Primjena rudarenja podataka na otvorenim podacima

Muše, Nikola

Undergraduate thesis / Završni rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Organization and Informatics / Sveučilište u Zagrebu, Fakultet organizacije i informatike**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:211:002187

Rights / Prava: Attribution-NonCommercial-NoDerivs 3.0 Unported/Imenovanje-Nekomercijalno-Bez prerada 3.0

Download date / Datum preuzimanja: **2024-03-29**

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Repository / Repozitorij:

Faculty of Organization and Informatics - Digital Repository

zir.nsk.hr

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Nikola Muše

PRIMJENA RUDARENJA PODATAKA NA OTVORENIM PODACIMA

ZAVRŠNI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Nikola Muše, 46192/17-R

Studij: Informacijski sustavi

PRIMJENA RUDARENJA PODATAKA NA OTVORENIM PODACIMA

ZAVRŠNI RAD

Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, rujan 2020.

Nikola Muše

Izjava o izvornosti

Izjavljujem da je moj završni rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Rudarenje podataka je proces kojim se izvlači znanje iz velikog skupa podataka. Veoma je korisno i može se koristiti i gotovim svim područjima gospodarstva. Rudarenjem podataka u radu se analizira otvoreni skup podataka „World Happiness Report“ - istraživanje stanja globalne sreće u 156 država svijeta, koje je proveo UN. Deskriptivnim modeliranjem, odnosno klasterskom i faktorskom analizom grupiraju se slične karakteristike država kako bi se kreirale grupe uz reduciranje dimenzije podataka zbog lakše analize. Grupiranjem država po sličnim karakteristikama dobivamo zanimljive podatke o tome koji atributi u kolikoj mjeri utječu na mjeru sreće za određenu skupinu država. Prediktivnim modeliranjem nastoji se predvidjeti mjera sreće za pojedinu državu namještanjem vrijednosti pojedinih atributa koji u određenoj mjeri utječu na krajnji rezultat. Primijenjena su dva prediktivna modela – stablo odlučivanja i neuronska mreža. Stablo odlučivanja jasnim vizualnim pregledom daje uvid u grananja i vrijednosti atributa te predviđa konačni rezultat. Neuronska mreža omogućava manipulaciju vrijednostima atributa te time daje rezultat predviđanja.

Ključne riječi: rudarenje podataka, otvoreni podaci, modeliranje podataka, World Happiness Report, deskriptivno modeliranje, prediktivno modeliranje

Sadržaj

1. Uvod 1
2. Metodologija 3
2.1. CRISP-DM faze i zadaće 4
2.1.1. Razumijevanje poslovanja 4
2.1.2. Razumijevanje podataka 7
2.1.3. Priprema podataka 8
2.1.4. Modeliranje 11
2.1.5. Evaluacija 13
2.1.6. Dostavljanje 15
3. Opis problema 17
3.1. Slična istraživanja 18
4. Opis i priprema podataka 20
5. Modeliranje podataka 26
5.1. Deskriptivni modeli 27
5.1.1. Klasterska analiza 27
5.1.2. Faktorska analiza 33
5.2. Prediktivni modeli 36
5.2.1. Stablo odlučivanja 36
5.2.2. Neuronska mreža 41
6. Interpretacija i evaluacija modela 47
7. Zaključak 49
8. Popis literature 50
9. Popis tablica 52
10. Popis slika 53

1. Uvod

Za početak, treba objasniti što je rudarenje podataka. To možemo opisati kao proces pronalaženja uzoraka u velikim skupovima podataka, koji za krajnji cilj ima izvući informacije i pretvoriti ih u razumljivu strukturu za daljnju upotrebu te otkriti znanje. Rudarenjem podataka mogu se prepoznati trendovi i uzorci u velikim skupovima podataka koje čovjek čistom opservacijom teško ili gotovo nikako ne može zapaziti. Koristi se u raznim područjima, od kojih su među popularnijim; marketing i targeting (prodavači koriste rudarenje podataka da bi bolje razumjeli svoje kupce, a ono im omogućava da bolje segmentiraju tržišne skupine i prilagođavaju ponude kako bi što učinkovitije proveli svoje ponude i promocije različitim kupcima), određivanje rizika kredita (banke predviđaju sposobnost dužnika da uzme i vrati dug što im zbog korištenja raznih demografskih i osobnih podataka dobro polazi za rukom), otkrivanje i prevencija prijevara (financijske institucije koriste rudarenje podataka kod detektiranja i zaustavljanja sumnjivih transakcija), zdravstvena bioinformatika (predviđaju se vjerojatnosti da će pacijent imati određene zdravstvene probleme na temelju određenih faktora rizika, odnosno demografskih, obiteljskih i genetskih podataka te se pronalaze potencijalna rješenja koja bi mogla zaustaviti ili barem ublažiti tegobe koje bi ih mogle zateći) te filtriranje neželjene pošte (sustavi mogu analizirati zajedničke karakteristike milijuna zlonamjernih poruka te identificirati nove, a osim identifikacije, može čak i ukloniti te poruke prije nego što uopće stignu do korisnika primatelja).[1] Ukratko, možemo zaključiti da je rudarenje podataka uvelike doprinijelo napretku širokog spektra grana gospodarstva.

U završnom radu analizira se skup podataka o mjeri sreće u svijetu te provode deskriptivne i prediktivne metode pomoću alata BigML. Cilj rada je uočiti utjecaj pojedinih faktora na mjeru sreće stanovnika pojedine države i utvrditi načine na koje se ona može poboljšati. Također, predviđa se mjera sreće pojedine države temeljem raznih faktora i utvrđuje koji faktori u kojoj mjeri utječu na predviđanje.

Završni rad je podijeljen na 7 poglavlja. Prvo poglavlje predstavlja uvod u temu te predstavljanje teme. Drugo poglavlje odnosi se na CRISP-DM metodologiju rudarenja podataka gdje su detaljno opisane faze izvođenja. U trećem poglavlju je dan detaljniji opis problema te su opisana još 2 istraživanja koja su tematikom povezana sa ovim radom. Četvrto poglavlje je posvećeno opisu i pripremi podataka gdje su opisani atributi podataka, prikazana je tablica skupa podataka, navedeni su tipovi podataka, njihove vrijednosti i distribucije te povezanost nekoliko atributa skupa podataka. U petom poglavlju objašnjeno je deskriptivno i

prediktivno modeliranje te je dan opis metoda koje će se koristiti. Peto poglavlje se dijeli na 2 potpoglavlja – deskriptivo i prediktivno modeliranje gdje se deskriptivno modeliranje dijeli na potpoglavlja faktorske i klasterske analize, a prediktivno modeliranje na potpoglavlja stabla odlučivanja i neuronske mreže. U navedenim poglavljima provedene su analize na skupu podataka s grafičkim prikazima i objašnjenjima. Šesto poglavlje daje interpretaciju i evaluaciju modela. Posljednje se poglavlje odnosi na zaključak, a na kraju su popis literature, popis tablica i popis slika.

Skup podataka koji se koristio u završnom radu preuzet je sa stranice kaggle.com pod nazivom „*World Happiness Report*“ koji predstavlja godišnje izdanje Ujedinjenih naroda i sadrži rangove mjera sreće temeljenih na raznim faktorima koji utječu na sreću i percepciji sreće stanovnika pojedine države.

