelligence/bi-reporting/ [03.08.2018]
[17] PWC (2017): Guide to Key Performance Indicators, [Internet] raspoloživo na: https://www.pwc.com/gx/en/audit-services/corporate-reporting/assets/pdfs/uk_kpi_guide.pdf [03.08.2018]

Menadžment srednje kategorija upravlja odjelima i ostalim radnim jedinicama poduzeća. Ovakvoj razini i dalje su potrebni podaci na agregiranoj razini, ali s dodatnom mogućnošću „drill-down“ operacije, odnosno mogućnošću spuštanja na niže, detaljiziranije razine pregleda podataka. Izvještaji ovoj razini menadžmenta najčešće dolaze u obliku pisanih izvještaja, zajedno s interaktivnim sistemima koji omogućuju daljnje manipuliranje podacima iz izvještaja. Upravo ovoj razini menadžmenta najviše mogu biti od koristi informacije dobivene tehnikama rudarenja podataka.

Na posljetku, voditelji timova, linijski menadžeri te zaposlenici trebaju pristup podacima na visokoj razini detalja. S obzirom da se odluke donose na dnevnoj razini, kašnjenje informacija mora biti svedeno na minimum. Korisnici ovih podataka često trebaju podatke ne starije od jednog Data, sata, a nekada i manje. Ova skupina korisnika također može imati koristi od tehnika rudarenja podataka kako bi se uočili trendovi i korelacije u podacima iz svakodnevnog poslovanja.[18]

2.2 Prikupljanje i skladištenje podataka

2.2.1. Prikupljanje podataka

Podatke koje organizacije prikupljaju u velikom broju i mogu im poslužiti za daljnju obradu prema načinu na koji su generirani možemo podijeliti na:[19]

1. Podatke interneta stvari *(eng. Internet of things, IoT)*
2. Podatke organizacije
3. Podatke iz znanstvenih istraživanja
4. Podatke s mreža (*eng. networking Data)*

Podaci interneta stvari podrazumijevaju setove podataka generirane od strane senzora. Perilice za rublje, rasvjeta, alarmi, GPS sustav, mobilni telefoni, hladnjaci itd. mogu poslužiti kao dobar primjer izvora ove vrste podataka.

[18] TechTarget, SearchBusiness Analytics (2016): Understanding benefits of business intelligence reporting, data mining, [Internet] raspoloživo na: https://searchbusinessanalytics.techtarget.com/feature/Understanding-benefits-of-business-intelligence-reporting-data-mining [03.08.2018]
[19] Karacan, H., Sirin, E. (2017): A review on Business Intelligence and Big Data, International Journal of Intelligent Systems and Applications in Engineering 2147-6799

Potencijal senzora leži u njihovoj mogućnosti prikupljanja podataka o fizičkom okruženju, koji nakon toga mogu biti analizirani ili kombinirani s drugim izvorima podataka kako bi se uočili uzorci.[20]

Podaci organizacije označavaju uglavnom strukturirane podatke upravljane sustavom za upravljanjem baza podataka *(eng. Relation Database Management System – RDBMS)*, a sačinjavaju ih podaci iz sustava za upravljanje odnosa s klijentima *(eng. Customer Relationship Management – CRM),* ERP (Enterprise Resource Planning) sustava, podaci o prodaji, podaci o financijama organizacije kao i podaci o proizvodnji.[21]

Znanstvena istraživanja također su domena koju fenomen obrade i prikupljanja velike količine podataka nije zaobišao, a sve veći značaj ima u granama kao što su društvene znanosti, bio informatika, medicina itd.. Iako je ova vrsta podataka najmanje zastupljena u domeni Poslovne inteligencije, kombinacijom s postojećim podacima tvrtke može poslužiti za bolje razumijevanje procesa poslovanja te tržišta na kojima organizacija posluje ili želi poslovati.[22]

Mreža, odnosno Internet bazna je infrastruktura za prijenos i dijeljenje podataka u gotovo svakom od aspekata ljudskog života. Pretrage na web-pretraživačima, aktivnosti na društvenim mrežama, ponašanje korisnika na internetskim stranicama i *clickstreamovi* jedni su od glavnih izvora ovakvih baza podataka. Broj na ovakav način generiranih podataka eksponencijalno raste iz godine u godinu, te zahtjeva sve snažniju mrežnu infrastrukturu. [23]

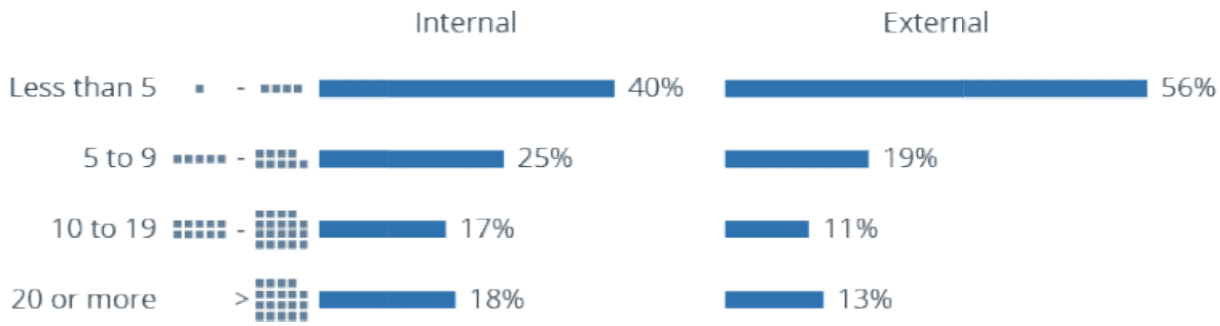
Izvore podataka koje organizacija koristi u procesu Poslovne inteligencije možemo podijeliti i na interne i eksterne izvore podataka. Interni podaci podrazumijevaju informacije generirane unutar kompanije, odnosno iz raznih odijela kao što su prodaja, financije, marketing i ljudski resursi, dok pod eksterne izvore podataka spadaju podaci iz raznih istraživanja i anketa ili sa društvenih mreža.[24]

---

[20] Harris, D. (2017): 4 Emerging Use Cases for IoT Data Analytics [Internet], raspoloživo na: https://www.softwareadvice.com/resources/iot-data-analytics-use-cases/ [04.08.2018]
[21] Karacan, H., Sirin, E. (2017): A review on Business Intelligence and Big Data, International Journal of Intelligent Systems and Applications in Engineering 2147-6799
[22] Sisense (2018): Data sources to improve your decision making [Internet], raspoloživo na: https://www.sisense.com/blog/free-data-sources-upgrade-business-decision-making/ [06.08.2018]
[23] Karacan, H., Sirin, E. (2017): A review on Business Intelligence and Big Data, International Journal of Intelligent Systems and Applications in Engineering 2147-6799
[24] Worthwhile (2016): 3 Keys to Managing Data and Maximizing Business Intelligence [Internet], raspoloživo na: https://worthwhile.com/blog/2017/02/20/data-business-intelligence/ [06.08.2018]

Prema istraživanju organizacije Clutch, BI analitičari generalno smatraju podatke iz internih izvora vrjednije od onih iz eksternih izvora, tako da 65% ispitanih analitičara smatra interne podatke korisnijima od eksternih, dok njih 70% tvrdi kako su upravo interni podaci najkorišteniji u analizama.[25]

Ipak, u svrhu postizanja izvrsnosti u pružanju podloge za donošenje odluka u poduzeću, potrebno je kombinirati podatke iz oba izvora. Rezultati istraživanja provedenog od strane organizacije BI-Survey provedenom nad 678 poslovnih subjekata sugeriraju kako srednja vrijednost broja izvora koje tvrtke koriste u procesu Poslovne inteligencije iznosi 5, stoga se može zaključiti kako ipak većina poslovnih subjekata koristi više izvora podataka pri analizi, dok se samo 6% poduzeća izjasnilo kako koriste isključivo jedan izvor podataka.



Slika 1 Grafički prikaz broja korištenih izvora podataka

Također potrebno je napomenuti pozitivnu korelaciju između veličine poslovnog subjekta te broja izvora podataka koje poduzeće koristi, posebno kod broja internih izvora podataka.

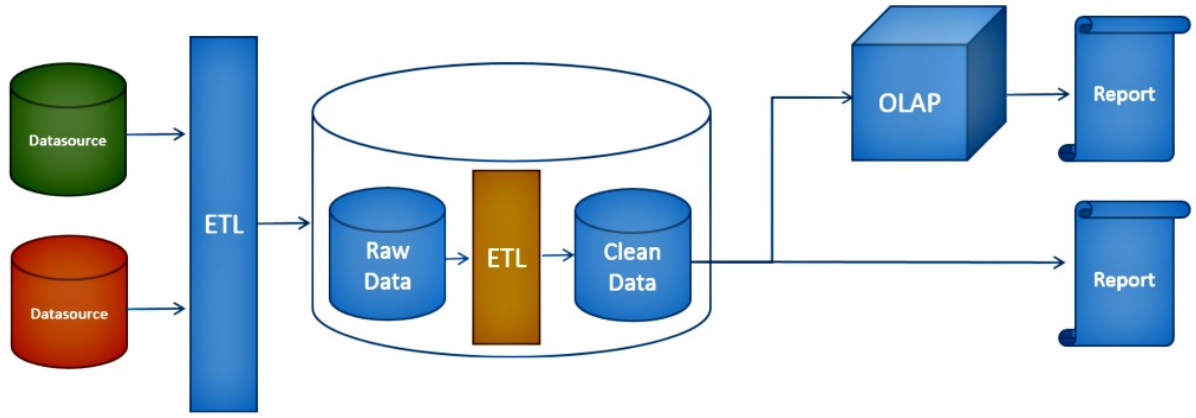
2.2.2 Skladištenje podataka

Tipična arhitektura skladištenja podataka u svrhu provođenja Poslovne inteligencije unutar poduzeća podrazumijeva tri sloja skladištenja podataka, i to:

1. „Primarne“ baze podataka *(eng. Primary data storage)*
2. Skladište podatka – *(eng. Data Warehouse)*
3. Baze podataka za provođenje analitike *(eng. Analytical Databases)*

[25] Technologydecisions (2016): Internal Data more useful to BI than external Data [Internet], raspoloživo na: https://www.technologydecisions.com.au/content/it-management/article/internal-data-more-useful-to-bi-than-external-data-1260850199, [06.08.2018]

Također, za uspješno funkcioniranje ovakve strukture između svakoga sloja potreban je ETL (Extraction-Transformation-Load) alat koji omogućava prebacivanje podataka u sljedeći sloj, i to u odgovarajućem obliku.[26]



Slika 2. Shematski prikaz skladištenja podataka

Izvor: Centreeurope.info

Sloj „primarnih" podataka podrazumijeva one podatke generirane u sustavima za podršku poslovanju u svom sirovom obliku. Ovakvi podaci bili su predmet prethodnog poglavlja ovog rada. Ovakvi podaci, kako bi se njihov format prilagodio onome pogodnome za skladištenje u Data warehouse, trebaju proći kroz ETL proces. ETL proces podrazumijeva povlačenje novonastalih podataka iz privatnih izvora, obično u „staging" bazu podataka, u kojoj se nad podacima provode operacije kao što su transformacija, homogenizacija i čišćenje podataka, te se zatim, u odgovarajućem obliku, prebacuju u skladište podataka. Ovakav proces visoko je standardiziran kako bi se poštovala poslovna pravila i integritet skladišta podataka.[27]

Skladište podataka je kolekcija podataka koji pružaju podršku u procesu donošenja odluka, te ima sljedeća obilježja:[28]

- Subjekto je orijentirano
- Integrirano je i konzistentno
- Prikazuje evoluciju kroz vrijeme i nije volatilno.

[26] CMBI (2015): Business Intelligence Data Storage Architecture [Internet], raspoloživo na: http://www.cmbi.com.au/5040_DataStorageArchitecture.html [06.08.2018]
[27] Simitsis, A., Vassiliadis, P. (2009): Extraction, transformation and loading, Database Encyclopedia 2009
[28] Golarelli F, Rizzi A. (2011): Data Warehouse Design: Modern Principles and Methodologies, CompRef8 039-1

Za skladišta podatka možemo ustanoviti da su subjektno orijentirana zbog toga jer su podaci oblikovani prema subjektima relevantnima za poduzeće kao što su kupci, proizvodi, prodaja i narudžbe. Integritet i konzistentnost skladišta podatka od visoke su važnosti, s obzirom da se skladište često puni iz različitih izvora odnosno operativnih baza podataka, pa je stoga nužno voditi brigu da format podataka iz različitih izvora odgovara onome odabranom za skladište podataka, kako bi se nad podacima uspješno mogli izvršavati upiti. Zahtjev za vremenskom evolucijom podataka u skladištu proizlazi iz činjenice kako se skladište podataka regularno osvježava podacima iz operativnih baza i nastavlja rasti, zadržavajući pri tom sve povijesne podatke. [29]

2.3 Preduvjeti uspješne implementacije u organizaciju

Uspješna implementacija ovakvog sustava u organizaciju podrazumijeva ispunjene uvjete na nekoliko različitih razina, i to na: [30]

- Organizacijskoj razini
- Procesnoj razini
- Tehnološkoj razini

Jedan od krucijalnih uvjeta na organizacijskoj razini projekta implementacije jest **preData podrška i financijsko sponzorstvo menadžmenta**. Prema istraživanju[31] podrška menadžmenta zauzima prvo mjesto po važnosti u uspjehu implementacije BI sustava u organizaciju. Ovakva vrsta podrške značajno utječe na sposobnost pribavljanja potrebnih operativnih resursa kao što su financijski i ljudski kapital. Dodatno, istraživanje[32] pokazuje kako je ovakva podrška znatno izraženija ukoliko dolazi od poslovne strane organizacije, a ne od one zadužene za IT rješenja. Ovo se može objasniti činjenicom kako je u ovom slučaju potreba za implementacijom sustava zaista prepoznata ukoliko ima snažnu podršku od strane sponzora zaduženo za poslovnu stranu organizacije.

[29] Golarelli F, Rizzi A. (2011): Data Warehouse Design: Modern Principles and Methodologies, CompRef8 039-1

[30] Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of computer information systems, vol. 50, no. 3, Spring, pp. 23-32.

[31] Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of computer information systems, vol. 50, no. 3, Spring, pp. 23-32.

[32] Watson, H., Annino, D. A., Wixom, B. H. "Current Practices in Data Warehousing," Journal of Information Systems Management, 18(1), 2001, 1-9

Jasna vizija i dobro definiran poslovni slučaj također su od krucijalne važnosti za uspjeh projekta. Implementacija ovakvog sustava u poduzeće trebala bi biti motivirana ostvarenjem dugoročnih strateških ciljeva, stoga kvalitetno razumijevanje poslovne vizije poduzeća i načina na koji bi implementacija sustava poslovne inteligencije doprinijela uspješnijem ostvarenju iste predstavlja ključno pitanje na koje poslovni slučaj mora dati odgovor.

Na procesnoj razini, veliki broj ispitanika istraživanja[33] ističe kako je važno da je vođa projekta implementacije **vidi sustav poslovne inteligencije iz strateške i organizacijske perspektive**, te izbjegava pretjerano fokusiranje na tehničke aspekte implementacije. **Prikladne vještine projektnog tima** također igraju značajnu ulogu u uspjehu implementacije ovakvog rješenja u organizaciju. Ovakve vještine uključuju tehničke te interpersonalne vještine članova tima.[34]

U okviru tehnološke razine kao krucijalne faktore većina ispitanih subjekata navodi **poslovno orijentiran, skalabilan i fleksibilan tehnološki okvir.**[35] Fleksibilan i skalabilan dizajn infrastrukture omogućava jednostavno proširenje sustava kako bi se efikasno odgovorilo na informacijske zahtjeve kao što su dodavanje dodatnih izvora podataka, atributa te dimenzija lako povezivanje s eksternim izvorima podataka te omogućilo odgovaranje na zahtjeve od strane dobavljača, zakonodavnih tijela i industrijskih standarda.[36]

Prema istraživanju[37] zadovoljenje uvjeta **održive kvalitete, točnosti i integriteta podataka** navedeno je od strane najviše ispitanih subjekata kao presudno za uspjeh projekta implementacije. Samo uz održivo prikupljanje čistih, konzistentnih, kvalitetnih i integriranih podataka moguće je ostvariti puni potencijal implementacije ovakvoga sustava u poduzeću.[38]

---

[33] Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of computer information systems, vol. 50, no. 3, Spring, pp. 23-32.
[34] Hawking P. and Sellitto C. (2010). Business Intelligence (BI) Critical Success Factors. ACIS 2010 Proceedings .
[35] Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of computer information systems, vol. 50, no. 3, Spring, pp. 23-32.
[36] Olszak, C & Ziemba, E. "Approach to Building and Implementing Business Intelligence Systems," Interdis-ciplinary Journal of Information, Knowledge, and Management, 2, 2007, 135-148.
[37] Jones M.C., Ramakrishnan T. and Sidorova A. (2012). Factors influencing business intelligence (BI) data collection strategies: An empirical investigation.Decision, Support Systems 52
[38] Hawking P. and Sellitto C. (2010). Business Intelligence (BI) Critical Success Factors. ACIS 2010 Proceedings .

