Analiza i rudarenje podataka o projektima financiranim kroz Erasmus+ program

Župančić, Erik

SVEUČILIŠTE U SPLITU

EKONOMSKI FAKULTET

DIPLOMSKI RAD

ANALIZA I RUDARENJE PODATAKA O PROJEKTIMA FINANCIRANIM KROZ ERASMUS+ PROGRAM

Mentor:

izv.prof.dr. sc. Maja Ćukušić

Student:

Erik Župančić

Split, kolovoz, 2017.

SADRŽAJ:

1. UVOD

1.1 Problem istraživanja

Izumom osobnog računala čovječanstvo je trajno prešlo iz industrijskog doba u doba informacija. Ovaj period ljudske povijesti karakterizira eksponencijalni rast količine i brzine dotoka informacija iz praktično svake sfere našeg života[1]. Svaki dan kreiramo 2.5 kvintilijuna bajtova podataka. Kako bi se bolje dočarala ova činjenica potrebno je spomenuti da je gotovo 90% svih podataka u svijetu stvoreno tek u posljednjih par godina[2].

Suvremena poduzeća nastoje stvoriti konkurentske prednosti na brojne načine. Jedan od njih je iskorištavanje ogromne količine sirovih podataka koji im stoje na raspolaganju, bilo iz vlastitih internih izvora, bilo iz eksternih izvora[3]. Ti podaci u svom standardnom obliku ne mogu pomoći poduzećima u procesu donošenja odluka, već je potrebno iz njih korištenjem odabranih metoda i tehnika ekstrahirati informacije.

Upravo rudarenje podataka (eng. *Data Mining*) predstavlja proces kojim poduzeća pretvaraju sirove podatke u korisne informacije[4]. Radi se o disciplini koja omogućava automatsku obradu velike količine podataka kako bi se identificirali obrasci i trendovi u podacima koji nadilaze jednostavne analize. Aktivnosti rudarenja podataka predstavljaju iterativan proces usmjeren prema analizi velikih količina podataka, s ciljem izdvajanja informacija i znanja koji se mogu pokazati potencijalno korisnim osobama zaduženim za donošenje odluka[5]. Rudarenje podataka 'iskorištava' računalne tehnologije kako bi uz pomoć sofisticiranih matematičkih algoritama i modela automatski otkrili uzorke u podacima, predvidjeli najvjerojatnije ishode i stvorili iskoristive informacije.

Iako se rudarenje podataka danas u najvećoj mjeri koristi u profitnim poduzećima s snažnim naglaskom na kupce kao što su prodaja, financije, komunikacije i marketinške organizacije, ono bilježi porast upotrebe i u neprofitnim i državnim institucijama. Jedna od takvih mogućih primjena je program Europske unije Erasmus+.

---

[1] SAS (2016): Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities, [Internet], raspoloživo na: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf , [09.05.2017]. str. 1.

[2] IBM (2017): What is big data?, [Internet], raspoloživo na: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html , [15.05.2017]]

[3] Witten, I. & Eibe, F. (2005): Data Mining:Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington, str.23.

[4] Investopedia (2017): Data Mining, [Internet], raspoloživo na: http://www.investopedia.com/terms/d/datamining.asp ,[15.05.2017]

[5] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 77.

Erasmus + je novi program za obrazovanje, usavršavanje i mlade koji će zamijeniti sedam (7) postojećih programa (Program za cjeloživotno učenje [Erasmus, Leonardo da Vinci, Comenius i Grundtvig], Mladi na djelu, Erasmus Mundus, Tempus, Alfa, Edulink i Program suradnje sa industrijaliziranim zemljama)[6]. Svake godine se iz nepovratnih fondova Europske unije financira niz projekata iz domene učenja kroz mobilnost, suradnje na inovacijama u visokom obrazovanju i reforma obrazovne politike. Iako se radi o ogromnoj količini financijskih sredstava, većina Erasmus+ projekata ipak ne rezultira željenim rezultatima, niti predstavljaju ispravnu praksu.

S obzirom da se na internetskim stranicama Erasmus+ programa nalaze javno dostupni podaci o provedenim projektima, strukturirani u obliku Excel radne knjige, moguće je primijeniti metode i tehnike rudarenja podataka kako bi se došlo do vrijednih informacija potrebnih za daljnje donošenje odluka.

1.2 Predmet istraživanja

Unutar ovog rada fokus će biti stavljen na analizu i rudarenje podataka i njihove primjene u poslovanju.. Pružiti će se teorijski uvid u procese analize i rudarenja podataka kroz njihove faze, od definicije problema, preko prikupljanja i pripreme podataka, izgradnje modela, pa sve do same isporuke informacija[7]. Samo rudarenje podataka usporediti će se sa njemu sličnim konceptima kao što su npr. OLAP i klasične statističke metode, kako bi utvrdili po čemu se ono razlikuje od istih[8]. Značajan dio pažnje biti će posvećen i teorijskoj analizi metodologija i tehnika koje stoje na raspolaganju za rudarenje podataka.

Drugi dio rada odnosi se na empirijsku primjenu prethodno izloženih koncepata, korištenjem odabranog računalnog alata za rudarenje podataka. Kao što je prethodno navedeno, Erasmus+ je novi program Europske unije kojim se financiraju razni projekti mobilnosti u području obrazovanja. Njihove godišnje statistike dostupne na *web* stranici omogućavaju primjenu metoda i tehnika rudarenja podataka kako bi se ekstrahirale interesantne informacije o samim projektima. U tom kontekstu cilj je dobiti informacije o karakteristikama projekata prema liniji financiranja, partnerima, koordinatorima i slično. Te informacije mogu biti od koristi svim prijaviteljima na projekte Erasmus+ programa kako bi donosili bolje odluke npr.

[6] Sveučilište u Zagrebu. Erasmus+:Opće Informacije, [Internet], raspoloživo na: www.unizg.hr/suradnja/medunarodna-suradnja/partnerstva/erasmus/

[7] Oracle (2015): What is data mining, [Internet], raspoloživo na: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046 ,[15.05.2017.]

[8] 24. Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str 80-81.

prilikom odabira vrste projekata na koji se žele prijaviti, uspostavljanja partnerstva s drugim institucijama ili odabira koordinatora projekta. Jedan od tih sudionika Erasmus+ programa kojem će informacije biti od koristi je i sami Ekonomski fakultet u Splitu.

1.3 Istraživačka pitanja

Unutar ovog rada želi se dati odgovor na sljedeći niz istraživačkih pitanja:

- Što je rudarenje podataka i kako ono može utjecati na poslovanje poduzeća?
- Kako izgleda proces rudarenja podataka i faze kroz koje ono prolazi?
- Koje nam metode i tehnike rudarenja stoje na raspolaganju?
- Po čemu se rudarenje podataka razlikuje od sličnih koncepata?
- Možemo li primijeniti metode i tehnika rudarenja podataka na empirijskom primjeru kako bi stvorili korisne informacije?
- Kakve su prosječne karakteristike projekata po liniji financiranja?
- Možemo li za većinu projekata reći da su uspješno odrađeni i da predstavljaju dobru praksu?
- Koje zemlje povlače najviše sredstava za financiranje projekata mobilnosti u obrazovanju?
- Koliki je udio projekata s partnerima u ukupnom broju projekata, te koji su najčešći partneri?
- Postoji li mogućnost korištenja tehnike rudarenja teksta (eng. *Text-mining)* kako bi se ekstrahirale korisne informacije iz detaljnih opisa projekata?
- Kakvi su trendovi u projektima prema godinama?
- Na koji način možemo prezentirati informacija stvorene prilikom rudarenja podataka?

1.4 Ciljevi istraživanja

Unutar ovog rada definirati će se pojmovi eksplorativne analize i rudarenja podataka, te pružiti uvid u faze kroz koje je potrebno proći tijekom jednog projekta analize i rudarenja podataka. Potrebno je i definirati karakteristike osnovnih metoda i tehnika koje stoje na raspolaganju, kao i moguće primjene istih. Nakon teorijskog osvrta na analizu podataka i rudarenje podataka uz usporedbu s njegovim sličnim konceptima, ukratko će se prezentirati alat koji će se koristiti u praktičnom dijelu rada.

Osnovni cilj bi bila upravo empirijska primjena odabranih metoda i tehnika rudarenja podataka na stvarnom setu podataka kako bi se iz istih ekstrahirale vrijedne informacije potrebne za poboljšanje procesa donošenja odluka.

Nakon što modeli budu izrađeni, te iz njih budu ekstrahirane informacije, dobiveni rezultati će biti prezentirani korištenjem odabranih tehnika vizualizacije podataka.

1.5 Metode istraživanja[9]

- **Induktivna metoda** kao sustavna primjena induktivnog načina zaključivanja kojim se na temelju analize pojedinačnih činjenica dolazi do zaključka o općem sudu, od zapažanja konrektnih pojedinačnih slučajeva do općih zaključaka.
- **Metoda analize** kao postupak znanstvenog istraživanja raščlanjivanjem složenih pojmova, sudova i zaključaka na njihove jednostavnije sastavne dijelove i elemente.
- **Metoda klasifikacije** kao sistematska i potpuna podjela općeg pojma na posebne, u okviru opsega pojma.
- **Metoda deskripcije** kao postupak jednostavnog opisivanja ili očitavanja činjenica, procesa i predmeta u prirodi i društvu te njihovih empirijskih potvrđivanja odnosa i veza, ali bez znanstvenog tumačenja i objašnjavanja.
- **Metoda kompilacije** kao postupak preuzimanja tuđih rezultata znanstvenoistraživačkog rada, odnosno tuđih opažanja, stavova, zaključaka i spoznaja.
- **Metoda rudarenja podataka** kao postupak analize velike količine podataka s ciljem izvlačenja korisnih informacija i znanja potrebnih za donošenje odluka.
- **Metoda rudarenja** kao postupak automatske analize podataka koji se nalaze u tekstu u obliku prirodnog jezika.
- **Vizualizacija podataka** kao postupak predstavljanja podataka u grafičkom formatu. Time komuniciranje informacija u podacima postaje jednostavnije i lakše za komuniciranje drugim korisnicima.

1.6 Doprinos istraživanja

Doprinos ovog istraživanja se manifestira u prikazu načina na koji se rudarenje podataka može koristiti za izvlačenje korisnih informacija iz sirovih podataka. Primijeniti će se metode i tehnike analize i rudarenja podataka kako bi se prikazali načini na koji se analiza i rudarenje podataka mogu primijeniti na stvarnim podacima. Pri tom će se ono vršiti na primjeru podataka Erasmus+ programa, na čijim je projektima mobilnosti i naš sam fakultet sudionik. Na taj način će se prikazati način na koji suvremeni poslovni subjekti mogu doći do korisnih informacija na jednostavan način, upotrebom intuitivnih alata za analizu i rudarenje podataka.

[9] UNIZD (2014): Metode znanstvenih itraživanja, [Internet], raspoloživo na: http://www.unizd.hr/portals/4/nastavni_mat/1_godina/metodologija/metode_znanstvenih_istrazivanja.pdf, [15.05.2017.]

1.7 Struktura diplomskog rada

Diplomski rad je strukturiran unutar sedam poglavlja.

Prvo poglavlje predstavlja uvodni dio rada, unutar kojeg se sažeto iznose opisani predmet i područje istraživanja. Osim toga, u ovom dijelu rada definirana su i postavljena istraživačka pitanja (hipoteze), kao i sami ciljevi istraživanja i doprinos koji se od njega očekuje.

Drugo poglavlje predstavlja kratki teoretski osvrt na poslovnu inteligenciju, koncept kojim se teži iz raspoloživih podataka izvući korisne informacije potrebne za donošenje pravovremenih i efikasnih poslovnih odluka. Također će se pružiti kratak uvid u podatke, informacije i znanje, njihove sličnosti i razlike, te njihovu važnost za suvremeni poslovni subjekt.

Treće poglavlje će biti fokusirano na sažetu teoriju koja stoji iza koncepta istraživačke (eng. *exploratory*) analize podataka. Analizirane će biti osnovne tehnike koje stoje na raspolaganju za 'rezimiranje' podataka i prezentaciju rezultata.

Četvrto poglavlje se odnosi na teorijsku obradu rudarenja podataka. U ovom dijelu definirati će se što je to uopće rudarenje podataka, te koji se modeli i metode rudarenja podataka mogu koristiti za izvlačenje informacija iz podataka. Pružiti će se kratak uvid u tipičan proces rudarenja podataka, te moguće primjene istog u rješavanju stvarnih poslovnih problema.

U petom poglavlju ukratko će se argumentirati zašto je odabran upravo taj alat za rudarenje podataka za realizaciju empirijskog dijela rada. Sažeto će se prikazati njegove osnovne karakteristike, kao i rad u samom alatu.

Šesto poglavlje je empirijski dio, i sama srž ovog rada. Cilj ovog dijela je primjena nekih od raspoloživih tehnika analize i rudarenja podataka na javno dostupnim podacima o projektima Erasmus+ programa kako bi se prikazala na praktičnom primjeru mogućnost izvlačenja informacija iz navedenih podataka.

Zadnje, sedmo poglavlje predstavlja sažeti prikaz rezultata, odnosno zaključaka do kojih se došlo prilikom izrade ovog rada.

2. POSLOVNA INTELIGENCIJA

Gotovo svake minute u poduzeću se donose odluke na svim razinama upravljanja koje će imati značajan utjecaj na njegove performanse. Te odluke pokreću organizacije. Donošenje dobre odluke u kritičnom trenutku može dovesti do efikasnije proizvodnje, profitabilnijeg poduzeća ili zadovoljnijih kupaca. Iako je danas na tržištu dostupna značajna količina potentnih alata za obradu podataka i njihovu distribuciju, još uvijek postoji veliki broj organizacija koje se prilikom donošenja odluka oslanjaju na instinkt, savjete drugih i prethodno iskazane prakse[10].

Povećanje memorijskih kapaciteta, smanjenje njihovih cijena i široka dostupnost internetske veze olakšalo organizacijama i individualnim korisnicima pohranu i distribuciju podataka. S obzirom da se radi o podacima prikupljenim iz različitih izvora, za očekivati je da će isti imati raznoliku strukturu, ali i sadržaj[11]. Mogućnost pristupa velikoj količini podataka predstavlja priliku za suvremena poduzeća da njihovom obradom stvore adekvatnu informacijsku podlogu za efektivno i pravovremeno donošenje odluka. Upravo to je problem kojim se bavi i nastoji riješiti poslovna inteligencija.

Poslovnu inteligenciju (eng. *Business Intelligence*) možemo definirati kao skup matematičkih modela i analitičkih metodologija koje koriste dostupne podatke za stvaranje informacija i znanja, korisnih za složene procese donošenje odluka.[12]

Druga definicija kaže da je poslovna inteligencija tehnologijom upravljan proces za analizu podataka i prezentiranje stvorenih informacija kako bi se pomoglo izvršnim direktorima, menadžerima i brojnim drugim krajnjim korisnicima donositi informirane poslovne odluke.[13]

Za poslovnu inteligenciju bi u suštini mogli reći da se radi o bilo kakvoj aktivnosti, alatu ili procesu koji se koristi za dobivanje releventnih informacija za podršku procesu donošenja odluka.[14]

S obzirom da posljednja definicija nigdje ne spominje tehnološki aspekt, moglo bi se lako zaključiti da je kontradiktorna prethodnim definicijama. Istina je da je moguće vršiti aktivnosti poslovne inteligencije i bez upotrebe suvremenih tehnoloških rješenja, ali takav

[10] Scheps, S. (2008): Business Intelligence for Dummies, Wiley Publishing Inc., Indianapolis, str. 9.
[11] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 9.
[12] Ibid., str. 9.
[13] TechTarget (2014): Business Intelligence, [Internet], raspoloživo na: http://searchdatamanagement.techtarget.com/definition/business-intelligence ,[16.07.2017.]
[14] Scheps, S. (2008): Business Intelligence for Dummies, Wiley Publishing Inc., Indianapolis, str. 11.

pristup se ne može nositi s ogromnom količinom digitalno pohranjenih podataka koji nam stoje na raspolaganju.

2.1 Pravodobno i efektivno donošenje odluka

Unutar kompleksnih organizacija gotovo da ne prođe trenutak bez da se donese neka odluka. Te odluke mogu biti više ili manje kritične za poslovanje, imati dugoročne ili kratkoročne efekte i uključivati zaposlene na različitim hijerarhijskim razinama[15]. Kvaliteta odluka i prednosti koje ona može donijeti organizaciji ovise o načinu na koji se odluke donose. Prethodno je spomenuto da veliki broj donositelja odluka koristi jednostavne i intuitivne metode koje su neprikladne za donošenje odluka u suvremenim, dinamičnim i nepredvidivim okruženjima.

S toga se razvijaju suvremeni sustavi za poslovnu inteligenciju kako bi donositelji odluka dobili pristup alatima i metodama za efektivno i pravovremeno odlučivanje.

Efektivno odlučivanje - Poslovna inteligencija primjenom rigoroznih analitičkih metoda omogućava stvaranje informacija na koje se donositelji odluka mogu bolje osloniti, te koje će u većoj mjeri odgovarati postavljenim poslovnim ciljevima[16].

Dobro dizajniran sustav poslovne inteligencije osigurava dosljednu i pouzdanu distribuciju informacija unutar organizacije. Na taj način sustav ne samo da smanjuje vrijeme potrebno za analizu i pripremu podataka, već putem različitih kontrolnih ploča olakšava krajnjim korisnicima izradu izvještaja[17].

Pravodobno odlučivanje - S obzirom da suvremena poduzeća posluju unutar dinamičnih i nepredvidljivih ekonomskih okruženja, koje karakterizira visok stupanj konkurencije, sposobnost brze adaptacije na promjene u okolini predstavlja jednu od važnijih konkurentskih prednosti. Prilikom donošenja odluka potrebno je analizirati nekoliko raspoloživih alternativnih strategija. U usporedbi s klasičnim pristupom donošenja odluka, upotreba modela i metodologija poslovne inteligencije omogućava nam da analiziramo veći broj alternativa unutar jednakog vremena. Ovakvi sustavi omogućavaju poduzećima da brže

[15] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 3-4.
[16] Ibid., str. 5.
[17] Martin A. et al. (2011): Better Decision Making with Proper Business Intelligence, [Internet], raspoloživo na: https://www.atkearney.com/documents/10192/247903/Better_Decision_Making_with_Proper_Business_Intelligence.pdf/e55e6880-ed1b-4b25-a0b6-33b94c0cc641, [16.06.2017.]

prikupe, analiziraju i transformiraju podatke u korisne informacije koje će se potom koristiti prilikom donošenja odluka za iskorištavanje novonastalih prilika u ekonomskom okruženju.

2.2 Podaci, informacije i znanje

Iako se često koriste kao sinonimi, pojmovi podatak i informacija označavaju različite pojave. Stoga je potrebno ukratko definirati pojmove podatka, informacije i znanja kako bi ih lakše razlikovali.

Najjednostavnije rečeno, podaci su činjenice o nečemu. U informatičkom smislu radi se o činjenicama pretvorenim u oblik koji je efikasan za obradu i distribuciju[18]. Podaci su po svojoj naravi sirovi, neoblikovani i neobrađeni. Sami po sebi podaci su beskorisni za donositelje odluka, ali predstavljaju važan input za stvaranje informacija[19]. To npr. mogu biti zapisi o prihodima i troškovima organizacije u određenom periodu.

Informacija je rezultat aktivnosti obrade podataka. Taj rezultat bi trebao biti smislen onima koji ga poslije koriste za određenu svrhu. To su podaci koji su točni, pravovremeni, specifični i organizirani s ciljem, prezentirani s značenjem i relevantnosti i mogu se iskoristiti za ostvarivanje određenog cilja[20]. Nadovezujući se na prethodni primjer, informacija koja može nastati obradom tih podataka je npr. da je u tom određenom periodu vremena došlo do smanjenja profita.

Kada se informacije iskoriste za donošenje odluka i provođenje odgovarajućih aktivnosti, ono postaje znanje. Stoga znanje možemo definirati kao informacije iskorištene u specifičnom području kako bi se riješili problemi, unaprijeđene s iskustvom i kompetencijama donositelja odluka[21].

[18] Vaughan, J. (2017): Data, [Internet], raspoloživo na: http://searchdatamanagement.techtarget.com/definition/data ,[16.07.2017.]
[19] Doyle, M. (2014): What is the Difference Between Data and Information?,[Internet],raspoloživo na http://www.business2community.com/strategy/difference-data-information-0967136#EO5oZjHX874qQ3ZD.97 ,[16.07.2017.]
[20] BusinessDictionary (2017): Information,[Internet],raspoloživo na: http://www.businessdictionary.com/definition/information.html ,[16.07.2017.]
[21] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 7.

3. EKSPLORATIVNA ANALIZA PODATAKA

3.1 Definicija eksplorativne analize podataka

Eksplorativna analiza podataka (eng. *exploratory data analysis*) predstavlja prvi korak u analizi podataka. Radi se o pristupu ili filozofiji analize podataka koji koristi brojne tehnike kako bi se stekao uvid u skup podataka, otkrila njegova struktura, izvukle bitne varijable, identificirali *outlieri* i testirale pretpostavke[22]. Neformalno bi mogli reći da se radi o bilo kojoj metodi uviđaja u podatke koja ne uključuje formalno statističko modeliranje i zaključivanje[23].

Podaci se u računalima najčešće pohranjuju u dvodimenzionalnom obliku, bilo u proračunskim tablicama, bilo u bazama podataka. U situaciji kada imamo veliku količinu podataka ovakav format zapisa nije primjeren za utvrđivanje bitnih karakteristika podataka. Tehnike eksplorativne analize nam služe za skrivanje određenih aspekata podataka, uz istovremeno naglašavanje željenih aspekata.

Iako se često koriste kao sinonimi, eksplorativna analiza i statistička vizualizacija se odnose na različite stvari. Klasična statistička vizualizacija podataka koristi skup tehnika kako bi ispitali pretpostavke o modelima koje podaci prate, dok eksplorativna analiza podataka ne postavlja pretpostavke o modelima[24].

Eksplorativna analiza podataka se uobičajeno klasificira na dva načina. S jedne strane metode mogu biti *graphical* ili *non-graphical*, dok s druge strane mogu biti *univariate* ili *multivariate*. *Non-graphical* metode uključuju izračunavanje deskriptivnih statistika, dok grafičke metode nastoje sažeti podatke u obliku dijagrama ili drugih slikovitih načina. Što se tiče druge podjele, *univariate* metode se odnose na istovremeno promatranje samo jedne varijable, dok *multivariate* istovremeno promatra dvije ili više varijable. Unakrsnom klasifikacijom na temelju ovih podjela možemo dobiti četiri osnovne vrste eksplorativne analize podataka: *graphical univariate, non-graphical univariate, graphical multivariate, non-graphical multivariate*[25].

[22] Engineering Statistics Handbook (2013): What is EDA?,[Internet],raspoloživo na: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm , [16.07.2017.]
[23] Carnegie Mellon University: Exploratory Data Analysis,[Internet],raspoloživo na: http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf ,[16.07.2017.]
[24] Engineering Statistics Handbook (2013): What is EDA?,[Internet],raspoloživo na: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm , [16.07.2017.]
[25] Carnegie Mellon University: Exploratory Data Analysis,[Internet],raspoloživo na: http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf ,[16.07.2017.]