2. Metodologija

Rudarenje podataka se može koristi u raznoraznim područjima, a osobe koje koriste rudarenje podataka ne moraju biti eksperti u znanosti o podacima, statistici ili programiranju. Zahvaljujući grafičkim sučeljima koja su savršeno "skrojena" točno po njihovoj mjeri, to mogu biti ljudi iz raznih područja koja se uopće ne moraju doticati sa znanosti o podacima ili statistikom te im je zbog toga omogućeno da u velikoj količini podataka otkrivaju nove informacije koji bi im mogle biti ključne za daljnje poslovanje, bilo da se radi o smanjenju ili prevenciji grešaka ili napretku i poboljšanju njihovog posla. [2]

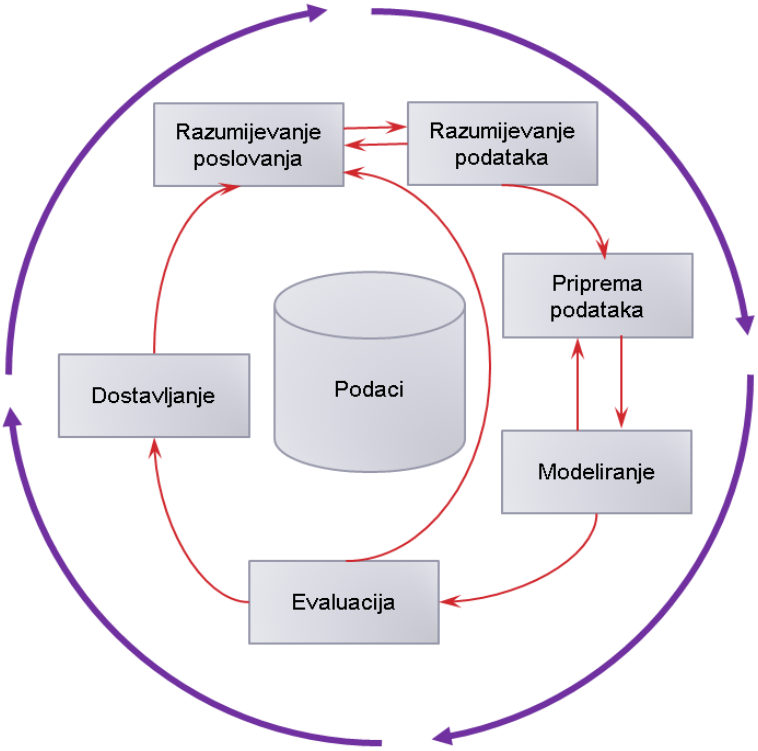
Jedan od najpopularnijih procesa rudarenja podataka je CRoss Industry Standard Process for Data Mining ili ukratko CRISP-DM. [3] Njegova popularnost korištenja je opravdana nekolicinom razloga ; osigurava opći okvir za smjernice i dokumentiranje iskustava, fleksibilan je tako da omogućuje uključivanje razlika za različite probleme poslovnog sektora i ustanova te omogućuje uključivanje razlika za različite podatke, pouzdan je i ponovljiv za ljude koji imaju malo znanja o rudarenju podataka.[3] Metodologija CRISP-DM omogućuje uporabu:

- standardnih procesa
- postojećih podataka
- softverskih tehnologija
- iskustva za pojedine situacije [3]

Osim toga, CRISP-DM nudi pomoć u planiranju i upravljanju projektima, pruža okvir za zapisivanje iskustva i omogućuje da se ponavljaju projekti te daje određenu „udobnost“ za nove korisnike jer reducira ovisnost o ljudima koji imaju veliko iskustvo i sposobnost rješavanja problema. [3] CRISP-DM je neutralan prema aplikacijama/industriji, a prednost je i to što je neutralan prema alatima.

2.1. CRISP-DM faze i zadaće

Na dijagramu na slici vidljivo je da se proces sastoji od 6 faza.



Slika 1: *CRISP-DM dijagram [3]*

2.1.1. Razumijevanje poslovanja

Ova početna faza se fokusira na razumijevanje ciljeva projekta i zahtjeva s poslovne perspektive, da se pretvori znanje u definiciju problema rudarenja podataka i stvori preliminarni plan dizajniran da postigne ciljeve. [3] Ovaj korak se dijeli na 4 „podkoraka“ koji se opet dijele na manje „podkorake“ :

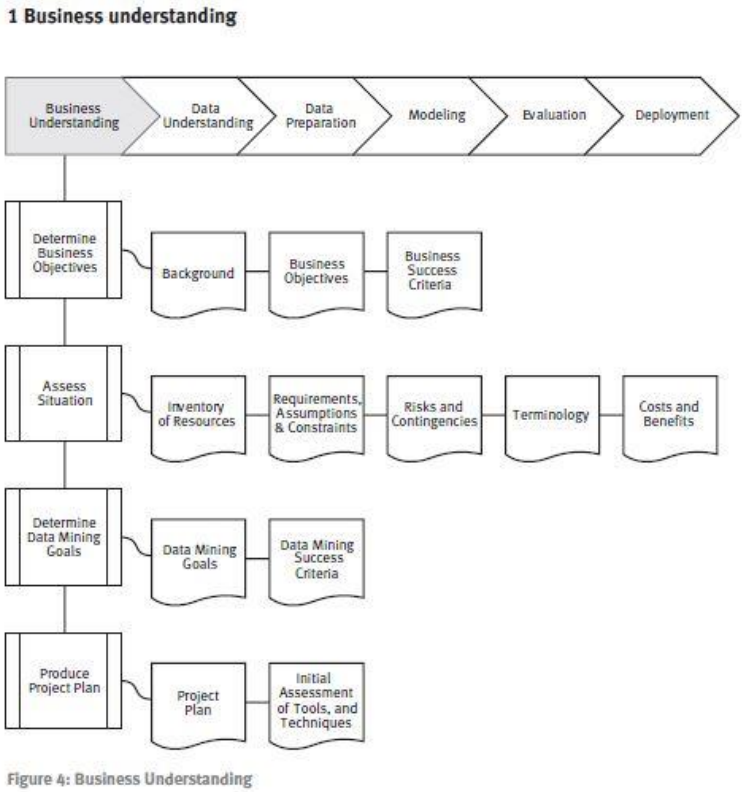
1) Određivanje poslovnih ciljeva - Prvo što treba je temeljito razumjeti (iz poslovne perspektive) što kupac stvarno želi. U projektima često ima mnogo zahtjeva i ograničenja koje se mora izbalansirati. Cilj je otkriti važne faktore koji u početku mogu utjecati na ishod projekta. Ovaj korak se ne smije zanemariti jer su moguće posljedice puno veći troškovi napora za traženje odgovora na pogrešna pitanja.
2) Zabilježiti sve poznate podatke o npr. Situaciji organizacije na početku projekta;
    - Pozadina - Opisati primarni cilj klijenta iz poslovne perspektive.

- Poslovni ciljevi - Kreirati plan za postizanje ciljeva rudarenja podataka odnosno postizanje poslovnih ciljeva. Plan bi trebao sadržavati korake koje treba izvesti tijekom ostatka projekta, uključujući odabir alata i tehnika
- Kriteriji poslovnih ciljeva - Opisati kriterije uspješnog/korisnog rezultata projekta (sa poslovnog stajališta). To može biti poprilično specifično i podložno mjerenju ili općenito. Općenite kriterije donosi netko što je onda subjektivno pa onda treba navesti tko je donio tu prosudbu.

3) Prosudba situacije – detaljnije utvrđivanje činjenica o svim resursima, ograničenjima, pretpostavkama i drugim čimbenicima koje bi trebalo u obzir pri određivanju cilja analize podataka i plana projekta. U ovom koraku bi trebalo proširiti prvi korak, „ući“ u detalje;
    - Pregled resursa - Navesti resurse koji su dostupni u projektu; osoblje, podatke, računalne resurse i softver
    - Zahtjevi, pretpostavke i ograničenja - Navesti sve zahtjeve projekta, uključujući raspored, razumljivost i kvalitetu rezultata, sigurnost te pravna pitanja. Treba navesti i pretpostavke koje je dao projekt ; to mogu biti pretpostavke koje se mogu provjeriti tijekom rudarenja podataka, ali mogu biti i neke „neprovjerljive“ pretpostavke o poslu o kojem se radi u projektu. Treba ispisati i ograničenja ; to mogu biti ograničenja poput dostupnosti resursa ili tehnološka ograničenja kao što je veličina skupa podataka koja je praktična za modeliranje
    - Rizici i mogući problemi – treba navesti rizike ili događaje koji bi mogli produljiti trajanje provođenja projekta ili ga čak skroz zaustaviti. Treba kreirati planove za slučaje da se dogodi neka krizna situacija i što će se učiniti ako se neki od problema ili neželjenih događaja dogode
    - Terminologija – sastaviti pojmovnik terminologije koja se odnosi na projekt. Izrađivanje navedenog pojmovnika je izvrstan način za prikupljanje znanja o poslovnim procesima za koji se projekt izvodi te za pojašnjavanje termina vezanih za samo rudarenje podataka
    - Troškovi i korist - kreiranje analize troškova - uspoređuje troškove projekta i potencijalne prihode ako je posao uspješan. Usporedba treba biti što specifičnija
4) Određivanje ciljeva rudarenja podataka;