3. RUDARENJE PODATAKA

3.1 Definicija rudarenja podataka

Rudarenje podataka može se promatrati kao prirodan rezultat evolucije informacijskih tehnologija. Industrije baza podataka te menadžmenta podataka evoluirale su u nekoliko kritičnih funkcionalnosti kao što su prikupljanje podataka i dizajn baza, menadžment podataka te naprednu analizu podataka, koja uključuje i sam koncept rudarenja podataka.[39]

Rudarenje podataka možemo definirati kao proces sadržan od seta tehnika i metoda kako bi se iz podataka izvukle vrijedne informacije, pronašli interesantni neočekivan i nepoznati uzorci u podacima, u svrhu boljeg predviđanja i kvalitetnijeg donošenja odluka. Ovakav proces može biti automatizirani ili (češće) polu-automatiziran. [40]

Dva su ekstrema kada su u pitanju strukturalni uzorci u podacima: oni koji mogu biti definirani kao crna kutija *(eng. black box)* te oni kod kojih je struktura vidljiva. Iz uzoraka s vidljivom strukturom rezultat jest jasno utvrđen set pravila po kojima je donesen određen zaključak, kao primjer strukture ovog tipa možemo navesti stabla odlučivanja te grupiranje *(eng.clustering)*. [41]

Tehnički govoreći, ciljevi rudarenja podataka mogu biti podijeljeni u dvije glavne skupine:[42]

- Potvrđivanje hipoteze korisnika
- Otkrivanje novih uzoraka u podacima u svrhu predviđanja ili opisivanja

Testiranje hipoteze može biti primijenjeno u onim situacijama kada korisnik odabire iznenađujuće uzorke koji se jasno ne naziru u velikom broju podataka. Kako je svaki od uzoraka predmet zasebnoga testiranja na signifikantnost, čest je slučaj da se u istraživanjima unutar rudarenja podataka simultano testira set hipoteza.[43]

---

[39] Han, J., Kamber, M., Pei, J. (2011): Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, Massachusetts

[40] Ahlemeyer-Stubbe, A. and S. Coleman, A practical guide to data mining for business and industry. 2014: John Wiley & Sons.

[41] Eibe, F., Hall, M., Witten, I. (2011): Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, Massachusetts

[42] Sristava, J. (2015): Understanding Linkage between Data Mining and Statistics, International Journal of Engineering Technology, Management and Applied Sciences, Volume 3, Issue 10, ISSN 2349-4476

[43] Sristava, J. (2015): Understanding Linkage between Data Mining and Statistics, International Journal of Engineering Technology, Management and Applied Sciences, Volume 3, Issue 10, ISSN 2349-4476

Tehnike predikcije uključuju korištenje nekih od varijabli ili polja u bazi podataka kako bi se prognozirale nepoznate ili buduće vrijednosti drugih varijabli koji su istraživaču od interesa, dok tehnike deskripcije pomažu u pronalasku uzoraka pogodnih za ljudsku interpretaciju i opis podataka. [44]

3.2. Rudarenje podataka i povezana područja

Na istraživanja o rudarenju podataka značajno je utjecao broj drugih polja kao što su strojno učenje i statistika.[45] U ovom dijelu rada biti će kratko definirane poveznice između navedenih disciplina i rudarenja podataka

- **Statistika** je jedan od fundamentalnih principa na kojima je izgrađeno rudarenje podataka kao tehnika i koncept. Sustavi statističkih analiza korišteni su od strane analitičara kako bi se detektirali neobični uzorci u podacima, te kako bi se uzorci objasnili korištenjem statističkih modela, kao što su linearni modeli. Statističke metode primjerene su za primjenu nad rezultatima rudarenja podataka, u svrhu provođenja usmjerenije analize.[46]
- **Strojno učenje** kao-pod disciplina znanosti o podacima fokusira se na dizajn algoritama koji mogu učiti na podacima i predviđati određene atribute istih..[47] Algoritam, nakon izvršenoga procesa učenja kao input uzima set podataka ili informaciju, a kao output odgovarajući rezultat, najčešće nedostajući atribut kojega se želi predvititi. Tehnike treniranja odnosno učenja algoritma možemo podijeliti na *supervised* i *unsupervised* tehnike. *Supervised* tehnike podrazumijevaju tehnike učenja iz primjera, gdje je *training dataset* sa primjerima predviđen kako bi algoritam na njemu definirao klase. Kada algoritam „nauči“ određena klasifikacijska pravila, sposoban je predviđati vrijednosti atributa na setovima sa kojima se još nije susreo, a imaju ista obilježja. Kod *unsupervised* tehnika učenja takav *training dataset* ne postoji, već sustav analizira dani set podataka kako bi uočio sličnosti i definirao podsetove podataka, slično statističkome *klasteriranju.*[48]

[44] Friedman J.H , (1998). Data mining and Statistics-What's the Connection, 29th Symposium on the interface
[45] Pujari. A (2001): Data Mining Techniques, Orient Blackswan London, stri. 16
[46] Hand, D (1999): Statistics and data mining: intersecting disciplines, ACM SIGKDD Explorations, Volume 1,
[47] Medium (2016): What is the relationship between machine learning and data mining? [Internet], raspoloživo na: https://medium.com/@xamat/what-s-the-relationship-between-machine-learning-and-data-mining-8c8675966615 [10.08.2018]
[48] Machine Learnig Matery (2016): Supervised and Unsupervised Machine Learning Algorithms [Internet] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [10.08.2018]

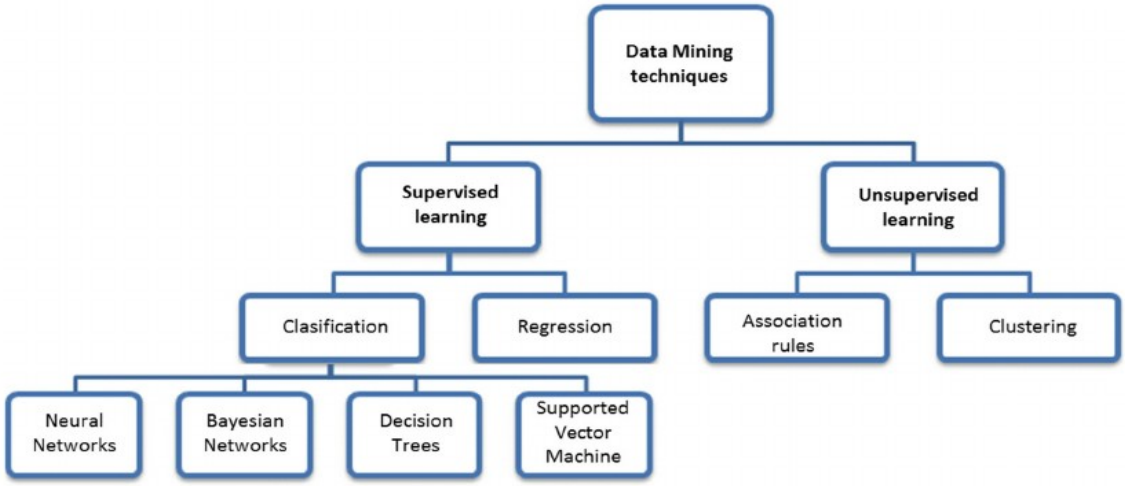
3.3 Tehnike rudarenja podataka

Tehnike unutar polja rudarenja podataka možemo podijeliti u dvije osnovne skupine:[49]

- *Supervised* tehnike učenja
- *Unsupervised* tehnike učenja

Supervised tehnike predstavljaju formalizaciju ideje učenja na primjerima. Kod *supervised* učenja, učeniku (u ovom slučaju algoritam) su Data dva seta podataka, *training* i *test* set. Ideja jest da sustav znanje prikupljeno na *training* setu podataka, primijeni u svrhu identificiranja željene varijable na *test* setu.

U *training* setu podataka svaki primjer, odnosno svaki redak predstavlja par koji se sastoji od atributa u službi „inputa", te pripadajuće željene „output" varijable, koja se u kontekstu rudarenja podatak naziva *label. Supervised* algoritam analizira *training* dataset, te u slučaju diskontinuiranog outputa formira *classifier* funkciju, dok u slučaju kada output predstavlja kontinuirana vrijednost, formira regresijsku funkciju. [50]



Slika 3. Pregled osnovnih tehnika rudarenja podataka
Izvor: https://www.researchgate.net/figure/Main-data-mining-techniques_fig2_270552309

Kod *unsupervised* tehnika učenja sustav prima set podataka kao input, no u ovom slučaju bez definirane output varijable, odnosno *labela.* Ova vrsta učenja bazira se na pronalaženju uzoraka u velikim količinama podataka.[51]

[49] Towards Data Science (2017): Supervised vs. Unsupervised Learning [Internet], dostupno na: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d [13.08.2018]
[50] Learned-Miller. E. (2014): Introduction to Supervised Learning, Department of Computer Science, University of Massachusets, Amherst (str. 2)
[51] Ghahramani, Z. (2014): Unsupervised Learning, Gatsby Computational Neuroscience Unit, University College London, UK

3.3.1 Grupiranje

Grupiranje *(eng. Clustering)* je tehnika razdjeljivanja seta podataka ili objekata u smislene i/ili korisne grupe (klastere). Ukoliko je cilj klasteriranja stvoriti smislene grupe, klasteri bi trebali biti formirani u odnosu na prirodnu strukturu podataka. U nekim slučajevima, proces grupiranja koristi se kao polazna točka za primjenu ostalih tehnika, kao što je npr. sažimanje (*eng. summarization).* [52]

Cilj grupiranja jest da je svaki od objekata unutar grupe što sličniji ostalim objektima unutar iste grupe, te što različitiji ili što manje povezan s objektima unutar ostalih grupa. Što je sličnost objekata unutar grupe veća, te razlika između grupa izraženija, grupiranje se smatra kvalitetnijim. Za pojašnjenje nekih od vrsta klastera poslužiti ćemo se pretpostavkom o distribuciji objekata odnosno opservacija na dvodimenzionalnom prostoru.

Klastere možemo smatrati **kvalitetno razdvojenima** u slučaju kada je udaljenost između bilo koja dva objekta iz različitih klastera veća nego udaljenost od bilo koje dvije točke unutar jednog klastera. [53]

Klasteri bazirani na prototipovima (eng. Prototype-based) podrazumijevaju setove podataka u kojima su podaci bliži prototipu koji definira određeni klaster. Za diskontinuirane vrijednosti, ovakva vrijednost naziva se *centroid*, a najčešće podrazumijeva srednju vrijednost svih točaka unutar klastera. Kod kategorijskih atributa ovakva vrijednost naziva se *medoid*, a označava najreprezentativniju točku u klasteru. [54]

Klasteri bazirani na gustoći (eng. Density-based) predstavljaju dio prostora na kojemu su objekti gusto raspoređeni, a okruženi su prostorom niske gustoće. Ovakva definicija klastera najprimjenjivija je u situacijama kada su klasteri nepravilnih oblika, kada su djelomično isprepleteni, te kada je prisutan visok broj *outliera.* [55]

---

[52] Stefanovski J. (2009): Data Mining – Clustering, Institute of Computing Sciences, Poznan University of Technology [Internet], dostupno na: http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf [14.08.2018]

[53] Kumar, V., Pang-Ning, T. (2018): Cluster Analysis:Basic Concepts and Algorithms [Internet], raspoloživo na: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf [14.08.2018]

[54] Kumar, V., Pang-Ning, T. (2018): Cluster Analysis:Basic Concepts and Algorithms [Internet], raspoloživo na: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf [14.08.2018]

[55] Helbing, D., Moise, I., Pournaras, E. (2014): Density-Based Clustering [Internet], raspoloživo na: https://www.ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-dam/documents/education/Spring2015/datascience/clustering2.pdf [14.08.2018]

3.3.2 Asocijacijska pravila

Tehnika rudarenja asocijacijskih pravila za svoj cilj ima pronalazak interesantnih asocijacijskih ili korelacijskih veza unutar velikog seta podataka. Rezultati dobiveni provođenjem ovakve tehnike nazivaju se asocijacijska pravila. Podrška pravila (*eng. Rule support)* i pouzdanost (*eng. confidence)* smatraju se glavnim mjerama interesantnosti (*eng. interestingness)* asocijacijskog pravila, pa se tako interesantnim pravilima smatraju ona koja zadovolje minimalne vrijednosti oba parametra, postavljene od strane područnih stručnjaka.[56]

Ovaj koncept nadalje biti će nadalje pojašnjen na primjeru rudarenja asocijacijskih pravila na transakcijskoj bazi podataka supermarketa, tehnici poznatijoj kao analiza košarice dobara (*eng. market basket Data analysis*)*,* u kojoj se analizom transakcijskih podataka o prodanim proizvodima pokušavaju pronaći pravila, kao npr. koji se proizvodi često kupuju zajedno.[57]

Tako u ovom kontekstu pokazatelj *podrška pravila* predstavlja udio transakcija koje sadrže predmete za koje je utvrđeno asocijativno pravilo u ukupnom broju transakcija. Pokazatelj *pouzdanosti* u ovom slučaju predstavlja omjer broja transakcija koje sadrže predmete za koje je utvrđeno asocijacijsko pravilo te broja transakcija koje sadržavaju jedan od predmeta. [58]

Razvijen je velik broj algoritama za podršku procesa rudarenja asocijativnih pravila. Potrebno je napomenuti kako bi svaki od algoritama trebao pronaći jednaka asocijacijska pravila, no međusobno se razlikuju po računalnoj efikasnosti te zahtjevima za radnom memorijom. Jedan od najkorištenijih algoritama jest Apriori algoritam, čiji se rad zasniva na dva koraka: .[59]

1. **Generiranje frekventnih setova predmeta –** odnosno prikupljanje setova podataka koji zadovoljavaju minimalnu vrijednost pokazatelja podrške pravila
2. **Generiranje pouzdanih asocijativnih pravila iz frekventnih setova –** Iteracijom po frekventnim setovima pronalaze se setovi koji zadovoljavaju minimalnu vrijednost pokazatelja pouzdanosti.