Svaka od navedenih kategorija analize se može dalje dijeliti ovisno o ulogama i vrstama (kategorijski ili kvantitativni) podataka.

3.2 Tehnike eksplorativne analize podatka

Postoji veliki broj različitih kvantitativnih i kvalitativnih tehnika koje se mogu koristiti prilikom eksplorativne analize podataka.

Srednje vrijednosti (eng. *central tendency*) - Iako se za eksplorativnu analizu podataka najčešće koriste različite tehnike vizualizacije podataka, potrebno je spomenuti i nekoliko osnovnih kvantitativnih tehnika. Jedna od osnovnih skupina kvantitativnih tehnika su mjere centralne tendencije, odnosno mjere srednjih vrijednosti. Najznačajniji predstavnik ove skupine je aritmetička sredina.

Aritmetička sredina (eng. *mean*) je onaj jednaki dio vrijednosti numeričkog obilježja koji otpada na jedan element skupa[26]. Kada bi numeričku vrijednost nekog obilježja ravnopravno raspodijelili na subjekte ispitivanja, dobili bi aritmetičku sredinu.

Medijan je ona srednja vrijednost koja statistički niz dijeli na dva jednaka dijela[27]. Kada podatke stavimo u poredani niz, vrijednost koja se nalazi u samoj sredini predstavlja medijan tog statističkog niza.

Mod je ona vrijednost obilježja koja se najčešće pojavljuje[28]. Iako se rijetko koristi, mod ukazuje na „vrh“ distribucije.

Mjerama disperzije (eng. *spread*) opisuje se brojčani stupanj varijabilnosti podataka. Neke od poznatijih mjera disperzije su:

Varijanca je srednje kvadratno odstupanje numeričkih vrijednosti obilježja od aritmetičke sredine. Korijenujemo li varijancu dobit ćemo **standardnu devijaciju.**

Koeficijent varijacije je omjer standardne devijacije i aritmetičke sredine pomnožen sa sto, tj. pokazuje koliki postotak vrijednosti AS iznosi vrijednost standardne devijacije.

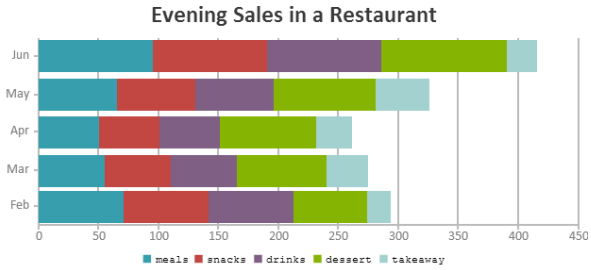
Interkvartil predstavlja mjeru raspona vrijednosti obilježja srednjih 50% jedinica u distribuciji, odnosno razliku između gornjeg i donjeg kvartila.[29] **Kvartili** su vrijednosti koje dijele distribuciju ili podatke na jednake četvrtine.

[26] Rozga, A.(2009): Statistika za ekonomiste, Ekonomski fakultet, Split, str. 39
[27] Ibid., str. 41.
[28] Ibid., str. 42

Prilikom eksplorativne analize podataka često se koriste različiti oblici vizualizacije podataka. Neki od njih su redom:

- **Bar Chart**

Stupčasti grafikon je jedna od najčešće korištenih tehnika vizualizacije podataka. Na jednostavan način omogućava usporedbu podataka i identificiranje ekstremnih vrijednosti. Radi se o efektivnom načinu za vizualizaciju numeričkih podataka prema različitim kategorijama[30]. Visinom i širinom stupca prikazani su numeričke vrijednosti obilježja[31].

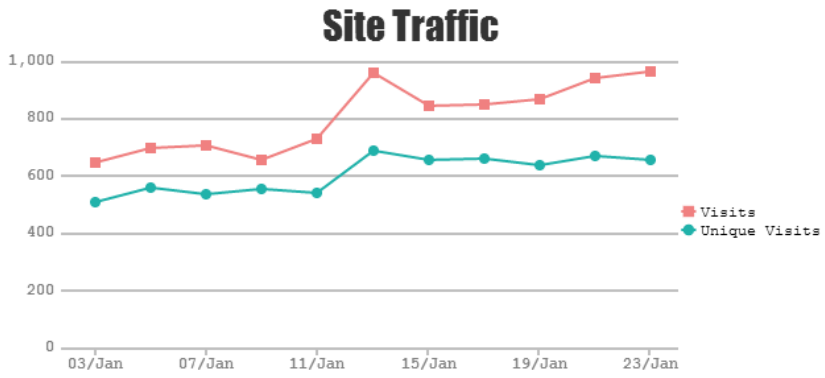


Slika 1 Primjer stupčastog dijagrama

Izvor: https://canvasjs.com/wp-content/uploads/2013/01/javascript_stacked_bar_chart.jpg

- **Linijski dijagram**

Linijski dijagram je također jedan od najčešće korištenih načina vizualizacije podataka. Koristi se za prikazivanje varijabli u nizu, kao što je npr. kretanje neke varijable kroz vrijeme.



Slika 2 Primjer linijskog dijagrama

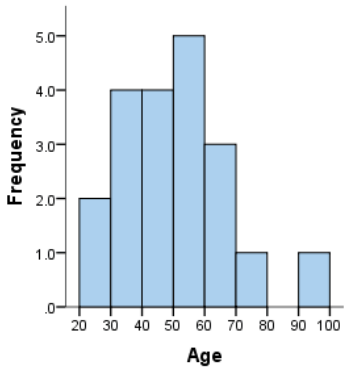
Izvor: https://canvasjs.com/wp-content/uploads/2013/01/html5_multiseries_line_chart.jpg

---

[29] Ibid., str. 58
[30] Hardin, M. et al. (2017): Which chart or graph is right for you?, [Internet], raspoloživo na: https://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf,[16.07.2017.]
[31] Statistics How To (2013): What is a Bar chart?,[Internet], raspoloživo na: http://www.statisticshowto.com/what-is-a-bar-chart/ ,[16.07.2017.]

- **Histogram**

Histogram je jedna od temeljnih tehnika vizualizacije. Visina svakog stupca histograma predstavlja frekvencije varijable, broj pojavljivanja pojedinih vrijednosti određene varijable[32]. Najčešće se koristi kako bi se utvrdila i prikazala distribucija skupa podataka. Informacije kojima rezultira histogram mogu se koristiti za ispitivanje distribucija, pronalaženje *outliera,* izračunavanje zaobljenosti i sl.[33].



Slika 3 Primjer histograma

Izvor: https://statistics.laerd.com/statistical-guides/img/histogram-1.png

- **Stem-and-leaf**

Stem-and-Leaf je relativno jednostavna supstitucija za histogram koja se najčešće izrađuje ručno. Koristi se kao i histogram za prikazivanje frekvencija pojedinih vrijednosti nekog obilježja. Prednost koju ova tehnika ima nad histogramom je zadržavanje originalnih podataka u svom prikazu[34].

stem	leaf
1	2 3
2	1 7
3	3 4 5 7
4	0 0 1

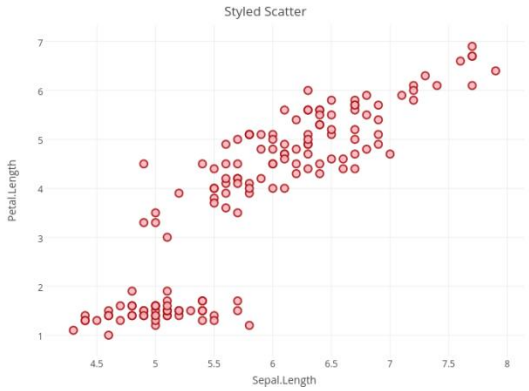
Slika 4 Primjer Stem-and-leaf tehnike

Izvor: http://www.purplemath.com/modules/stats/stemleaf01.gif

[32]Carnegie Mellon University: Exploratory Data Analysis,[Internet],raspoloživo na: http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf ,[16.07.2017.], str. 72.
[33] Laerd (2013): Histograms, [Internet],raspoloživo na:https://statistics.laerd.com/statistical-guides/understanding-histograms.php ,[16.07.2017.]
[34] Purplemath (2017): Stem-and-Leaf Plots, [Internet], raspoloživo na: http://www.purplemath.com/modules/stemleaf.htm ,[16.07.2017.]

- **Dijagram rasipanja**

Dijagram rasipanja (eng. *scatter plot*) koristi se kada se želi istražiti veza između dva različita obilježja neke pojave. Pri tom omogućava jednostavno vizualno identificiranje trendova, koncentracija vrijednosti kao i *outliera*. Često se koristi kao pomoć prilikom identificiranja daljnjih koraka istraživanja[35]. Svakom točkom na dijagramu rasipanja prikazane su vrijednosti dva numerička obilježja za jednu opažanje.



Slika 5 Primjer dijagrama rasipanja

Izvor: https://plot.ly/~RPlotBot/4326/styled-scatter.png

- **Toplinske mape**

Toplinska mapa je način vizualizacije podataka unutar kojeg se koristi intenzitet boje kako bi se izrazila vrijednost numeričkog obilježja[36]. Može se koristiti bilo kao samostalna tehnika, bilo u kombinaciji s drugim tehnikama vizualizacije podataka.

	Reports		Analysis					
	Param'tized	Static	Scorecards	Dashboards	Simple	Moderately complex	Complex	Predictive
Tibco	38	28	20	22	50	23	18	19
Oracle	70	39	21	23	36	22	17	17
Qliktech	55	37	34	29	36	24	19	17
Tableau	51	37	27	29	38	28	16	18
Infor	54	63	15	18	32	12	7	2
Board	57	42	30	23	29	17	15	20
IDS Scheer	58	42	22	21	20	18	16	14
LogiXML	61	42	20	24	21	18	16	15
Jaspersoft	65	60	18	19	21	18	12	5
Microsoft	42	37	24	18	29	19	9	7
SAS	53	49	17	12	21	16	14	7
ArcPlan	54	48	23	22	19	11	13	15
Panorama	28	45	18	17	43	21	20	9
Targit	42	39	20	14	37	18	11	19
Actuate	46	44	9	9	21	11	8	2
SAP	39	40	14	13	21	17	16	7
IBM	52	41	7	8	21	10	9	8
Information Builders	58	42	8	14	19	12	7	5
MicroStrategy	47	44	20	18	41	21	17	1

Slika 6 Primjer toplinske mape

Izvor: https://i.stack.imgur.com/cGSob.png

[35] Hardin, M. et al. (2017): Which chart or graph is right for you?, [Internet], raspoloživo na: https://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf,[16.07.2017.], str. 10.
[36] Skupina autora (2011): Višedimenzijski informacijski sustavi, Ekonomski fakultet, Split str. 162

4. RUDARENJE PODATAKA

4.1 Definicija rudarenja podataka

Postoji više razloga zbog kojih se čovječanstvo danas suočava s informacijskom krizom, odnosno pretrpanosti s podacima. Jeftini memorijski prostor omogućava neselektivno pohranjivanje ogromne količine podataka koja se danas mjeri u petabajtima ili exabajtima[37]. U suvremenom društvu praktično svaki automatizirani sustav generira neku vrstu podataka, bilo za potrebe dijagnoze, bilo za potrebe analize[38]. Gotovo svaka osoba dnevno korištenjem internet pretraživača pristupa nekim od nekoliko milijardi dokumenata koji su dostupni putem *World Wide Weba*. Pritom se svaka njihova akcija, svaki pokret, svaki klik pohranjuje u obliku podataka. Razvoj i pristupačnost senzornih tehnologija dovodi do fenomena koji nazivamo *Internet of Things.* Veliki dio suvremenih uređaja i predmeta, od mobilnih telefona pa do automobilskih guma, opremljeni su sa senzorima koji prikupljaju i odašilju podatke na pohranu. Korporacije pohranjuju podatke o svojim zaposlenicima, klijentima, transakcijama i slično.

Kako se količina podataka povećava neumoljivo, tako se i razmjer onoga što čovjek razumije smanjuje. Može se reći da se raskorak između stvaranja i razumijevanja podataka neprestano povećava[39]. S obzirom da se podaci pohranjuju s ciljem daljnje analize i generiranja korisnih informacija za potrebe odlučivanja, nužno je postojanje metodologija i tehnika kojim će se taj cilj i ostvariti.

Aktivnosti rudarenja podataka predstavljaju iterativni proces usmjeren prema analizi velikih količina podataka s ciljem izvlačenja korisnih informacija i znanja, koji se mogu pokazati korisnim za rješavanje problema i donošenje odluka[40].

Rudarenje podataka moguće je definirati i kao istraživanje i analizu velikih količina podataka pomoću automatskih ili poluautomatskih metoda s ciljem otkrivanja smislenih pravilnosti[41].

S obzirom na osnovni cilj, moguće je reći da postoje dvije osnovne vrste rudarenja i analize podataka:[42]

---

[37] Witten, I. H. et al. (2011): Data mining: Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington, str. 4.
[38] Aggarwal, C.C. (2015): Data Mining: The Textbook, Springer, London, str. 1.
[39] Witten, I. H. et al. (2011): Data mining: Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington. str. 4.
[40] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 78.
[41] Pejić, M. (2005): Rudarenje podataka u bankarstvu, Sveučilište u Zagrebu, Ekonomski fakultet.

- **Interpretacija** – Osnovni cilj ovih aktivnosti rudarenja podataka jest identificiranje uobičajenih obrazaca u podacima i njihovo iskazivanje kroz skup pravila i kriterija koji su lako razumljivi krajnjim korisnicima.
- **Predviđanje** – Druga kategorija aktivnosti rudarenja podataka nastoji predvidjeti vrijednost koju će slučajna varijabla poprimiti u budućnosti, kao i vjerojatnost da se ta vrijednost stvarno i ostvari.

4.2 Modeli i metode rudarenja podataka

Postoji veliki broj raznolikih metoda i tehnika koje se mogu koristiti prilikom rudarenja podataka. Neke od tih tehnika spadaju pod klasične statističke tehnike, dok je dio njih jedinstven za rudarenje podataka. U nastavku će biti ukratko opisane neke od najpopularnijih tehnika.

Asocijacija (eng. *Association*) – Asocijacijska pravila, poznata i kao pravila sklonosti, koriste se za identificiranje zanimljivih i ponavljajućih povezanosti između grupe zapisa u setu podataka[43]. Ova tehnika koristi niz *if* i *then* izjava kako bi pomogla otkriti vezu između naizgled nepovezanih podataka u relacijskoj bazi podataka ili nekom drugom izvoru[44]. Ova tehnike se primjerice može koristiti za utvrđivanje proizvoda koji se često kupuju zajedno, kao i vjerojatnosti da će doći do te kupnje.

Korelacija (eng. *Correlation*) – Korelacija je jedna od najjednostavnijih tehnika koje se koriste prilikom rudarenja podataka. Osnovni cilj korelacijskog rudarenja je otkrivanje interesantnih i neuobičajenih zavisnosti između velikog broja varijabli[45]. Koristi se kao brz i lagan način za utvrđivanje u kakvoj se interakciji određeni podaci nalaze[46]. Pokazatelj koji dobijemo provođenjem korelacijskog rudarenja naziva se koeficijent korelacije. Taj koeficijent ukazuje na jačinu i smjer veze između dvije varijable.

Grupiranje (eng. *Clustering*) - Ova skupina tehnika nastoji definiranjem prikladnih pokazatelja i uvođenjem pojmova udaljenosti i sličnost između parova promatranja

---

[42] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 78.
[43] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 93.
[44] Rouse, M. (2011): Association Rules, [Internet], raspoloživo na: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining, [16.07.2017.]
[45] Hero, A. (2013): Correlation Mining in Massive Data, [Internet], raspoloživo na: http://www.eecs.umich.edu/eecs/pdfs/events/2711.pdf ,[16.07.2017.]
[46] North, M. (2012): Data Mining for the Masses, [Internet], raspoloživo na: https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf , [15.05.2017.], str. 67.

identificirati homogene grupe opažanja koje nazivamo grozdovima (eng. *Cluster*)[47]. Radi se o algoritmu baziranom na udaljenosti koji dijeli podatke u unaprijed određeni broj grozdova (pod uvjetom da postoji dovoljno jedinstvenih slučajeva)[48]. Rezultat primjene ove tehnike bi trebao biti određeni broj skupina unutar kojih se nalaze podaci koji su istovremeno međusobno slični, ali također i različiti od podataka koji se nalaze u drugim skupinama.

Grupe objekata koje dijele zajednička svojstva imaju bitnu ulogu u načinu na koji ljudi analiziraju i opisuju svijet. To nam omogućava lakše razumijevanje određenih pojava. Biolozi su kroz dugi period vremena nastojali izraditi taksonomiju svih živih bića. Taj pothvat po mnogo čemu nalikuje grupiranju, ali ono što ih u suštini dijeli jest automatizacija[49].

Grupiranje također možemo koristiti kako bi apstrahirali individualne podatke koji se nalaze u tim grupama. Te grupe potom možemo koristiti kako bi njihovom daljnjom analizom donijeli zaključke o individualnim podacima koji se nalaze unutar njih.

Linearna regresija (eng. *Linear Regression*) – Cilj ove metode jest da se na temelju raspoloživih nezavisnih varijabli predvidi vrijednost ciljne varijable za svako opažanje[50]. Zamisao je da se ciljna varijabla izrazi kao linearna kombinacija nezavisnih varijabli s predodređenim ponderima[51]. Svoj cilj ostvaruje tako da pronađe vrijednosti regresijskih koeficijenata za koje će regresijska funkcija najbolje odgovarati skupu podataka koji nam stoji na raspolaganju[52]. Općenito se radi o vrijednosti regresijskih koeficijenata za koje se minimizira određena mjera pogreške, kao što je na primjer zbroj kvadrata grešaka.

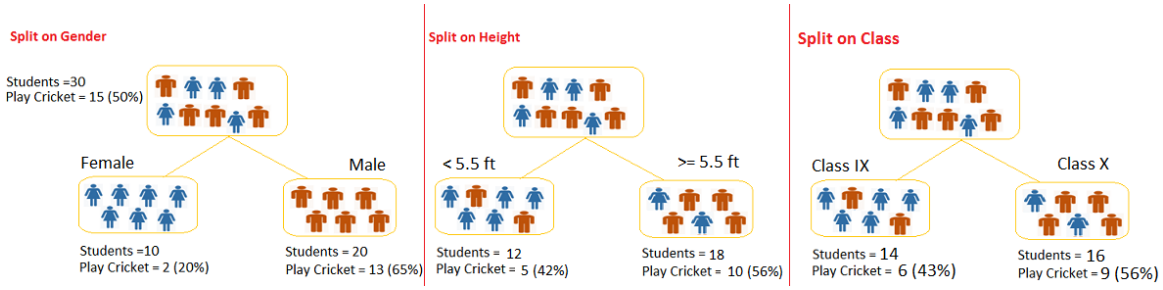
Regresija se koristi kako bi se ostvarila dva cilja. Prvi cilj je naglašavanje i tumačenje zavisnosti ciljne varijable o drugim, nezavisnim varijablama. Drugi cilj je predviđanje buduće

[47] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 293.
[48] Oracle (2017): k-Means, [Internet], raspoloživo na: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_kmeans.htm#DMCON057 ,[16.07.2017.]
[49] University of Minnessota Twin Cities: Cluster Analysis: Basic Concepts and Algorithms, [Internet], raspoloživo na: https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf, str. 488.
[50] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 93.
[51] Witten, I. H. et al. (2011): Data mining: Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington, str. 120.
[52] Singh, N. et al. (2012): Data Mining with Regression Technique, Journal of Information Systems and Communication, Volume 3, Issue 1, str. 200.

vrijednosti ciljne varijable temeljeno na identificiranoj funkcionalnoj zavisnosti i budućoj vrijednosti nezavisnih varijabli[53].

Najjednostavniji oblik regresije je kada postoji linearna zavisnost jedne varijable o jednoj nezavisnoj varijabli. Postoji veliki broj različitih regresijskih modela koje je potrebno spomenuti, kao što su npr. multivarijatna regresija. Standardna multivarijatna regresija istovremeno uzima u obzir sve nezavisne varijable. Kod *stepwise* i hijerarhijske regresije radi se o procesu unutar kojeg se analiziraju pojedinačne nezavisne varijable te se odabiru samo one koje najbolje odgovaraju modelu[54].

Stabla odlučivanja (eng. *Decision Trees*) - Stablo odlučivanja je vrsta *supervised learning* algoritma koji se najčešće koristi kod klasifikacijskih problema. Kod ove tehnike dijeli se populacija ili uzorak na dva ili više homogena skupa prema određenoj varijabli[55]. Sastoji se od čvorova i listova koji predstavljaju varijablu prema kojoj je moguće najbolje podijeliti određenu grupu podataka.



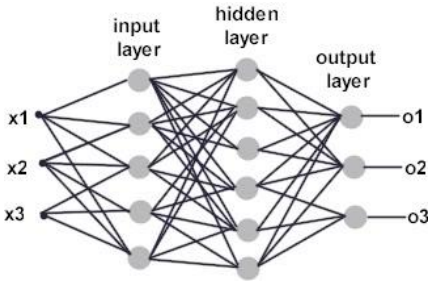
Slika 7 Stablo odlučivanja

Izvor: https://www.analyticsvidhya.com/wp-content/uploads/2015/01/Decision_Tree_Algorithm1.png

Neuronske mreže (eng. *Neural Networks*) – Ovom tehnikom imitira se rad ljudskog mozga kako bi se korištenjem umjetnih neurona međusobno usporedili atributi. Obrađivanjem vrijednosti atributa i stvaranjem čvorova spojenih neuronima, ova tehnika omogućava predviđanje čak i u uvjetima neizvjesnosti[56]. Neuronske mreže se uobičajeno sastoje od više slojeva. Svaki sloj se sastoji od određenog broja međusobno povezanih čvorova koji sadržavaju aktivacijsku funkciju. Ulazni sloj komunicira podatke, odnosno obrasce,

[53] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 154.
[54] Chapple, M. (2016): Defining the Regression Statistical Model, [Internet], raspoloživo na: https://www.thoughtco.com/regression-1019655 ,[16.07.2017.]
[55] Analytics Vidhya (2016): A Complete Tutorial on Tree Based Modeling from Scratch, [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one , [16.07.2017.]
[56] North, M. (2012): Data Mining for the Masses, [Internet], raspoloživo na: https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf , [15.05.2017.], str. 186.

skrivenom sloju koji je zadužen za njihovu obradu sustavom ponderiranih veza. Nakon što su podaci obrađeni šalju se izlaznom sloju. Svaka neuronska mreža ima sebi pridružena pravila učenja. U prvom ciklusu obrade podataka sustav neuronskih mreža nagađa obrazac tih podataka. Nakon toga potrebno je sagledati koliko taj rezultat odstupa od stvarnog i provodi potrebne korekcije[57].