- Cilj Data Mininga - navođenje ciljeva projekta u tehničkom pogledu, za razliku od poslovnih ciljeva koji navodi ciljeve u poslovnom pogledu, Primjer : poslovni cilj je povećati prodaju novina postojećim kupcima dok je cilj rudarenja podataka predvidjeti koliko će proizvoda kupiti kupac s obzirom na dane informacije o povijesti kupovine za zadnje tri godine, demografske podatke i cijenu proizvoda
- Kriteriji uspješnosti Data Mininga – npr. određena razina preciznosti predviđanja

5) Izrađivanje plana projekta – opisati plan za realizaciju rudarenja podataka i time postizanja poslovnih ciljeva. Plan bi trebao specificirati korake koji će se izvoditi tijekom projekta, alate i tehnike.
    - Projektni plan – navesti faze koje treba izvesti u projektu. Treba navesti i trajanje, potrebne resurse, ulaze i izlaze. Plan treba sadržavati detaljni plan za svaku fazu. Na kraju svake faze treba napraviti pregled napretka i postignuća, preporučeno je i ažurirati projektni plan.
    - Početni alati i tehnike – odabir alata za rudarenje podataka koje podržava razne metode za različite faze procesa. Važno je dobro procijeniti alate i tehnike rano u projektu jer odabir istih može bitno utjecati na cijeli projekt. [3][4]



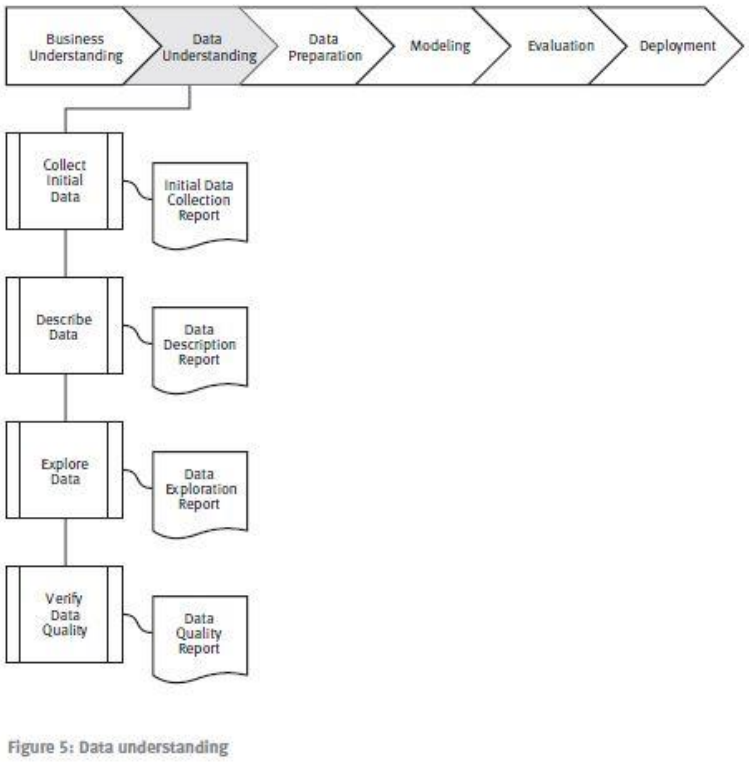
Slika 2: *Razumijevanja poslovanja [4]*

2.1.2. Razumijevanje podataka

Faza razumijevanja podataka počinje s prikupljanjem početnih podataka i nastavlja se s aktivnostima upoznavanja s podacima, da se identificiraju problemi kvalitete, da se otkrije početni uvid u podatke, ili da se prepoznaju zanimljivi podskupovi koji čine hipoteze u skrivene informacije. [3] Ovaj korak se dijeli na 4 „podkoraka“ koji se opet dijele na manje „podkorake“:

1) Prikupite početne podatke – nabavljaju se podatci navedeni u projektnim resursima, te se učitavaju u alate;
    - Navode se prikupljeni skupovi podataka zajedno sa njihovim lokacijama, metodama korištenim za njihovo prikupljanje, i problemima (ako postoje)
2) Opisivanje podataka – ispituju se svojstva prikupljenih podataka i radi se izvještaj
    - Izvješće o početnom skupu podataka – opisuju se podaci, uključujući format podataka, količinu podataka i sve druge osobine podataka koje su otkrivene. Treba procijeniti je li podaci zadovoljavaju zahtjeve projekta
3) Istraživanje podataka - Ovaj zadatak se bavi pitanjima rudarenja podataka koristeći vizualizaciju, ispitivanja i izvještavanja, koje se mogu izravno odnositi na ciljeve rudarenja podataka te također mogu doprinijeti ili detaljizirati opis podataka.
    - Izvješće o istraživanju podataka – opisuju se rezultati ovog koraka, uključujući prve zaključke i njihov utjecaj na ostatak projekta. Ako ima smisla, mogu se uključiti i grafovi kojima se naznačuju karakteristike podataka kojima se onda sugeriraju daljnja ispitivanja zanimljivih podskupova podataka
4) Provjera kvalitete podataka – ispituje se kvaliteta podataka; odgovara se na pitanja kao što je :
    - Jesu li podatci kompletni?
    - Jesu li podatci točni, sadrže li pogreške i koliko često ako sadrže?
    - Postoje li nedostajuće vrijednosti u podacima, gdje se nalaze i koliko su česte?

    Provjera kvalitete sadrži:
    - Izvješće o kvaliteti podataka – navode se rezultati provjere kvalitete podataka, ako problemi postoje, navesti potencijalna rješenja. Rješenja problema kvalitete podataka uglavnom jako ovise o poznavanju posla. [3][4]



Slika 3: *Razumijevanje podataka [4]*

2.1.3. Priprema podataka

Faza pripreme podataka obuhvaća sve aktivnosti da se konstruira konačni skup podataka (podaci koji će se ubaciti u alat za modeliranje) iz početnih sirovih podataka. Zadaće pripreme podataka će se vjerojatno provoditi više puta ponovljeno, i ne u nekom prethodno propisanom redoslijedu. Zadaci koji su tu uključeni su izbor tablica, slogova i atributa kao i transformacija i čišćenje podataka za alate za modeliranje. [3] Ovaj korak se dijeli na 5 „podkoraka“ koji se opet dijele na manje „podkorake“ :

1) Priprema podataka
    - Skup/ovi podataka – proizveden/i u fazi pripreme podataka, koristit će se za modeliranje
    - Opis skupa podataka – opisivanje skupova podataka koji će se koristiti za modeliranje
2) Izbor podataka – treba izabrati podatke koji će se koristiti za analizu. Kriteriji za izbor su: relevantnost za ciljeve rudarenja podataka, kvaliteta i tehnička ograničenja poput ograničenja količine podataka ili vrsta podataka. Potrebno je imati na umu da izbor podataka obuhvaća odabir atributa (stupaca) kao i odabir zapisa (redaka) u tablici.