[56] Tudor, I. (2008), Association rule mining as a data mining technique, BULETINULuniversitatii Petrol-Gaze din Ploiesti*,* vol. LX, str. 50
[57] Towards Data Science (2017): A Gentle Introduction on Market Basket Analysis—Association Rules [Internet], raspoloživo na: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce [14.08.2018]
[58] Informationbuilders (2015): Explanation of the Market Basket Model, [Internet], raspoloživo na: https://infocenter.informationbuilders.com/wf80/index.jsp?topic=%2Fpubdocs%2FRStat16%2Fsource%2Ftopic49.htm [15.08.2018]
[59] Tudor, I. (2008), Association rule mining as a data mining technique, BULETINULuniversitatii Petrol-Gaze din Ploiesti*,* vol. LX, str. 50

3.3.3 Regresija

Regresijska analiza najpopularnija je metoda numeričke predikcije, odnosno predikcije numeričke kontinuirane vrijednosti, a regresijski model definira veze između jedne ili više nezavisnih ili prediktorskih varijabli i jedne zavisne, odnosno ciljne varijable. Regresijska analiza idealan je odabir za provedbu predikcije onda kada su sve nezavisne varijable izražene u kontinuiranim vrijednostima.[60]

U kontekstu rudarenja podataka nezavisne varijable su atributi koji opisuju određeno opažanje, odnosno redak u bazi, dok ciljna varijabla predstavlja vrijednost koju želimo predvidjeti. Važno je napomenuti kako se većina tipova modela kao što su stabla odlučivanja i neuronske mreže mogu primijeniti na zadacima regresije kao i na onima klasifikacije (predviđanja diskontinuiranih vrijednosti)[61]

Cilj regresije jest utvrditi vrijednost parametara regresijske funkcije kako bi definirali funkciju koja najbolje odgovara pruženom setu podataka.

$$y = F(x, \theta) + e$$

Gore navedena jednadžba objašnjava odnose između ovakvih podataka u simbolima, odnosno predstavlja regresiju kao proces procjene vrijednosti diskontinuirane varijable (y) kao funkciju (F) jedne ili više nezavisnih varijabli (x_1, x_2, …, x_n), seta parametara (θ_1, θ_2, …, θ_n) te greške (e). Proces učenja regresijskog modela uključuje pronalaženje onih vrijednosti parametara koji će rezultirati najmanjom greškom. [62]

Postoje različite skupine regresijskih funkcija kao i različiti načini mjerenja greške, a pod skupine funkcija podrazumijevamo sljedeće: [63]

- **Linearna regresija**
- **Multivarijacijska linearna regresija**
- **Nelinearna regresija**
- **Multivarijacijska nelinearna regresija**

[60] Masoud Yaghini (2010): Data Mining: Prediction – Regression Analysis [Internet], raspoloživo na: http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_882/DM_05_07_Regression%20Analysis.pdf [15.08.2018]
[61] Halili, F., Rustemi, A. (2016): Predictive Modeling: Data Mining Regression Technique Applied in a Prototype, Department of Informatics, State University of Tetovo, Macedonia, str. 210
[62] Chauhan R.K., Shringar, R. Singh, N. (2012): Data Mining with Regression Technique, Journal of Information Systems and Communication, Volume 3, str. 199-202
[63] Halili, F., Rustemi, A. (2016): Predictive Modeling: Data Mining Regression Technique Applied in a Prototype, Department of Informatics, State University of Tetovo, Macedonia, str. 210

3.3.4 Klasifikacijske tehnike

Metode klasifikacije sastoje se od predviđanja određenog ishoda baziranome na danome inputu. Za razliku od regresijskih metoda, kod kojih je cilj predviđanja diskontinuirana numerička varijabla, klasifikacijske metode za cilj imaju predviđanje diskontinuirane varijable, odnosno klase ili *labela.*[64]

Kako bi se predvidio ishod, algoritam procesuira *training* set podataka koji se sastoji od atributa i pripadajućih ishoda, odnosno ciljne varijable. Algoritam pokušava otkriti veze između atributa koje bi omogućile predviđanje ciljne varijable, proces koji se definira kao „učenje" algoritma. U sljedećem koraku algoritmu se daje prethodno neviđeni set podatka, koji sadrži jednak set atributa, uz izostanak ciljne varijable. Algoritam tada analizira dani set i na temelju poznatih atributa izrađuje predikciju.[65]

Training set

Age	Heart rate	Blood pressure	Heart problem
65	78	150/70	Yes
37	83	112/76	No
71	67	108/65	No

Prediction set

Age	Heart rate	Blood pressure	Heart problem
43	98	147/89	?
65	58	106/63	?
84	77	150/65	?

Slika 4. Primjer training i prediction seta podataka

Izvor: https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

U tipičnom procesu klasifikacije podataka, evaluacijska metrika uključena je u procesu učenja algoritma, kao i u procesu testiranja algoritma. U procesu učenja algoritma rezultati se evaluiraju u svrhu optimizacije klasifikacijskog algoritma, dok se u testnoj fazi evaluacijska metrika koristi za mjerenje efektivnosti algoritma pri testiranju s neviđenim podacima. Metrike evaluacije klasifikacijskih tehnika uključuju:[66]

- **Točnost (*eng. accuracy*)** kao omjer točnih klasifikacija i ukupnog broja ispitanih objekata

[64] MLmastery (2017): Difference Between Classification and Regression in Machine Learning [Internet], raspoloživo na: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ [15.08.2018]
[65] Viana. L, Voznika, F. (2014): Data Mining Classification [Internet], raspoloživo na: https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf, [15.08.2018]
[66] Hossin, M., Sulaiman, M.N (2015): A Review on Evaluation Metrics For Data Classification Evaluations

- **Stopu pogreške (*eng. Error rate*)** kao omjer netočnih klasifikacija i ukupnog broja ispitanih objekata
- **Preciznost (*eng. Precision)*** kao omjer točno predviđenih pozitivnih objekata i ukupnog broja pozitivno predviđenih objekata.
- **Opoziv (*eng. Recall)*** kao omjer točno predviđenih pozitivnih objekata i ukupnog broja točno predviđenih objekata.

Neke od klasifikacijskih tehnika u kontekstu rudarenja podataka uključuju:

- **Stabla odlučivanja –** Tehnika formiranja *flowcharta* koji svojim grananjima podsjeća na strukturu stabla, kod kojega je svaki od čvorova testira objekt prema određenom atributu, svaka od grana predstavlja rezultat testa, a svaki od listova predstavlja *label***,** odnosno ciljnu varijablu testa. Prednost ovakve metode je u jednostavnosti razumijevanja i interpretacije, te u primjenjivosti na numeričke i kategoričke vrijednosti.[67]
- **K-najbliži susjedi (*eng. k-nearest neighbor*) –** Metoda za klasificiranje objekata bazirana na najbližim primjerima iz training seta u prostoru. Ovakav tip učenja bazira se na učenju po slučaju (*eng. instance based learning)* odnosno lijenog učenja (eng. *lazy learning)*. Spada u najjednostavnije algoritme klasifikacije, a funkcionira po principu dijeljenja prostora *label* varijablama iz training seta. Prostor se dodjeljuje onoj *label* vrijednosti čija je vrijednost najbrojnija između k-najbližih susjeda. Parametar k u ovom slučaju označava koliko će susjednih vrijednosti pri klasifikaciji algoritam analizirati. [68]
- **Neuronske mreže** jest tehnika modelirana prema sintetiziranom procesu učenja u kognitivnim sustavima te neurološkoj funkciji mozga. Prvi korak uključuje dizajn specifične mrežne arhitekture, sačinjene od određenog broja „slojeva“ i „neurona“. Mreža je zatim izložena procesu učenja, u kojem „neuroni“ provode iterativan proces pronalaženja optimalnih težinskih faktora, odnosno onih koji prouzrokuju minimalnu vrijednost greške.[69]

[67] Han J., Kamber, M. (2001): Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, USA, str 72
[68] Aggarwal, N., Gupta, M. (2010): Classification Techniques Analysis, UIET Punjab University Chandigarh, NCCI 2010, 19-20
[69] Aggarwal, N., Gupta, M. (2010): Classification Techniques Analysis, UIET Punjab University Chandigarh, NCCI 2010, 19-20

3.4 Metodologija rudarenja podataka

3.4.1 Potreba za metodologijom

Najčešće navedene probleme kod provođenja projekata rudarenja podataka u kontekstu poslovne inteligencije u poduzeću sačinjava nekoliko stavki:

- Nedovoljno razumijevanje projekta zbog visoke kompleksnosti projekta
- Nedovoljno prepoznavanje poslovne inteligencije kao inicijative aktivne koja povezuje gotovo sve organizacijske dijelove
- Nedostatak iterativne razvojne metode
- Neprikladna struktura razvojnog tima
- Previše zavisnosti o međusobno nepovezanim metodama i alatima

Sve gore navedene problematične stavke mogu se sažeti u jedinstven problem – **nedostatak standardizirane metodologije** za provođenje ovakve vrste projekata.[70]

U prethodnim dijelovima rada rudarenje podataka definirano je kao kompleksan proces koji za uspješnu provedbu zahtjeva različite alate i ljudske vještine. Standardiziran procesni model, odnosno metodologija pomogla bi u razumijevanju te upravljanju interakcijama u svim fazama procesa.

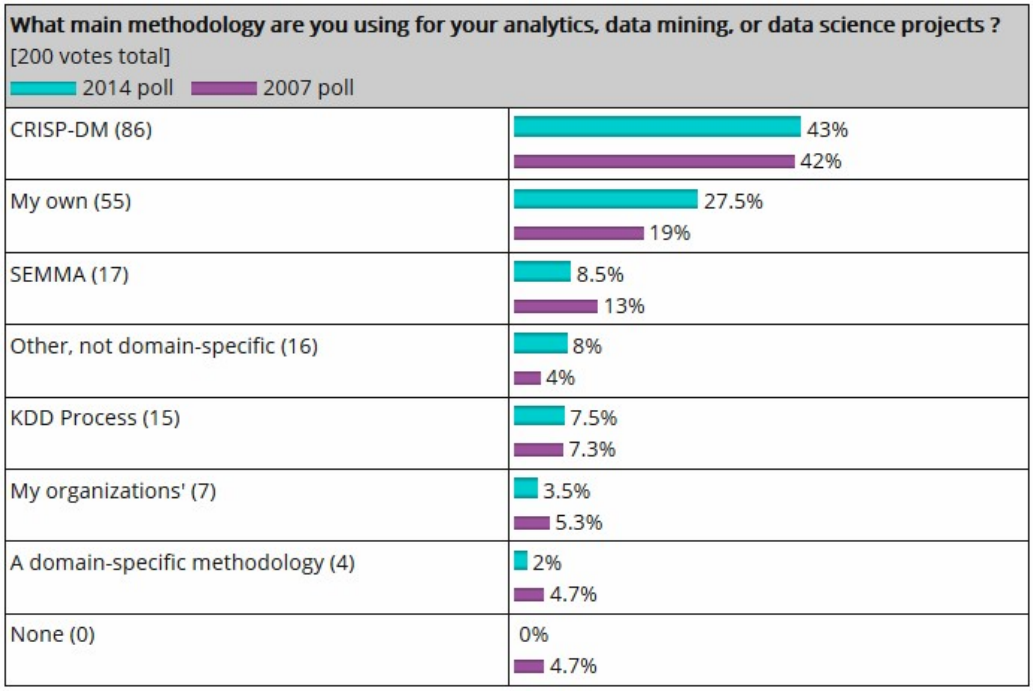
Usvajanje standardizirane metodologije imalo bi pozitivan utjecaj na:[71]

- **Tržište** – model može služiti kao referentna točka za raspravljanje i pojašnjavanje problema svim zainteresiranim osobama u projektu, osobito na strani korisnika ili naručitelja rješenja. Osim toga standardan model omogućio bih lakšu usporedbu alternativa koje poslovnim organizacijama stoje na raspolaganju, te razumnija očekivanja od razvoja i rezultata provedbe ovakvih projekata.
- **Ponuditelje ovakvih rješenja** – Standardiziran model smanjio bi potrebu obrazovanja klijenata o čestim problemima kod provedbe projekata rudarenja podataka. Također, fokus sa pitanja treba li ovakav sustav uopće implementirat u organizaciju bio bi prebačen na pitanje na koji način rudarenje podataka može riješiti određeni poslovni problem.

[70] Gonzales-Aranda, P et al. (2008): Towards a Methodology for Data Mining Project Development: The Importance of Abstraction, Universidad Politenica de Madrid, Madrid, Spain, str. 4
[71] Hipp, J., Wirth, R. (2015): CRISP-DM: Towards a Standard Process Model for Data Mining, DaimlerChrysler Research & Technology FT3/KL

- **Analitičare i eksperte na polju rudarenja podataka** – Znatno bi se olakšalo praćenje rezultata projekta kao i izrada dokumentacije, što je naročito važno za polje rudarenja podataka, gdje se tim često razlikuje od stručnjaka različitih vještina i pozadina.

Ovakav problem prati organizacija KDnuggets, a usporedba rezultata istraživanja o korištenju metodologija u provođenju projekata rudarenja podataka iz 2007 i 2014 godine pokazuje iznenađujuće stabilne rezultate [72]



Slika 5. Rezultati istraživanja o korištenju metodologija

Izvor: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Iz rezultata je vidljivo kako CRISP-DM metodologija i dalje uvjerljivo zauzima prvo mjesto. Zanimljivo je primijetiti kako je porastao broj uporabe vlastite metodologije, što se djelomično može objasniti nedostatkom metodologije prikladne za nove izazove na polju rudarenja podatak i općenito znanosti o podacima. [73]

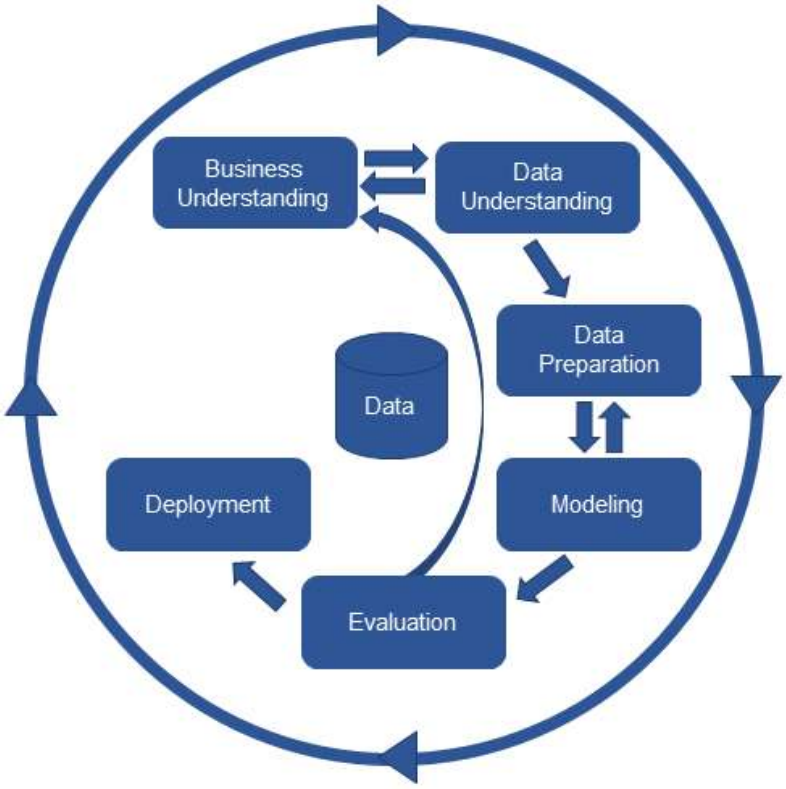
[72] KDnuggets (2014): CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Internet], raspoloživo na: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[73] KDnuggets (2014): CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Internet], raspoloživo na: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

U nastavku rada CRISP-DM metodologija biti će kratko analizirana, uz pobliže pojašnjenje svake od pripadajućih faza procesa.

3.4.2 CRISP-DM

CRISP-DM (*Cross-industry standard process for Data Mining)* metodologija će u nastavku ovog rada biti definirana kroz pobliže pojašnjavanje svake od 6 sastavnih faza.