Slika 8 Neuronske mreže

Izvor: http://i702.photobucket.com/albums/ww29/emiter40/viseslojni_perceptron.jpg

Text-mining – Korištenjem ove tehnike analiziraju se podaci koji se nalaze u obliku prirodnog jezika, odnosno teksta. Ova tehnika može pomoći organizaciji da izvuče vrijedne informacije o poslovanju iz različitih elektronički pohranjenih dokumenata, elektroničke pošte, komentara na društvenim mrežama i slično[58]. Rezultat provođenja ove metode je pretvaranje nestrukturiranih tekstualnih podataka u strukturirane numeričke podatke nad kojima se mogu koristiti tehnike vizualizacije i analize podataka. Može se koristiti za automatsku obradu otvorenih pitanja u anketama, identificiranje ključnih riječi u dokumentima, utvrđivanje sličnosti dokumenata i sl.

Slika 9 Text mining - Word Cloud

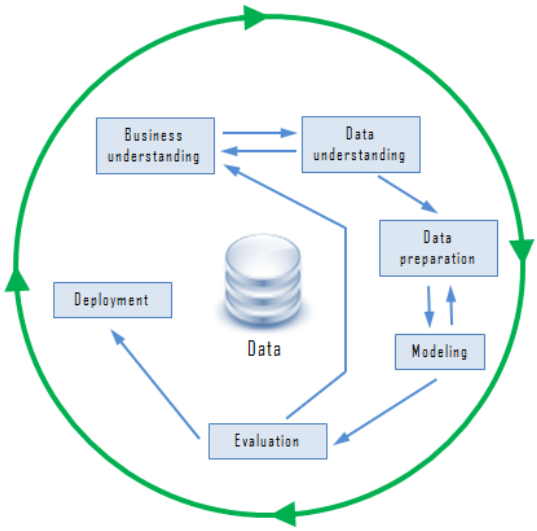
Izvor: https://www.meaningcloud.com/wp-content/uploads/2015/09/Text-Analytics.jpg

[57] University of Wisconsin-Madison: A basic introduction to Neural Networks, [Internet], raspoloživo na http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html ,[16.07.2017.]
[58] Rouse, M. (2013): Text Mining, [Internet], raspoloživo na: http://searchbusinessanalytics.techtarget.com/definition/text-mining ,[16.07.2017.]

4.3 Proces rudarenja podataka

Praćenjem strukturiranog pristupa rudarenju podataka značajno se povećava vjerojatnost uspjeha projekta, uz potencijalno smanjenje troškova[59]. Kako bi se sustavno provodila analiza i rudarenje podataka, potrebno je slijediti opći proces[60]. Iako postoji veći broj metodologija koje se mogu koristiti za provođenje projekata rudarenja podataka, u praksi se ipak najčešće koristi CRISP (*Cross-Industry Standard Process for Data Mining*) metodologija. Prema podacima iz 2015. godine ova se metodologija koristila u otprilike 43% zabilježenih projekata rudarenja podataka[61].

CRISP-DM metodologiju možemo proučavati dvojako. S jedne strane je možemo promatrati kao metodologiju kojom su opisane uobičajene faze projekta rudarenja podataka, kao i radni zadaci koje je potrebno izvršiti unutar svake faze. S druge strane CRISP-DM možemo promatrati kao procesni model kojim se pruža uvid u životni ciklus projekta rudarenja podataka.



Slika 10 CRISP-DM proces

Izvor: http://www.zentut.com/wp-content/uploads/2012/10/CRISP-DM.png

Radi se o cikličkom procesu unutar kojeg se prolazi kroz šest narednih faza: razumijevanje poslovanja, razumijevanje podataka, priprema podataka, izgradnja modela, testiranje i evaluacija, te isporuka.

[59] Stanton, J. (2012): An Introduction to Data Science, [Internet], raspoloživo na: https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf , [19.05.2017.], str. 22.
[60] Olsen, D. L. I Dursun, D. (2008): Advanced Data Mining Techniques, Springer, str 9.
[61] http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Potrebno je naglasiti da se radi o idealiziranom slijedu aktivnosti. Često se u praksi događa da se radni zadaci izvršavaju u različitom redoslijedu, budu preskočeni ili se ne ponavljaju[62]. Stoga možemo zaključiti da se radi o relativno fleksibilnoj metodologiji koja se može prilagoditi potrebama konkretnog projekta ili korisnika. U nastavku će se pružiti kratak uvid u svaku pojedinačnu fazu ovog procesa.

4.3.1 Razumijevanje poslovnog problema (eng. *Business Understanding*)

Prije nego što se započne s analizom i rudarenjem podataka potrebno je utvrditi razloge, odnosno ciljeve zbog kojih se ono uopće provodi. Unutar ove faze potrebno je proći kroz četiri ključne aktivnosti[63].

Određivanje poslovnih ciljeva – Unutar ovog koraka pokušava se utvrditi razlog provođenja projekta iz poslovne perspektive. Drugim riječima, potrebno je saznati koji se problemi žele riješiti korištenjem rudarenja podataka. Formalnim određivanjem poslovnih ciljeva smanjuje se vjerojatnost nepodudaranje poslovnih problema koji se žele riješiti i rezultata analize i rudarenja podataka. Kako bi bili sigurni da rezultati projekta odgovaraju određenim poslovnim ciljevima, potrebno je definirati niz mjerljivih pokazatelja uspješnosti projekta.

Procjena situacije – Prethodno određene poslovne ciljeve nije moguće promatrati izolirano od situacije u kojoj se poduzeće nalazi. Stoga je u ovom koraku potrebno detaljnije analizirati sami poslovni problem koji se želi riješiti i identificirati faktore koji na njega djeluju. Tako je potrebno izraditi popis resursa koji organizaciji stoje na raspolaganju, od samih ljudskih resursa, do podataka. Bitno je i identificiranje različitih pretpostavki i ograničenja kao što su npr. vremenski rokovi, zakonske obveze, sigurnosne obveze i sl. Značajan dio pažnje potrebno je posvetiti i identifikaciji rizika koji mogu naštetiti završetku projekta, te izradi planova za njihovo suzbijanje. Čak i da projekt u potpunosti riješi probleme s kojima se organizacija suočava, postavlja se pitanje njegove isplativosti. Kako bi se utvrdila isplativost projekta provodi se *cost-benefit* analiza kojom se u odnos stavljaju očekivane koristi od projekta s očekivanim troškovima.

Određivanje ciljeva rudarenja podataka – Prethodno određene poslovne ciljeve potrebno je pretvoriti u ciljeve rudarenja podataka. Poslovni ciljevi su općenite naravi i koriste poslovnu terminologiju, dok su ciljevima rudarenja podataka precizno određeni željeni

[62] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding ,[17.07.2017.]
[63] Brown, M. S. (2014): Business Understanding, [Internet],raspoloživo na: http://www.dummies.com/programming/big-data/phase-1-of-the-crisp-dm-process-model-business-understanding/ [17.07.2017.]

rezultati s kojima će biti ostvareni poslovni ciljevi. Unutar ovog koraka definiraju se modeli, izvještaji i načini prezentacije rezultata kojima se žele ostvariti poslovni ciljevi.

Izrada projektnog plana – Izrada detaljnog vremenskog rasporeda projektnih aktivnosti, korak po korak, kako bi se olakšale realizacija i nadzor projekta. Pri tom je potrebno definirati potrebne resurse za svaku aktivnost, te očekivane rezultate aktivnosti.

4.3.2 Razumijevanje podataka (eng. *Data Understanding*)

Nakon što su određeni poslovni ciljevi i ciljevi rudarenja podataka, potrebno je prijeći na podatke. Unutar prethodnog koraka definirani su potrebni resursi, među koje pripadaju i sami podaci[64]. Kako bi se uopće izvršile aktivnosti rudarenja podataka potreban je pristup različitim podacima. Stoga je unutar ove faze potrebno odabrati i prikupiti podatke iz različitih izvora, te utvrditi da li prikupljeni podaci odgovaraju postavljenim ciljevima[65]. Detaljniji uvid u svaku aktivnost ovog koraka biti će pružen u narednom dijelu.

Prikupljanje inicijalnih podataka – Podaci koji su potrebni za analizu i rudarenje podataka mogu se pronaći u velikom broju raznolikih izvora. Ti izvori se najčešće dijele na[66]:

- Postojeći podaci (eng. *Existing data*) – Radi se o skupu podataka kojim provoditelji projekata rudarenja već raspolažu. To su najčešće podaci o transakcijama, web logovi i sl.
- Kupljeni podaci (eng. *Purchased data*) – Postoji veliki broj organizacija specijaliziranih za prikupljanje podataka kod kojih je moguće preuzeti ili kupiti podatke potrebne za ostvarivanje poslovnih ciljeva.
- Dodatni podaci (eng. *Additional data*) – Ukoliko ni kombinacija primarnih i sekundarnih podataka nije dovoljna, organizacija se može odlučiti na dodatno prikupljanje podataka putem anketa, intervjua i drugih metoda.

Nakon što su podaci prikupljeni potrebno je dokumenti njihove izvore, opisati metode pomoću kojih su prikupljeni i sve probleme do kojih je pritom došlo.

---

[64] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding ,[17.07.2017.]
[65] Olsen, D. L. I Dursun, D. (2008): Advanced Data Mining Techniques, Springer, str. 12.
[66] IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf ,[17.07.2017.], str. 13.

Opisivanje podataka – Iako postoji više načina za opisivanje podataka, većina ih se ipak fokusira na kvantitativne i kvalitativne aspekte. Osnovne karakteristike na koje bi se pritom trebalo osvrnuti su[67]:

- **Količina podataka** – Potrebno je odabrati količinu podataka koja će se koristiti, odnosno hoće li se koristiti cijeli skup podataka ili njegov podskup. Naime, iako se korištenjem cjelokupnog skupa podataka donose precizniji zaključci, veća količina podataka može znatno produljiti očekivano vrijeme obrade podataka.
- **Vrsta podataka** – Prikupljeni podaci mogu pripadati različitim vrstama, stoga je potrebno odrediti formate u kojima se nalaze pojedini podaci. Vrsta podataka kojom se raspolaže određuje koje se tehnike rudarenja mogu koristiti nad njima.
- **Shema kodiranja** – Nad nenumeričkim podacima u njihovom standardnom obliku nije moguće provoditi rudarenje podataka. Stoga je potrebno svakoj vrijednosti nenumeričkog obilježja pridružiti numeričku vrijednost koja će je predstavljati. Taj se postupak naziva kodiranje podataka.

Istraživanje podataka – Tijekom ove faze nastoji se dati odgovor na istraživačka pitanja korištenjem tehnika vizualizacije podataka i izvještavanja[68]. Svakoj varijabli ispituje se njena distribucija i vrijednosti[69]. Eksplorativna analiza provodi se s ciljem upoznavanja s podacima i uočavanja njihovih nedostataka.

Verificiranje kvalitete podataka – Podaci su u svom standardnom obliku rijetko pogodni za direktnu primjenu u analizi i rudarenju. Ti podaci često sadržavaju veliki broj nedostataka kao što su npr. nedostajuće vrijednosti, nekonzistentni sustavi kodiranja ili pogrešno uneseni podaci. Sve uočene nedostatke i mogućnost njihovog uklanjanja potrebno je dokumentirati u izvještaju o kvaliteti podataka.

4.3.3 Priprema podataka (eng. *Data Preparation*)

U prethodnoj su fazi prikupljeni relevantni podaci, opisane njihove karakteristike, provedena preliminarna istraživanja i verificirana njihova kvaliteta. Sljedeći koraci se odnose na odabir i pripremu podataka za samu analizu.

---

[67] IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf ,[17.07.2017.], str. 14-15.

[68] Ibid., str. 16.

[69] Brown, M.S. (2014): Data Understanding, [Internet], raspoloživo na: http://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/ ,[17.07.2017.]

Odabir podataka – U ovom koraku je potrebno odabrati podatke za koje se smatra da će u najvećoj mjeri omogućiti ostvarivanje postavljenih ciljeva, vodeći pritom računa o njihovom formatu i kvaliteti.[70]. Pri odabiru podataka moguće se koristiti s dva različita pristupa. Prvim pristupom se odabire podskup, odnosno uzorak, kojim je obuhvaćen samo dio jedinki promatranja. Drugim pristupom se pak biraju stupci, odnosno obilježja kojima se opisuju jedinke u recima.

Čišćenje podataka – Radi se o skupini aktivnosti usmjerenih prema poboljšanju kvalitete podataka na razinu potrebnu za analizu[71]. Najčešći problemi koji se mogu pojaviti, kao i mogući načini njihovog rješavanja prikazani su u sljedećoj tablici[72].

Tablica 1 Problemi u skupu podataka

Problem	**Moguće rješenje**
Nedostajuće vrijednosti (eng. *Missing data*)	Isključivanje redaka ili stupaca Popunjavanje praznina pretpostavljenim vrijednostima
Pogreške u podacima (eng. *Data errors*)	Isključivanje redaka ili stupaca Korištenje logičkog razmišljana za ručno uklanjanje pogrešaka
Nekonzistentnost kodiranja (eng. *Coding inconsistencies*)	Odlučiti se za jedan sustav kodiranja i potom pretvoriti i zamijeniti vrijednosti
Nedostajući metapodaci (eng. *Missing metadata*)	

Izvor: Izrada autora u Wordu

Kontrukcija podataka – Postoji mogućnost da će prije same analize biti potrebno iz postojećih podataka izvesti nove atribute ili generirati potpuno nove podatke[73]. S jedne strane

[70] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding ,[17.07.2017.]
[71] Ibid.
[72]IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na:ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf ,[17.07.2017.], str. 20.
[73] Brown, M.S. (2014): Data Preparation, [Internet], raspoloživo na:http://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-preparation/[17.07.2017.]

postoje izvedeni atributi koji se stvaraju na temelju jednog ili više postojećih atributa[74]. Na primjer izvedeni atribut profit moguće je generirati oduzimanjem dvaju postojećih atributa, prihoda i troškova. druge strane može se pojavit potreba za generiranjem potpuno novih podataka. Postoje situacije u kojima je potrebno imati zapise o kupcu, unatoč tome što taj kupac nije imao kupovina ili narudžbi u prošloj godini[75].

Integracija podataka – Radi se o postupku spajanja podataka iz različitih baza podataka, tablica, zapisa ili drugih izvora. U tom kontekstu moguće je govoriti o samom spajanju podataka iz različitih tablica (spajanje više tablica koje sadrže različite informacije o istim objektima), ili agregaciji (zbrajanjem) podataka iz više tablica[76].

4.3.4 Modeliranje

Nakon što je skup podataka doveden u željeno stanje, prelazi se na sljedeći korak. Modeliranje podataka je korak unutar kojeg se koristi softver za rudarenje podataka kako bi se došlo do željenih rezultata[77].

Odabir tehnika modeliranja – Tehnike koje su prikazane u jednom od prethodnih poglavlja predstavljaju samo dio tehnika koje je moguće koristiti prilikom rudarenja podataka. Svaka tehnika ima određene pretpostavke koje je potrebno poštivati, najčešće u vidu kvalitete i formata podataka. Vodeći računa o tome potrebno je odabrati jednu ili više tehnika s kojima će se ostvariti postavljeni ciljevi.

Dizajniranje testova – Prije izgradnje modela potrebno je definirati način kojim će se testirati kvaliteta i vrijednost samog modela. Najčešći pristup dizajniranu testova svodi se na dijeljenje podataka u dvije skupine. Pritom se jedna skupina podataka koristi za izgradnju modela, dok se druga koristi za njegovo testiranje[78].

Izgradnja modela – Korak izgradnje modela predstavlja srž cjelokupnog procesa rudarenja podataka. Model se najčešće izrađuje kombiniranjem različitih operatora raspoloživih u alatu za rudarenje. Prilikom korištenja svake tehnike rudarenja podataka potrebno je definirati

[74] IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf,[17.07.2017], str. 21.
[75] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding,[17.07.2017.]
[76] Ibid.
[77] Olsen, D. L. I Dursun, D. (2008): Advanced Data Mining Techniques, Springer, str. 15.
[78] Brown, M.S. (2014): Modeling, [Internet], raspoloživo na:http://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/, [17.07.2017.]

inicijalne vrijednosti njenih parametara[79]. Odabrane inicijalne parametre potrebno je dokumentirati u izvještaju. Osim samih parametara, u ovom koraku potrebno je detaljno opisati model koji će se koristiti, varijable koje su korištene i način na koji je on interpretiran.[80]

Procjena modela – Na samom kraju potrebno ocijeniti izgrađeni model s tehničkog i poslovnog stajališta na temelju prethodno određenih kriterija uspjeha, testova i vlastitog znanja[81]. Ukoliko se pokaže da model ne zadovoljava definirane kriterije, moguće je izmijeniti postavke parametara kako bi se dobio bolji rezultat.

4.3.5 Evaluacija podatkovnog proizvoda

Prethodni napori u evaluaciji modela odnosili su se na tehničke faktore kao što su preciznost modela. Za razliku od toga, u ovom dijelu je potrebno procijeniti u kojoj mjeri model zadovoljava poslovne ciljeve i utvrđivanje poslovnih razloga njegovih nedostataka[82]. U tom kontekstu pojavljuju se dva problema. Prvi problem se odnosi na prepoznavanje poslovnih vrijednosti koje se nalaze u uzorcima podataka dobivenim rudarenjem. Drugi problem se odnosi na odabir adekvatnih alata za vizualizaciju podataka.

Nakon što se donio sud o podudarnosti rezultata s poslovnim ciljevima, i nakon što su rezultati prezentirani, potrebno je obratiti pozornost na sami proces rudarenja podataka. Analizom samog procesa nastoji se uočiti njegovi nedostaci, u cilju identifikacije načina za njihovo rješavanje u daljnjim ciklusima rudarenja.

Na samom kraju potrebno je odlučiti koji će biti sljedeći korak. Ovisno o rezultatima evaluacije moguće je pokrenuti novu iteraciju projekta unutar koje se želi poboljšati podatkovni proizvod, ili ipak zatvoriti projekt i isporučiti proizvod.

4.3.6 Isporuka podatkovnog proizvoda

Isporuka stvorenog modela predstavlja posljednju fazu ciklusa rudarenja podataka. U ovoj fazi se koriste novostvoreni uvidi kako bi se riješili poslovni problemi i stvorile promjene

[79] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding,[17.07.2017.]
[80] Brown, M.S. (2014): Modeling, [Internet], raspoloživo na:http://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/, [17.07.2017.]
[81] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding,[17.07.2017.]
[82] SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding,[17.07.2017.]

unutar organizacije[83]. Isporuka podatkovnog proizvoda je uvjerljivo najvažnija faza, jer nije bitno koliko je izvanredan model koji se kreirao, ukoliko nikada ne dođe do njegove praktične uporabe.

Prvi korak u ovoj fazi je izrada strategije za isporuku, odnosno primjenu, stvorenog podatkovnog proizvoda. Ukoliko se rezultati projekta rudarenja koriste u svakodnevnom poslovanju potrebno je izraditi plan njihove kontrole i održavanja.

Na samom kraju potrebno je izraditi završni izvještaj. U njemu su sadržani svi izvještaji i dokumenti izrađeni u prethodnim fazama, uz dodatak konačnog pregleda samog projekta i njegovih rezultata[84].

4.4 Poslovne primjene rudarenja podataka

Metode analize i rudarenja podataka moguće je primijeniti u velikom broju različitih djelatnosti kako bi se došlo do vrijednih saznanja. Ukratko će biti navedene samo neke od potencijalnih primjena rudarenja podataka.

Customer Relationship Management – S obzirom da se radi o djelatnosti fokusiranoj na vjernosti kupaca i dugoročnim odnosima, potrebna je adekvatna informacijska podloga za upravljanje odnosima s kupcima[85]. Rudarenje podataka omogućava donositeljima odluka mogućnost identificiranja jednostavnog obrazaca ponašanja kupaca, na temelju kojeg mogu lakše prilagoditi svoj marketinški mix pojedincima ili segmentima.

Analiza košarice (eng. *Basket Analysis*) – Primjena tehnika rudarenja podataka za identificiranje proizvoda i sadržaja koji se često prodaju zajedno. Služi se kako bi se bolje oblikovali stvarni i virtualni izlozi, ali i za izradu atraktivnih paketa (eng. *bundle*) proizvoda.

Predviđanje prodaje – Mogućnost primjene prediktivnih tehnika rudarenja podataka kako bi se na temelju povijesnih podataka utvrdila buduća prodaja.

Otkrivanje prijevara – Radi se o problemu s kojim se suočava veći broj industrija, kao što su banke, osiguravajuća društva, državne agencije i drugi. Korištenjem tehnika rudarenja

[83] IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf,[17.07.2017] , str. 35.
[84] Brown, M.S. (2014): Deployment, [Internet], raspoloživo na: http://www.dummies.com/programming/big-data/phase-6-of-the-crisp-dm-process-model-deployment/, [17.07.2017.]
[85] Rouse, M. (2014): Relationship Marketing, [Internet], raspoloživo na: http://searchcrm.techtarget.com/definition/relationship-marketing .[17.07.2017.]

podataka moguća je identifikacija faktora i obrazaca koji mogu dovesti do prijevare[86]. To omogućava poduzimanje pravovremenih akcija za suzbijanje prijevara i minimizaciju troškova.

Procjena rizika – Analiza rizika koristi se za procjenu rizika povezanog s budućim odlukama. Često se koristi u financijskom sektoru za potrebe odobravanja zajmova klijentima, na temelju njihovih karakteristika.

Rudarenje teksta – Tehnike rudarenja podataka moguće je koristiti na različitim dokumentima, knjigama, člancima, *e-mailovima* i sl., u svrhu njihove komparacije, klasifikacije i izvlačenja korisnih informacija[87].

Rudarenje weba – Kako bi bolje prilagodili izgled i sadržaj internet trgovina potrebama korisnika, možemo provesti analizu *clickstreamova* – slijeda posjećenih stranica i odluka donesenih od strane posjetitelja.

[86] StatSoft (2016): Fraud Detection, [Internet], raspoloživo na: http://www.statsoft.com/Textbook/Fraud-Detection ,[17.07.2017.]
[87] Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester, str. 82.

5. ALATI ZA RUDARENJE PODATAKA

5.1 Kriteriji odabira alata za analizu i rudarenje podataka

Prije nego se započne s samom analizom i rudarenjem podataka, potrebno je odabrati alate koji će se koristiti u tu svrhu.

Kriterij kojem je ovom prilikom pridružen najveći ponder je dostupnost alata. Stoga se prednost daje *open source* alatima, ili alatima čija je licenca dostupna autoru rada. Drugi značajan kriterij je postojanje dostupne detaljne dokumentacije o alatu. Alat treba imati dostupne dokumente ili video materijale kojima je podrobno objašnjen rad u njemu.

Na temelju prethodno navedenih kriterija odabrana su dva alata. Prvi alat koji će se koristiti za eksplorativnu analizu i vizualizaciju podataka je **Tableau Desktop**. Tableau omogućava poslovnim korisnicima da na interaktivan i intuitivan način pristupaju, pripreme i analiziraju podatke bez potrebe za kodiranjem[88].