- Razlozi za uključivanje/isključivanje – navesti podatke koji će biti uključeni/isključeni i razlog zašto

3) Čišćenje i formatiranje podataka – povećati kvalitetu podataka na razinu koja je zahtijevana u metodama analize. To može biti odabir kvalitetnih podskupina podataka, dodavanje „default“ ili zadanih vrijednosti ili procjena nedostajućih podataka modeliranjem
    - Izvješće o čišćenju podataka – opisivanje odluke i radnji poduzetih da bi se riješili problemi sa kvalitetom podataka prijavljeni tijekom koraka „Provjera kvalitete podataka“ u prethodnoj fazi („Razumijevanje podataka“).
    - Reformatiranje podataka (ako je potrebno) - Ukoliko je potrebno, treba transformirati podatke zbog mogućeg utjecaja na rezultate analize.
4) Konstruirajte podatke - uključuje konstruktivne radnje pripreme podataka poput kreiranja izvedenih atributa ili novih zapisa ili transformiranih vrijednosti za postojeće atribute.
    - Izvedeni atributi - novi atributi koji se grade od jednog ili više postojećih atributa u istom zapisu
    - Novi zapisi – opisivanje razloga kreiranja novih zapisa
5) Integriranje podataka – metode kojima se informacije kombiniraju iz više baza podataka, tablica ili zapisa kako bi se stvorili novi zapisi ili vrijednosti
    - Povezani podaci - spajanje dviju ili više tablica koje imaju različite informacije o istim objektima. Primjer: trgovački lanac ima jednu tablicu s informacijama o općim karakteristikama svake trgovine (npr. Površina, vrsta trgovačkog centra), drugu tablicu sa podacima o prodaji (npr. Dobit, postotna promjena prodaje u odnosu na prethodnu godinu) i treću s informacijama o demografiji okolice. Svaka od ovih tablica sadrži po jedan zapis za svaku trgovinu. Te se tablice mogu spojiti u novu tablicu s jednim zapisom za svaku trgovinu, kombinirajući polja iz izvornih tablica. Spojeni podaci također obuhvaćaju agregaciju. Agregacija se odnosi na operacije u kojima se nove vrijednosti izračunavaju sažimanjem podataka iz više zapisa i / ili tablica. Na primjer, pretvaranje tablice kupnji kupaca u kojoj postoji jedan zapis za svaku kupnju u novu tablicu u kojoj postoji jedan zapis za svakog kupca, s poljima kao što su broj kupnji, prosječni iznos kupnje, postotak plaćanja kreditnom karticom, postotak proizvoda koji su na akciji itd. [3][4]

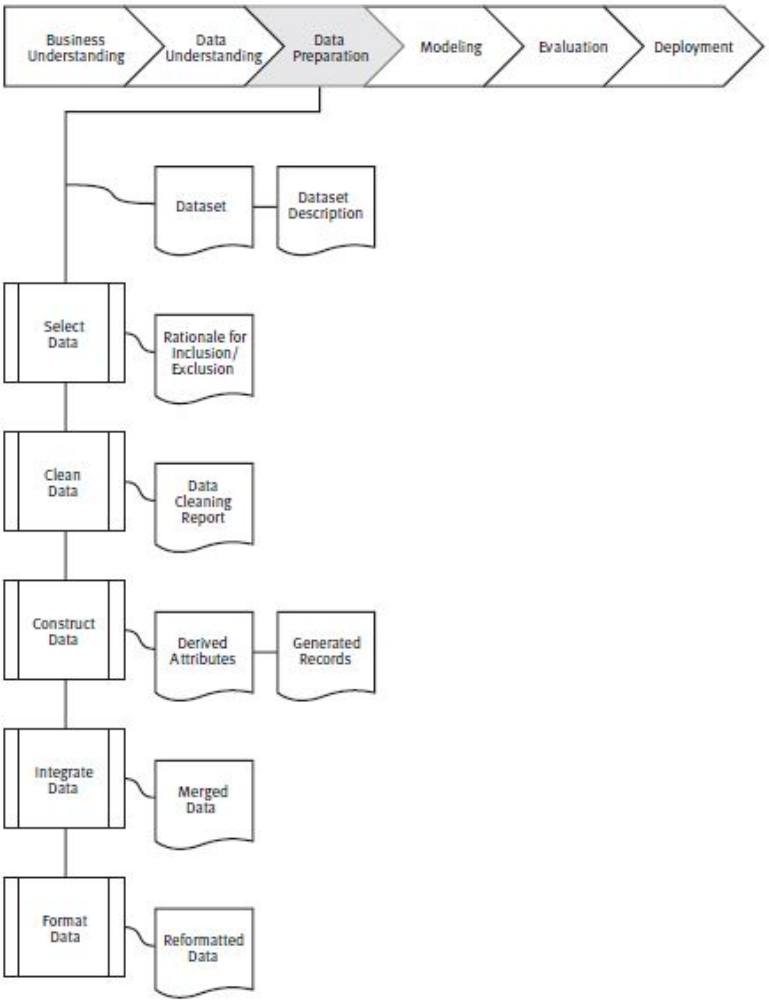


Figure 6: Data preparation

Slika 4: *Priprema podataka [4]*

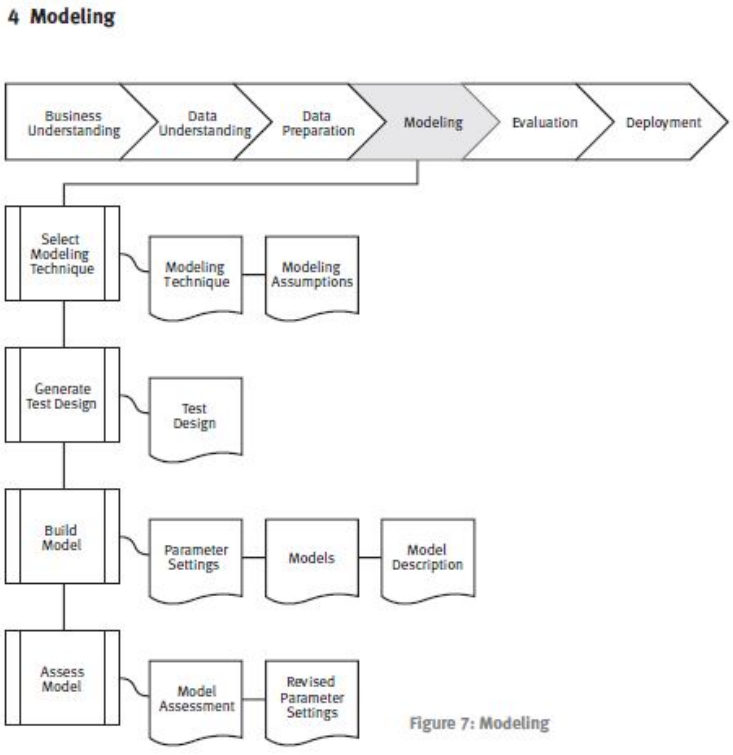
2.1.4. Modeliranje

Različite tehnike modeliranja se odabiru i primjenjuju, i njihovi parametri se kalibriraju za optimalne vrijednosti. Tipično, postoji nekoliko tehnika za isti tip problema rudarenja podataka. Zbog toga je često nužno (za različite tehnike), vratiti se natrag u fazu pripreme podataka. [3] Ovaj korak se dijeli na 4 „podkoraka“ koji se opet dijele na manje „podkorake“ :

1) Odaberite tehniku modeliranja - prvi korak u modeliranju je odabiranje tehnike modeliranja koja će se koristiti. Iako ste možda već odabrali alat tijekom faze razumijevanja poslovanja, ovaj se zadatak odnosi na određenu tehniku modeliranja. Ako se primjenjuje više tehnika, izvedite ovaj zadatak zasebno za svaku tehniku.
    - Tehnike modeliranja – dokumentiranje tehnike modeliranja koja će se koristiti
    - Pretpostavke modeliranja - mnoge tehnike modeliranja daju određene pretpostavke o podacima – npr. da svi atributi imaju jednolike distribucije, da nema vrijednosti koje nedostaju, atributi klase su simbolični itd. Sve takve pretpostavke treba zabilježiti.
2) Izradite probni dizajn - Prije nego što se model izgradi, mora se generirati procedura ili mehanizam za testiranje kvalitete i valjanosti modela. Na primjer, u nadziranim zadacima rudarenja podataka, poput klasifikacije, uobičajeno je koristiti stopu pogrešaka kao mjere kvalitete za modele rudarenja podataka. Zbog toga se skup podataka obično odvaja u skup za izradu i skup za testiranje, napravi se model na skupu za izradu i procjenjuje kvaliteta na zasebnom skupu za testiranje.
    - Testni dizajn – Opisuje se planirani plan za izradu, testiranje i ocjenu modela. Primarna komponenta plana je određivanje načina kako podijeliti skup podataka u skup podataka u skup za obuku, skup za testiranje i skup za validaciju odnosno procjenu kvalitete
3) Izgradite model – pokreće se alat za modeliranje na pripremljenom skupu podataka da bi stvorio modeli (ili više njih)
    - Postavljanje parametara – u puno slučajeva postoji velik broj parametara koji se mogu prilagoditi. Navode se parametri i njihove vrijednosti kao i obrazloženje za odabir postavki parametara
    - Modeli
    - Opisi modela – opisuju se napravljeni modeli. Potrebno je objasniti interpretaciju modela i eventualne poteškoće