Slika 6 CRISP-DM proces

Izvor: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

Slika 6 prikazuje iterativnu prirodu CRISP-DM procesa. Iterativnim ovaj proces čini činjenica da rezultati pojedinih faza ponekad mogu zahtijevati vraćanje na prethodne faze projekta. Na primjer, rezultat faze modeliranje može zahtijevati vraćanje procesa na prethodnu fazu zbog potrebe za dodavanjem dodatnih podataka, ili pripremanja podataka na drugačiji način.[74]

[74] PAM Analytics (2014): CRISP-DM Methodology, [Internet], raspoloživo na: http://www.pamanalytics.com/downloads/The%20CRISP-DM%20Methodology.pdf [20.08.2018]

3.4.2.a Razumijevanje poslovnog problema (eng. Business Understanding)

Inicijalna faza fokusira se na razumijevanje ciljeva projekta iz poslovne perspektive. Na ovakav način definirani ciljevi prebacuju se u kontekst rudarenja podataka, a zatim se izrađuje preliminarni projektni plan dizajniran u svrhu ostvarenja definiranih ciljeva. [75]

Tablica 2 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Određivanje poslovnih ciljeva	*Poslovni kontekst* *Poslovni ciljevi* *Kriteriji poslovnog uspjeha*
Procjena situacije	*Zahtjevi, pretpostavke i ograničenja resursa* *Rizici i nepredviđene situacije* *Terminologija* *Cost-benefit analiza*
Određivanje ciljeva rudarenja podataka	*Ciljevi rudarenja podataka* *Kriteriji uspješnosti rudarenja podataka*
Izrada projektnoga plana	*Projektni plan* *Inicijalna procjena potrebnih alata i tehnika*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

Ova faza krucijalna je za uspjeh ukupnog projekta, poslovne ciljeve potrebno je točno definirati kako se velika količina resursa i vremena ne bi uložila u dobivanje točnog odgovora na krivo postavljena pitanja. Također, potrebno je definirati jasne kriterije uspjeha projekta, poglavito za one tehničkoga karaktera u kontekstu rudarenja podataka.[76]

3.4.2.b Razumijevanje podataka (eng. *Data understanding*)

Faza razumijevanja podataka uključuje pobliže sagledavanje podataka raspoloživih za provođenje rudarenja. Ovo je kritičan korak u izbjegavanju neočekivanih problema tijekom sljedeće faze - pripremanja podataka, koja je vremenski obično najzahtjevnija.[77]

Važno je napomenuti kako je u tijeku ove faze moguće uočiti probleme koji zahtijevaju povratak u prethodnu fazu, kako bi se preispitalo razumijevanje poslovnog problema ili preoblikovao plan.[78]

[75] Hipp, J., Wirth, R. (2015): CRISP-DM: Towards a Standard Process Model for Data Mining, DaimlerChrysler Research & Technology FT3/KL

[76] Smart Vision (2016): What is the CRISP – DM methodology?, [Internet], raspoloživo na: https://www.sv-europe.com/crisp-dm-methodology/ [20.08.2018]

[77] IBM (2016): Data Understanding Overview [Internet], raspoloživo na: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_data_understanding_phase.htm, [20.08.2018]

Tablica 3 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Prikupljanje inicijalnih podataka	*Izvještaj o prikupljanju inicijalnih podataka*
Opisivanje podataka	*Izvještaj o opisivanju podataka*
Istraživanje podataka	*Izvještaj o istraživanju podataka*
Potvrđivanje kvalitete podatka	*Izvještaj o kvaliteti podataka*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

Faza započinje prikupljanjem inicijalne kolekcije podataka, a u nastavku se provode akcije za daljnje upoznavanje s podacima, identificiranje problema u smislu kvalitete podataka, otkrivanje prvih uvida u podatke, te detektiraju interesantni podsetovi kako bi se formirale hipoteze u svrhu pronalaženja skrivenih informacija. [79]

3.4.2.c Pripremanje podataka (eng. *Data preparation*)

Pripremanje podataka jedan je od najvažnijih te često vremenski najiscrpnijih aspekata rudarenja podataka. Procjenjuje se da priprema podataka obično uzima 50-70% napora i vremena u projektu. Ukoliko se prethodnim fazama razumijevanja poslovnog problema i podataka pridodalo dovoljno vremena, ovo vrijeme može biti umanjeno, no i dalje je potrebno uložiti popriličan broj resursa u pripremu podataka za rudarenje. [80]

Tablica 4 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Odabir podataka	*Izvještaj uključivanja/isključivanja podataka*
Čišćenje podataka	*Izvještaj o čišćenju podataka*
Konstrukcija podataka	*Derivirani atributi* *Generirani redci*
Integriranje podatka	*Spojeni setovi podataka*
Formatiranje podatka	*Formatirani podaci* *Dataset* *Opis dataseta*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

[78] Dummies (2015): Phase 2 of the CRISP-DM process model: Data understanding [Internet], raspoloživo na: https://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/ [20.08.2018]
[79] Data Science Central (2016): CRISP DM – A Standard Methodology to Ensure a Good Outcome [Internet], raspoloživo na: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome [20.08.2018]
[80] IBM (2016): IBM SPSS Modeler CRISP-DM Guide [Internet], raspoloživo na: https://inseaddataanalytics.github.io/INSEADatalytics/CRISP_DM.pdf, [20.08.2018]

Pri odabiru podataka važno je voditi brigu o tome koji su nam podaci zaista potrebni za provođenje analize, te imati na umu i tehnička ograničenja, kao što su broj redaka koje alat koji koristimo može analizirati, te jeli format podataka odgovara potrebama analize. [81]

Odabrane podatke potrebno je očistiti, što podrazumijeva korištenje tehnika kao što su isključivanje onih redaka sa problematičnim ili nedostajućim vrijednostima, zamjena istih s default vrijednostima, ili korištenje nekih od naprednijih tehnika modeliranja, kao što su predviđanja nedostajućih vrijednosti. [82]

Također, moguće je da će za provedbu sljedeće faze procesa biti potrebni dodatni podaci, pa se u ovoj fazi provodi konstruiranje novih atributa ili redaka. Za ilustraciju derivacije novih atributa poslužiti ćemo se primjerom, pa tako atribut „površina“ može biti generiran množenjem atributa „dužina“ i „visina“. Dodavanje novih redaka također možemo opravdati potrebom za određenim redcima za kojima inače ne postoji potreba u transakcijskom sustavu, kao npr. kupac koji tijekom prošle godine nije kupio niti jedan proizvod. [83]

3.4.3.d Modeliranje

U fazi modeliranja, odabiru se i primjenjuju različite metode rudarenja podatka, te se njihovi parametri podešavaju na optimalne vrijednosti. Obično postoji nekoliko različitih tehnika za isti tip problema rudarenja podataka. Neke tehnike mogu imati specifične zahtjeve u pogledu oblika podataka, stoga je vraćanje u prethodnu fazu pripreme podataka neophodno kako bi se podaci doveli u oblik pogodan za primjenu odabrane tehnike.[84]

U nastavku rada će, kao što je to bio slučaj s prethodnim fazama procesa, u tabličnom obliku biti prezentirani zadaci unutar faze te pripadajući outputi.

---

[81] Dummies (2015): Phase 3 of the CRISP-DM process model: Data understanding [Internet], raspoloživo na: https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-understanding/ [20.08.2018]

[82] Singular (2016): CRISP-DM Phase III: Data Preparation. Data analysis and features selection [Internet], raspoloživo na: https://data.sngular.team/en/art/51/crisp-dm-phase-iii-data-preparation-data-analysis-and-features-selection [20.08.2018]

[83] Smart Vision (2016): Data Preparation [Internet], raspoloživo na: https://www.sv-europe.com/data-preparation [20.08.2018]

[84] Smart Vision (2016): What is the CRISP – DM methodology?, [Internet], raspoloživo na: https://www.sv-europe.com/crisp-dm-methodology/ [20.08.2018]

Tablica 5 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Odabir tehnike modeliranja	*Tehnika modeliranja* *Pretpostavke o modeliranju*
Generiranje dizajna testa	*Dizajn testa*
Izrada modela	*Postavke parametara* *Modeli* *Opisi modela*
Procjena modela	*Procjena modela* *Procjena parametara*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

Prije same izgradnje modela, važno je generirati proceduru ili mehanizam testiranja kvalitete i validnosti modela. Npr. kod *supervised* tehnika rudarenja podataka kao što je klasifikacija, često se koriste stope pogreške (eng. *error rates)*, stoga je dijeljenje dataseta na set za učenje i set za testiranje neophodno.[85]

3.4.3.e Evaluacija

U ovoj fazi projekta, izrađeni model odnosno modeli čine se kao kvalitetni sa stajališta analize podataka. Prije nastavka u završnu fazu, odnosno prezentaciju rezultata, potrebno je detaljno evaluirati model te pregledati sve korake poduzete u svrhu izgradnje istoga, kako bi se osiguralo da model na ispravan način ostvaruje poslovne ciljeve. Ključni cilj jest odrediti postoji li poslovni problem koji nije dostatno adresiran. Na kraju ove faze, trebala bi se donijeti odluka o korištenju rezultata rudarenja podataka.[86]

Tablica 6 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Evaluacija rezultata	*Procjena rezultata rudarenja podataka* *Procjena zadovoljenja poslovnih kriterija* *Odobreni modeli*
Procjena procesa	*Procjena Procesa*
Određivanje sljedećih koraka	*Popis mogućih sljedećih koraka* *Odluka o sljedećim koracima*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

Prethodni koraci evaluacije bili su vezani uz preciznost i generalnu ispravnost modela. Evaluacijski koraci u ovoj metodologiji podrazumijevaju procjenu razine na kojoj dizajnirani

[85] CRISP-DM (2014): Modeling [Internet], raspoloživo na: http://crisp-dm.eu/modelling/ [20.08.2018]
[86] Smart Vision (2016): What is the CRISP – DM methodology?, [Internet], raspoloživo na: https://www.sv-europe.com/crisp-dm-methodology/ [20.08.2018]

model zadovoljava poslovne ciljeve, te se pokušava definirati poslovni razlog zbog kojeg je ovakav model uspješan. Rezultati rudarenja podataka također sadržavaju modele koji nužno nisu povezani s originalnim poslovnim ciljevima, ali postoji mogućnost da baš oni ukažu na dodatne izazove, informacije ili smjernice za daljnji razvoj.[87]

3.4.3.f Implementacija

Kreiranje i evaluacija modela ne znači ujedno i kraj projekta. Čak i kada je svrha modela povećati obim znanja o podacima, stvoreno znanje treba biti organizirano i prezentirano na način da ga korisnik može upotrijebiti. Ova faza često uključuje primjenu „živih" modela unutar organizacijskog procesa donošenja odluka – npr. personalizacija web-stranica u *real-time-u* ili automatski obnavljajući proces *scoringa* u marketinškim bazama podataka. Ovisno o zahtjevima, faza implementacije može biti jednostavna kao što je generiranje izvještaja, ili kompleksan proces implementacije ponavljajućeg procesa koji se proteže kroz organizaciju.[88]

Tablica 6 Fazni zadaci i pripadajući outputi

Generički zadaci	Outputi
Planiranje implementacije	*Plan implementacije*
Praćenje i održavanje plana	*Praćenje i održavanje plana*
Izrada završnog izvještaja	*Završni izvještaj* *Završna prezentacija*
Osvrt na projekt	*Iskustvo* *Dokumentacija*

Izvor: The Modeling Agency, https://www.the-modeling-agency.com/crisp-dm.pdf

[87] CRISP-DM (2014): Evaluation [Internet], raspoloživo na: http://crisp-dm.eu/modelling/ [20.08.2018]
[88] The Modeling Agency (2015): CRISP-DM 1.0 [Internet], raspoloživo na: https://www.the-modeling-agency.com/crisp-dm.pdf [20.08.2018]

4. ANALIZA I RUDARNJE PODATAKA NA PRIMJERU

U nastavku rada neke od prethodno definiranih tehnika i koncepata biti će primjene na odabranom setu podataka. Praktičan dio ovog rada, odnosno proces primjene rudarenja podatka slijediti će, koliko god je to moguće, prethodno definiranu CRISP-DM metodologiju.

Iz ukupnoga Yelp seta podataka – za provedbu praktičnoga dijela ovoga rada odabrana su dva seta, i to:

- Set podataka o poslovnim subjektima
- Set podatka o recenzijama

Neki od navedenih setova podataka biti će analizirani kao zasebne cjeline, dok će ostali biti međusobno pripajani u svrhu provođenja analize.

Prije provođenja same analize objasniti će se odabrani alati za podršku manipulaciji podataka te primjeni tehnika rudarenja podataka.

4.1 Odabrani alati

Za izradu praktičnoga dijela ovog rada korišteno je nekoliko alata, a glavni kriteriji odabira istih bili su primjenjivost za obradu velike količine podataka, dostupnost te autorova upoznatost s istima.

4.1.1 Infrastruktura

S obzirom na veličinu odabranog seta podataka, te posljedičnim zahtjevima za radnom memorijom i brojem threadova grafičke kartice, za potrebe izrade praktičnog dijela ovoga rada korištena je virtualna mašina (eng. *virtual machine*) pružatelja IaaS (*Infrastructure as a Service*) usluga Paperspace**.**

Paperspace pruža usluge iznajmljivanja računalnih resursa u oblaku na gotovo svim platformama, uključujući virtualne mašine, high-end workstatione za animacijska studija, gotove platforme za provođenje strojnoga učenja, te virtualne mašine pogodne za igranje računalnih igara.[89]

[89] Paperspace (2018): About Paperspace [Internet], raspoloživo na: https://www.paperspace.com/about, [21.08.2018]

Konfiguracija virtualne mašine korištene za izradu ovog rada je sljedeća:

- NVIDIA Quadro P4000 with 1792 CUDA cores.
- 8 x CPU
- 8 GB GPU dedicated
- 30GB RAM

4.1.2. Manipulacija podacima i primjena tehnika rudarenja podataka

Kao alat za podršku manipulaciji i rudarenju podataka odabran je programski jezik Python 3.6. Python je programski jezik opće namjene, interpretiran i visoke razine. Zbog svoje jednostavne sintakse, izrazito čitljivoga koda te velikoga broja librarya, primijenjen je u raznim domenama kao što su razvoj web aplikacija, analizi podataka, edukaciji, izradi grafičkih sučelja, razvoju softvera te razvoju brojnih aplikacija poslovne primjene kao što su ERP i E-commerce sustavi. [90]

Neki od glavnih librarya i aplikacija korištenih u praktičnom dijelu ovog rada uključuju:

- Pandas – open source library koji omogućava visoko performansne i lako razumljive strukture podataka te alate za analizu.[91]
- Scikit-learn – jednostavan i efikasan library za podršku rudarenju podataka i strojnom učenju. [92]
- Jupyter Notebook – Open source web aplikacija koja omogućava korisniku kreiranje i dijeljenje dokumenata.[93] Ovakve „bilježnice“ korištene su u ovom radu kod zadataka analize podataka zbog mogućnosti izvršenja koda „u hodu“, za razliku od tradicionalnih skripta koje je potrebno pokretati kao cjelovite.

Ovaj alat također je izabran zbog dostupnosti, visokog stupnja primjenjivosti na problem istraživanja te upoznatosti autora s istim.

[90] Python Software Foundation (2018): Applications for Python [Internet], raspoloživo na: https://www.python.org/about/apps/ [21.08.2018]
[91] Pandas (2018): About Pandas [Internet], raspoloživo na: https://pandas.pydata.org/ [21.08.2018]
[92] Scikit-learn (2018) About [Internet], raspoloživo na: http://scikit-learn.org/stable/ [21.08.2018]
[93] Project Jupyter (2018) Notebook [Internet], raspoloživo na: http://jupyter.org/ [21.08.2018]

4.1.3 Vizualizacija podataka

Kao glavni alat za podršku vizualizaciji podataka u ovom radu biti će korišten Tableau Desktop. Tableau je alat za podršku provođenju procesa analize podataka, sa naglaskom na njihovu vizualizaciju, pa je tako smatran industrijskim „zlatnim standardom“ u domeni vizualne analitike podataka.[94]

4.2 Analiza i rudarenje podatka nad odabranim datasetovima

4.2.1 Razumijevanje poslovnog problema

Yelp je jedan od najpopularnijih online servisa za recenziranje i pronalaženje lokalnih uslužnih djelatnosti. Prema podacima organizacija Datanyze, prema globalnome tržišnom udjelu u grani servisa za recenziranje, Yelp zauzima 11. Mjesto s 1.94% udjela, zaostajući tako za vodećim Google Reviews servisom za 1.32% postotna poena.[95]

Korisnici servisa ocjenjuju poslovne subjekte u tekstualnom obliku – pišući tekstualne recenzije, te dodatno kvantificiraju svoje utiske dodjeljivanjem ocjene od 1 do 5.

Prosječan broj zvjezdica dodijeljenih određenom subjektu unutar servisa vrlo je važan indikator njegove popularnosti i kvalitete, a pripadajuće tekstualne recenzije mogu biti od visoke važnosti pri donošenju odluke korisnika servisa.