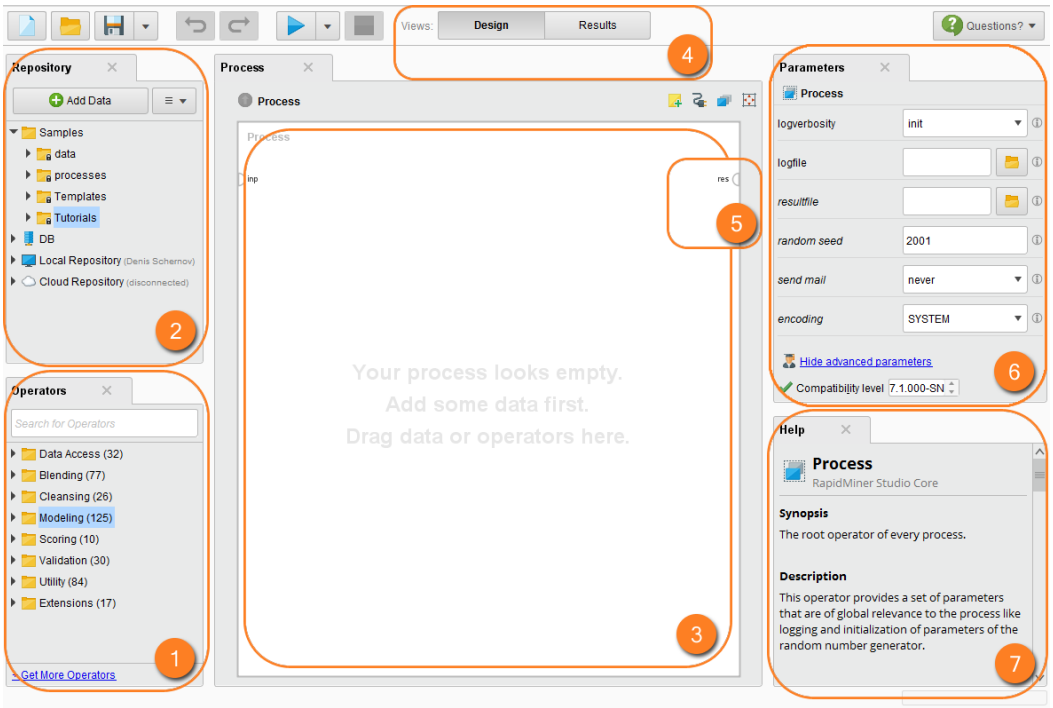
O kvaliteti ovog alata najbolje govori podatak da se već pet godina pojavljuje u Gartnerovom magičnom kvadrantu kao jedan od tri *leadera* u kategoriji alata za poslovnu inteligenciju. S obzirom na njegovu iznimnu zastupljenost u stvarnom poslovanju, na Internetu je moguće pronaći veliki broj uputa za rad u samom alatu.

Drugi alat koristit će se za samo rudarenje podataka. **RapidMiner Studio** je snažan alat za brzu izradu interpretativnih i prediktivnih analitika. Radi se o besplatnom, *all-in-one* alatu koji uključuje stotine različitih algoritama za pripremu podataka i strojno učenje. Alat je zbog svoje fleksibilnosti prigodan za korisnike različitih tehnoloških vještina i znanja. Tako da korisnici mogu birati između samostalnog pisanja koda i dostupnih unaprijed izrađenih algoritama.

[88] Gartner (2017): Magic Quadrant for Business Intelligence and Analytics Platforms, [Internet], raspoloživo na: https://www.gartner.com/doc/reprints?id=1-3TYE0CD&ct=170221&st=sb ,[24.08.2017.]

5.2 Sučelje i rad u alatu za analizu i rudarenje podataka

Ukratko će se prikazati sučelje svakog od odabranih alata. Na sljedećoj slici prikazano je sučelje alata RapidMiner Studio koje se sastoji od sedam bitnih dijelova.

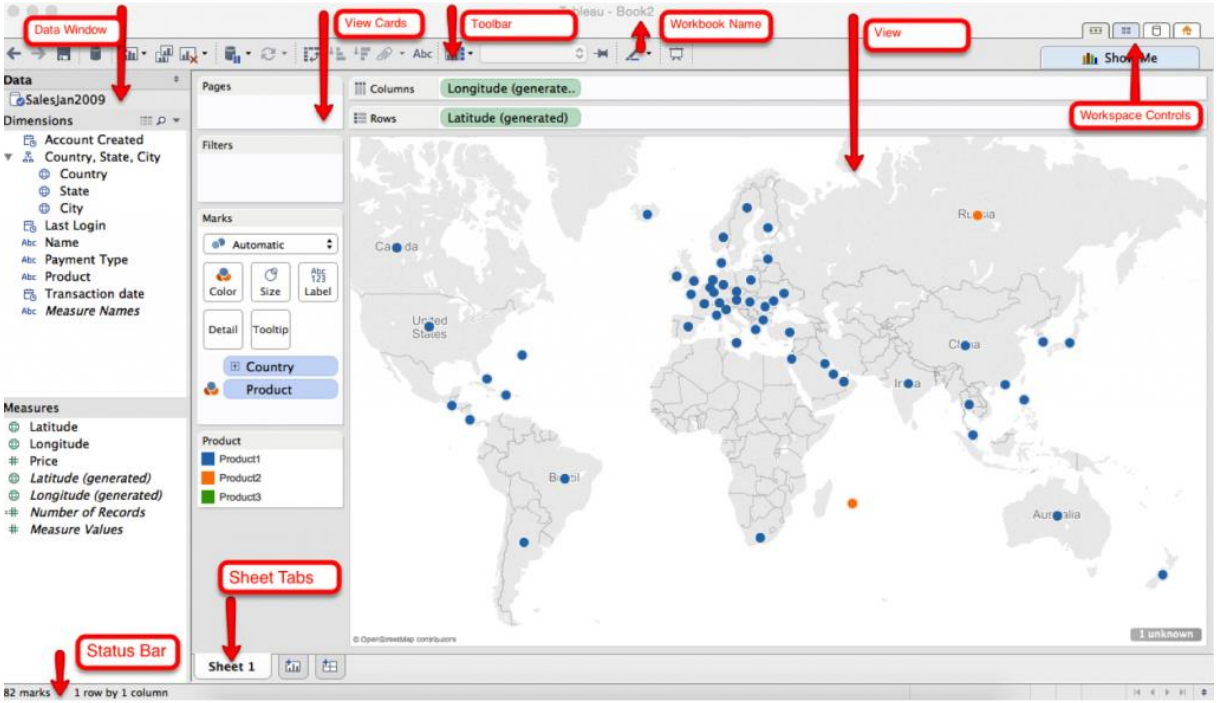


Slika 11 Sučelje alata RapidMiner Studio

Izvor: https://docs.rapidminer.com/studio/getting-started/img/studio-callouts.png

1. Operatori (eng. *Operators*) – U ovom prozoru nalaze se unaprijed dostupni operatori koji sadrže algoritme za pripremu i rudarenje podataka. Svaki operator predstavlja jedan građevni blok, čijim će se međusobnim povezivanjem pomoću jednostavne *drag-and-drop* tehnike izraditi model.
2. Repozitorij (eng. *Repository*) – Mjesto na kojem se pristupa pohranjenim RapidMiner modelima ili prethodno uspostavljenim vezama s podacima.
3. Procesna ploča (eng. *Process Panel*) – Prostor na kojem se *drag-and-drop* tehnikom izrađuje procesni model.
4. Pogledi (eng. *Views*) – Koristi se za odabir između trenutačnog pogleda unutar kojeg se dizajnira model i rezultata pokretanja tog istog modela.
5. Priključci (eng. *Ports*) – Svaki model treba imati definiran ulaz (veza na podatke) i izlaz (rezultati).
6. Parametri (eng. *Parameters*) – Postavke kojima se izmjenjuje ponašanje odabranog operatora.
7. Pomoć (eng. *Help*) – Prozor unutar kojeg je objašnjen odabrani operator.

Slijedeće na redu je sučelje alata Tableau Desktop, prikazano na slici.



Slika 12 Sučelje alata Tableau Desktop

Izvor: http://bi-impact.com/alpha/wp-content/uploads/2016/02/Untitled-1024x591.png

1. Podaci (eng. *Data Window*) – U ovom prozoru nalaze se mjere i dimenzije koje želimo vizualno analizirati.
2. Kartice (eng. *View Cards*) – Povlačenjem dimenzija ili mjera na neku od ovih kartica rezultira vizualnim prikazom podataka. Tako je moguće povući podatke na stupac, redak, veličinu, boju i sl.
3. Alatna traka (eng. *Toolbar*) – Na ovom području pristupa se različitim alatima koji stoje na raspolaganju.
4. Ime radne knjige (eng. *Workbook Name*) – Ime radne knjige u kojoj se trenutačno radi.
5. Pogled (eng. *View*) – Okvir unutar kojeg se nalazi vizualni prikaz podataka.
6. Kontrole radnog prostora (eng. *Workspace Controls*)
7. Stranice (eng. *Sheets*) – Svaka radna knjiga može se sastojati od više stranica kroz koje je ovdje moguće manevrirati.
8. Traka stanja (eng. *Status Bar*) – Pruža uvid u broj stupaca, redaka i podataka u trenutačnom prikazu.

6. ANALIZA I RUDARENJE PODATAKA NA PRIMJERU

Prilikom izrade i prezentacije rezultata praktičnog primjera koristiti će se CRISP-DM metodologija za rudarenje podataka. S obzirom da je ona u svom standardnom obliku prilagođenija za velike projekte, ovdje će se okvirno pratiti njene smjernice prema fazama.

6.1 Razumijevanje poslovnog problema

Kako bi uopće bilo moguće započeti s praktičnim primjerom, potrebno se kratko osvrnuti na sami subjekt ovog istraživanja, odnosno na Erasmus+ projekte.

Erasmus + je novi program za obrazovanje, usavršavanje i mlade koji zamjenjuje sedam (7) postojećih programa (Program za cjeloživotno učenje [Erasmus, Leonardo da Vinci, Comenius i Grundtvig], Mladi na djelu, Erasmus Mundus, Tempus, Alfa, Edulink i Program suradnje sa industrijaliziranim zemljama)[89].

Program je usmjeren prema jačanju znanja i vještina te zapošljivosti europskih građana, kao i unaprjeđivanju obrazovanja, osposobljavanja te rada u području mladih i sporta. Posebno je usmjeren povezivanju obrazovanja, osposobljavanja i sektora mladih s poslovnim sektorom, te je otvoren za njihove zajedničke projekte.

Da bi se ostvarili ciljevi ovog programa, provode se tri ključne aktivnosti[90]:

- KA 1 – *Learning Mobility* – Projekti mobilnosti pojedinaca u svrhu učenja i usavršavanja. Pod tim se podrazumijeva mobilnost učenika, studenata, vježbenika, mladih ljudi, volontera, nastavnog i nenastavnog osoblja i brojnih drugih, u inozemstvu.
- KA 2 – *Suradnja* – Suradnja u svrhu inovacija i razmjene dobrih praksi (suradnja s gospodarskim subjektima). Cilj ovih programa je ostvarivanje transnacionalnih strateških partnerstava usmjerenih na razvoj inicijativa u jednom ili više područja obrazovanja, osposobljavanja ili promicanja inovacija. Osim toga želi se poticati i udruživanje znanja između ustanova visokog obrazovanja i poduzeća.
- KA 3 – *Policy Reform* – Otvorena metoda potpore koordinacije obrazovne politike EU i bolonjskom procesu, provedba strategije „Europe 2020“.

[89] Sveučilište u Zagrebu (2016): Erasmus+:Opće Informacije , [Internet], raspoloživo na: www.unizg.hr/suradnja/medunarodna-suradnja/partnerstva/erasmus/ , [15.05.2017.]
[90] Sveučilište u Zagrebu (2016): Erasmus+:Opće Informacije , [Internet], raspoloživo na: http://www.unizg.hr/suradnja/medunarodna-suradnja/partnerstva/erasmus/ , [15.05.2017.]

Suštinski cilj koji se želi ostvariti u ovom dijelu rada je ispitivanje mogućnosti praktične primjene metoda i tehnika analize i rudarenja podataka kako bi se iz dostupnih godišnjih statistika Erasmus+ programa mogle izvući korisne informacije.

6.2 Razumijevanje i priprema podataka

Podaci na kojima će biti rađena analiza podataka su sekundarni, prethodno prikupljeni podaci dostupni na stranicama Erasmus+ programa. Tablica koju je moguće preuzeti s njihove stranice sadrži opće informacije o projektima realiziranim između 2014. i 2017. godine. Radi se o impozantnom skupu podataka u kojem se nalaze vrijednosti 251. obilježja za sveukupno 61427. projekata. Za svaki projekt zabilježeno je njegovo ime, vrsta ključnih aktivnosti, vrsta pod aktivnosti, dodijeljena financijska sredstva, status, opis projekta i sl.

Iako se radi o relativno sređenom i strukturiranom skupu podataka, potrebno ga je prije same obrade dodatno doraditi i pročistiti. Prvi problem koji je potrebno riješiti je redundantnost stupaca, odnosno atributa. Naime, skup podataka sadrži veliki broj stupaca koji se odnose na podatke o partnerskim institucijama kao što su adrese i adrese web stranica. S obzirom da se radi o podacima koji nisu relevantni za analizu, moguće ih je ukloniti kako bi se smanjila količina podataka i time olakšala računalna obrada podataka. Drugi problem se odnosi na postojanje malog postotka nedostajućih vrijednosti za određene projekte. S obzirom da se radi o kvalitativnim, kategorijskim varijablama, problem nedostajućih vrijednosti riješiti će se filtriranjem projekata koji ih sadrže.

Iz početnog skupa podataka moguće je izvesti određena svojstva projekata koja mogu biti zanimljiva za analizu. Tako na primjer iz podataka o partnerskim institucijama možemo izračunati broj partnera koji su sudjelovali na pojedinom projektu.

Prethodno je spomenuto da skup podataka sadrži određeni broj kvalitativnih, kategorijskih varijabli. Nad ovakvim tipom varijabli nije moguće primijeniti određeni broj tehnika rudarenja podataka, stoga ih je potrebno kodirati. Kodiranje je postupak dodjeljivanja numeričkih vrijednosti kvalitativnim varijablama.

Uspješno izvršavanje navedenih aktivnosti trebalo bi rezultirati skupom podataka koji je adekvatan za daljnju obradu.

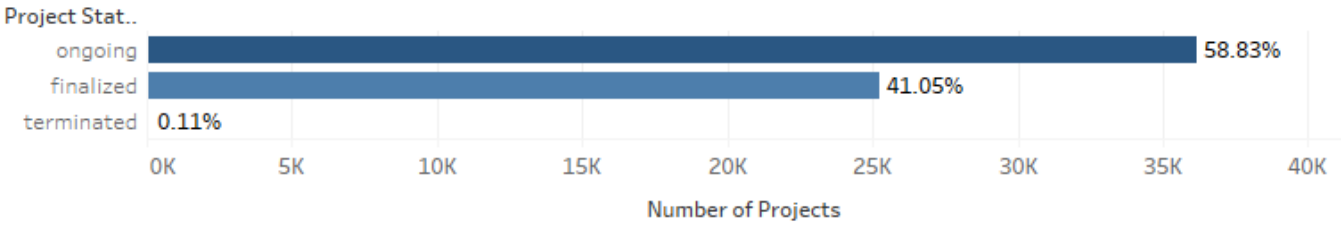
Rezultat izvršavanja prethodno navedenih aktivnosti jest očišćeni i prilagođeni skup podataka, adekvatan za provođenje daljnjih analiza. Sastoji se od svega 32 stupca u kojima se nalaze podaci o svakom pojedinom projektu.

1. *Project Identifier* – Predstavlja jedinstveni identifikator svakog pojedinačnog projekta.
2. *Key Action* – Predstavlja kvalitativnu varijablu kojom je definirana vrsta ključne aktivnosti pojedinog projekta. Svi projekti Erasmus+ programa spadaju pod tri vrste ključnih aktivnosti: *Learning Mobility*, *Cooperation*, *Policy Reform*.
3. *Key Action Code* – Radi se o numeričkoj, kodiranoj vrijednosti kvalitativne varijable *Key Action*, koja se kreće u rasponu od 1 do 3.
4. *Action Type* – Predstavlja kvalitativnu varijablu s kojom se definira podvrsta aktivnosti svakog projekta.
5. *Action Type Code* – Radi se o numeričkoj, kodiranoj vrijednosti kvalitativne varijable *Action Type*, te se kreće u rasponu od 1 do 39.
6. *Call Year* – Vremenska varijabla s kojom se definira početna godina projekta.
7. *Project Title* – Nominalna varijabla koja sadrži ime svakog pojedinog projekta.
8. *Topics* – Nominalna varijabla kojom je definirano tematsko područje svakog pojedinog projekta.
9. *Project Summary* – Svaka ćelija u ovom stupcu sadrži relativno detaljan, tekstualni opsi svakog pojedinog projekta.
10. *Project Status* – Kvalitativna varijabla s kojom se opisuje trenutačni status projekta koji može biti u tijeku, završen ili okončan.
11. *Project Status Code* – Kodirana vrijednost kvalitativne varijable *Project Status*. Vrijednosti se kreću od 1 do 3, od tekućih projekata, do okončanih projekata.
12. *EU Grant award in euros* – Količina novčanih sredstava koja je dodijeljena pojedinom projektu.
13. *Good Practice* – Kvalitativna varijabla koja može poprimiti dvije vrijednosti, a određuje smatra li se praksa provođenja projekta dobrom, ili ne.
14. *Good Practice Code* – Kodiranja vrijednost varijable Good Practice. Lošoj praksi dodijeljena je vrijednost 2, dobroj praksi 1.
15. *Success Story* – Kvalitativna varijabla koja može poprimiti dvije vrijednosti, a određuje smatra li se projekt uspješnim, ili ne.
16. *Succes Story Code* – Kodiranja vrijednost varijable Good Practice. Neuspješnim projektima dodijeljena je vrijednost 2, uspješnim 1.
17. *Number of Partners* – Ovaj stupac sadrži podatke o broju partnera koji sudjeluju na svakom pojedinom projektu.
18. *Number of Participants* - Ovaj stupac sadrži podatke o broju sudionika na svakom pojedinom projektu.

19. *Coordinating Organisation Name* – Nominalna varijabla kojom je određeno ime organzacije koja provodi projekt.
20. *Coordinating Organisation Type* – Kvalitativna varijabla kojom je određena vrsta organizacije koja provodi projekt.
21. *Coordinating Organisaton Type Code* – Kodirana vrijednost kvalitativne varijable *Coordinating Organisation Type*.
22. *Coordinator's Country* – Kvalitativna varijabla kojom je određena država kojoj pripada organizacija koja provodi projekt.

6.3 Eksplorativna analiza podataka

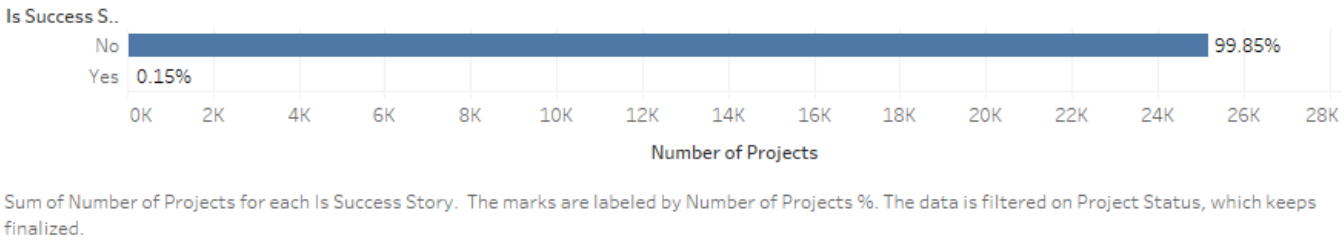
U ovom dijelu provoditi će se eksplorativna analiza kako bi se stekao bolji uvid u raspoložive podatke i informacije koje oni skrivaju. Prva informacija koja može biti interesantna donositeljima odluka predstavlja udio projekata prema statusu. Tri su moguća stanja u kojima projekt može biti: u tijeku, završen i otkazan. Najveći dio zabilježenih projekata (58.83%) još je uvijek u tijeku. Završeni projekti čine drugi veliki udio (41.05%) ukupnog broja projekata. Preostali projekti spadaju u kategoriju otkazanih ili terminiranih projekata, te je njihov udio gotovo zanemariv (0.11%). O statusu projekta potrebno je voditi računa prilikom analize drugih varijabli kao što je na primjer uspješnost projekta.



Slika 13 Broj projekata prema statusu

Izvor: Izrada autora u alatu Tableau

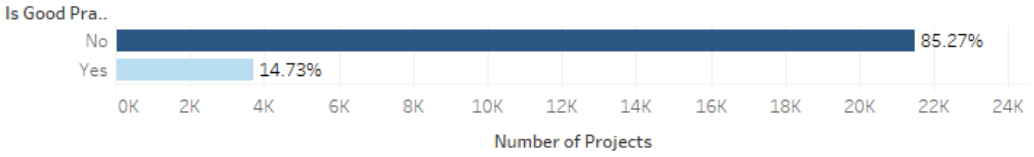
Unutar dostupnih podataka o projektima Erasmus+ programa moguće je pronaći binarnu varijablu kojom je određeno smatra li se projekt uspješnim ili ne. S obzirom na veliki broj projekata koji se još uvijek realiziraju, potrebno izvršiti filtraciju skupa podataka. Dobiveni podskup sadrži samo podatke o završenim projektima. Provođenje eksplorativne analize rezultira informacijom da se čak 99.85% projekata završenih smatra neuspješnim. S obzirom da se radi o iznimno visokom postotku, upitan je kredibilitet ovog podatka.



Slika 14 Broj završenih projekata prema uspješnosti

Izvor: izrada autora u alatu Tableau

Godišnji podaci o projektima sadržavaju i binarnu varijablu kojom se definira da li se ukupna praksa provođenja projekta smatra dobrom, ili ne. Na temelju podskupa dovršenih projekata možemo primijetiti da manji udio projekata pripada grupi dobre prakse (14.73%).