4) Ocijenite model – podatkovni znanstvenik, stručnjak za podatke ili stručnjak za rudarenje podataka interpretira modele prema svom znanju, kriterijima uspješnosti rudarenja podataka i željenom dizajnu testa, zatim prosuđuje primjenu tehnika modeliranja i otkrivanja. Nakon toga kontaktira poslovnog analitičara i stručnjake za domenu kako bi razmotrili rezultate rudarenja podataka u poslovnom kontekstu. Stručnjak za rudarenje podatka zatim rangira modele procjenjujući po kriterijima ocjenjivanja. Koliko god je moguće, pokušava uzeti u obzir poslovne ciljeve. U većini projekata se primjenjuje jedna tehnika više od jednog puta ali se i generira rezultate rudarenja sa različitim tehnikama. Potrebno je i usporediti sve rezultate prema kriterijima ocjenjivanja.

- Ocjena modela – sažimaju se rezultati ovog koraka, navode se kvalitete generiranih modela (npr. U smislu točnosti) i rangira se kvaliteta jednih u odnosu na druge
- Postavke revidiranih parametara – revidiraju se postavke parametara prema procjeni modela i prilagođava ih se za sljedeće pokretanje u koraku izrade modela. Ponavlja se izrada i procjena modela sve dok se čvrsto vjeruje da su pronađeni najbolji model/i. Potrebno je dokumentirati sve takve revizije i procjene. [3][4]



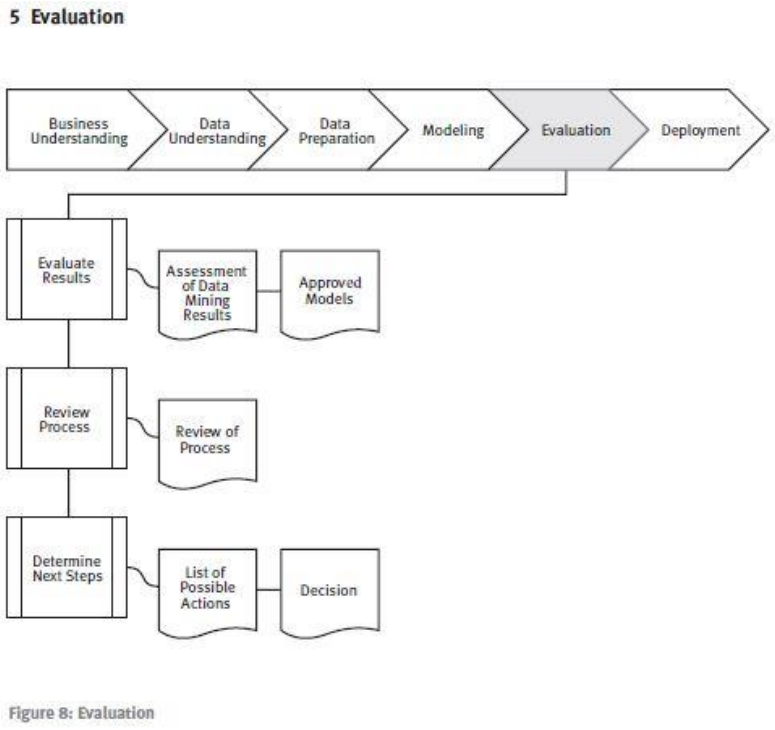
Slika 5: *Modeliranje [4]*

2.1.5. Evaluacija

U ovoj fazi projekta već se izgradio model (ili više njih) za koje se čini da ima visoku kvalitetu, s perspektive analize podataka. Prije nastavka na konačnu dostavu modela, važno je temeljitije vrednovati model, i pregledati sve korake koji su doveli do konstrukcije modela, da bude sigurno da taj model dobro udovoljava poslovnim ciljevima. Ključni cilj je odrediti da li postoje neki važni poslovni problemi koji još nisu razmatrani. Na kraju faze, treba se donijeti odluka o korištenju rezultata. [3] Ovaj korak se dijeli na 3 „podkoraka“ koji se opet dijele na manje „podkorake“ :

1) Evaluacija rezultata – prethodni koraci evaluacije su se bavili čimbenicima kao što je preciznost modela, dok ovaj korak procjenjuje u kojoj mjeri model ispunjava poslovne ciljeve i nastoji utvrditi postoji li neki poslovni razlog zašto ovaj model nije dobar. Druga opcija je testiranje modela na testnim aplikacijama u stvarnim problemima, ako vremenska i proračunska ograničenja dopuštaju. Štoviše, evaluacijom se također procjenjuju i drugi generirani rezultati rudarenja podataka. Rezultati rudarenja podataka uključuju modele koji su nužno povezani s izvornim poslovnim ciljevima i svim ostalim nalazima koji nisu nužno povezani sa izvornim poslovnim ciljevima, ali mogu otkriti i dodatne izazove, informacije ili savjete za buduće smjernice.
    - Ocjena rezultata rudarenja podataka u odnosu na kriterije poslovne uspješnosti – sažimaju se procjene u smislu kriterija poslovnog uspjeha uključujući konačnu izjavu o tome ispunjava li projekt inicijalne poslovne ciljeve
    - Provjereni modeli – nakon procjene modela s obzirom na kriterije poslovnog uspjeha, generirani modeli koji udovoljavaju odabranim kriterijima postaju provjereni modeli
2) Pregledati proces – već sada bi proizvedeni model(i) trebali biti zadovoljavajući odnosno zadovoljavati poslovne potrebe. Potrebno je napraviti temeljitiji pregled postoji li neki važni faktor ili zadatak koji je bio slučajno ispušten ili previđen za vrijeme rudarenja podataka.
    - Izvještaj o pregledu procesa – treba sumirati postupak i istaknuti aktivnosti koje su propuštene ili ako treba neke ponoviti
3) Odrediti sljedeće korake – ovisno o rezultatima procjene i pregledu procesa, projektni tim odlučuje kako nastaviti. Tim odlučuje treba li završiti projekt i prijeći na implementaciju, pokrenuti daljnje iteracije ili krenuti na nove projekte. Ovaj korak uključuje analiziranje preostalih resursa i budžet što može utjecati na daljnje odluke.