Iz gore navedenoga možemo zaključiti kako je uz pripadajuću tekstualnu recenziju od izrazite važnosti dodijeliti odgovarajuću ocjenu poslovnog subjekta. Cilj ovog projekta jest izraditi model pogodan za klasifikaciju napisane recenzije kao one pozitivnog ili negativnog karaktera, te predviđanje pripadajuće ocijene. Uspješno implementiran model, trebao bi uz visoku razinu točnosti na temelju teksta recenzije odrediti radi li se o pozitivnoj ili negativnoj recenziji.

[94] Tableau (2018): What is Tableau? [Internet], raspoloživo na: https://www.tableau.com/products/what-is-tableau [21.08.2018]
[95] Datanyze (2018): Yelp market sharea and competitors analysis [Internet], raspoloživo na: https://www.datanyze.com/market-share/orm/yelp-market-share [22.08.2018]

4.2.2 Razumijevanje i priprema podatka

Za provedbu ovakve analize, potrebni su nam podaci iz dataseta o poslovnim subjektima, te podaci iz dataseta o recenzijama. Podaci o poslovnim subjektima u ovom kontekstu biti će isključivo deskriptivne naravi, odnosno iako neće biti direktno uključeni u model podataka, informacije o subjektima za koje su napisane pripadajuće recenzije pomoći će dodatnom razumijevanju dobivenog modela, te upotpuniti proces implementacije, odnosno dostave rezultata.

Prije početka procesa istraživanja, iz dokumentacije o datasetu[96] izvući će se podaci o broju i obliku pripadajućih atributa.

Dataset o **poslovnim subjektima** sadrži sljedeće atribute (s pripadajućim formatom):

- *business_id* – ID poslovnog subjekta (22 char string)
- *name*– Ime poslovnog subjekta (string)
- *address* – Adresa poslovnog subjekta (string)
- *city* – Grad u kojem se nalazi poslovni subjekt (string)
- *state* – Država u kojoj se nalazi poslovni subjekt (string)
- *postal code* – Poštanski kod subjekta (string)
- *latitude* – Geografska širina lokacije objekta (float)
- *longitude* – Geografska dužina lokacija objekta
- *stars* **–** Prosječna ocjena poslovnog subjekta (float)
- *review_count* **-** Broj recenzija poslovnog subjekta (integer)
- *is_open* – Jeli poslovni subjekt trenutno posluje (integer, 0 ako ne posluje, 1 ako posluje)
- *attributes* – Atributi poslovnog subjekta (json object)

Dataset o **recenzijama** sadrži sljedeće atribute (s pripadajućim formatom):

- *review_id* – ID recenzije (string)
- *user_id* – ID korisnika koji je napisao recenziju (string)

[96] Yelp (2018): Yelp Dataset Documentation [Internet], raspoloživo na: https://www.yelp.com/dataset/documentation/main [22.08.2018]

- *business_id* – ID poslovnog subjekta na kojega se recenzija odnosi (string)
- *stars* – Ocjena dodijeljena uz recenziju (integer)
- *date* – Datum unosa recenzije (string, YYYY-MM-D.D.)
- *text* – Tekst recenzije (string)
- *useful* – broj ocjena recenzije kao korisne od strane drugih korisnika (integer)
- *funny* – broj ocjena recenzije kao smiješne od strane drugih korisnika (integer)
- *cool* – broj ocjena recenzije kao cool od strane drugih korisnika (integer)

Proces istraživanja podataka započinjemo izvlačenjem jednostavnih karakteristika dataseta, prije čega je dataset potrebno importirati u strukturu podataka unutar python librarya Pandas. Ovakve struktura naziva se *data frame* pa će shodno tome svaki od objekata uz pripadajuće ime, imati i dodatak imenu *df.* Za početak importiramo dataset o poslovnim subjektima, formata JSON i veličine 126 MB.

```
In [1]: import pandas as pd

In [2]: business_df = pd.read_json("business.json", lines =True)

In [4]: business_df.describe()

Out[4]:
```

	is_open	latitude	longitude	review_count	stars
count	174567.000000	174566.000000	174566.000000	174567.000000	174567.000000
mean	0.840376	38.627312	-92.679009	30.137059	3.632196
std	0.366258	5.389012	26.240079	98.208174	1.003739
min	0.000000	-36.086009	-142.466650	3.000000	1.000000
25%	1.000000	33.631550	-112.125879	4.000000	3.000000
50%	1.000000	36.144257	-89.410128	8.000000	3.500000
75%	1.000000	43.606181	-79.657609	23.000000	4.500000
max	1.000000	89.999314	115.086769	7361.000000	5.000000

Slika 7 Importiranje i deksriptivna statistika nad datasetom o poslovnim subjektima

Izvor: Izrada autora

Primjenom metode *describe* za rezultat dobivamo rezultate nekih od metoda deskriptivne statistike. Output u vidu tablice daje nam i neke neupotrebljive podatke, no i neke korisne informacije o strukturi dataseta, uključujući:

- Broj ukupnih redaka u datasetu iznosi 174 567.
- Srednja vrijednost atributa is_open iznosi 0.84. S obzirom da je atribut binarnog tipa, odnosno vrijednost 0 označava da poslovni subjekt trenutno ne posluje, a 1 ukazuje na aktivnost poslovanja subjekta, možemo zaključiti da je 84% poslovnih subjekata u datasetu aktivno posluje.

- Prosječan broj recenzija po poslovnom subjektu jest 30, no ovaj pokazatelj možemo uzeti s rezervom s obzirom na visoku standardnu devijaciju (98,2).
- Prosječna ocjena poslovnog subjekta iznosi 3,63 (uz standardnu devijaciju od 1,0).
- Medijan, odnosno pozicijska srednja vrijednost niza ocjena subjekata iznosi 4,5.
- Minimalan broj recenzije koje se odnose na pojedini objekt iznosi 3.
- Maksimalan broj recenzija koje se odnose na pojedini objekt iznosi 7361.

Isti proces importiranja biti će proveden i nad datasetom o recenzijama, također JSON formata, veličine 3,9 GB.

```
In [5]: review_df = pd.read_json("review.json", lines=True)

In [9]: len(review_df)
Out[9]: 5261669
```

Slika 8 Importiranje i izračun broja redaka dataseta o recenzijama

Izvor: Izrada autora

Nakon importiranja dataseta, primjenom funkcije *len* nad importiranim Data frameom *review_df,* doznajemo kako je broj recenzija unutar dataseta 5 261 669.

```
In [10]: review_df.head()
```

Out[10]:

	business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	0W4lkclzZThpx3V65bVgig	0	2016-05-28	0	v0i_UHJMo_hPBq9bxWvW4w	5	Love the staff, love the meat, love the place....	0	bv2nCi5Qv5vroFiqKGopiw
1	AEx2SYEUJmTxVVB18LICwA	0	2016-05-28	0	vkVSCC7xljjrAI4UGfnKEQ	5	Super simple place but amazing nonetheless. It...	0	bv2nCi5Qv5vroFiqKGopiw
2	VR6GpWIda3SfvPC-lg9H3w	0	2016-05-28	0	n6QzIUObkYshz4dz2QRJTw	5	Small unassuming place that changes their menu...	0	bv2nCi5Qv5vroFiqKGopiw
3	CKC0-MOWMqoeWf6s-szl8g	0	2016-05-28	0	MV3CcKScW05u5LVfF6ok0g	5	Lester's is located in a beautiful neighborhoo...	0	bv2nCi5Qv5vroFiqKGopiw
4	ACFtxLv8pGrrxMm6EgjreA	0	2016-05-28	0	IXvOzsEMYtiJI0CARmj77Q	4	Love coming here. Yes the place always needs t...	0	bv2nCi5Qv5vroFiqKGopiw

Slika 9 Prvih 5 redaka dataseta o recenzijama

Izvor: Izrada autora

Kao što je vidljivo na slici 9, primjenom metode *head,* prikazali smo prvih 5 redaka dataseta, kako bi provjerili ispravnost formata importiranih podataka.

U svrhu daljnjeg kvalitetnijeg razumijevanja podatka korištenjem metoda vizualizacije podataka koristiti će se alat Tableau.

Set podataka o recenzijama zbog velikih je vrijednosti unutar pojedinih stupaca problematičan, pa je za uspješnu vizualizaciju iz istoga potrebno ukloniti cijeli stupac,

odnosno atribut *text* koji sadrži tekstove recenzija. Uklanjanje ovog stupca neće imati efekta na kvalitetu vizualizacije podataka, kako taj atribut svejedno ne bi bio korišten pri vizualizaciji.

```
In [10]: review_tableau = review_df.drop(columns=["text"])

In [11]: review_tableau.keys()

Out[11]: Index(['business_id', 'cool', 'date', 'funny', 'review_id', 'stars', 'useful',
                'user_id'],
               dtype='object')
```

Slika 10 Uklanjanje stupca s tekstualnim recenzijama

Izvor: izrada autora

Metodom *drop* ispustili smo atribut, odnosno stupac koji sadrži recenzije u punom tekstualnom obliku, te novonastali data frame spremili pod imenom *review_tableau*. Nakon ispuštanja, metodom *keys* izlistali smo nazive stupaca novonastaloga seta podataka, kako bi se uvjerili da je stupac *text* zaista ispušten, te da su ostali željeni atributi prisutni.

Set podataka o poslovnim subjektima također sadrži atribute nerelevantne za grafičku analizu podataka, pa stoga stvaramo novi set *business_tableau,* na način da iz data framea *business_df* uzimamo samo one atribute potrebne za provođenje analize (*business_id, city, state, is_open, longidute, latitude).*

```
In [13]: business_tableau = business_df[["business_id","city","state","is_open", "longitude", "latitude"]]

In [14]: review_business_tableau = pd.merge(business_tableau, review_tableau, on="business_id", how="inner")

In [15]: review_business_tableau.keys()

Out[15]: Index(['business_id', 'city', 'state', 'is_open', 'longitude', 'latitude',
                'cool', 'date', 'funny', 'review_id', 'stars', 'useful', 'user_id'],
               dtype='object')

In [36]: review_business_tableau.to_csv("review_business_tableau.csv")
```

Slika 11 Ispuštanje atributa i spajanje setova

Izvor: izrada autora

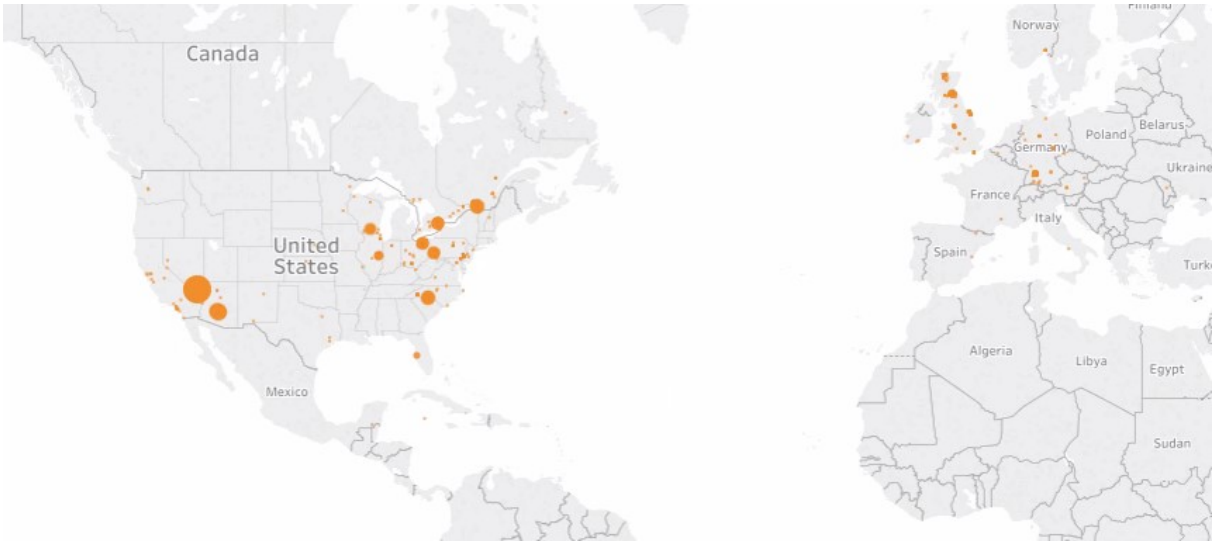
Nakon ispuštanja atributa, novonastali setovi *review_tableau* i *business_tableau* spojiti će se u jedinstven set *review_business_tableau.* Metoda *merge* omogućuje nam spajanje setova na određenom atributu. Ovakav način spajanja podsjeća na onaj u SQL jeziku, gdje se povezivanje vrši preko primarnog, odnosno stranog ključa u drugoj tablici. U ovom slučaju zajednički atribut jest *business_id*, pa će se shodno tome preko njega vršiti spajanje tablica.

Kao vrstu spajanja odabiramo *inner*, što podrazumijeva da će novonastali dataset sadržavati samo one retke čija je vrijednost prisutna u obe tablice, ovakva vrsta spajanja u SQL sintaksi poznata je kao *inner join.*

Nakon pripajanja, provjeravamo sadržava li novonastali dataset sve željene atribute. Nakon provjere atributa, metodom *to_csv* eksportiramo set podataka kao csv (Comma Separated Values) datoteku u svrhu importiranja i daljnje obrade u alatu Tableau.

Kao sljedeći korak grafički će se prikazati distribucija recenzija po mjestu njihova geografksoga nastanka, odnosno geografske lokacija objekta na kojega se recenzija odnosi.

Alat tableau podržava rad s vrijednostima geografske širine i dužine, pa stoga kao atribute za prikaz postavljamo vrijednosti atributa *latitude* i *longitude,* a kao mjeru broj recenzija koje se odnose na određenu lokaciju.

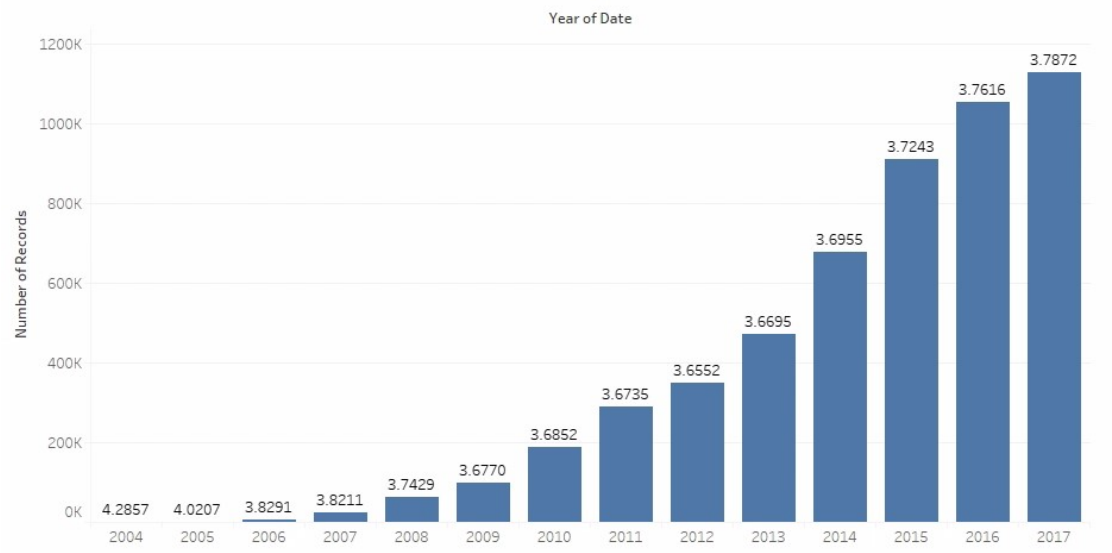


Slika 12 Geografska distribucija recenzija

Izvor: izrada autora u alatu Tableau

Oznake u obliku točkica na karti prikazuju lokacije poslovnih subjekata za koje su recenzije napisane, a veličina svake od oznaka predstavlja broj recenzija napisanih za poslovne subjekte na toj lokaciji.