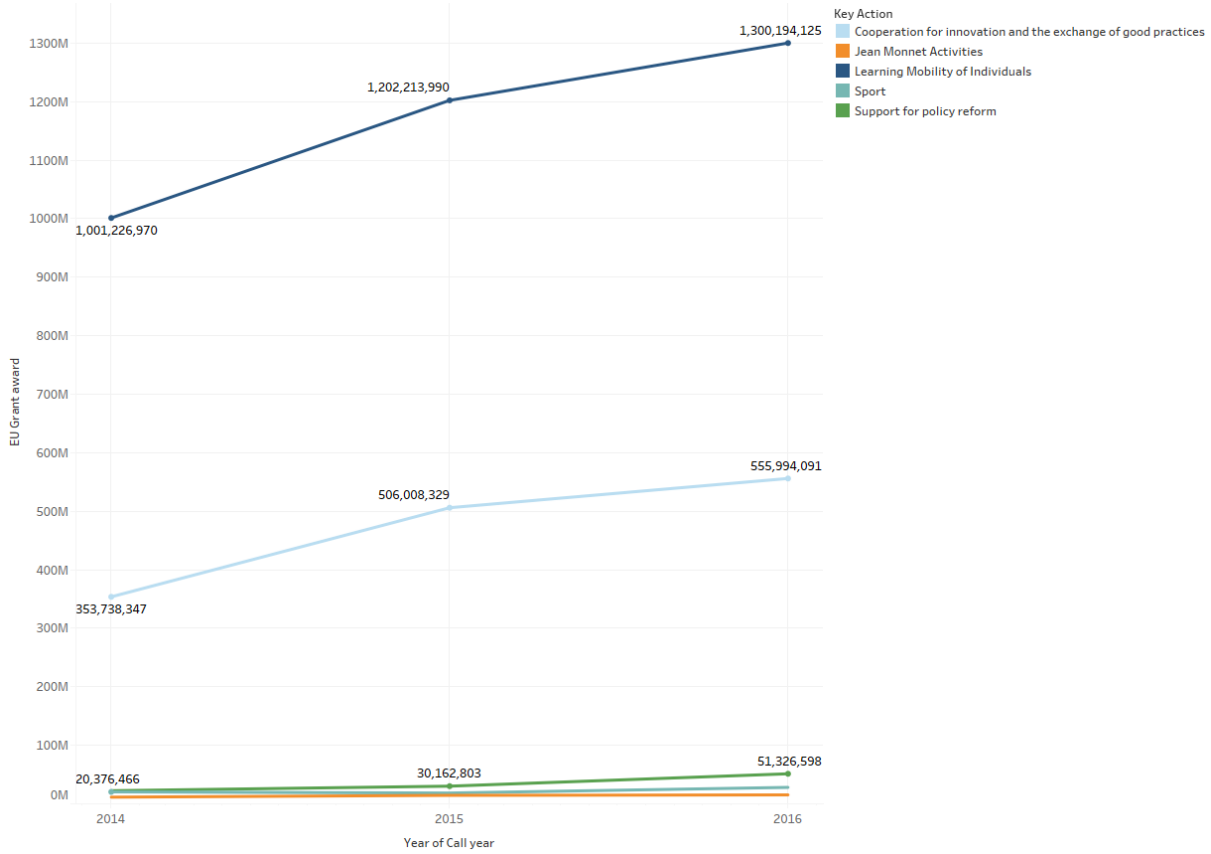
Sum of Number of Projects for each Is Good Practice. Color shows sum of Number of Projects. The marks are labeled by Number of Projects %. The data is filtered on Project Status, which keeps finalized.

Slika 15 Broj završenih projekata prema uspješnosti prakse

Izvor: izrada autora u alatu Tableau

6.3.1 Financiranje projekata po godinama

Podaci dostupni na stranicama Erasmus+ programa sadrže informacije o količini dodijeljenih financijskih sredstava prema godini. S grafičkog prikaza moguće je primijetiti da količina financijskih sredstava za svaku od ključnih aktivnosti postepeno raste od 2014. prema 2016. godini. Najviše financijskih sredstava dodjeljno je projektima mobilnosti, te isti ostvaruju najveći rast koji u 2015. iznosi 20.07%, dok u 2016. usporava i iznosi 8.15%. Znatno ispod njih nalaze se projekti kooperacije u inovaciji. Njihov porast u 2015. u odnosu na 2014. iznosi 43.04%, te usporava u 2016. na 9.87%.

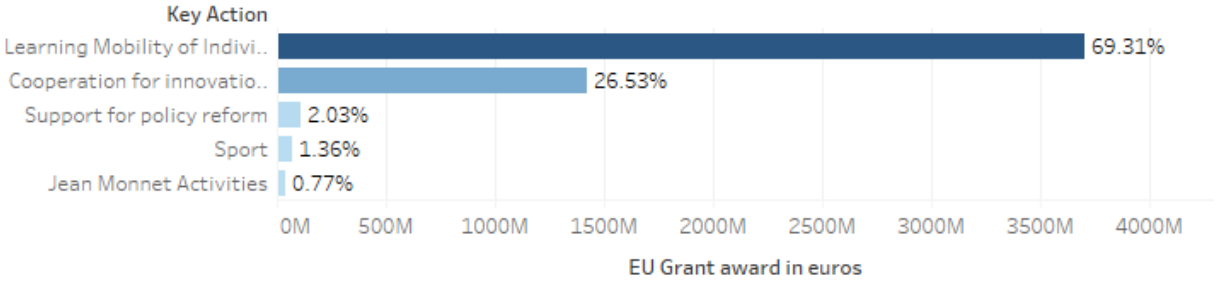


Slika 16 Financijska sredstva po godinama

Izvor: izrada autora u alatu Tableau

6.3.2 Financiranje projekata prema ključnim aktivnostima

Moguće je primijetiti da financijska sredstva za projekte u najvećoj mjeri (69.31%) otpadaju na projekte učenja kroz mobilnost. Osim njih, značajan udio sredstava otpada na suradnju na inovacijama u visokom obrazovanju (26.53%). Na ostale tri ključne aktivnosti otpada znatno manji dio sredstava.



Slika 17 Financijska sredstva prema ključnim aktivnostima

Izvor: izrada autora u alatu Tableau

U sljedećoj tablici nalaze se podaci o apsolutnom i relativnom broju projekata, apsolutnoj i relativnoj količini financijskih sredstava, kao i prosječna vrijednost istih.

Tablica 2 Broj projekata i količina financijskih sredstava prema ključnim aktivnostima

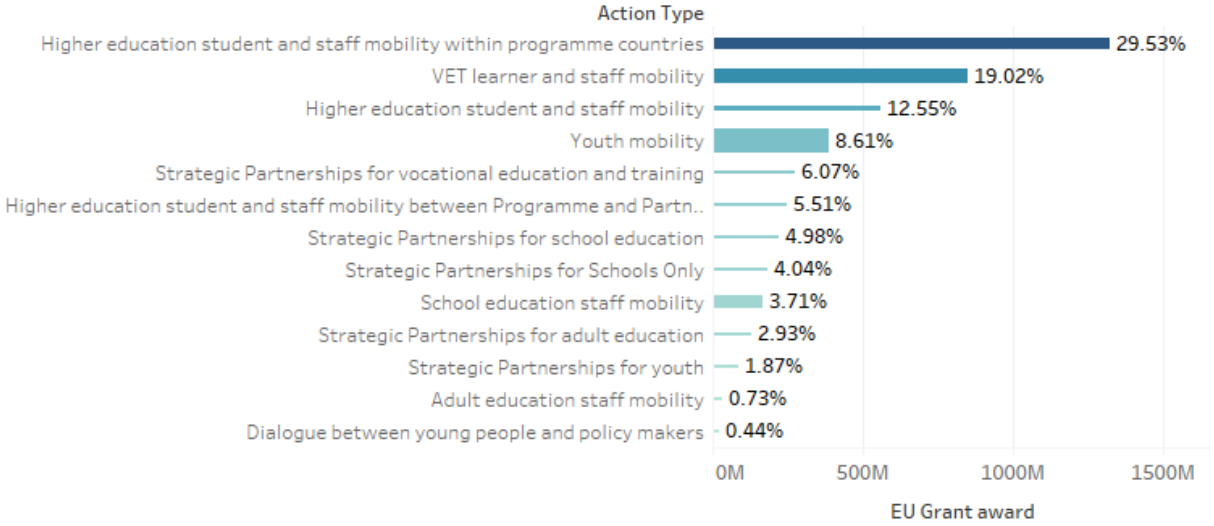
Key Action	Number of Projects	Number of Projects %	EU Grant award	Avg. EU Grant award	EU Grant award %
Learning Mobility of Ind..	52,566	85.57%	3,705,994,777	70,504	69.31%
Cooperation for innovat..	6,730	10.96%	1,418,744,030	210,809	26.53%
Support for policy refor..	1,152	1.88%	108,578,604	94,252	2.03%
Sport	237	0.39%	72,854,677	307,404	1.36%
Jean Monnet Activities	742	1.21%	41,043,913	55,315	0.77%

Number of Projects, Number of Projects %, EU Grant award, Avg. EU Grant award and EU Grant award % broken down by Key Action.

Izvor: izrada autora u alatu Tableau

Moguće je primijetiti da iako projekti učenja kroz mobilnost čine 85.57% ukupnog broja projekata, njihov udio u ukupnim financijskim sredstvima je svega 69.31%. S druge strane projekti suradnje na inovacijama u visokom obrazovanju čine 10.96% ukupnih projekata, ali na njih otpada 26.53% financijskih sredstava više. Razlog leži u činjenici da prosječna financijska sredstva koja se dodjeljuju projektima suradnje iznose 210.809 €, dok za projekte mobilnosti ona iznose svega 70.504 €. Projekti koji se odnose na organiziranje i održavanje sportskih događanja prosjeku izvlače najviše financijskih sredstava (307.404 €), ali je broj istih relativno mali (1.21%). „Jean Monnet“ aktivnosti nalaze se na zadnjem mjestu i po broju projekata (1.21%), i po financijskim sredstvima (0.77%).

6.3.3 Financiranje projekata prema podaktivnostima



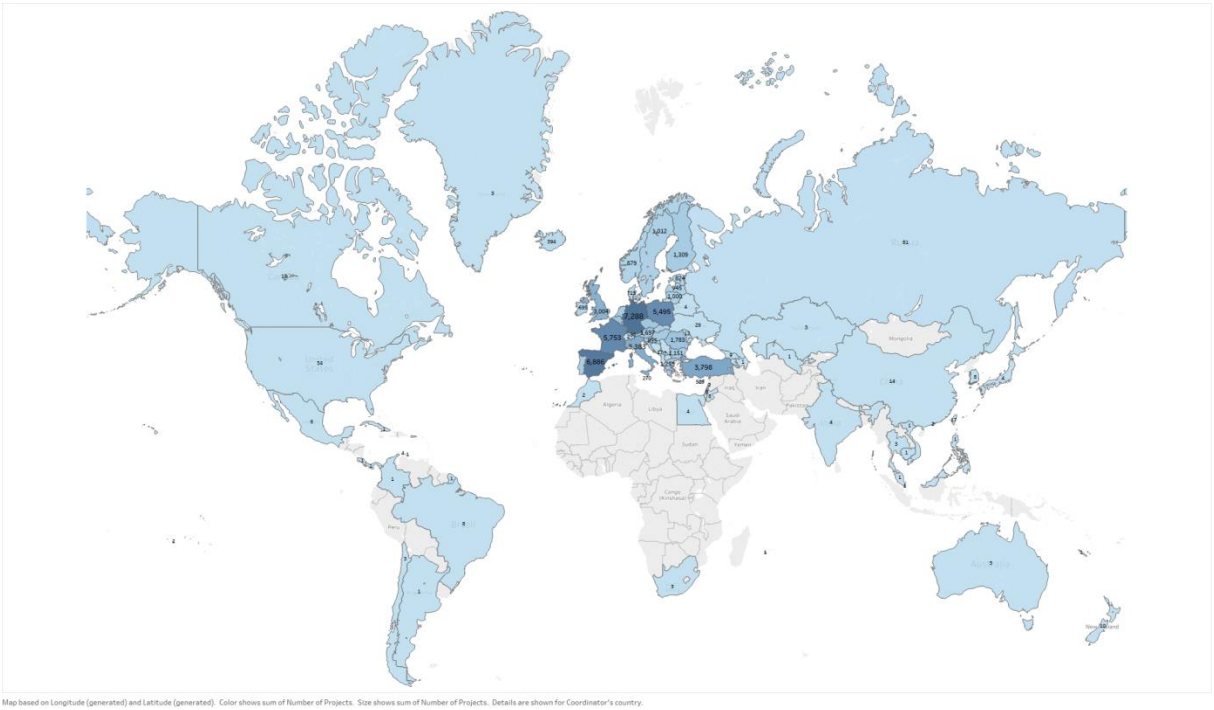
Slika 18 Financijska sredstva prema podaktivnostima

Izvor: izrada autora u alatu Tableau

Tri ključne aktivnosti Erasmus+ programa predstavljaju širi pojam kojim je obuhvaćen veći broj aktivnosti kojima se realiziraju ciljevi programa. Zanimljiv podatak govori o količini financijskih sredstava dodijeljenih pojedinim vrstama aktivnosti. Na horizontalnoj osi vizualizacije podataka nalazi se količina sredstava u eurima, dok se na vertikalnoj osi nalaze različite vrste aktivnosti. Duljinom stupca za pojedinu aktivnost određena je količina financijskih sredstava, dok širina stupca pak upućuje na broj projekata. Moguće je primijetiti da se s najvećom količinom sredstava (29.53%) financiraju projekti mobilnosti studenata i djelatnika u visokom obrazovanju unutar članica Erasmus+ programa. Na ove aktivnosti otpada gotovo 1.25 milijardi Eura. Značajan udio imaju i projekti mobilnosti strukovnih studenata i osoblja (19.02%), mobilnost studenata i osoblja u visokom obrazovanju (12.55%) i mobilnost mladih (8.61%). Iako na mobilnost mladih otpada gotovo četiri puta manje financijskih sredstava nego na vodeću vrstu aktivnosti, projekti ove vrste su najzastupljeniji (18,043). Projekti mobilnosti visokog obrazovanja unutar članica programa su tek peti prema zastupljenosti (10,000).

6.3.4 Broj projekata i veličina financijskih sredstava prema koordinatorima

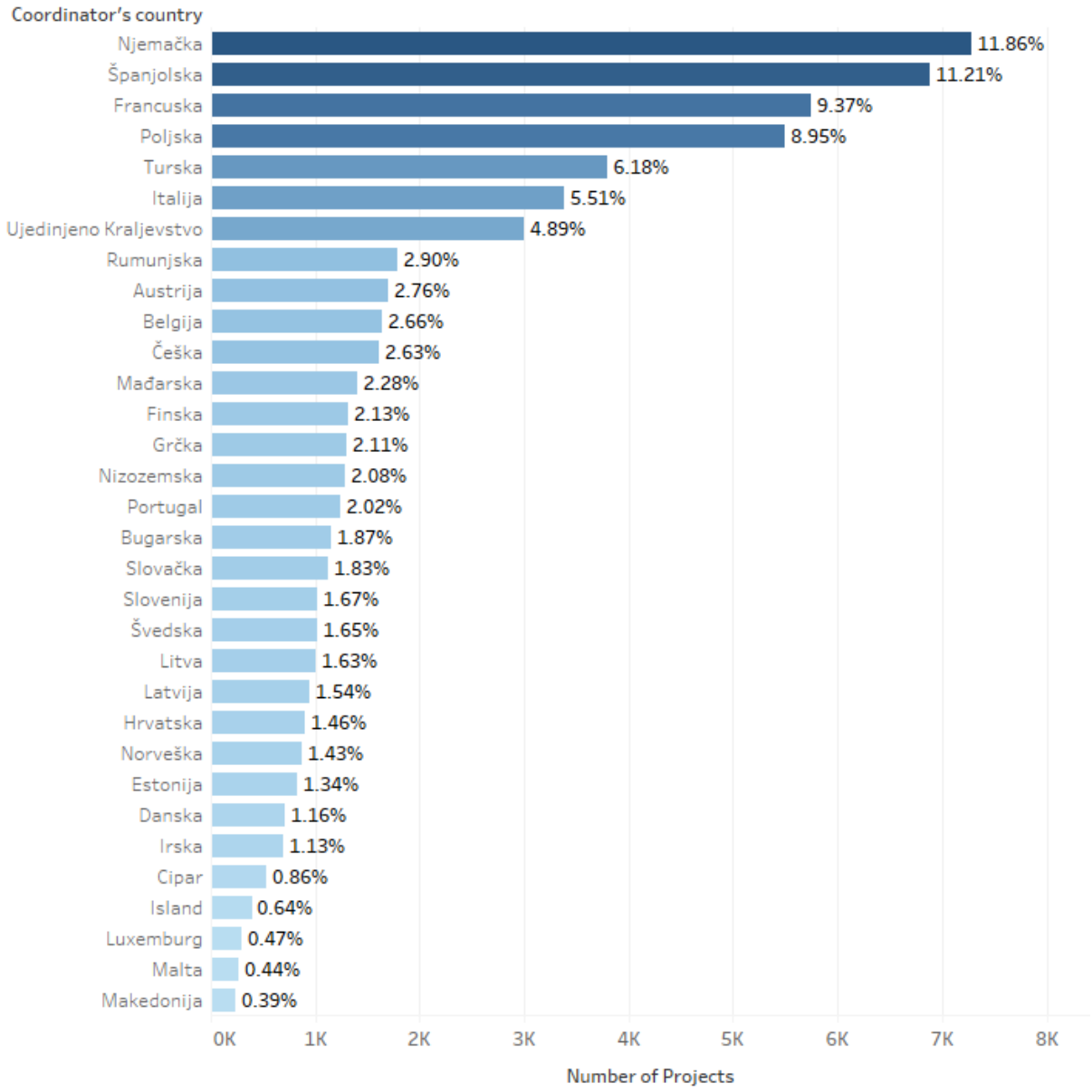
Za svaki projekt koji se provodi unutar Erasmus+ programa postoji organizacija koordinator. Korištenjem jednostavne kombinacije toplinskih mapa i karte svijeta moguće je prikazati na intuitivan način zemlje čije su institucije koordinirale najviše projekata. Logično je i očekivano da se najveći broj projekata upravo koordinira od strane zemalja Europske Unije, prvenstveno Njemačke, Španjolske, Francuske i Italije. Značajan broj projekata koordiniran je i od strane turskih organizacija.



Slika 19 Toplinska mapa broja projekata prema zemlji koordinatora

Izvor: izrada autora u alatu Tableau

Iako vizualno atraktivan, ovim prikazom otežano je rangiranje zemalja prema broju projekata, kao i sagledavanje detalja o njihovim udjelima u ukupnoj sumi projekata. Stoga se potrebno osvrnuti na klasični stupčasti dijagram prikazan na sljedećoj slici. Na apscisi se nalazi broj projekata koordiniranih od strane zemalja koje se pak nalaze na osi ordinate. Kako bi se izbjegla pretrpanost dijagrama, prikazani rezultati se odnose samo na 32 zemlje s najvećim brojem realiziranih projekata. Moguće je primijetiti da su samo četiri zemlje koordinirale više od 5000 projekata, te da njihovi projekti zajedno čine više od 40% ukupnog broja istih. Na prvom mjestu nalazi se Njemačka (11.86%) koju slijede Španjolska (11.21%), Francuska (9.37%) i Poljska (8.95%). Hrvatska se nalazi na 23. mjestu, s 895 realiziranih projekata (1.46%). Time se svrstava u skupinu zemalja kao što su Slovenija, Litva, Latvija, Estonija i Norveška.

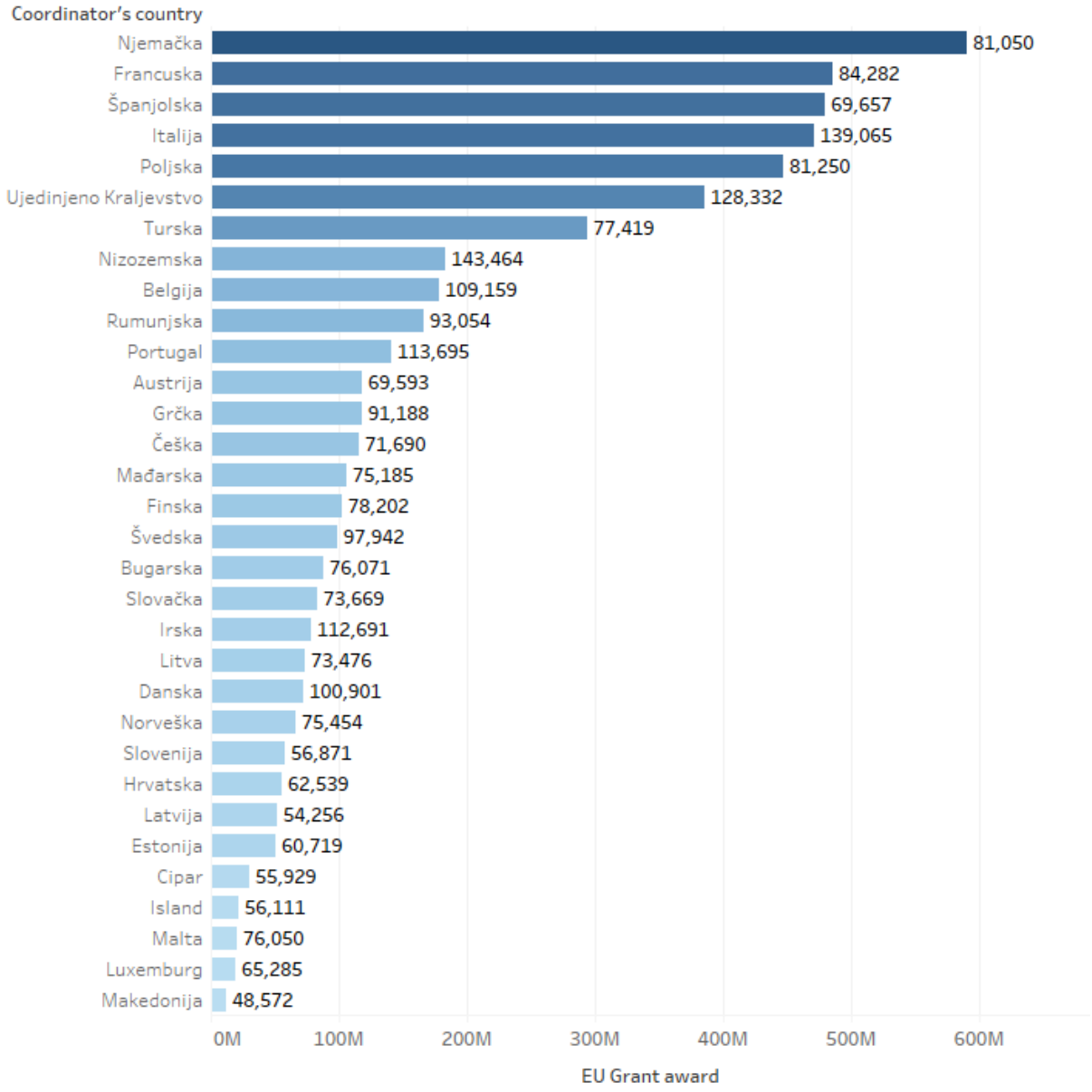


Slika 20 Broj projekata prema zemlji koordinatora

Izvor: izrada autora u alatu Tableau

Iako je broj projekata bitan faktor, potrebno je i proučiti kolika su financijska sredstva povukle pojedine zemlje. Prvih pet zemalja ostalo je gotovo nepromijenjeno. Njemačka je još uvijek na prvom mjestu, s gotovo 600 mil. € povučenih sredstava. Prosječna dodijeljena sredstva za svaki projekt iznose 81050 €. Francuska i Španjolska uspjele su povući oko 500 mil. € svaka, s tim da je Francuska povukla zanemarivo veći broj sredstava unatoč činjenici da je njezin udio u ukupnom broju projekata za 2 p.p. manji od Španjolskog. Razlog leži u činjenici da su Francuski projekti u prosjeku povlačili 84282 €, dok su Španjolski povlačili 69657€. Slična situacija je i između Italije i Poljske. U usporedbi s brojem realiziranih

projekata, Hrvatska je pala s 23. na 25. mjesto. Ukupna sredstva kojom su financirani projekti koordinirani od strane hrvatskih organizacija i institucija iznose oko 55 mil. €, dok je prosječna vrijednost 62539€.



Sum of EU Grant award for each Coordinator's country. Color shows sum of EU Grant award. The marks are labeled by average of EU Grant award. The data is filtered on sum of Number of Projects, which includes values greater than or equal to 100.