- Lista mogućih daljnjih akcija i odluka – navesti potencijalne daljnje radnje zajedno s razlozima za i protiv za svaku opciju. [3][4]



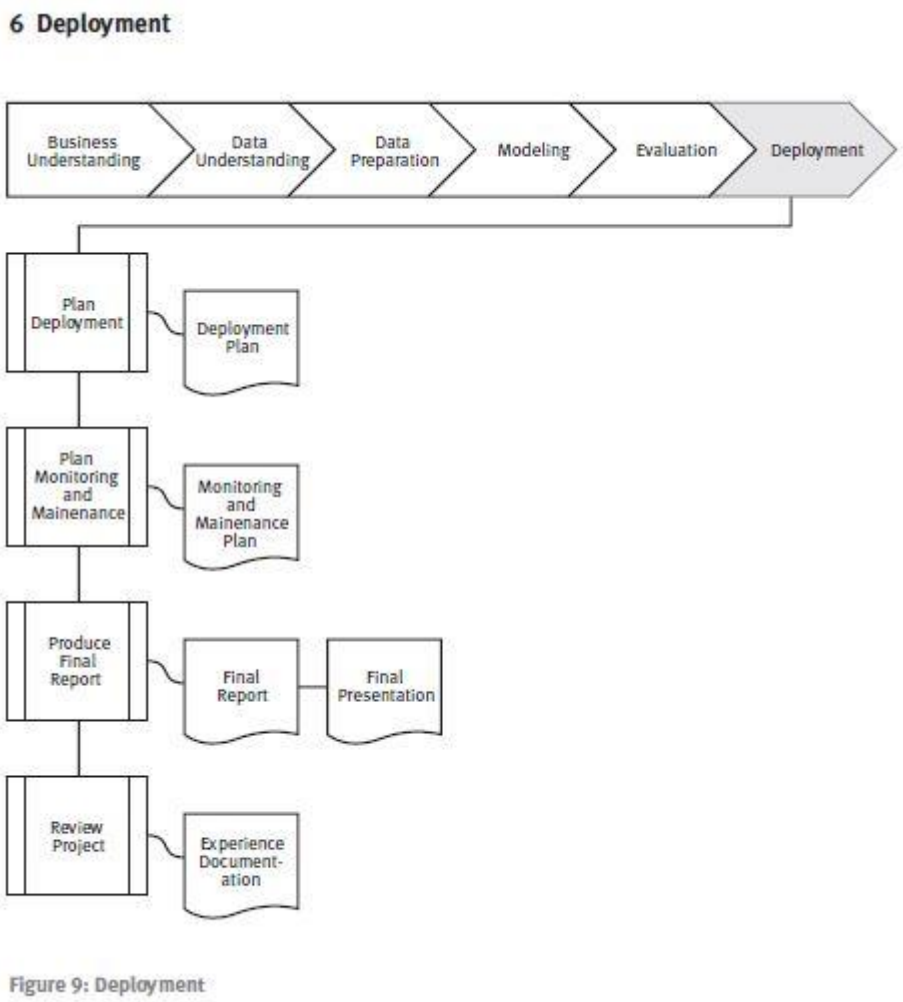
Slika 6: *Evaluacija [4]*

2.1.6. Dostavljanje

Stvaranje modela općenito nije kraj projekta. Čak i kada je svrha modela samo znanje, njega će trebati organizirati i prezentirati na način da ga mogu koristiti korisnici. Faza diseminacije može biti jednostavna (izvješća), ili kompleksnija (implementacija ponovljenog procesa rudarenja podataka). U mnogim slučajevima će korisnik, a ne analitičar podataka, biti taj koji će provoditi diseminaciju. U svakom slučaju će korisnik morati dobro razumjeti proces da stvarno osigura upotrebu kreiranih modela.

1) Planirati diseminaciju - Ovaj zadatak uzima rezultate evaluacije i određuje strategiju diseminacije. Ako je utvrđen opći postupak za stvaranje relevantnih modela, ovaj je postupak ovdje dokumentiran za kasniju primjenu.
    - Plan diseminacije – sumira se strategija diseminacije, uključuju se potrebni koraci i kako ih izvesti
2) Planirati nadzor i održavanje - nadzor i održavanje su važne stavke ako rezultat rudarenja podataka postane dio svakodnevnog poslovanja i njegovog okruženja. Pažljiva priprema strategije održavanja pomaže u izbjegavanju dugih razdoblja nepravilne upotrebe rezultata rudarenja podataka. U svrhu nadziranja diseminacije rezultata za rudarenja podataka, projekt treba detaljan plan nadziranja procesa koji uzima u obzir specifičnu vrstu diseminacije.
    - Plan nadzora i održavanja – sumira se strategija za nadzor i održavanje, uključujući potrebne korake i objašnjenje kako ih izvesti
3) Izraditi konačni izvještaj - na kraju projekta, projektni tim sastavlja završno izvješće. Ovisno o planu diseminacije, ovo izvješće može biti samo sažetak projekta i iskustava na projektu (ako do sada već nisu dokumentirani) ili može biti konačna i sveobuhvatna prezentacija rezultata rudarenja podataka.
    - Konačni izvještaj - konačno pisano izvješće o utjecaju rudarenja podataka. Uključuje sve prethodne rezultate, sažimanje i organiziranje rezultata.
    - Konačna prezentacija - često će se održati sastanak na kraju projekta na kojem se naručitelju prezentiraju rezultati
4) Pregledati projekt – procjena što je prošlo u redu, a što nije, što je dobro učinjeno i što se treba poboljšati
    - Dokumentacija iskustva - sažeti važna iskustva stečena tijekom projekta. To mogu biti zamke, varljivi pristupi ili savjeti za odabir najprikladnijih tehnika rudarenja podataka u sličnim situacijama. U idealnim projektima,

dokumentiranje iskustava također obuhvaća sva izvješća koja su napisali pojedini članovi projekta tijekom prethodnih faza projekta. [3][4]



Slika 7: *Diseminacija [4]*

3. Opis problema

Kao što je navedeno u uvodnom poglavlju, u završnom radu analizira se skup podataka o istraživanju mjere sreće pojedine države. Podaci koji se analiziraju prikupljeni su 2019. godine temeljem istraživanja koje je pokrenula organizacija UN. Rangovi mjere sreće temelje se na atributu „Ladder“ gdje su ispitanici dužni zamisliti ljestvicu gdje 10 označava najbolji mogući život za njih, a 0 najgori mogući. Zatim rangiraju svoj trenutni život brojem od 0 do 10. Osim opisanog atributa, na rezultate istraživanja utjecali su i ostali faktori u većoj ili manjoj mjeri. [5]

Istraživanje se provodi od 2012. godine, a podaci proizlaze iz 14 područja poput: poslovanja i ekonomije, građanskog angažmana, komunikacije i tehnologije, problemima različitosti, obrazovanja i obitelji, emocija, okoliša i energije, hrane i zaštite, vlade i politike, zakona, reda i sigurnosti, zdravlja, religije i etike, prijevoza i posla. [5]

Deskriptivnim modeliranjem, odnosno klasterskom i faktorskom analizom nastoje se grupirati slične karakteristike država kako bi se kreirale grupe uz reduciranje dimenzije podataka zbog lakše analize. Grupiranjem država po sličnim karakteristikama dobivamo zanimljive podatke o tome koji atributi u kolikoj mjeri utječu na mjeru sreće za određenu skupinu država.

Prediktivnim modeliranjem nastoji se predvidjeti mjera sreće za pojedinu državu namještanjem vrijednosti pojedinih atributa koji u određenoj mjeri utječu na krajnji rezultat. Primijenjena su dva prediktivna modela – stablo odlučivanja i neuronska mreža. Stablo odlučivanja jasnim vizualnim pregledom daje uvid u grananja i vrijednosti atributa te predviđa konačni rezultat. Neuronska mreža omogućava manipulaciju vrijednostima atributa te time daje rezultat predviđanja.