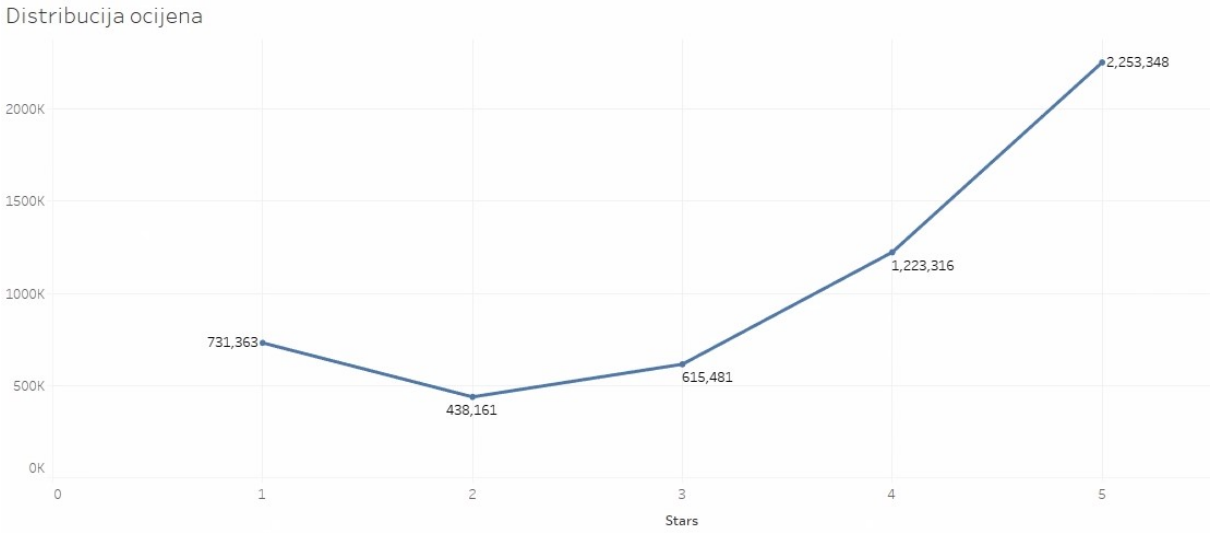
U sljedećemu koraku, biti će prikazana distribucija napisanih recenzija po vremenskoj varijabli, odnosno po godinama nastanka.



Slika 13 Distribucija recenzija po godinama

Izvor: izrada autora u alatu Tableau

Iz grafičkoga prikaza je vidljivo kako je broj recenzija u datasetu rastao od prve zabilježene recenzije u 2004-toj godini sve do posljednje zabilježene u 2017-toj godini. Brojevi na vrhu svakoga od stupaca prikazuju prosječnu vrijednost ocjena dodijeljenih uz recenzije u pripadajućoj godini. Ovakav pokazatelj je kroz godine relativno stabilan, odnosno ne pokazuje stabilan rast kao ni pad srednje vrijednosti ocjena.



Slika 14 Distribucija recenzija po ocjenama

Izvor: izrada autora u alatu Tableau

Izrađeni graf prikazuje nam distribuciju recenzija po dodijeljenim ocjenama.

Vidljivo je kako ocjene nisu pravilno raspoređene, a broj recenzija po ocjenama je sljedeći:

- 731 363 recenzija ima pripadajuću ocjenu 1.
- 438 161 recenzija ima pripadajuću ocjenu 2.
- 615 481 recenzija ima pripadajuću ocjenu 3.
- 1 223 316 recenzija ima pripadajuću ocjenu 4.
- 2 253 348 recenzija ima pripadajuću ocjenu 5.

Iz ovakvih podataka lako je uvidjeti kako distribucija nije jednaka – razlika između najčešće ocjene (ocjena 5 – 2 253 348 recenzija) i ocjene s najmanjim brojem pojavljivanja (ocjena 2 – 438 161 recenzija) značajna je. Ovakva nepravilna distribucija ocjena stvara problem u fazi učenja algoritma. Ovakav problem adresirati će se, zbog praktičnih razloga, u sljedećoj fazi ovoga rada.

4.2.3 Modeliranje, evaluacija i isporuka rezultata

Nakon faze razumijevanja i pripreme podataka, slijedi faza modeliranja, u kojoj će tehnike rudarenja podataka i strojnog učenja primijeniti na pripremljenome setu podatka. Već definirani cilj istraživanja jest razviti algoritam koji će na temelju teksta recenzije predvidjeti pripadajuću ocjenu, odnosno predvidjeti radi li se o recenziji pozitivnoga ili negativnoga karaktera.

Za provedbu ovakvoga zadatka, biti će paralelno korištene dvije najpopularnije metode klasifikacije teksta *Naive Bayes Classification* i *Support Vector Classification*,[97] a evaluacijom oba modela zaključiti će se s kolikom točnošću svaki od modela klasificira recenzije.

Kao prvi korak provođenja metode, potrebno je odabrati varijable koje ćemo uključiti u model. Za izradu modela potrebna su nam dva atributa

[97] Analytics Vidhya (2018): A Comprehensive Guide to Understand and Implement Text Classification in Python [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/ [26.08.2018]

- *Text* - tekstovi recenzija o poslovnim subjektima. Ovaj atribut poslužiti će nam kao input za model, odnosno na temelju vrijednosti ovog atributa biti će klasificirana ciljna varijabla. Ova varijabla u modelu označiti će se sa x.
- *Stars* – Označava ocjenu pridodanu uz napisanu tekstualnu recenziju. U kontekstu ovog modela predstavlja ciljnu, odnosno *label* varijablu, te će se u modelu označiti će sa y.

U prethodnome koraku procesa, ustanovljena je nejednaka distribucija frekvencija ocjena u datasetu, odnosno nebalansirane vrijednosti broja ocjena – situacija u kontekstu strojnog učenja poznata kao problem nebalansiranosti klasa *(*eng. *class imbalance problem*). Tri su moguća rješenja spomenutog problema:[98]

1. Uzorkovanje podataka *(*eng. *Data sampling*) – predstavlja modificiranje seta podataka na način da se proizvede set podataka s jednakom distribucijom klasa.
2. Modifikacija algoritma (*eng. Algorithmic modification)* – podrazumijeva modificiranje algoritma u svrhu prilagođavanja istog na rad s nebalansiranim klasama.
3. Troškovno-osjetljivo učenje (eng. *Cost-sensitive learning*) – Ovakva vrsta rješenja može se primijeniti na razini podataka, na razini algoritma, ili kombinirano na obe razine, na način da se poveća „cijena“, odnosno vrijednost greške modela ukoliko model krivo predvidi rezultat klasifikacije klase koja je prisutnija u modelu, te se na taj način pokuša smanjiti ukupna greška modela.

S obzirom na prirodu podataka u predmetnom datasetu, u ovom radu problemu nebalansiranosti klasa doskočiti će se tehnikom uzorkovanja podataka. S obzirom da se najmanje zastupljena klasa - ocjena 2 ima frekvenciju 438 161, formirati će se set podataka koji će se sastojati od svih vrijednosti s klasom „2“, te uzoraka ostalih klasa odgovarajuće jednake veličine, odnosno 438 161.

Kako bi se tehnika uzorkovanja podataka primijenila na predmetnom datasetu ovoga rada, kao prvi korak podatke je potrebno prebaciti u odgovarajuću strukturu, u ovom slučaju u *listu* (eng. *list*).

[98] Lopez, V. et.al (2013): An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information Sciences Volume 250, str 113 - 141

```
In [65]: stars = review_df["stars"].tolist()

In [39]: reviews = review_df["text"].tolist()

In [66]: type(reviews)

Out[66]: list

In [67]: type(stars)
Out[67]: list
```

Slika 15 Prebacivanje vrijednosti odabranih stupaca u listu

Izvor: Izrada autora

Na slici 15 prikazan je postupak prebacivanja vrijednosti u strukturu liste, pa tako lista *reviews* sadrži sve pripadajuće recenzije atributa *text,* dok lista *stars* sadrži sve vrijednosti ocjena poslovnih subjekata, poredanih po redu uključivanja u listu, što podrazumijeva da recenzija i pripadajuća ocjena imaju jednak indeks liste, tako da npr. ocjena s indeksom liste 2 odgovara recenziji s indeksom liste 2.

U sljedećemu koraku potrebno je odrediti najmanju frekvenciju distribucije ocjena, kako će nam, u svrhu ravnomjerne distribucije ocjena, služiti kao veličina broja vrijednosti za sve klase.

```
In [68]: from collections import Counter

In [44]: frequencies = Counter(stars)

In [45]: frequencies

Out[45]: Counter({5: 2253348, 4: 1223316, 3: 615481, 1: 731363, 2: 438161})

In [61]: frequencies.most_common()
Out[61]: [(5, 2253348), (4, 1223316), (1, 731363), (3, 615481), (2, 438161)]

In [57]: class_limit = frequency_counter.most_common()[-1][1]

In [58]: class_limit
Out[58]: 438161
```

Slika 16 Izračun najniže frekvencije

Izvor: Izrada autora

Iz librarya collections importirana je funkcija Counter. Funkcija kao ulazne parametre uzima listu vrijednosti, a kao output daje *dictionary* u kojemu su ključevi distinktivne vrijednosti koje se pojavljuju u listi, a pripadajuće vrijednosti broj pojavljivanja istih vrijednosti u listi. U ovom slučaju, output jest *dictionary* prigodno nazvan *frequencies* u kojemu su ključevi vrijednosti ocjena (1, 2, 3, 4, 5), a pripadajuće vrijednosti broj pojavljivanja svake od vrijednosti unutar liste.

Provedbom metode *most_common* nad counter dictionaryem *frequencies* kao rezultat dobivamo listu *tupleova*, svaki *tuple* jest par ocjene te pripadajuće broja pojavljivanja vrijednosti poredanih po veličini, od vrijednosti s najvećim brojem pojavljivanja do one s najnižim brojem. Lista tupleova pogodna je za pristupanje vrijednostima preko indeksa liste, pa tako se zadnjemu tupleu pristupamo indeksom [-1], a drugoj vrijednosti unutar odabranoga tuplea, koja označava broj pojavljivanja, indeksom [1]. Na ovaj način unutar varijable *class_limit* spremili smo broj 438 161, koji će u sljedećemu koraku poslužiti kako bi se izradio uzorak svih ocjena, odnosno klasa, jednakoga broja, te se na taj način izradio set podataka s jednakim broj primjera za svaku od klasa.

Nakon dobivenoga broja uzoraka, potrebno je pripremiti nove varijable – liste u kojima će se spremati novonastali set podataka s jednakim brojem klasa.

```
In [87]: count_additions = {star: 0 for star in frequencies.keys()}

In [88]: count_additions
Out[88]: {5: 0, 4: 0, 3: 0, 1: 0, 2: 0}

In [71]: reviews_balanced = []

In [72]: stars_balanced = []

In [74]: for i, y in enumerate(stars):
             if count_additions[y] < class_limit:
                 stars_balanced.append(y)
                 reviews_balanced.append(reviews[i])
                 count_additions[y] += 1

In [85]: Counter(stars_balanced)
Out[85]: Counter({5: 438161, 4: 438161, 3: 438161, 1: 438161, 2: 438161})
```

Slika 17 Uzorkovanje

Izvor: Izrada autora

Varijabla *count_additions* tipa je *dictionary*, a kreirana je u svrhu zaustavljanja *for* petlje, dok su varijable *reviews_balanced* i *stars_balanced* tipa *list,* a upravo će se u njih pohranjivati vrijednosti uzorkovanoga seta podataka.

For petljom prelazimo preko vrijednosti originalne liste ocjena, gdje *i* predstavlja indeks, odnosno poziciju objekta unutar liste, dok *y* predstavlja objekt unutar liste. Tako se svaka vrijednost (u ovom slučaju ocjena) preko koje se prelazi dodaje u listu *stars_balanced,* u listu *reviews_balanced* se preko indeksa liste dodaje odgovarajuća tekstualna recenzija, a u varijabli *count_additions* dodaje se vrijednost 1 onom ključu koji odgovara ocjeni koja je doData u *stars_balanced* varijablu. Ovaj proces se nastavlja sve dok je zadovoljen uvjet

postavljen operatorom *if,* odnosno dok je broj ocjena i pripadajućih recenzija dodijeljen u novi set manji od maksimalne vrijednosti za svaku klasu - 438 161. Još jednom, pozivom funkcije *counter* nad novonastalom listom *stars_balanced* provjerava se jesu li ocjene distribuirane na željeni način.

Sljedeći korak u procesu jest transformiranje recenzija iz tekstualnog oblika u oblik pogodan za input modela. U strojnom učenju, modeli kao input ne primaju tekstualni oblik, već ga je potrebno transformirati u numeričke vrijednosti – vektore. [99]

Neke od tehnika koje korisnicima stoje na raspolaganju pri transformiranju tekstualnih vrijednosti u vektore uključuju:[100]

- Broj riječi (eng. *Word Count*) – alat za provođenje ove tehnike jest *CountVectorizer*, a podrazumijeva spremanje svih riječi koji se pojavljuju u tekstovima, a zatim u obliku vektora bilježi broj puta koji se riječi pojavljuju u pojedinom tekstu..
- Frekvencije riječi (eng. *Word Frequencies)* provode se alatom *TF-IDF Vectorizer (Term Frequency – Inverse Document).* TF-IDF Vectorizer također zabilježava sve riječi u tekstovima te pripadajući broj njihovih ponavljanja, no pridodaje manju težinu onim riječima koje se često pojavljuju u svim tekstvoima (npr. „the“, „a“).

U ovom radu u svrhu vektoriziranja teksta biti će korišten TF-IDF Vectorizer, a funkcija za provođenje istoga implementirana u scikit-learn, python library za podršku strojnom učenju.

```
In [76]: from sklearn.feature_extraction.text import TfidfVectorizer
In [77]: vectorizer = TfidfVectorizer(ngram_range=(1,2))
In [78]: vectors = vectorizer.fit_transform(reviews_balanced)
```

Slika 18 Konverzija teksta u vektore

Izvor: Izrada autora

Pri konverziji važno je odrediti broj *n-gramova* po kojima će se vektorizirati tekstovi. Unigramovi (N=1) promatraju isključivo riječi kao jedan objekt, bigramovi (N = 2) povezuju

[99] Towards Data Science (2017): Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK [Internet], raspoloživo na: https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a [25.08.2018]
[100] Machine Learning Mastery (2017): How to Prepare Text Data for Machine Learning with scikit-learn [Internet], raspoloživo na:

dvije riječi, trigramovi tri (N = 3) itd. Ukoliko su n-gramovi pre kratki, postoji šansa da će se zanemariti bitne razlike, naročito u kontekstu teksta, dok prevelik broj n-gramova često vodi gubljenju generalnog znanja, te fokusiranjem na pojedinačne slučajeve.[101] Kao veličina n-gramova u ovom radu odabrana je kombinacija unigramova i bigramova.

Nakon uspješne transformacije recenzija u tekstualnom obliku u vektore, set podataka potrebno je podijeliti na *training* i *test* set, od kojih training set služi algoritmu za proces učenja, dok test set koristimo u svrhu evaluacije modela.

```
In [42]: from sklearn.model_selection import train_test_split

In [43]: review_train, review_test, star_train, star_test = train_test_split(vectors, stars_balanced, test_size=0.3)
```

Slika 19 Dijeljenje seta podataka

Izvor: Izrada autora

U ovom radu, set podataka biti će podijeljen na način da će 70% podataka pripasi *training* setu, a preostalih 30% podataka biti će dio *test* seta. Ovakvu podjelu podataka definiramo podešavanjem parametra *test_size* na vrijednost 0.3. Dijeljenje dataseta predstavlja posljednji korak pripreme podataka prije njihovoga ulaska u model.