Slika 21 Financijska sredstva prema zemljama koordinatorima

Izvor: izrada autora u alatu Tableau

Koordinatori projekata Erasmus+ programa čine širok spektar različitih institucija koje pripadaju različitim sektorima. S toga od značaja može biti i informacija o tome koje vrste institucija najčešće koordiniraju projekte.

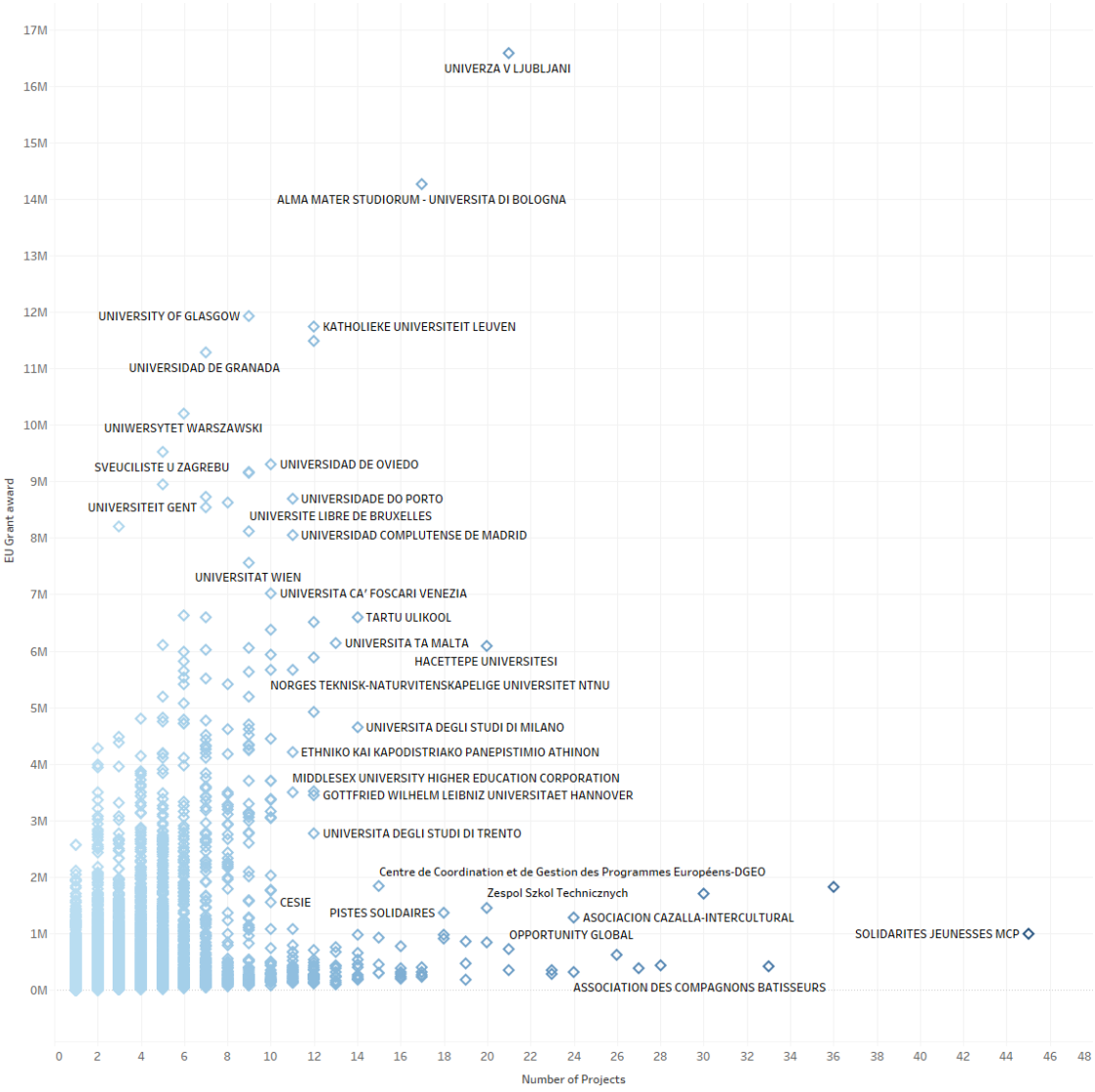
Na sljedećem prikazu prikazane su različite vrste organizacija i njihovi udjeli u ukupnom broju projekata. Moguće je primijetiti da najveći broj projekata (oko 24%) ima nedefiniranu vrstu djelatnosti. Najveći udio prema tome imaju nevladine organizacije/asocijacije/društvena poduzeća. Njihov udio u projektima mobilnosti iznosi 13.55%, a sumirani udjeli u kooperaciji i reformama politike iznose oko 2%. Na drugom mjestu nalaze se institucije visokoškolskog obrazovanja s 13.50% udjela u projektima mobilnosti, i svega 2% udjela u ostalim projektima. Slijede ih strukovne škole sekundarne razine (12.48%) i opće obrazovanje druge razine (7.54%). Sveukupno ove vrste institucija čine preko 40% ukupnog broja projekata.

	Key Action		
Coordinating organisation type	Support for polic..	Cooperation for in..	Learning Mobility ..
Other	0.39%	2.12%	21.69%
Non-governmental organisation/association/social enterprise	0.58%	1.73%	13.55%
Higher education institution (tertiary level)	0.05%	1.55%	13.50%
School/Institute/Educational centre – Vocational Training (secondary level)	0.00%	0.78%	11.70%
School/Institute/Educational centre – General education (secondary level)	0.01%	1.83%	6.68%
School/Institute/Educational centre – General education (primary level)		0.62%	3.92%
School/Institute/Educational centre – Vocational Training (tertiary level)	0.00%	0.19%	3.58%
Local Public body	0.06%	0.25%	1.70%
Foundation	0.09%	0.24%	1.45%
School/Institute/Educational centre – Adult education	0.00%	0.29%	1.28%
Group of young people active in youth work	0.01%	0.03%	1.37%
European NGO	0.22%	0.07%	0.63%
Non-Profit making cultural organizations	0.02%	0.07%	0.79%
Regional Public body	0.03%	0.17%	0.55%
Small and medium sized enterprise	0.00%	0.26%	0.35%
School/Institute/Educational centre – General education (pre-primary level)		0.10%	0.49%
National Public body	0.13%	0.08%	0.28%
Civil Society Organisation	0.04%	0.07%	0.34%
Accreditation, certification or qualification body	0.03%	0.07%	0.34%
Research Institute/Centre	0.01%	0.15%	0.07%
Social partner or other representative of working life (chambers of commerc..	0.02%	0.07%	0.14%
Sport club		0.00%	0.16%

Slika 22 Udio projekata prema vrsti organizacije

Izvor: izrada autora u alatu Tableau

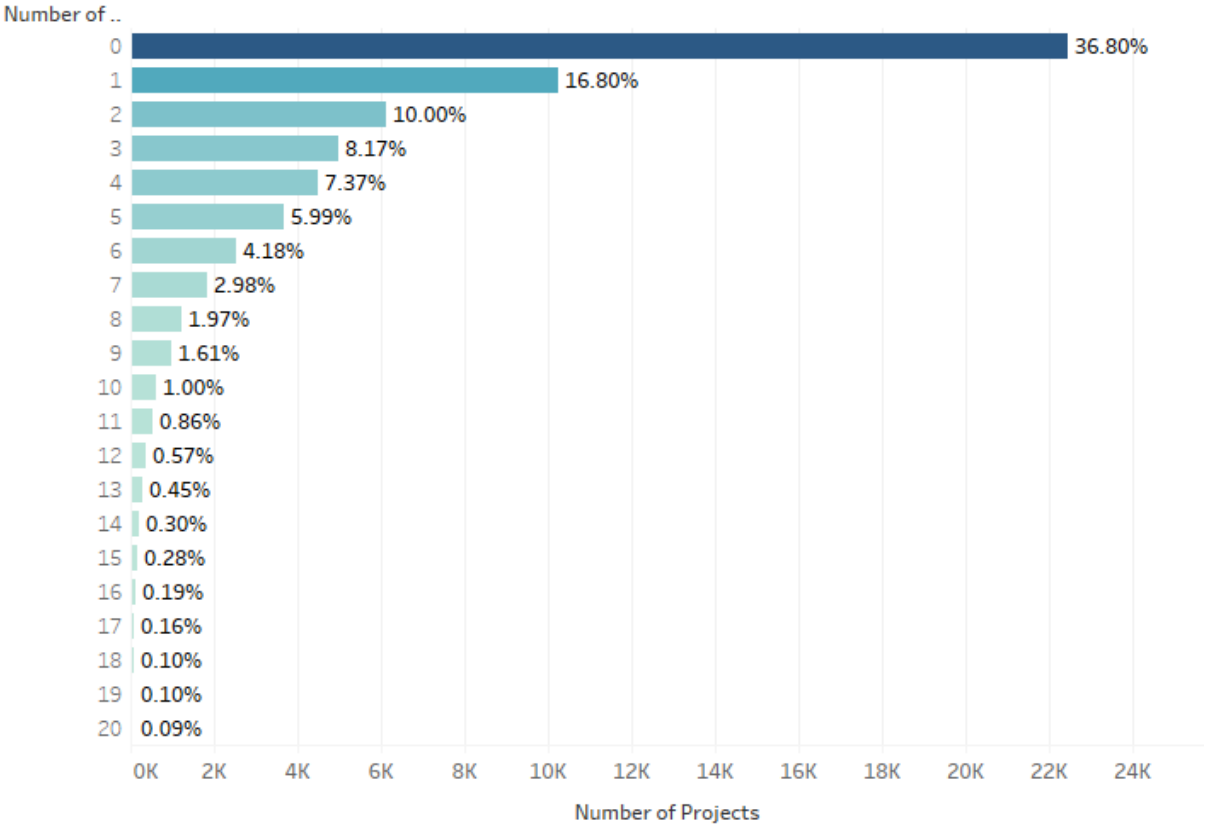
Korištenjem dijagrama rasipanja moguće je prikazati podatke o uspješnosti pojedinih organizacija u pogledu broja projekata i financijskih sredstava. Najveću količinu sredstava povukli su Sveučilište u Ljubljani (16.5 mil. €, 21 proj.), Sveučilište u Bolonji (14.25 mil. €, 17 proj.), Sveučilište u Glasgow (12 mil. €, 9 proj.) Sveučilište u Leuvenu (11.8 mil. €, 12 proj.). S druge strane imamo institucije koje su realizirale veliki broj projekata kao što su Solidarités Jeunesses (0.9 mil. €, 45 proj.), Idrima Neolaias (1.8 mil.€, 36 proj.), Hrvatska škola Outward Bound (0.4 mil €, 33 proj.) i Association Des Compagnons Batisseurs (0.3 mil., 24 proj.). Interesantno je da su organizacije koje se realizirale najveći broj projekata većinom nedržavne institucije, dok sveučilišta prevladavaju među organizacijama koje se povukle najveći broj sredstava.



Slika 23 Dijagram rasipanja

Izvor: izrada autora u alatu Tableau

6.3.5 Projekti prema partnerstvima



Slika 24 Projekti prema broju partnera

Izvor: izrada autora u alatu Tableau

Određeni projekti nisu koordinirani samostalno od strane organizacije koordinatora, već u njegovom provođenju sudjeluju i druge partnerske institucije. Na temelju prethodnog prikaza moguće je primijetiti da projekti bez partnera čine najveći udio (36.80%), te da ih je sveukupno oko 22000. Značajan udio imaju i projekti s jednim partnerom (16.80%) koji obuhvaćaju oko 10000 projekata, projekti s dva partnera (10.00%) s oko 6000 projekata i projekti s tri (8.17%) i projekti s četiri partnera (7.37%).

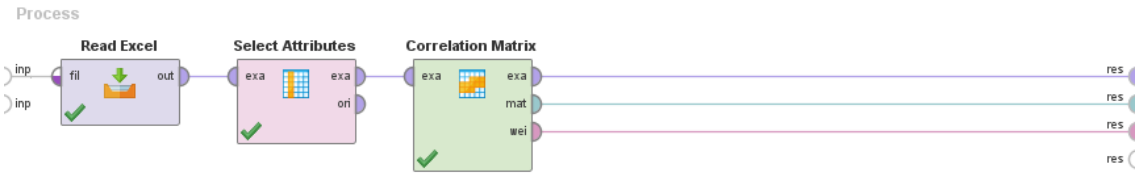
Logično bi bilo zaključiti da će s povećanjem broja partnerskih institucija doći do smanjenja broja projekata. Tu pretpostavku potvrđuje vizualni prikaz podataka, koji ukazuje da povećanjem broja partnera, broj projekata teži nuli.

6.4 Modeliranje, evaluacija i isporuka

Modeliranje predstavlja ključni korak CRISP-DM procesa unutar kojeg se podaci upotrebljavaju kako bi se pronašli korisni obrasci. Sastoji se od relativno jednostavnog slijeda aktivnosti, unutar kojeg je prvo potrebno odabrati tehnike modeliranja, potom izgraditi i procijeniti model odabranog podskupa podataka. Izrađeni model će se ukratko i evaluirati, odnosno biti će objašnjeni njegovi rezultati.

6.4.1 Korelacija (eng. *Correlation*)

Koeficijent korelacije predstavlja broj između -1 i +1 kojim se mjeri stupanj povezanosti između dva atributa. Pozitivna vrijednost koeficijenta korelacije ukazuje na to da se povećanjem vrijednosti jedne varijable, povećava i vrijednost druge. Negativna vrijednost koeficijenta korelacije ukazuje na to da se povećanjem vrijednosti jedne varijable, smanjuje vrijednost druge, i obratno. Zbog svoje jednostavnosti radi se o jednoj od najčešće korištenih tehnika rudarenja podataka.



Slika 25 Korelacijski Model

Izvor: Izrada autora u alatu RapidMiner Studio

Moguće je primijetiti da se korelacijski model sastoji od samo tri operatora. Prvim operatorom se učitavaju podaci iz Excel datoteke, drugim se biraju atributi između kojih se žele identificirati koeficijenti korelacije, dok se zadnjim operatorom izrađuje matrica korelacije.

Attributes	Key Action	Action Type	Call year	Topics	EU Grant ...	Good Practice	Success Story	Number of partners	Coordinating organisation type
Key Action	1	0.250	0.021	-0.126	-0.157	0.014	0.024	-0.164	0.010
Action Type	0.250	1	-0.035	0.103	0.158	-0.014	0.033	-0.067	0.088
Call year	0.021	-0.035	1	0.110	0.020	-0.216	-0.028	-0.030	0.046
Topics	-0.126	0.103	0.110	1	0.156	-0.050	0.013	0.077	0.028
EU Grant award in euros	-0.157	0.158	0.020	0.156	1	-0.006	0.000	0.143	-0.061
Good Practice	0.014	-0.014	-0.216	-0.050	-0.006	1	0.087	0.036	-0.002
Success Story	0.024	0.033	-0.028	0.013	0.000	0.087	1	-0.004	0.012
Number of partners	-0.164	-0.067	-0.030	0.077	0.143	0.036	-0.004	1	0.101
Coordinating organisation type	0.010	0.088	0.046	0.028	-0.061	-0.002	0.012	0.101	1

Slika 26 Korelacijska Matrica

Izvor: Izrada autora u alatu RapidMiner Studio

S obzirom da vrijednosti su procijenjene vrijednosti korelacijskog koeficijenta za svaki par varijabli između -0.4 i +0.4, moguće je zaključiti da ne postoji međusobna povezanost između analiziranih varijabli.

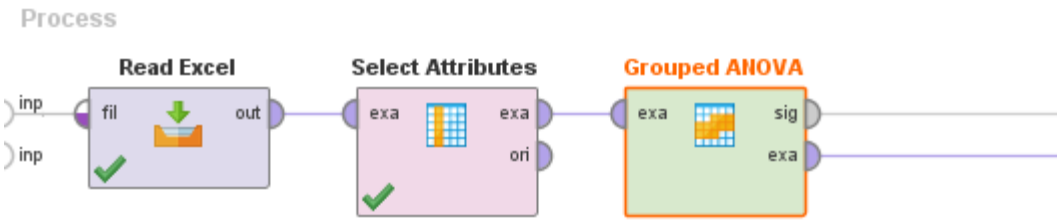
6.4.2 Analiza varijance ANOVA

ANOVA je tehnika rudarenja podataka kojom se uspoređuju aritmetičke sredine više uzoraka i donosi se zaključak o postojanju ili ne postojanju razlika između sredina više populacija. Radi se o tehnici prikladnoj za dostupni skup podataka jer kombinira nominalne varijable s numeričkim. Moguće je npr. ispitati da li se može kao istina prihvatiti pretpostavka da vrsta ključne aktivnosti ne djeluje značajno na prosječna financijska sredstva. Postavljaju se sljedeće hipoteze.

$$H_0 \dots \dots \sigma_A^2 = 0$$

$$H_1 \dots \dots \sigma_A^2 \neq 0$$

Model koji će se pritom koristiti u alatu izgleda kako je prikazano na sljedećoj slici.



Slika 27 ANOVA Model

Izvor:Izrada autora u alatu RapidMiner Studio

Pokretanjem izrađenog modela rezultirati će sljedećim ispisom:

Source	Square Sums	DF	Mean Squares	F	Prob
Between	40409353778390.640	4	10102338444597.6...	290.355	0.000
Residuals	347687139639282.060	9993	34793069112.307		
Total	388096493417672.700	9997			

Slika 28 ANOVA Rezultati - Ključne aktivnosti i financijska sredstva

Izvor: izrada autora u alatu RapidMiner Studio

Testiranje značajnosti djelovanja faktora može se vršiti F-testom[91]. S obzirom da je empirijska vrijednost $F^* = 290.355$ veća od tablične vrijednosti $F^{0.05}_{(4;9993)} = 2.372$ pri razini

[91] Rozga, A.(2009): Statistika za ekonomiste, Ekonomski fakultet, Split, str. 168

signifikantnosti od 5%, prihvaća se alternativna hipoteza da je djelovanje vrste ključne aktivnosti na odobrena financijska sredstva statistički značajno. Odnosno može se reći da postoji statistički značajna razlika u prosječnim financijskim sredstvima dodijeljenim pojedinim vrstama aktivnosti.

Isti model možemo iskoristiti za ispitivanje pretpostavke da zemlja koordinatora ne djeluje statistički značajno na prosječna financijska sredstva. Pri tom se postavljaju jednake hipoteze kao i u prethodnom primjeru.

$$H_0 \dots \dots \sigma_A^2 = 0$$

$$H_1 \dots \dots \sigma_A^2 \neq 0$$

Izmjenom odabranih atributa i ponovnim pokretanjem modela rezultira sljedećim ispisom:

Source	Square Sums	DF	Mean Squares	F	Prob
Between	9603655357193...	58	1655802647792...	46.263	0.000
Residuals	3557263338759...	9939	35790958232.811		
Total	4517628874478...	9997			

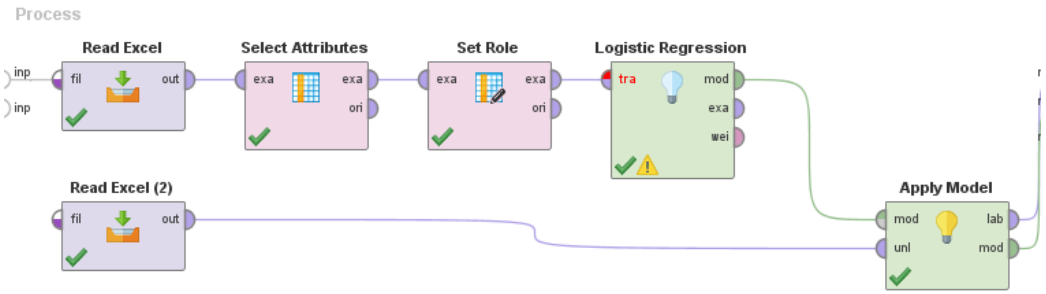
Slika 29 ANOVA Rezultati - Zemlja koordinatora i financijska sredstva

Izvor: Izrada autora u alatu RapidMiner Studio

S obzirom da je empirijska vrijednost $F^* = 46.263$ veća od tablične vrijednosti $F^{0.05}_{(4;9993)} = 1.325$ pri razini signifikantnosti od 5%, prihvaća se alternativna hipoteza da je djelovanje zemlje koordinatora na odobrena financijska sredstva statistički značajno. Odnosno može se reći da postoji statistički značajna razlika u prosječnim financijskim sredstvima dodijeljenim pojedinim zemljama koordinatorima.

6.4.3 Logistička regresija

Logistička regresija koristi se za opisivanje veze između zavisne binarne varijable i jedne ili više nominalnih, ordinalnih ili intervalnih nezavisnih varijabli[92]. Drugim riječima logističkom regresijom previđa se s određenom sigurnosti realizacija nekog događaja. Pri tom su potrebna dva skupa podataka. Prvi skup podataka sadržava vrijednosti i zavisnih i nezavisnih varijabli za određeni broj opažanja i koristi se za učenje modela. Drugi skup podataka sadrži samo vrijednosti nezavisnih varijabli kod kojih se želi predvidjeti vrijednost zavisne varijable.



Slika 30 Model logističke regresije

Izvor: Izrada autora u alatu RapidMiner Studio

Pokretanjem modela u alatu dolazi se do nekoliko rezultata. Prvi rezultat predstavlja tablicu u kojoj se nalaze koeficijenti logističke regresije, standardne greške, z-vrijednosti i p-vrijednosti, za svaki pojavni oblik odabranih nezavisnih varijabli. Dio koeficijenta logističke regresije prikazan u sljedećoj tablici.

Attribute	Coefficie...	Std. Co...	Std. Error	z-Val...	p-Val...
Number of partners	0.025	0.098	0.009	2.679	0.007
Key Action.Support for policy reform	-0.170	-0.170	0.322	-0.527	0.598
Key Action.Sport	-11.520	-11.520	119.372	-0.097	0.923
Key Action.Jean Monnet Activities	-0.630	-0.630	0.727	-0.866	0.387
Key Action.Cooperation for innovation and the exchange...	-0.643	-0.643	0.173	-3.714	0.000
Intercept	-3.345	-3.262	0.204	-16.423	0
EU Grant award in euros	0.000	0.039	0.000	0.901	0.368
Coordinating organisation type.Twinning committee	-10.967	-10.967	237.141	-0.046	0.963

Slika 31 Koeficijenti logističke regresije

Izvor: Izrada autora u alatu RapidMiner Studio

[92] StatisticsSolutions (2017): What is Logistic Regression?,[Internet], raspoloživo na: http://www.statisticssolutions.com/what-is-logistic-regression/, [20.08.2017.]

Drugi dio rezultata odnosi se na previđanje uspješnosti prakse svakog pojedinog projekta na temelju vrijednosti nezavisnih varijabli. Osim što se previđa hoće li projektna praksa biti uspješna, prikazana je i vjerojatnost da se to stvarno ostvari.