3.1. Slična istraživanja

Zbog boljeg razumijevanja i podrobnije analize ove teme dobro je proširiti ovaj projekt sa sličnim. U istraživanju „THE ASSOCIATION BETWEEN THE NATIONAL INTELLECTUAL CAPITAL COMPONENTS AND THE QUALITY OF LIFE“ koje su proveli Doc. dr. sc. Dijana Oreški, Doc. dr. sc. Igor Pihir i Doc. dr. sc. Irena Kedmenec, studija ispituje povezanost nacionalnog intelektualnog kapitala i kvalitete života u 24 članice Europske unije. Na prvom mjestu ukupne kvalitete života se nalazi Austrija, a na posljednjem Rumunjska. [6] Kvaliteta života pokazuje najveću pozitivnu korelaciju sa indeksom financijskog kapitala. [6] Ovim istraživanjem su doneseni bitni zaključci koji bi bitno mogli utjecati na kvalitetu života, uočene su komponente koje su već dovoljno razvijene te ne bi imale nekakav bitan utjecaj na ukupnu kvalitetu života, a uočene su i one komponente koje još uvijek imaju prostora za poboljšanje (pozitivno bi utjecale, poboljšale kvalitetu života). [6]

Što se tiče ljudskog kapitala, stopa pismenosti u EU je jako visoka u svim državama što znači da daljnje ulaganje u to područje neće imati neki značajan utjecaj na poboljšanje kvalitete života. [6] Komponente koje, s druge strane, imaju prostora za napredak su obuka zaposlenika i izdaci za obrazovanje. [6]

Ako pričamo o kapitalu marketa i procesnom kapitalu, najveća korelacija je zabilježena između varijable kvalitete života i varijabli slike odnosno ugleda zemlje, transparentnosti države, učinkovitosti vlade i dostupnosti kapitala. [6] U ovom završnom radu će također biti govora o korelacijama između komponenti koje potencijalno uzrokuju poboljšanje mjere sreće. [6] Zabilježeno je i da nema veze između kvalitete života i otvorenosti prema stranim kulturama i izvozu dobara. [6] Ista stvar je i sa brojem pretplatnika na telefonske usluge zbog toga što je postotak navedenog relativno visok u svim promatranim državama. [6]

Ono što je interesantno je to da su sve komponente „renewal“ kapitala pozitivno povezane sa kvalitetom života. [6] U to spadaju R&D aktivnosti (Research and Development - istraživanje i razvoj), što bi mogla biti „posljedica“ činjenice da države sa većim BDP-om po stanovniku imaju više resursa za R&D aktivnosti te isto tako države sa većim BDP-om po stanovniku imaju zadovoljnije državljane. [6]

U istraživanju „NATIONAL INTELLECTUAL CAPITAL OF EU-15 COUNTRIES FROM 1995 TO 2011“ koje su provele Doc. dr. sc. Dijana Oreški, i Doc. dr. sc. Irena Kedmenec, analiziraju se trendovi u nacionalnom intelektualnom kapitalu EU-15 između 1995. i 2011. Ovo istraživanje ima malo starije podatke od prethodnog, što će nam dobro doći jer će

se pokazati kako su se predikcije iz starijeg istraživanja potvrdile, odnosno pokazale točnim u novijem. U ovom se istraživanju također radi o nacionalnom intelektualnom kapitalu ali u kontekstu redukcije disproporcije u razvoju članica Europske unije (sa naglaskom na razvojne inpute, ne na sam razvoj).

S obzirom na rezultate istraživanja, autorice zaključuju da države EU-15 imaju visoku razinu ljudskog kapitala. [7] Međutim, da bi ljudski kapital stvorio dodanu vrijednost mora biti podržan od kapitala marketa i procesnog kapitala (u kojima je bilo dosta prostora za napredak). Što se tiče procesnog kapitala, zaključeno je kako bi države trebale raditi na učinkovitosti vlade, što su u novijem istraživanju i potvrdile kroz korelaciju kvalitete života i varijable učinkovitosti vlade. [7] Ako se priča o kapitalu marketa, autorice su zaključile da bi države trebale poraditi i na povećanju transparentnosti vladinih politika, što se ponovo potvrđuje u novom istraživanju (naime, u novijem radu je pokazano kako je korelacija između varijable kvalitete života i varijable transparentnosti vladinih djelovanja bitna, koeficijent korelacije između te dvije varijable je 0,719). [6] [7]

Nadalje, u ovom radu se navodi kako je „renewal“ kapital (koji obuhvaća istraživanja i razvoj) iznimno važan jer je usmjeren na stvaranje dugoročnih konkurentskih prednosti. Objašnjavaju kako je to važan faktor za jačanje ljudskog kapitala, što ima izravan utjecaj na financijski kapital. [7] To je prepoznala i Europska unija te je počela raditi na strategiji za stvaranje „innovation-friendly“ okruženja koje bi dovelo do ukupnog rasta. [7]

4. Opis i priprema podataka

U završnom radu analizira se skup podataka pod nazivom „*World Happiness Report*“, istraživanje stanja globalne sreće koje svrstava 156 zemalja svijeta prema mjeri sreće i zadovoljstva koju percipiraju njihovi građani. Podaci su prikupljeni tijekom 2019. godine te navode atribute koji u određenoj mjeri utječu na mjeru sreće pojedine države. Skup podataka preuzet je sa stranice kaggle.com, a u popisu literature naveden je i link na isti.

Otvoreni podaci su podaci kojima svatko može pristupiti, koristiti ih i dijeliti. [8] Glavni cilj istraživačkog rada [9] je otkriti glavne aspekte otvorenih podataka i kvalitete otvorenih podataka. Da bi to postigli, autori su definirani cilj razložili na sljedeća dva cilja; utvrditi aspekte koje treba uzeti u obzir prilikom definiranja otvorenih podataka i utvrditi pokazatelje kvalitete prema utvrđenim aspektima otvorenih podataka. Autori istraživačkog rada su utvrdili sljedeće aspekte otvorenih podataka; cijena(neznačajni trošak), dostupnost, višestruka iskoristivost, mogućnost redistribucije, ograničenja na autorska prava i patente, strukturiranost, zahtijevanje otvorene licence, nemogućnost identificiranja podataka bez prethodne de-identifikacije, kreirano za vrijeme poslovnih operacija, pripadanje poreznom obvezniku, mogućnost preuzimanja i jednostavnost ubacivanja u software. Autori su također utvrdili i sljedeće pokazatelje kvalitete otvorenih podataka: konzistentnost, kompletnost, točnost, jedinstvenost, dostupnost i vidljivost, iskoristivost i razumljivost, pravovremenost, vrijednost, detaljnost. [9]

Tablica 1 prikazuje popis podataka nekoliko država, njih 16 od 156 te rang za pojedini atribut. Primjerice, umjesto konkretne vrijednosti za GDP per capita za pojedinu državu naveden je rang gdje „1“ označava državu sa najvišim rangom za GDP per capita. Ukratko će se objasniti atributi skupa podataka.

Country (region) - naziv države

Ladder – Central Ladder je mjera zadovoljstva životom, mjera sreće. Mjeri zadovoljstvo života na način da najprije ispita kako stanovnici zamišljaju svoj život u najboljim mogućim uvjetima i opisuju svoje želje za budućnost.

SD of Ladder – standardna devijacija atributa Ladder

Positive affect – mjera pozitivnih utjecaja (sreća,smijeh) prethodnog dana

Negative affect – mjera negativnih utjecaja (stres,tuga) prethodnog dana

Social support – označava u kojoj mjeri je društvena potpora pridonijela mjeri sreće