Kao što smo prethodno naveli, u svrhu klasifikacije recenzija koristiti će se dva različita modela. Prvi korišteni klasifikator jest *Naive Bayes*, a nakon importiranja funkcije modela iz librarya *scikit-learn*, potrebno mu je dodijeliti ulazne parametre, odnosno testne setove.

```
In [68]: from sklearn.naive_bayes import MultinomialNB
         classifier_NB = MultinomialNB()

In [69]: classifier_NB.fit(review_train, star_train)
Out[69]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

Slika 20 Treniranje Naive Bayes klasifikatora

Izvor: Izrada autora

Nakon dodavanja ulaznih parametara Naive Bayes klasifikatoru, metodom *fit* pokrećemo treniranje klasifikatora. Uspješno izvršen proces treniranja klasifikatora omogućuje nam konačno pokretanje modela na testnom setu podataka.

[101] Analytics Vidyha (2018): Ultimate guide to deal with Text Data (using Python) – for Data Scientists & Engineers [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/ [25.08.2018]

```
In [71]: predictions_NB = classifier_NB.predict(review_test)
         print(list(predictions_NB[:20]))
         star_test[:20]

         [4, 5, 4, 2, 3, 3, 4, 1, 5, 1, 3, 1, 2, 4, 2, 5, 4, 4, 2, 3]

Out[71]: [4, 5, 3, 1, 2, 3, 5, 1, 5, 2, 2, 5, 1, 5, 3, 5, 4, 4, 2, 4]
```

Slika 21 Rezultati klasifikacije Naive Bayes algoritmom

Izvor: Izrada autora

Pokretanjem modela algoritam prelazi preko recenzija iz testnog seta recenzija *review_test*, te svaku od analiziranih recenzija svrstava u jednu od klasa – ocjene 1 do 5. Prva lista u outputu predstavlja prvih dvadeset recenzija , odnosno procjenu predviđenu od strane algoritma, dok druga lista predstavlja stvarne vrijednosti recenzija.

Iako je ovakav prikaz predviđenih i stvarnih ocjena iz seta podataka služi kao zgodan prikaz outputa algoritma, za ocjenjivanje klasifikacijskoga modela potrebno se poslužiti metrikama za evaluaciju ovakvih algoritama.

```
In [81]: from sklearn.metrics import accuracy_score
         accuracy_score(star_test, predictions_NB)

Out[81]: 0.5437575757575758
```

Slika 22 Izračun točnosti modela

Izvor: Izrada autora

Točnost algoritma prethodno je definirana u ovom radu kao omjer točnih klasifikacija i ukupnoga broja ispitanih objekata. Točnost Naive Bayes klasifikacijskog algoritma u ovom primjeru iznosi 0.5437. Ovaj parametar možemo interpretirati na način da je algoritam od ukupnog broja klasificiranih recenzija točno klasificirao njih 54,37%.

U nastavku rada isti će proces biti proveden LinearSVC (Linear Support Vector Classifier) algoritmom. Algoritam LinearSVC koristiti će jednake ulazne vrijednosti kao i Naive Bayes. Prema istraživanjima, LinearSVC klasifikacijski algoritam obično pokazuje bolje performanse u zadacima tekstualne klasifikacije s većim datasetovima (preko 10 000 redaka), te u radu s dužim tekstualnim dokumentima, iako proces istoga traje znatno duže, te zahtjeva više računalnih resursa.[102]

[102] Hassan, S., Rafi,. M, Shaikh, M.(2012): Comparing SVM and Naïve Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment [Internet], raspoloživo na: https://arxiv.org/ftp/arxiv/papers/1202/1202.4063.pdf [26.08.2018]

```
In [77]: from sklearn.svm import LinearSVC
         classifier_SVC = LinearSVC()

In [78]: classifier_SVC.fit(review_train, star_train)

Out[78]: LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True,
              intercept_scaling=1, loss='squared_hinge', max_iter=1000,
              multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
              verbose=0)

In [79]: predictions_SVC = classifier_SVC.predict(review_test)
         print(list(predictions_SVC[:20]))
         print(star_test[:20])

         [4, 5, 3, 1, 2, 3, 5, 1, 5, 1, 3, 1, 1, 2, 3, 5, 4, 4, 2, 3]
         [4, 5, 3, 1, 2, 3, 5, 1, 5, 2, 2, 5, 1, 5, 3, 5, 4, 4, 2, 4]

In [80]: from sklearn.metrics import accuracy_score
         accuracy_score(star_test, predictions_SVC)

Out[80]: 0.6111757575757576
```

Slika 23 Treniranje, rezultati klasifikacije i evaluacija LinearSVC algoritma

Izvor: Izrada autora

Kao što se iz slike 23 može vidjeti, proces treniranja te pokretanja modela jednak je onom u radu s Naive Bayes klasifikatorom. Vrijednost pokazatelja točnosti modela LinearSVC iznosi 0.6112, a ovaj parametar također možemo interpretirati na način da je algoritam od ukupnog broja klasificiranih recenzija točno klasificirao njih 61,12%, što ukazuje na bolju performansu LinearSVC algoritma u odnosu na Naive Bayes za 6,75 p.p..

U sljedećemu koraku remodelirati ćemo, odnosno pojednostavniti problem klasifikacije. Kao novi cilj modela definiramo klasificiranje recenzije kao pozitivnoga ili negativnog karaktera. Za provođenje ovakve analize, sve ocjene „1“, „2“ i „3“ transformirati će se u vrijednost „0“, koja označava ocjenu negativnog karaktera, a ocjene „4“ i „5“ u „1“, odnosno ocjenu pozitivnoga karaktera.

```
In [84]: stars_simplified = []
         for star in stars:
             if star in [1,2,3]:
                 stars_simplified.append(0)
             else:
                 stars_simplified.append(1)
```

Slika 24 Transformiranje seta ocjena

Izvor: Izrada autora

Korištenjem *for* petlje prelazimo po vrijednostima u listi *stars*, te testiramo njihovu vrijednost – ukoliko je vrijednost ocjene „1“, „2“ ili „3“, u novu listu *stars_simplified* dodajemo vrijednost „0“, a ako je vrijednost „4“ ili „5“, dodajemo vrijednost 1.

Ostatak procesa pripreme podataka, treniranja i pokretanja algoritma jednak je kao kod klasifikacije sa više klasa.

```
In [126]: from sklearn.naive_bayes import MultinomialNB
          classifier_NB_simplified = MultinomialNB()

In [127]: classifier_NB_simplified.fit(review_train_simplified, star_train_simplified)
Out[127]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)

In [128]: predictions_NB_simplified = classifier_NB_simplified.predict(review_test_simplified)
          print(list(predictions_NB_simplified[:20]))
          star_test_simplified[:20]
          [0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0]
Out[128]: [0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0]

In [129]: from sklearn.metrics import accuracy_score
          accuracy_score(star_test_simplified, predictions_NB_simplified)
Out[129]: 0.8474
```

Slika 25 Treniranje, prikaz outputa i evaluacija Naive Bayes Algoritma

Izvor: Izrada autora

Iz slike 25 možemo vidjeti kako model pokazuje znatno bolje performanse otkada je problem klasifikacije pojednostavljen, pa tako vrijednost parametra točnosti iznosi 0.8474, odnosno algoritam ispravno klasificira recenziju u 84.74% slučajeva.

```
In [136]: from sklearn.svm import LinearSVC
          classifier_SVC_simplified = LinearSVC()

In [137]: classifier_SVC_simplified.fit(review_train_simplified, star_train_simplified)
Out[137]: LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True,
               intercept_scaling=1, loss='squared_hinge', max_iter=1000,
               multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
               verbose=0)

In [138]: predictions_SVC_simplified = classifier_SVC_simplified.predict(review_test_simplified)
          print(list(predictions_SVC_simplified[:20]))
          star_test_simplified[:20]
          [0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0]
Out[138]: [0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0]