Row No.	predictio...	confidence(No)	confidence(Yes)	Project Iden...	Key Action C...	Key Action	Action Type ...	Action Type	Call year
1	No	0.984	0.016	2015-1-CY01...	1	Learning Mob...	9	Higher educa...	2015
2	Yes	0.890	0.110	2014-1-ES01...	1	Learning Mob...	1	School educa...	2014
3	No	0.945	0.055	2016-1-DE04...	1	Learning Mob...	5	Youth mobility	2016
4	No	0.964	0.036	2014-1-SK01...	2	Cooperation f...	11	Strategic Part...	2014
5	No	0.941	0.059	2015-1-NO01...	1	Learning Mob...	2	VET learner a...	2015
6	No	0.947	0.053	2014-2-DE04...	1	Learning Mob...	5	Youth mobility	2014
7	No	0.929	0.071	2016-1-NL02...	1	Learning Mob...	5	Youth mobility	2016
8	No	0.939	0.061	2014-1-CZ01...	1	Learning Mob...	5	Youth mobility	2014
9	Yes	0.885	0.115	2014-1-FR01...	1	Learning Mob...	3	Higher educa...	2014
10	Yes	0.887	0.113	2014-2-PT02...	1	Learning Mob...	5	Youth mobility	2014
11	No	0.968	0.032	2014-1-TR01...	1	Learning Mob...	3	Higher educa...	2014
12	No	0.926	0.074	2014-1-FI01-...	1	Learning Mob...	1	School educa...	2014
13	Yes	0.887	0.113	2014-2-EE01...	1	Learning Mob...	5	Youth mobility	2014
14	No	0.956	0.044	2015-2-NL02...	3	Support for p...	27	Dialogue bet...	2015
15	No	0.937	0.063	2015-1-IT01-...	1	Learning Mob...	2	VET learner a...	2015

Slika 32 Predviđanje uspješnosti projektne prakse

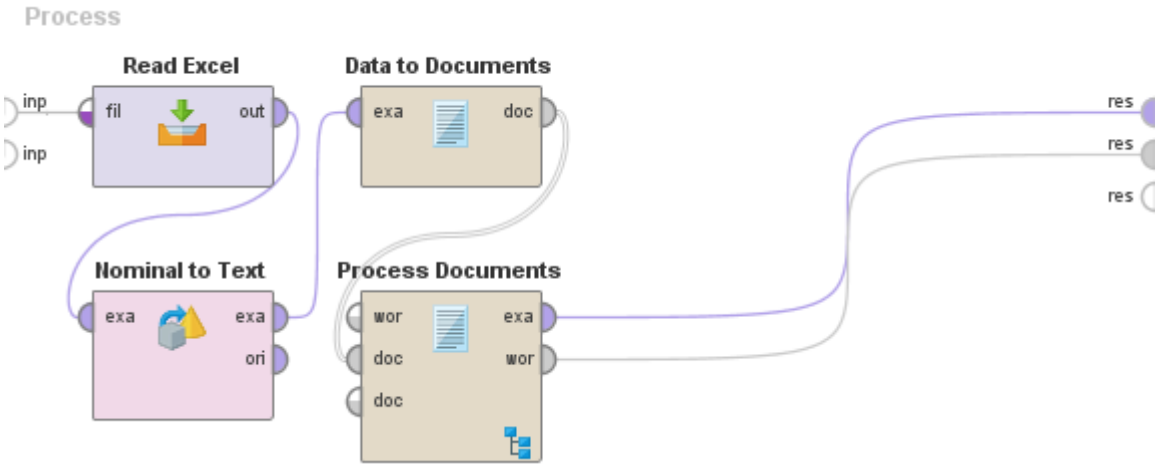
Izvor: Izrada autora u alatu RapidMiner Studio

Radi se o korisnoj tehnici pomoću koje donositelji odluka mogu procijeniti npr. uspješnost projekta, uspješnost njegove prakse ili neki drugi binarni događaj na temelju dostupnih vrijednosti nezavisnih varijabli. U tablici su prikazana predviđanja uspješnosti projektne prakse na temelju nekoliko nezavisnih varijabli kao što su broj partnerskih institucija, vrsta organizacije koordinatora, iznosa financijskih sredstava i sl.

6.4.4 Rudarenje teksta (eng. *Text Mining*)

Unutar analiziranog skupa podataka nalazi se stupac koji sadrži sažeti opis svakog realiziranog projekta. S obzirom da se radi o iznimno velikom broju projekata, postupak 'ručnog' analiziranja njihovog sadržaja zahtijevao bi znatnu količinu ljudskih i vremenskih resursa. Srećom moguće je koristiti tehnike rudarenja teksta kako bi se cjelokupni proces automatizirao.

Osnovni informacijski sadržaj do kojeg možemo doći primjenom tehnika rudarenja teksta jest učestalost pojave ključnih riječi u dokumentima. Kako bi ostvarili svoj cilj, potrebno je izgraditi model u alatu za rudarenje podataka.

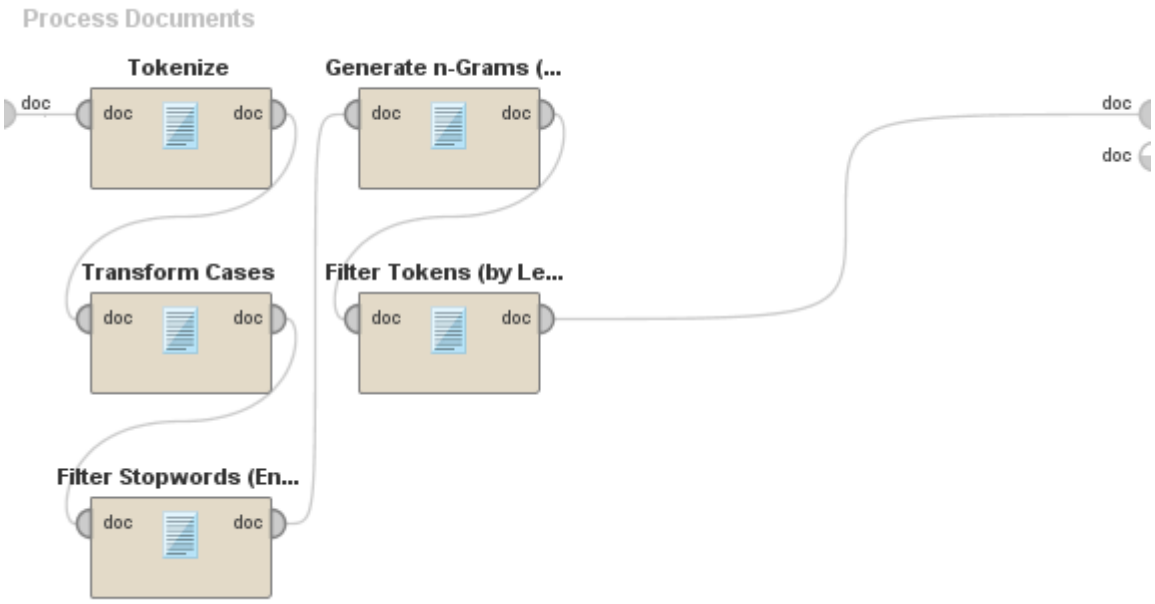


Slika 33 Text mining model – Ključne riječi

Izvor: izrada autora u alatu RapidMiner Studio

Model se sastoji od četiri operatora. Prvi operator *Read Excel* koristi se za učitavanje podataka iz .xls datoteke. Sljedeći u nizu je operator *Nominal to Text,* koji kao što mu samo ime kaže transformira vrstu podataka iz nominalne u tekstualnu. Treći operator, *Data to Documents*, koristi se za pretvaranje svake pojedine ćelije u zasebni dokument. Zadnji operator, *Process Documents,* koristi se za obradu dokumenata. Potrebno je istaknuti posljednji operator sam po sebi nije u mogućnosti obraditi dokumente, već on služi kao svojevrsni spremnik za operatore obrade teksta.

Kao što je moguće primijetiti na sljedećoj slici, unutar glavnog operatora nalazi se nekolicina međusobno povezanih operatora kojima se realizira obrada teksta.



Slika 34 Text mining model – Obrada dokumenata

Izvor: izrada autora u alatu RapidMiner Studio

Prvim operatorom se izvršava proces tokenizacije (eng. *Tokenize*). Time se niz nestrukturiranih tekstualnih podataka raščlanjuje u riječi ili druge smislene elemente koje nazivamo tokenima. Osnovni razlog upotrebe ovog operatora je identifikacija smislenih ključnih riječi.

S obzirom da računalo pravi razliku između istih riječi napisanih velikim i malim slovima, potrebno je upotrijebiti operator *Transform Cases*. Ovim operatorom su sva slova sadržana unutar dokumenata pretvorena u mala.

U svakom jeziku postoje riječi koje se učestalo koriste kao što su prijedlozi, članci i zamjenice. Prisutnost tih riječi u dokumentima ne samo da je redundantna, već može i otežati provođenje analize teksta. Stoga je potrebno iskoristiti operator *Filter Stopwords* koji će automatski izvršiti postupak isključivanja tih riječi iz analize.

Osim samih ključnih riječi, za analizu može biti zanimljivo koje se od njih često pojavljuju zajedno. Korištenjem *Generate n-Gram* operatora dobiti će se parovi riječi koje se koriste zajedno. Zadnjim operatorom se filtriraju tokeni koji sadrže samo jedno slovo.

Pokretanjem prethodno prikazanog modela rezultira tablicom u kojoj se nalaze ključne riječi, ukupan broj njihovih pojava, kao i broj dokumenata u kojima se pojavljuju. Unutar ove tablice prikazan je samo manji broj ključnih riječi, sortiranih prema ukupnom broju

pojavljivanja. Moguće je primijetiti da su unutar ključnih riječi zalutale određene besmislene riječi. S obzirom da se i radi o relativno nepreglednom prikazu rezultata, taj problem će biti otklonjen prilikom vizualizacije rezultata.

Word	Total Occurences	Document Occurences
x_d	4399	254
project	1903	462
d_x	1530	186
d_x_d	1530	186
x_d_x	1530	186
school	1025	196
students	963	165
people	792	208
youth	763	162
education	749	327
european	708	234
training	703	208
skills	695	244
work	674	242

Slika 35 Frekvencije ključnih riječi

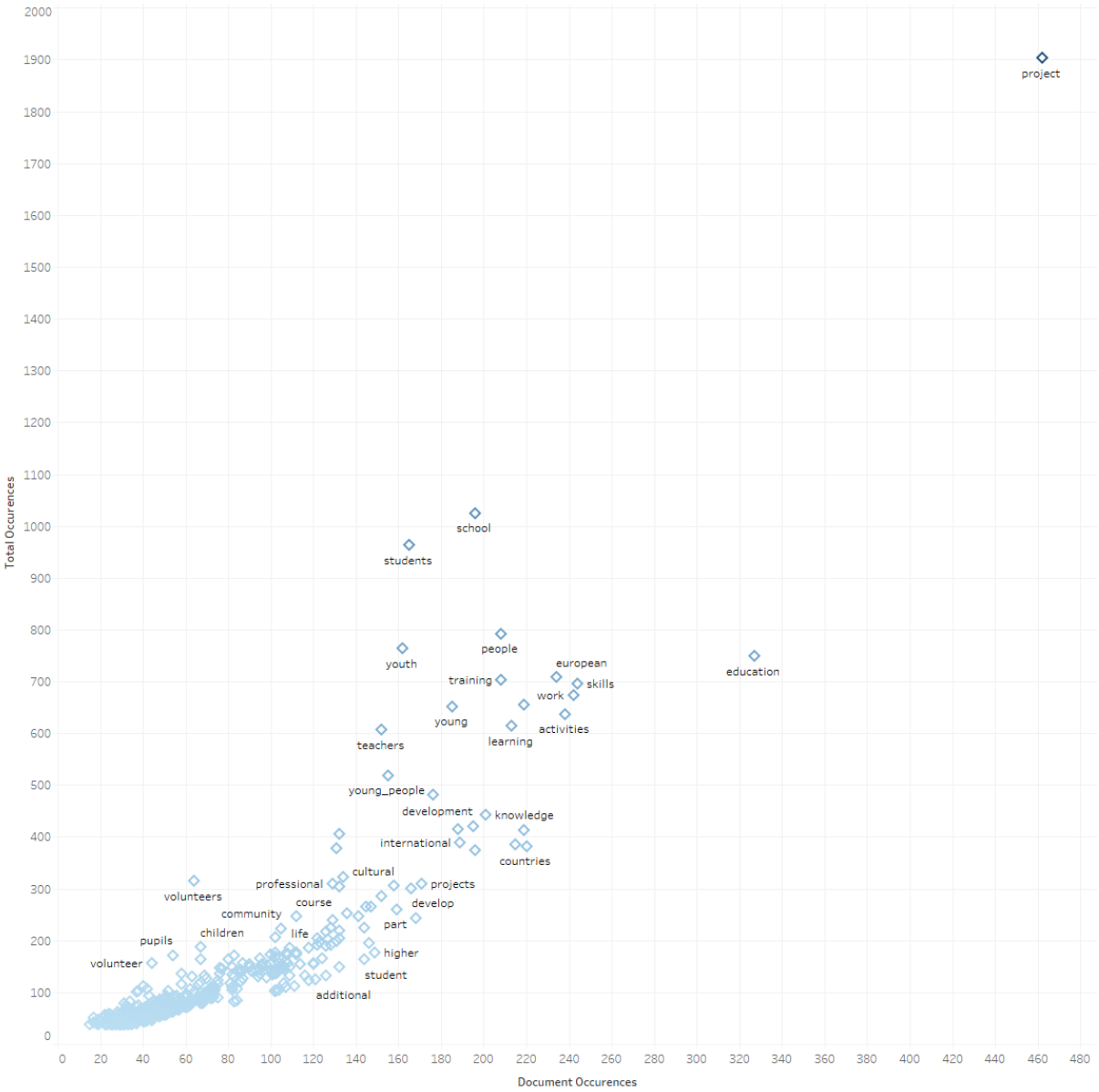
Izvor: izrada autora u alatu RapidMiner Studio

Dobiveni rezultati biti će eksportirani u Excel datoteku, kako bi se nad njima mogla vršiti vizualizacija u alatu Tableau. Jedan od popularnijih načina prikaza rezultata rudarenja teksta je tzv. *Word Cloud*. Radi se o prikazu sastavljenom od riječi korištenih unutar odabranih dokumenata, u kome veličina riječi označuje njenu frekvenciju ili važnost. Moguće je primijetiti da se najčešće pojavljuju riječi kao što su *project, participants, students, european, young_people* i sl. Iako je moguće brzo identificirati riječi koje se najčešće pojavljuju, nije moguće utvrditi koliko se puta pojavljuju, niti unutar koliko dokumenata se pojavljuju.

Slika 36 Word Cloud

Izvor: izrada autora u alatu Tableau

S obzirom na to potrebno je upotrijebiti drugačiji način vizualizacije podataka, kao što je npr. dijagram rasipanja. Na apscisi dijagrama rasipanja nalazi se broj dokumenata u kojima se riječ pojavljivala, dok se na ordinati nalazi ukupan broj pojavljivanja. Moguće je primijetiti da najveći *outlier* predstavlja riječ projekt, koju je moguće pronaći ukupno 1903 puta unutar 462 različita dokumenta. Značajan udio pojavljivanja imaju i riječi kao što su škola, obrazovanje, europsko, vještine, aktivnosti, ljudi, trening, studenti i slično.

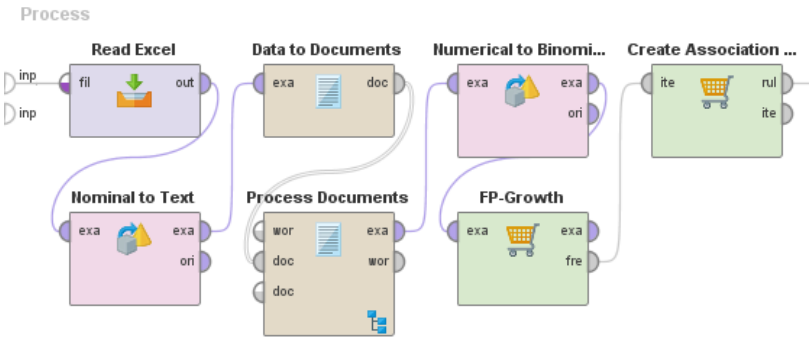


Slika 37 Dijagram rasipanja ključnih riječi

Izvor: izrada autora u alatu Tableau

6.4.5 Asocijacijska pravila

Pravila asocijacije koriste *if* / *then* izjave kako bi otkrili veze između naizgled nepovezanih podataka. Primjer pravila asocijacije bi bio: „Ako kupac kupi jaje, 80% je vjerojatno da će također kupiti i mlijeko“[93]. Kako bi se utvrdila pravila asocijacije između ključnih riječi koje se nalaze unutar dokumenata, potrebno je proširiti model za *FP-Growth i Create Association Rules* operatore.



Slika 38 Text mining model – Asocijacijska pravila

Izvor: izrada autora u alatu RapidMiner Studio

S obzirom na iznimno veliku količinu generiranih pravila, u tablici je prikazano samo njih par

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
1	study	education	0.080	0.800	0.982	-0.120	0.014	1.221	1.723
2	team	skills	0.144	0.800	0.969	-0.216	0.056	1.636	2.555
3	improving	skills	0.120	0.800	0.974	-0.180	0.047	1.636	2.555
4	study	skills	0.080	0.800	0.982	-0.120	0.031	1.636	2.555
5	outcomes	skills	0.072	0.800	0.983	-0.108	0.028	1.636	2.555
6	approach	european	0.088	0.800	0.980	-0.132	0.036	1.706	2.655
7	participant	participants	0.056	0.800	0.987	-0.084	0.025	1.823	2.806
8	basis	experience	0.064	0.800	0.985	-0.096	0.033	2.037	3.036
9	project, vocational	education	0.104	0.800	0.977	-0.156	0.019	1.221	1.723
10	project, weeks	education	0.056	0.800	0.987	-0.084	0.010	1.221	1.723
11	project, enable	skills	0.072	0.800	0.983	-0.108	0.028	1.636	2.555
12	project, potential	skills	0.064	0.800	0.985	-0.096	0.025	1.636	2.555
13	project, successful	work	0.056	0.800	0.987	-0.084	0.022	1.650	2.575
14	project, weeks	work	0.056	0.800	0.987	-0.084	0.022	1.650	2.575
15	project, daily	activities	0.088	0.800	0.980	-0.132	0.036	1.677	2.615

Slika 39 Asocijacijska pravila

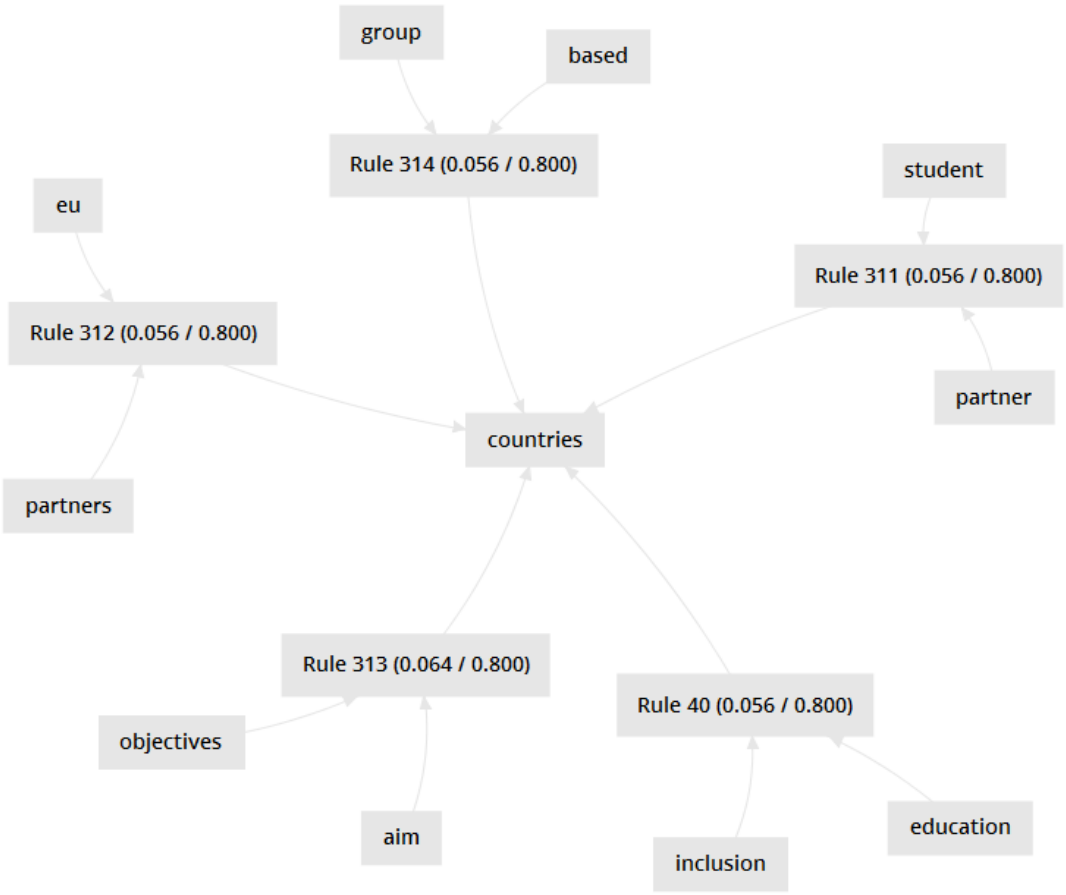
Izvor: izrada autora u alatu RapidMiner Studio

U prvom od sveukupno deset stupaca nalazi se broj asocijacijskog pravila. U drugom stupcu nalazi se ključna riječ prethodnik (eng. *antecedent*). Ona predstavlja riječ koja se nalazi u

[93] RapidMiner: Create Association Rules, [Internet], raspoloživo na: https://docs.rapidminer.com/studio/operators/modeling/associations/create_association_rules.html, [16.08.2017.]

dokumentu. U trećem stupcu nalaze se ključne riječi zaključci (eng. *conclusion*). Zaključci predstavljaju riječi ili skup riječi koje se pronalaze u kombinaciji s prethodnikom. Zadnji bitni stupci su *support* i *confidence*. *Support* je pokazatelj kojim ukazuje koliko često se predmeti pojavljuju zajedno, dok *confidence* ukazuje na to koliki je broj puta *if then* izjava bila istinita.

Grafički je moguće prikazati pravila asocijacije koja se vežu uz određenog prethodnika. Tako se na sljedećoj slici može pronaći vizualni prikaz asocijacijskih pravila vezanih uz riječ zemlje (eng. *countries*).

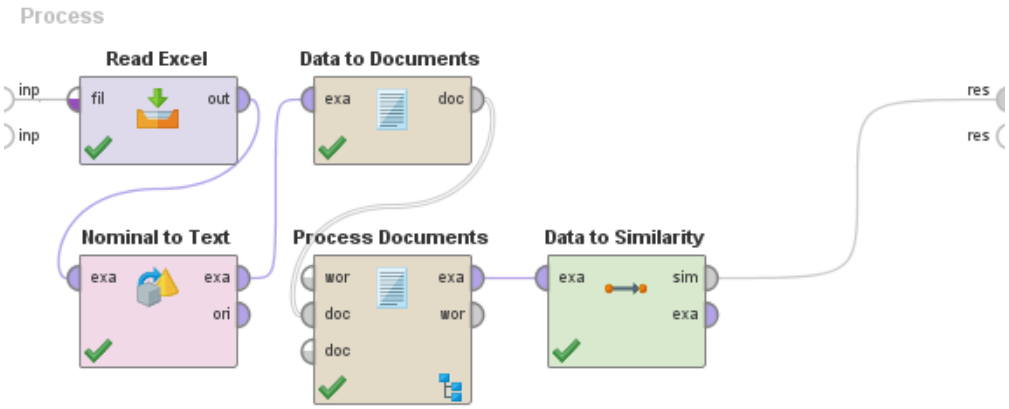


Slika 40 Vizualni prikaz asocijacijskih pravila za ključnu riječ *countries*

Izvor: izrada autora u alatu RapidMiner Studio

6.4.6 Sličnost između dokumenata

Ukoliko donositelji odluka žele utvrditi u koliko se mjeri opis određenog projekta poklapa s drugim, moguće je primjenom jednostavnog operatora doći do željenih informacija. U ovom slučaju su operatori iz prethodnog dijela zamijenjeni s operatorom *Data to Similarity*.