In [139]: accuracy_score(star_test_simplified, predictions_SVC_simplified)
Out[139]: 0.8996
```

Slika 26 Treniranje, prikaz i evaluacija LinearSVC algoritma

Izvor: Izrada autora

Nakon ponavljanja jednakoga postupka za LinearSVC model, može se izračunati kako je vrijednost pripadajućeg parametra točnosti 0,8996, odnosno algoritam u 89,96% (≈90%) slučajeva točno klasificira recenziju kao pozitivnu ili negativnu.

U sljedećemu koraku, usporediti će se performanse implementiranih modela.

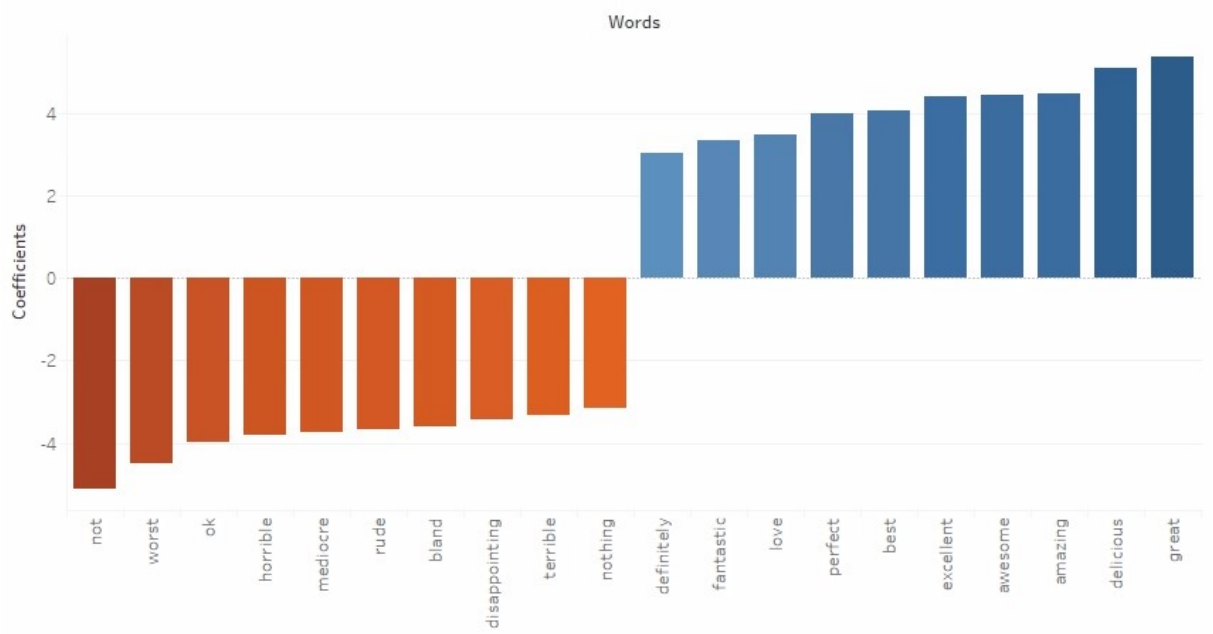
Tablica 7 Točnost modela

Korišteni algoritam / Vrsta klasifikacije	Naive Bayes	LinearSVC
Klasifikacija ocjene	54,37%.	61,12%,
Pozitivna/negativna klasifikacija	84.74%	89,96%

Izvor: Izračun autora

Iz tablice 7 može se zaključiti kako LinearSVC algoritam pokazuje bolje performanse, odnosno veću točnost klasificiranja recenzija, kako u slučaju klasifikacije recenzija po ocjenama, tako i u slučaju klasifikacije istih na one pozitivnoga i negativnoga karaktera.

Kako bi se bolje razumjeli kriteriji klasificiranja implementiranih algoritama, grafički će se prikazati 10 riječi s najvišim negativnim težinskim faktorima, odnosno 10 riječi s najvećim pozitivnim težinskim faktorom,. Ovakve težinske faktore izvući ćemo iz modela s najvećim parametrom točnosti, odnosno LinearSVC modela s pojednostavljenim problemom klasifikacije. Nakon izvlačenja parametara isti će se, zajedno s pripadajućim riječima, eksportirati u alat Tableau, kako bi se izradio grafički prikaz.

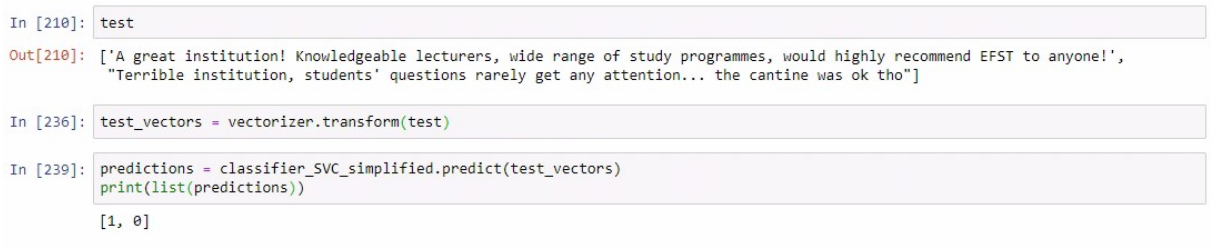


Slika 27 Prikaz riječi s najvišim vrijednostima težinskih faktora

Izvor: Izrada autora u alatu Tableau

Kao posljednji korak, demonstrirati će se funkcionalnost implementiranoga modela, pa tako modelu na klasifikaciju dajemo sljedeće dvije rečenice:

- „A great institution! Knowledgeable lecturers, wide range of study programmes, would highly recommend EFST to everyone!“
- „Terrible institution, students' questions rarely get any attention... the cantine was ok tho“



Slika 28 Primjer funkcionalnosti modela

Izvor: Izrada autora

Istrenirani algoritam pridodavanjem vrijednosti „1“ prvom primjeru prepoznaje ga kao recenziju pozitivnoga karaktera, dok je druga rečenica klasificirana kao „0“, odnosno kao recenzija negativnoga karaktera.

5. ZAKLJUČAK

U praktičnome dijelu ovoga rada proveden je niz operacija nad podacima web servisa za recenziranje usluga Yelp u svrhu razvoja algoritma za klasifikaciju recenzija korisnika. Čitav proces slijedio je CRISP-DM metodologiju. Iako se faze projekta jesu izvodile predviđenim redom, uočavanjem novonastalih problema često se pojavljivala potreba za vraćanje u prethodne korake procesa, što potvrđuje teoretski opis CRISP-DM procesa te zavisnost pripadajućih faza.

Izračunom distribucije frekvencija vrijednosti recenzija u odabranom setu podataka, dolazimo do zaključka o neravnomjernoj distribuciji, pa tako je najniži broj onih recenzija s pripadajućom ocjenom „2“ – 438 161, a najviši broj onih s pripadajućom ocjenom „5“ – 2 253 348. Algoritmu je, kako bi se smanjila pristranost u klasifikaciji, kao ulazni parametar važno dostaviti set podataka s jednakim brojem primjera za svaku klasu, pa se tako odabrani set podataka uzorkuje na način da sadrži jednak broj primjera za svaku od 5 različitih klasa.

U svrhu provođenja klasifikacije recenzija prema ocjenama korištena su dva algoritma, Naive Bayes i LinearSVC. Vrijednost pripadajućih pokazatelja točnosti klasifikacije algoritma pokazuje kako Naive Bayes postiže točnost od 54,37%, dok algoritam LinearSVC točno klasificira recenziju u 61,12% slučajeva. Iako točnost klasifikacije od 61,12% daleko premašuje onu postignutu nasumičnom klasifikacijom od 20%, ipak se ne može zaključiti kako se radi o pouzdanom rješenju spremnom za implementaciju.

Pojednostavljenjem problema na klasifikaciju recenzija na one pozitivnog i negativnog karaktera, točnost modela se znatno povećava, tako algoritam Naive Bayes postiže točnost od 84.74%, a vrijednost parametra za LinearSVC iznosi 89,96%. Razina točnosti od približno 90% sugerira pouzdanost modela, te spremnost istoga za implementaciju u poslovanje.

Rezultati točnosti u skladu su s ostalim istraživanja klasifikacije teksta, gdje algoritam LinearSVC u pravilu postiže višu točnost ukoliko se provodi nad setom podataka s relativno velikim brojem primjera.

U svrhu boljega razumijevanja procesa klasifikacije algoritma, grafički su prikazane riječi s najvišom negativnom, kao i one s najvišom pozitivnom vrijednošću koeficijenata, a ne čudi kako riječima poput „worst“, „horrible“ i „mediocre“ pripadaju visoke negativne, a riječima „great“, „amazing“ i „delicious“ visoke pozitivne vrijednosti koeficijenata.

Iz provedenoga istraživanja može se zaključiti kako je moguće izraditi kvalitetan algoritam za klasifikaciju recenzija na one pozitivnog i negativnog karaktera. Primjene ovakvoga modela su brojne, pa ga je tako npr. moguće implementirati u sustav koji automatski prikuplja recenzije o odabranom poslovnom subjektu ili proizvodu putem društvenih mreža ili ostalih online kanala, te automatskom klasifikacijom istih izrađuje izvještaje o stajalištima i mišljenjima korisnika u odabranom periodu.

LITERATURA

1. Aggarwal, N., Gupta, M. (2010): Classification Techniques Analysis, UIET Punjab University Chandigarh, CCI 2010, 19-20
2. Ahlemeyer-Stubbe, A. and S. Coleman, A practical guide to data mining for business and industry. 2014: John iley & Sons.
3. Analytics Vidhya (2018): A Comprehensive Guide to Understand and Implement Text Classification in Python [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/ [26.08.2018]
4. Analytics Vidyha (2018): Ultimate guide to deal with Text Data (using Python) – for Data Scientists & Engineers [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-ext-data-predictive-python/ [25.08.2018]
5. Brightlocal (2016): Local Consumer Survey, [Internet], raspoloživo na: https://www.brightlocal.com/learn/local-consumer-review-survey/ [23.9.2017]
6. Chauhan R.K., Shringar, R. Singh, N. (2012): Data Mining with Regression Technique, Journal of Information ystems and Communication, Volume 3
7. CMBI (2015): Business Intelligence Data Storage Architecture [Internet], raspoloživo na: http://www.cmbi.com.au/5040_DataStorageArchitecture.html [06.08.2018]
8. CRISP-DM (2014): Evaluation [Internet], raspoloživo na: http://crisp-dm.eu/modelling/ [20.08.2018]
9. Data Science Central (2016): CRISP DM – A Standard Methodology to Ensure a Good Outcome [Internet], raspoloživo na: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure--good-outcome [20.08.2018]
10. Datanyze (2018): Yelp market sharea and competitors analysis [Internet], raspoloživo na: https://www.datanyze.com/market-share/orm/yelp-market-share [22.08.2018]
11. Dummies (2015): Phase 2 of the CRISP-DM process model: Data understanding [Internet], raspoloživo na: https://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/ [20.08.2018]

12. Dummies (2015): Phase 3 of the CRISP-DM process model: Data understanding [Internet], raspoloživo na: https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-understanding/ [20.08.2018]
13. ECS (2015): OLAP operations [Internet], raspoloživo na: http://athena.ecs.csus.edu/~olap/olap/OLAPoperations.php [01.08.2018]
14. Eibe, F., Hall, M., Witten, I. (2011): Data Mining: Practical Machine Learning Tools and Techniques, 3rd dition, Morgan Kaufmann, Massachusetts
15. Finannces Online (2017): What Is the Purpose of Business Intelligence in a Business? [Internet], raspoloživo na: https://financesonline.com/purpose-business-intelligence-business/ [01.08.2018]
16. Friedman J.H , (1998). Data mining and Statistics-What's the Connection, 29th Symposium on the Interface
17. Ghahramani, Z. (2014): Unsupervised Learning, Gatsby Computational Neuroscience Unit, University College London, UK
18. Golarelli F, Rizzi A. (2011): Data Warehouse Design: Modern Principles and Methodologies, CompRef8 039-1
19. Gonzales-Aranda, P et al. (2008): Towards a Methodology for Data Mining Project Development: The Importance of Abstraction, Universidad Politenica de Madrid, Madrid, Spain,
20. Halili, F., Rustemi, A. (2016): Predictive Modeling: Data Mining Regression Technique Applied in a rototype, Department of Informatics, State University of Tetovo, Macedonia
21. Han, J., Kamber, M., Pei, J. (2011): Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, Massachusetts
22. Hand, D (1999): Statistics and data mining: intersecting disciplines, ACM SIGKDD Explorations, Volume 1,
23. Harris, D. (2017): 4 Emerging Use Cases for IoT Data Analytics [Internet], raspoloživo na: https://www.softwareadvice.com/resources/iot-data-analytics-use-cases/ [04.08.2018]
24. Hassan, S., Rafi,. M, Shaikh, M.(2012): Comparing SVM and Naïve Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment [Internet], raspoloživo na: https://arxiv.org/ftp/arxiv/papers/1202/1202.4063.pdf [26.08.2018]

25. Hawking P. and Sellitto C. (2010). Business Intelligence (BI) Critical Success Factors. ACIS 2010 Proceedings
26. Helbing, D., Moise, I., Pournaras, E. (2014): Density-Based Clustering [Internet], raspoloživo na: https://www.ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-am/documents/education/Spring2015/datascience/clustering2.pdf [14.08.2018]
27. Hipp, J., Wirth, R. (2015): CRISP-DM: Towards a Standard Process Model for Data Mining, DaimlerChrysler Research & Technology FT3/KL
28. Hossin, M., Sulaiman, M.N (2015): A Review on Evaluation Metrics For Data Classification Evaluations
29. IBM (2016): Data Understanding Overview [Internet], raspoloživo na: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_data_undestanding_phase.html, [20.08.2018]
30. IBM (2016): IBM SPSS Modeler CRISP-DM Guide [Internet], raspoloživo na: https://inseaddataanalytics.github.io/INSEADatalytics/CRISP_DM.pdf, [20.08.2018]
31. Informationbuilders (2015): Explanation of the Market Basket Model, [Internet], raspoloživo na: https://infocenter.informationbuilders.com/wf80/index.jsp?topic=%2Fpubdocs%2FRStat16%2Fsource%2Ftopic9.html [15.08.2018]
32. Jones M.C., Ramakrishnan T. and Sidorova A. (2012). Factors influencing business intelligence (BI) data Collection strategies: An empirical investigation.Decision, Support Systems 52
33. Karacan, H., Sirin, E. (2017): A review on Business Intelligence and Big Data, International Journal of Intelligent Systems and Applications in Engineering 2147-6799
34. KDnuggets (2014): CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Internet], raspoloživo na: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-ining-data-science-projects.html
35. Kumar, V., Pang-Ning, T. (2018): Cluster Analysis:Basic Concepts and Algorithms [Internet], raspoloživo na: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf [14.08.2018]

36. Learned-Miller. E. (2014): Introduction to Supervised Learning, Department of Computer Science, University f Massachusets, Amherst
37. Lopez, V. et.al (2013): An insight into classification with imbalanced data: Empirical results and current trends in using data intrinsic characteristics, Information Sciences Volume 250
38. Machine Learnig Matery (2016): Supervised and Unsupervised Machine Learning Algorithms [Internet] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [10.08.2018]
39. Machine Learning Mastery (2017): How to Prepare Text Data for Machine Learning with scikit-learn [Internet], raspoloživo na: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/
40. Maheshwari, A. (2014): Business Inteligence and Data Mining, Business Expert Press, New York; str. 3.
41. Masoud Yaghini (2010): Data Mining: Prediction – Regression Analysis [Internet], raspoloživo na: http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_882/DM_05_07_Regression%20Analysis.pdf [15.08.2018]
42. Medium (2016): What is the relationship between machine learning and data mining? [Internet], raspoloživo na: https://medium.com/@xamat/what-s-the-relationship-between-machine-learning-and-data-mining-c8675966615 [10.08.2018]
43. MLmastery (2017): Difference Between Classification and Regression in Machine Learning [Internet], raspoloživo na: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ [15.08.2018]
44. OLAP.com (2016): OLAP for Multidimensional Analysis [Internet], raspoloživo na : http://olap.com/olap-efinition/ [01.08.2018]
45. OLAP.com (2017): What is Business Intelligence (BI)? [Internet], raspoloživo na: http://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/ [31.07.2018]
46. Olszak, C & Ziemba, E. "Approach to Building and Implementing Business Intelligence Systems," Interdisciplinary Journal of Information, Knowledge, and Management, 2, 2007, 135-148.

47. PAM Analytics (2014): CRISP-DM Methodology, [Internet], raspoloživo na: http://www.pamanalytics.com/downloads/The%20CRISP-DM%20Methodology.pdf [20.08.2018]
48. Pandas (2018): About Pandas [Internet], raspoloživo na: https://pandas.pydata.org/ [21.08.2018]
49. Paperspace (2018): About Paperspace [Internet], raspoloživo na: https://www.paperspace.com/about,]21.08.2018]
50. Passioned Group (2017): BI Reporting, [Internet] raspoloživo na: https://www.passionned.com/business-ntelligence/bi-reporting/ [03.08.2018]
51. Project Jupyter (2018) Notebook [Internet], raspoloživo na: http://jupyter.org/ [21.08.2018]
52. Pujari. A (2001): Data Mining Techniques, Orient Blackswan, London
53. PWC (2017): Guide to Key Performance Indicators, [Internet] raspoloživo na: ttps://www.pwc.com/gx/en/audit-services/corporate-reporting/assets/pdfs/uk_kpi_guide.pdf [03.08.2018]
54. Python Software Foundation (2018): Applications for Python [Internet], raspoloživo na: https://www.python.org/about/apps/ [21.08.2018]
55. SAS (2016): Data Mining From A to Z: How to Discover Insights and Drive Better Opportunites, [Internet], raspoloživo na: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-04937.pdf, [19.9.2017]
56. Scikit-learn (2018) About [Internet], raspoloživo na: http://scikit-learn.org/stable/ [21.08.2018]
57. Simitsis, A., Vassiliadis, P. (2009): Extraction, transformation and loading, Database Encyclopedia 2009
58. Singular (2016): CRISP-DM Phase III: Data Preparation. Data analysis and features selection [Internet], raspoloživo na: https://data.sngular.team/en/art/51/crisp-dm-phase-iii-data-preparation-data-analysis-and-features-selection [20.08.2018]
59. Sisense (2018): Data sources to improve your decision making [Internet], raspoloživo na: ttps://www.sisense.com/blog/free-data-sources-upgrade-business-decision-making/ [06.08.2018]
60. Smart Vision (2016): Data Preparation [Internet], raspoloživo na: https://www.sv-europe.com/data-preparation [20.08.2018]

61. Smart Vision (2016): What is the CRISP – DM methodology?, [Internet], raspoloživo na: https://www.sv-europe.com/crisp-dm-methodology/ [20.08.2018]
62. Sristava, J. (2015): Understanding Linkage between Data Mining and Statistics, International Journal of Engineering Technology, Management and Applied Sciences, Volume 3, Issue 10, ISSN 2349-4476
63. Stefanovski J. (2009): Data Mining – Clustering, Institute of Computing Sciences, Poznan University of Technology [Internet], dostupno na: http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf [14.08.2018]
64. Tableau (2018): What is Tableau? [Internet], raspoloživo na: https://www.tableau.com/products/what-is-ableau [21.08.2018]
65. Technologydecisions (2016): Internal Data more useful to BI than external Data [Internet], raspoloživo na: https://www.technologydecisions.com.au/content/it-management/article/internal-data-more-useful-to-bi-than-xternal-data-1260850199, [06.08.2018]
66. Technopedia (2015): What does Query mean? [Internet], raspoloživo na https://www.techopedia.com/definition/5736/query [03.08.2018]
67. TechTarget, SearchBusiness Analytics (2016): Understanding benefits of business intelligence reporting, data mining, [Internet] raspoloživo na: https://searchbusinessanalytics.techtarget.com/feature/Understanding-benefits-of-business-intelligence-reporting-data-mining [03.08.2018]
68. TechTarget: SearchSQLServer (2014): What is Query? [Internet], raspoloživo na: https://searchsqlserver.techtarget.com/definition/query [03.08.2018]
69. TechTerms (2016): OLAP Definition [Internet], raspoloživo na: https://techterms.com/definition/olap [01.08.2018]
70. The Modeling Agency (2015): CRISP-DM 1.0 [Internet], raspoloživo na: https://www.the-modeling-gency.com/crisp-dm.pdf [20.08.2018]
71. Towards Data Science (2017): A Gentle Introduction on Market Basket Analysis— Association Rules Internet], raspoloživo na: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-ssociation-rules-fa4b986a40ce [14.08.2018]
72. Towards Data Science (2017): Machine Learning, NLP: Text Classification using scikit-learn, python and LTK [Internet], raspoloživo na: https://towardsdatascience.com/machine-learning-nlp-text-classification-using-cikit-learn-python-and-nltk-c52b92a7c73a [25.08.2018]

73. Towards Data Science (2017): Supervised vs. Unsupervised Learning [Internet], dostupno na: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d [13.08.2018]
74. Tudor, I. (2008), Association rule mining as a data mining technique, BULETINULuniversitatii Petrol-Gaze in Ploiesti, vol. LX,
75. Viana. L, Voznika, F. (2014): Data Mining Classification [Internet], raspoloživo na: https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf, [15.08.2018]
76. Watson, H., Annino, D. A., Wixom, B. H. "Current Practices in Data Warehousing," Journal of Information Science
77. Worthwhile (2016): 3 Keys to Managing Data and Maximizing Business Intelligence [Internet], raspoloživo a: https://worthwhile.com/blog/2017/02/20/data-business-intelligence/ [06.08.2018]
78. Yelp (2017): About us, [Internet], raspoloživo na: https://www.yelp.com/about, [19.9.2017]
79. Yelp (2017): Yelp Dataset Challenge, [Internet], raspoloživo na: https://www.yelp.com/dataset/challenge, [19.9.2017]
80. Yelp (2017): Yelp Open Dataset, [Internet], raspoloživo na: https://www.yelp.com/dataset, [19.9.2017]
81. Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of omputer information systems, vol. 50, no. 3, Spring, pp. 23-32.
82. Investopedia (2017): Data Mining, [Internet], raspoloživo na: http://www.investopedia.com/terms/d/datamining.asp [19.9.2017]
83. AS (2017): Data Mining: What it is and why it matters, [Internet], raspoloživo na: https://www.sas.com/en_us/insights/analytics/data-mining.html, [19.9.2017].

POPIS SLIKA I TABLICA

SAŽETAK

U današnjem svijetu, gotovo da i ne postoji poduzeće koje ne koristi podatke u svrhu unaprjeđenja poslovanja, pa stoga i ne čudi kako se upravo podaci smatraju najvrjednijim resursom današnjice. Podaci u svom sirovom obliku ne predstavljaju vrijednost za poduzeće, stoga kako bi se iskoristio njihov potencijal, potrebno je primijeniti odgovarajuće tehnike rudarenja podataka.

Poslovna inteligencija kao koncept objedinjuje strategije i tehnologije korištene od strane poslovnih subjekata za analizu poslovnih podataka, s rudarenjem podataka kao jednom sastavnih funkcija. CRISP-DM predstavlja najpopularniji metodološki okvir za provođenje projekata rudarenja podataka, a sastoji se od 6 faza koje uključuju razumijevanje poslovnog problema, razumijevanje podataka, pripremu podataka, modeliranje, evaluaciju modela te isporuku odnosno implementaciju rezultata.

Klasifikacija kao tehnika rudarenja podataka biti će, uz praćenje CRISP-DM metodološkoga okvira primijenjena na setu podataka web servisa za recenziranje usluga Yelp, a biti će evaluirani i različiti klasifikacijski algoritmi.

Ključne riječi: Poslovna inteligencija, rudarenje podataka, klasifikacija, CRISP-DM, Yelp

SUMMARY

In today's world, it is almost impossible to find a business that does not utilitze data in order to enhance its performance, therefore it is no wonder data is regarded as the most valuable resource to a company. Data in its raw form does not provide value to its owner, accordingly, an appropriate set of techniques has to be applied in order to exploit its potential.

Business intelligence (BI) comprises the strategies and technologies used by enterprises for the data analysis of business information, with data mining as one of its essential functions. CRISP-DM represents the most widely used structured approach for planning and execution of a data mining project. CRISP-DM methodology consists of six phases, including business understanding, data understanding, data preparation, modeling, evaluation and deployment.

Following CRISP-DM methodology framework, classification techniques are applied on a dataset consisted of reviews from Yelp user review platform, together with a performance evaluation of various classification algorithms.

Keywords: Business intelligence, data mining, classification, CRISP-DM, Yelp