Slika 41 Text mining model – Sličnost

Izvor: izrada autora u alatu RapidMiner Studio

Nakon što se navedeni model pokrene, dolazi se do rezultata prikazanih u sljedećoj tablici.

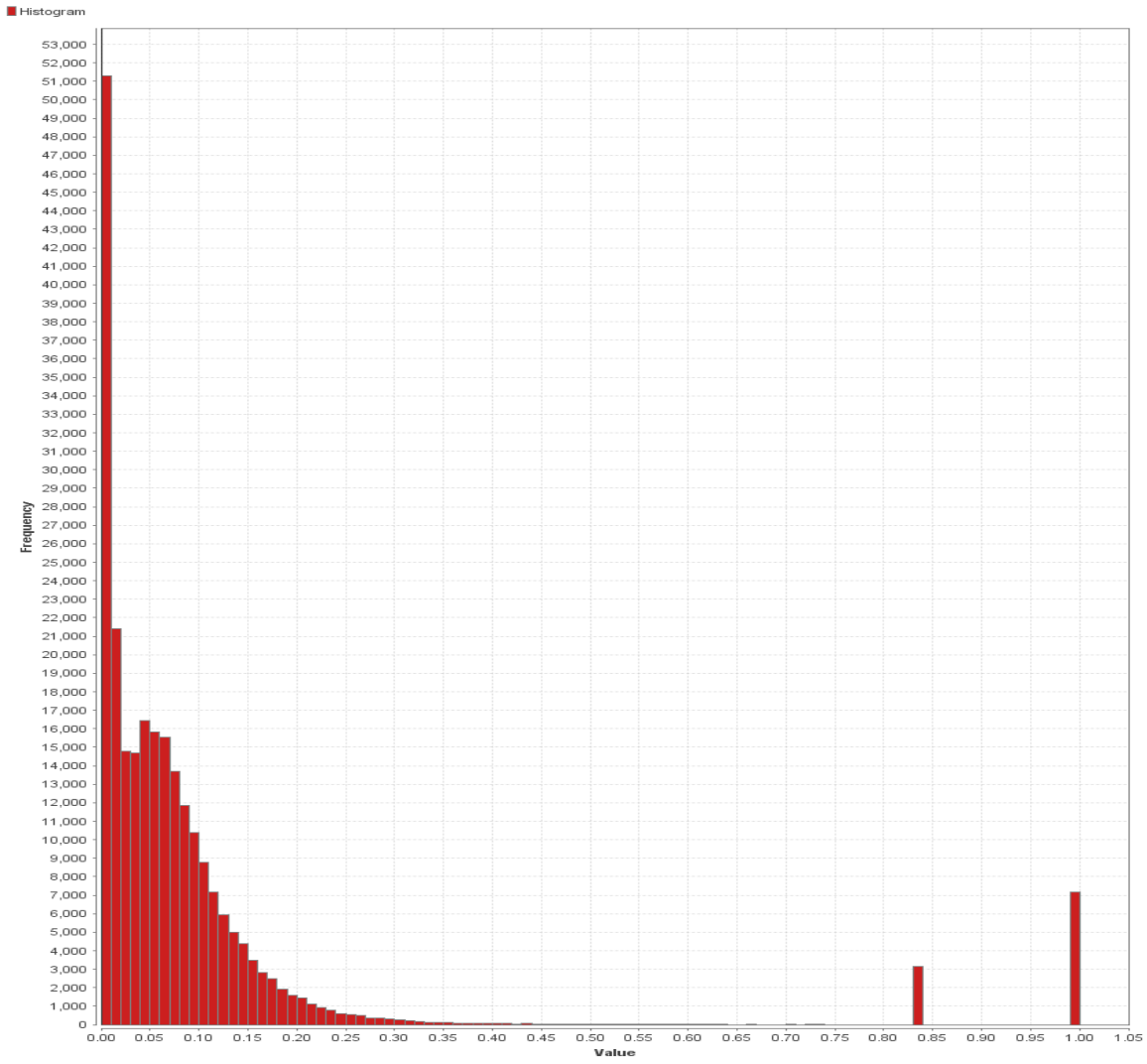
First	Second	Similarity
1.0	2.0	0.005
1.0	3.0	0.085
1.0	4.0	0.141
1.0	5.0	0.150
1.0	6.0	0.129
1.0	7.0	0.118
1.0	8.0	0.195
1.0	9.0	0.218
1.0	10.0	0.005
1.0	11.0	0.020
1.0	12.0	0.043
1.0	13.0	0.048

Slika 42 Tablični prikaz sličnosti između dokumenata

Izvor: izrada autora u alatu RapidMiner Studio

Potrebno je naglasiti da se ovdje nalazi samo dio rezultata, s obzirom da se radi o iznimno velikom broju kombinacija. U prvom i drugom stupcu nalaze se redoslijedni broj stupaca koji se uspoređuju, dok se u trećem stupcu nalazi postotak sličnosti između njih. Tako npr. prvi redak govori da je postotak sličnosti između prvog i drugog dokumenta svega 0.05%. S

obzirom da tablica sadrži nekoliko desetaka tisuća kombinacija, mogućnost njene interpretacije jest ograničena. Kako bi se dobila bolja slika o sveukupnoj situaciji, upotrijebit će se histogram. Moguće je primijetiti da većina parova dokumenata (51000) ima sličnost 0%, odnosno da se radi o manje više jedinstvenim opisima projekata. Broj parova teži prema nuli s porastom postotka sličnosti oko 40%. Interesantno je da postoji oko 7000 parova dokumenata koji su savršene kopije (100% sličnost). Pronalaskom dokumenata kojima sličnost iznosi 100% ukazuje na to da se radi o projektima kojima nisu detaljno definirani opisi već sadrže samo: „*This is a higher education student and staff mobility project, please consult the website of the organisation to obtain additional details.*“.



Slika 43 Histogram sličnosti dokumenata

Izvor: izrada autora u alatu RapidMiner Studio

7. ZAKLJUČAK

Primjenom iznesenih teorijskih koncepata analize i rudarenja podataka nad odabranim podskupu podataka o Erasmus+ programima dolazi se do određenog broja korisnih informacija. Za dolazak do željenih rezultata labavo se slijedila CRISP-DM metodologija.

Istraživačkom ili eksplorativnom analizom odabranih podataka prezentirane su najčešće karakteristike projekata. Započevši od trenutačnog statusa moguće je primijetiti da projekti u tijeku čine veći dio ukupnih projekata (58.83%), dok ostatak otpada na završene projekte. Unutar završenih projekata donesen je zaključak da se većina projekata (99.85%) smatra neuspješnima. Kredibilitet ovog podatka je upitan, te postoji mogućnost da osobe zadužene za izradu statistika ne raspolažu potpunim informacijama o ovom pitanju. Malo optimističniji podatak govori da se iskazana praksa provođenja projekata većim dijelom smatra dobrom (85.27%).

Gledajući s aspekta ključnih aktivnosti, moguće je primijetiti da u ukupnom broju projekata najveći udio otpada na projekte učenja kroz mobilnost (85.57%). Prosječna financijska vrijednost dodijeljena tim projektima iznosi 70 504 €. Drugi veliki udio u ukupnom broju projekata otpada na projekte suradnje za inovacije i razmjenu dobre prakse (10.96%). Iako se radi o znatno manjem broju projekata, prosječna financijska sredstva koja su im dodijeljena iznose 210 809 €. Na ostale tri ključne aktivnosti otpada manje od 5% ukupnih financijskih sredstava i broja projekata.

Projekti su u najvećoj mjeri koordinirani od strane zemalja kao što su Njemačka (11.86%), Španjolska (11.21%), Francuska (9.37%), Poljska (8.95%), Turska (5.51%) i Italija (5.51%). Slična situacija je i kod ukupne količine prikupljenih sredstava gdje uz male razlike u rangovima dominiraju iste zemlje. Projekti koje su koordinirali hrvatske organizacije iznose svega 1.46% ukupnih projekata, dok njihova ukupna financijska sredstva iznose oko 55 mil. €. Osim najuspješnijih zemalja koordinatora, analizom su utvrđene i najuspješnije organizacije, kao njihova vrsta.

Što se tiče partnerstava, moguće je primijetiti da je većina projekata (36.80%) realizirana samostalno od strane jedne zemlje. Jednog partnera imalo je 16.80% projekata, dok dva partnera 8.17% projekata. Povećanjem broja partnera broj realiziranih projekata teži nuli. Najčešća su partnerstva između Poljske i Njemačke (0.69%), Njemačke i Ujedinjenog Kraljevstva (0.60%), Španjolske i Italije (0.39%), Poljske i Španjolske (0.38%) i Poljske i Italije (0.30%).

Nakon izvršene eksplorativne analize prelazi se na primjenu tehnika rudarenja podataka nad odabranim podskupom. Prva tehnika koja se pritom koristila je klasična korelacijska matrica. Rezultati upućuju na nepostojanje statistički značajne povezanosti između devet odabranih varijabli. Tako se npr. može zaključiti da vrsta ključne aktivnosti i financijska sredstva nisu povezana.

Razlog ovakvih rezultata leži u činjenici da korelacija ne daje najbolje rezultate kada se kombiniraju numeričke i nominalne varijable. Stoga se u sljedećem koraku koristi analiza varijance kojom se ispituju određene informacije dobivene kod eksplorativne analize. Zaključci do kojih se dolazi navode da je djelovanje ključnih aktivnosti i zemlje koordinatora na dodijeljena financijska sredstva statistički značajno.

S obzirom da se unutar dostupnih podataka o projektima nalazi kratki opis svakog, provodi se i rudarenje teksta. Primjenom tehnika rudarenja teksta došlo se do interesantnih informacija o ključnim riječima koje se najčešće pojavljuju, unutar koliko dokumenata se pojavljuju, te koje se riječi često pojavljuju zajedno. Provedena je i analiza sličnosti dokumenata kojom se došlo do zaključka da je većima projektnih opisa unikatna, ali i da postoji određeni broj (7000 parova) dokumenata koji koriste predefinirani opis.

Prezentirani rezultati prikazuju kako se na relativno jednostavan način iz dostupnih podataka može doći do korisnih skrivenih informacija. Pri tom je potrebno paziti da su dostupni podaci prigodni za primjenu metoda, kako ne bi postojala sumnja u kredibilitet stvorenih informacija.

LITERATURA:

1. Aggarwal, C.C. (2015): Data Mining: The Textbook, Springer, London.
2. Analytics Vidhya (2016): A Complete Tutorial on Tree Based Modeling from Scratch, [Internet], raspoloživo na: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one , [16.07.2017.]
3. Booz Allen Hamilton (2013): The Field Guide to Data Science, [Internet], raspoloživo na: https://www.boozallen.com/s/insight/publication/field-guide-to-data-science.html , [19.05.2017.]
4. Bramer, M. (2016): Principles of Data Mining, Springer-Verlag London Ltd., London.
5. Brown, M.S. (2014): Data Mining for Dummies, John Wiley & Sons, Inc., New Jersey.
6. Brown, M.S. (2014): Data Understanding, [Internet], raspoloživo na: http://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/ ,[17.07.2017.]
7. Brown, M.S. (2014): Modeling, [Internet], raspoloživo na:http://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/, [17.07.2017.]
8. Caffo, B. (2015): Regression Models for Data Science, [Internet], raspoloživo na: http://leanpub.com/regmods , [15.05.2017.]
9. Caffo, B. (2016): Statistical inference for data science, [Internet], raspoloživo na: http://leanpub.com/LittleInferenceBook , [19.05.2017.]
10. Carnegie Mellon University: Exploratory Data Analysis,[Internet],raspoloživo na: http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf ,[16.07.2017.]
11. Chambers, M. et al. (2017): Breaking Data Science Open: How Open Data Science s Eating the World, [Internet], raspoloživo na: http://go.continuum.io/download-ebook-breaking-data-science-open/ , [19.05.2017.]
12. Cielen, D. et al. (2016): Introducing Data Science, Manning Publications Co., New York.
13. Doyle, M. (2014): What is the Difference Between Data and Information?,[Internet], raspoloživo na http://www.business2community.com/strategy/difference-data-information-0967136#EO5oZjHX874qQ3ZD.97 ,[16.07.2017.]

14. Engineering Statistics Handbook (2013): What is EDA?,[Internet],raspoloživo na: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm , [16.07.2017.]
15. Fayyad, U. et al. (1996): From Data Mining to Knowledge Discovery in Databases, [Internet], raspoloživo na: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131 , [15.05.2017.].
16. Gartner (2017): Magic Quadrant for Business Intelligence and Analytics Platforms, [Internet], raspoloživo na: https://www.gartner.com/doc/reprints?id=1-3TYE0CD&ct=170221&st=sb ,[24.08.2017.]
17. Hardin, M. et al. (2017): Which chart or graph is right for you?, [Internet], raspoloživo na: https://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf,[16.07.2017.]
18. Hastie, T. et al. (2008): Elements of Statistical Learning, [Internet], raspoloživo na: https://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf , [19.05.2017.]
19. Hero, A. (2013): Correlation Mining in Massive Data, [Internet], raspoloživo na: http://www.eecs.umich.edu/eecs/pdfs/events/2711.pdf ,[16.07.2017.]
20. IBM (2011): IBM SPSS Modeler CRISP-DM Guide,[Internet],raspoloživo na: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf ,[17.07.2017.]
21. IBM (2017). What is big data? Internet. Raspoloživo na: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
22. Investopedia (2017): Data Mining, [Internet], raspoloživo na: http://www.investopedia.com/terms/d/datamining.asp ,[15.05.2017]
23. Laerd (2013): Histograms, [Internet],raspoloživo na:https://statistics.laerd.com/statistical-guides/understanding-histograms.php ,[16.07.2017.]
24. Leek, J. (2015): The Elements of Data Analytic Style, [Internet], raspoloživo na: http://leanpub.com/datastyle ,[19.05.2017]
25. Marban, O. et al. (2009): A Data Mining & Knowledge Discovery Process Model, [Internet], raspoloživo na: http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf , [19.05.2017.]

26. Martin A. et al. (2011): Better Decision Making with Proper Business Intelligence, [Internet], raspoloživo na: https://www.atkearney.com/documents/10192/247903/Better_Decision_Making_with_Proper_Business_Intelligence.pdf/e55e6880-ed1b-4b25-a0b6-33b94c0cc641, [16.06.2017.]
27. Mayo, M. (2016): Data Science Basics: Data Mining vs. Statistics, [Internet], raspoloživo na: http://www.kdnuggets.com/2016/09/data-science-basics-data-mining-statistics.html, [17.07.2017.]
28. North, M. (2012): Data Mining for the Masses, [Internet], raspoloživo na: https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf , [15.05.2017.]
29. Olsen, D. L. I Dursun, D. (2008): Advanced Data Mining Techniques, Springer.
30. Oracle (2015): What is data mining: [Internet], Raspčoloživo na: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON046 [15.05.2017.]
31. Pejić, M. (2005): Rudarenje podataka u bankarstvu, Sveučilište u Zagrebu, Ekonomski fakultet.
32. Peng, R.D. (2016): Exploratory Data Analysis with R, [Internet], raspoloživo na: http://leanpub.com/exdata , [19.05.2017.]
33. Pierson, L. (2017): Data Science for Dummies, John Wiley & Sons, Inc, New Jersey.
34. Provost, F. & Fawcett, T. (2013): Data Science for Business: What You Need to KnowAbout Data Mining and Data-Analytic Thinking, O'Reilly Media Inc., Sebastopol.
35. Purplemath (2017): Stem-and-Leaf Plots, [Internet], raspoloživo na: http://www.purplemath.com/modules/stemleaf.htm ,[16.07.2017.]
36. RapidMiner: Create Association Rules, [Internet], raspoloživo na: https://docs.rapidminer.com/studio/operators/modeling/associations/create_association_rules.html, [16.08.2017.]
37. Ratner, B. (2011): Statistical and Machine-Learning Data Mining, Taylor & Francis Group, Boca Raton.
38. Rouse, M. (2013): Text Mining, [Internet], raspoloživo na: http://searchbusinessanalytics.techtarget.com/definition/text-mining ,[16.07.2017.]
39. Rouse, M. (2014): Association Rules, [Internet], raspoloživo na: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining, [16.07.2017.]

40. Rouse, M. (2014): Relationship Marketing, [Internet], raspoloživo na: http://searchcrm.techtarget.com/definition/relationship-marketing .[17.07.2017.]
41. Rozga, A.(2009): Statistika za ekonomiste, Ekonomski fakultet, Split.
42. Sammut, C. & Webb, G. I. (2017): Encyclopedia of Machine Learning and Data Mining, [Internet], raspoloživo na: https://link.springer.com/referencework/10.1007%2F978-1-4899-7502-7 , [19.05.2017.]
43. SAS (2015): Basic Concepts in Research and Data Analysis, [Internet], raspoloživo na: https://support.sas.com/publishing/pubcat/chaps/59814.pdf , [19.05.2017.]
44. SAS (2016): Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities, [Internet], raspoloživo na: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf ,[15.05.2017.]
45. Scheps, S. (2008): Business Intelligence for Dummies, Wiley Publishing Inc., Indianapolis.
46. Singh, N. et al. (2012): Data Mining with Regression Technique, Journal of Information Systems and Communication, Volume 3, Issue 1, str. 200.
47. Skupina autora (2011): Višedimenzijski informacijski sustavi, Ekonomski fakultet, Split str. 162
48. SmartVision (2016): What is the CRISP-DM methodology?,[Internet], raspoloživo na: http://www.sv-europe.com/crisp-dm-methodology/#businessunderstanding ,[17.07.2017.]
49. Stanton, J. (2012): An Introduction to Data Science, [Internet], raspoloživo na: https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf, [19.05.2017.]
50. Statistics How To (2013): What is a Bar chart?,[Internet], raspoloživo na: http://www.statisticshowto.com/what-is-a-bar-chart/ ,[16.07.2017.]
51. StatisticsSolutions (2017): What is Logistic Regression?,[Internet], raspoloživo na: http://www.statisticssolutions.com/what-is-logistic-regression/, [20.08.2017.]
52. StatSoft (2016): Fraud Detection, [Internet], raspoloživo na: http://www.statsoft.com/Textbook/Fraud-Detection ,[17.07.2017.]
53. Sveučilište u Zagrebu (2016): Erasmus+:Opće Informacije , [Internet], raspoloživo na: www.unizg.hr/suradnja/medunarodna-suradnja/partnerstva/erasmus/ , [15.05.2017.]

54. TechTarget (2014): Business Intelligence, [Internet], raspoloživo na: http://searchdatamanagement.techtarget.com/definition/business-intelligence ,[16.07.2017.]
55. Tuffery, S. (2011): Data Mining and Statistics for Decision Making, John Wiley & Sons, Ltd., Chichester.
56. University of Minnessota Twin Cities: Cluster Analysis: Basic Concepts and Algorithms, [Internet], raspoloživo na: https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
57. University of Wisconsin-Madison: A basic introduction to Neural Networks, [Internet], raspoloživo na http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html ,[16.07.2017.]
58. Vaughan J. (2017): Data, [Internet], raspoloživo na: http://searchdatamanagement.techtarget.com/definition/data ,[16.07.2017.]
59. Vercellis, C. (2009): Business Intelligence: Data mining and Optimization for Decision Making, John Wiley & Sons, Ltd., Chichester.
60. Witten, I. H. et al. (2011): Data mining: Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington.
61. Zaiane, O.R. (): Introduction to Data Mining, [Internet], raspoloživo na: https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ , [19.05.2017.]

POPIS SLIKA I TABLICA:

SAŽETAK

Suvremena poduzeća traže način na koji će korištenjem ogromnih količina dostupnih podataka steći konkurentske prednosti. Jedan od načina za ostvarenje tog cilja je korištenje tehnika rudarenja podataka. Aktivnosti rudarenja podataka (eng. *Data Mining*) predstavljaju iterativni proces usmjeren prema analizi velikih količina podataka s ciljem izvlačenja korisnih informacija i znanja, koji se mogu pokazati korisnim za rješavanje problema i donošenje odluka. Pri tom se koristi širokom lepezom alata kao što su asocijacije, grupiranja i stabla odlučivanja, kako bi postiglo svoje ciljeve.

Prema CRISP-DM metodologiji, uobičajeni proces rudarenja podataka sastoji se od šest iterativnih koraka. Prvi korak je razumijevanje poslovanja gdje je cilj shvatiti što se želi rudarenjem podataka postići iz poslovne perspektive. Druga faza, razumijevanje podataka, odnosi se na prikupljanje i opće sagledavanje podataka. Treća faza, priprema podataka, svodi se na pretvaranje podataka u oblik pogodan za analizu i rudarenje. Četvrta faza, modeliranje, sastoji se od odabira tehnika rudarenja, stvaranje modela rudarenja podataka i pokretanja tog modela. Peta faza, procjena, utvrđuje se stupanj zadovoljavanja poslovnih ciljeva, Posljednja faza, implementacija, identificira način na koji će se stvoreni rezultati koristiti.

Ti su teorijski koncepti praktično primijenjeni na primjeru skupa podataka o općim karakteristikama Erasmus+ projekata. Provođenjem eksplorativne analize na primjeru dolazi se do rezultata o učestalim karakteristikama projekata, dok se dubljim rudarenjem podataka otkrivaju skriveni obrasci u podacima.

Ključne riječi: Informacije, rudarenje podataka, analiza, CRISP-DM, Erasmus+

SUMMARY

Modern corporations seek for a way to use huge amounts of available data to accrue competitive advantages. One way of doing it is by using data mining. The term *data mining* refers to the overall process consisting of data gathering and analysis, development of inductive learning models and adoption of practical decisions and consequent actions bade on the knowledge acquired. It uses wide array of tools such as associations, clustering and decision trees, to achieve its goals.

Usual data mining process according to the CRISP-DM methodology consists of six iterative phases. First phase is business understanding where the goal is to understand what's expected to accomplish from a business perspective. Second phase, data understanding, requires acquiring the data listed in the project resources and analyse it. Third phase, data preparation, is where the decided data is being prepared for analysis. Fourth phase, modelling, consists of choosing modelling techniques, creating a data mining model, and running that model. Fifth phase, evaluation, assesses the degree to which the model meets business objectives. Last phase, deployment, identifies the way to use the produced results.

Those theoretical concepts can be practically applied on data set consisting of general information about Erasmus+ projects. Making the exploratory analysis on example set results with common characteristics of projects, while making a deeper, data mining analysis results with uncovering the hidden patterns in data.

Keywords: Information, data mining, analysis, CRISP-DM, Erasmus+