Freedom – označava u kojoj je mjeri sloboda pridonijela mjeri sreće

Corruption – mjera u kojoj je percepcija korupcije pridonijela mjeri sreće

Generosity - mjera u kojoj je velikodušnost pridonijela mjeri sreće

Log of GDP per capita – mjera u kojoj je GDP per capita pridonio mjeri sreće

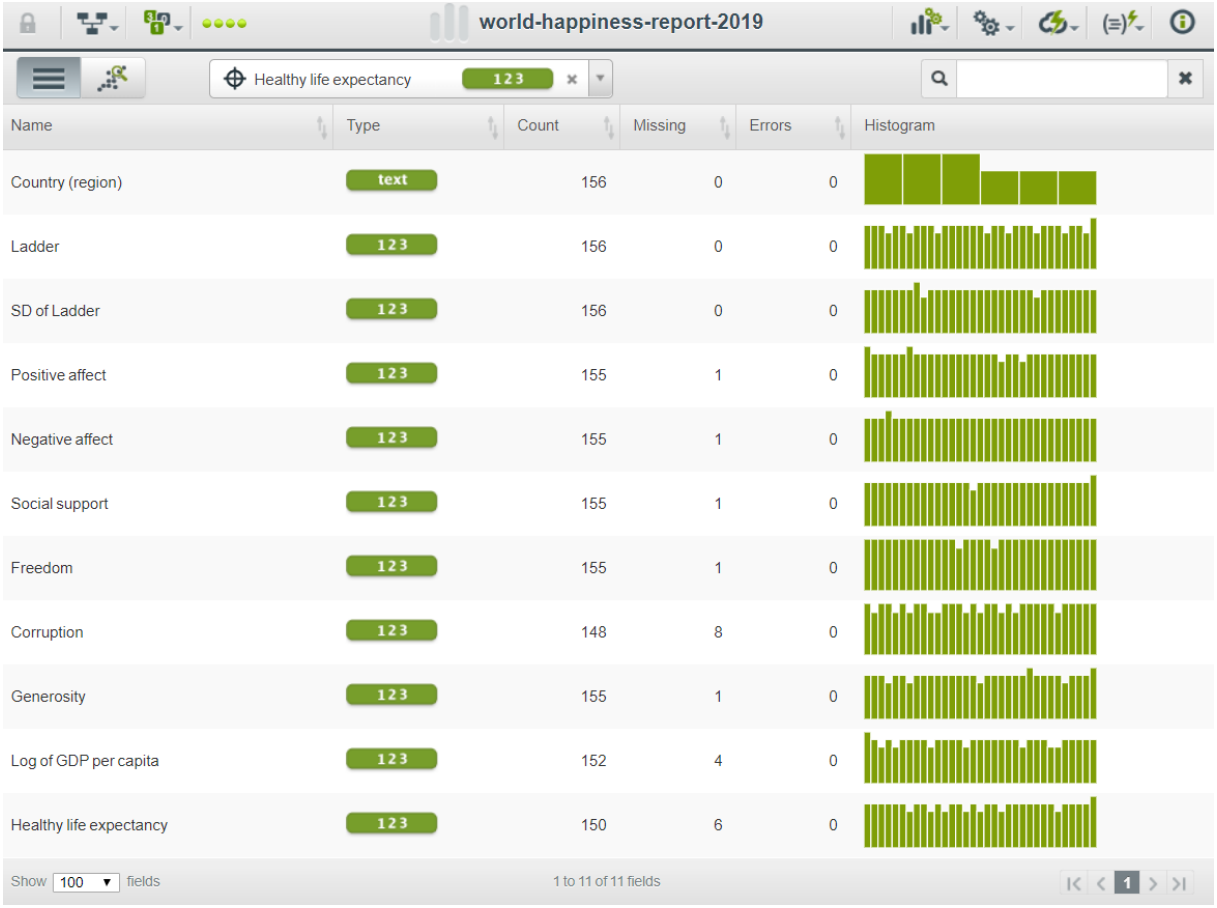
Healthy life expectancy – mjera u kojoj je očekivano trajanje života pridonijelo mjeri sreće

Tablica 1: *Popis podataka*

Country (region)	Ladder	SD of Ladder	Positive affect	Negative affect	Social support	Freedom	Corruption	Generosity	Log of GDP per capita	Healthy life expectancy
Finland	1	4	41	10	2	5	4	47	22	27
Denmark	2	13	24	26	4	6	3	22	14	23
Norway	3	8	16	29	3	3	8	11	7	12
Iceland	4	9	3	3	1	7	45	3	15	13
Netherlands	5	1	12	25	15	19	12	7	12	18
Switzerland	6	11	44	21	13	11	7	16	8	4
Sweden	7	18	34	8	25	10	6	17	13	17
New Zealand	8	15	22	12	5	8	5	8	26	14
Canada	9	23	18	49	20	9	11	14	19	8
Austria	10	10	64	24	31	26	19	25	16	15
Australia	11	26	47	37	7	17	13	6	18	10
Costa Rica	12	62	4	87	42	16	58	75	67	28
Israel	13	14	104	69	38	93	74	24	31	11
Luxembourg	14	3	62	19	27	28	9	30	2	16
United Kingdom	15	16	52	42	9	63	15	4	23	24
…										
South Sudan	156	140	127	152	148	154	61	85	140	143

Izvor [5]

U alatu BigML uvezeni su podaci te je kreiran skup podataka. Slika 8 prikazuje skup podataka u alatu sa navedenim vrijednostima, atributima, tipovima podataka, nedostajućim vrijednostima i prikazom distribucije.



Slika 8: *Skup podataka (Izvor: vlastita izrada)*

Skup podataka se sastoji od jednog tekstualnog atributa „Country“ (text) te 10 numeričkih atributa (123). Kako je ranije navedeno, skup podataka sadrži 156 država, odnosno instanci. Svaki atribut ima uniformnu distribuciju iz razloga što podaci prikazuju rangove, a ne konkretne vrijednosti. Nedostajućih vrijednosti u skupu podataka ima nekolicina, gdje ih najviše nedostaje za atribut „Corruption“.



Slika 9: *Korelacija između atributa „Ladder“ i „Healthy life expectancy“ (Izvor: vlastita izrada)*

Pomoću opcije „Scatterplot“ prikazana je povezanost nekoliko atributa skupa podataka. Promotrimo Pearson (r) i Spearman (p) vrijednosti. Ako je apsolutna vrijednost Pearson i Spearman bliža nuli, to je manja povezanost među podacima, a ako je bliža jedinici, to je povezanost među podacima veća. Uočavamo da je apsolutna vrijednost Pearson i Spearman bliža jedinici što označava da je povezanost među podacima veća ako promatramo atribute „Healthy life expectancy“ i „Ladder“.



Slika 10: *Korelacija između atributa „Corruption“ i „Healthy life expectancy“ (Izvor: vlastita izrada)*

Ovdje vidimo da je povezanost između atributa „Healthy life expectancy“ i atributa „Corruption“ slaba. Dakle, da se zaključiti da povezanost između korumpiranosti države i očekivanja zdravog života nije jaka. Na primjer, Italija je rangirana 128. po korupciji, a po očekivanju zdravog života je 7. S druge strane, Somalija je rangirana 16. po korupciji, ali po očekivanju zdravog života je 144. Stoga, nema smisla tražiti povezanost između ta dva atributa.

5. Modeliranje podataka

U ovom poglavlju opisane su metode za deskriptivno i prediktivno modeliranje podataka. Uz opis metoda, razrađene su i tehnike koristeći besplatnu platformu BigML za analizu, modeliranje i vizualizaciju odabranog skupa podataka.

Usporedba deskriptivnih i prediktivnih metoda rudarenja podataka [10]

Tablica 2: *Usporedba deskriptivnih i prediktivnih metoda rudarenja podataka*

Usporedba	Deskriptivno rudarenje podataka	Prediktivno rudarenje podataka
Opis	Utvrđuje što se dogodilo u prošlosti analizirajući skup podataka.	Opisuje što bi se moglo dogoditi u budućnosti analizirajući povijesne podatke.
Zahtjevi	Agregacija podataka i rudarenje podataka	Statistika i metode predviđanja
Tip pristupa	Reaktivno	Proaktivno
Preciznost	Pruža ispravne podatke	Kreira rezultate koji ne osiguravaju točnost
Metode praktične analize	Standardni izvještaji, upiti, ad-hoc izvještaji	Predviđanja, simulacije, upozorenja

Izvor [10]

U tablici 2 dana je usporedba značajki deskriptivnih i prediktivnih metoda rudarenja podataka kako bi se okvirno stekao dojam o razlikama u metodama. Obje metode su detaljnije objašnjenje u sljedećim poglavljima.

5.1. Deskriptivni modeli

Deskriptivna analiza ili statistika sažima „sirove“ podatke, opisuje karakteristike skupa podataka i prikazuje ih u obliku koji je razumljiv ljudima. [11] Deskriptivno rudarenje podataka se često koristi kako bi se prikazale veze i odnosi među podacima, uzorci među podacima, frekvencija podataka i sl. [10] Važna značajka deskriptivnog modeliranja podataka je i grupiranje podataka u skupine temeljem sličnih obilježja što olakšava razumijevanje podataka, kontrolu i sažimanje. [10] Provedbom deskriptivne analize skupa podataka, dobivaju se odgovori na pitanja: što se dogodilo, gdje je nastao problem, kolika je frekvencija problema. [10] U završnom radu su definirane i razrađene dvije tehnike deskriptivnog modeliranja podataka: klasterska analiza i faktorska analiza.

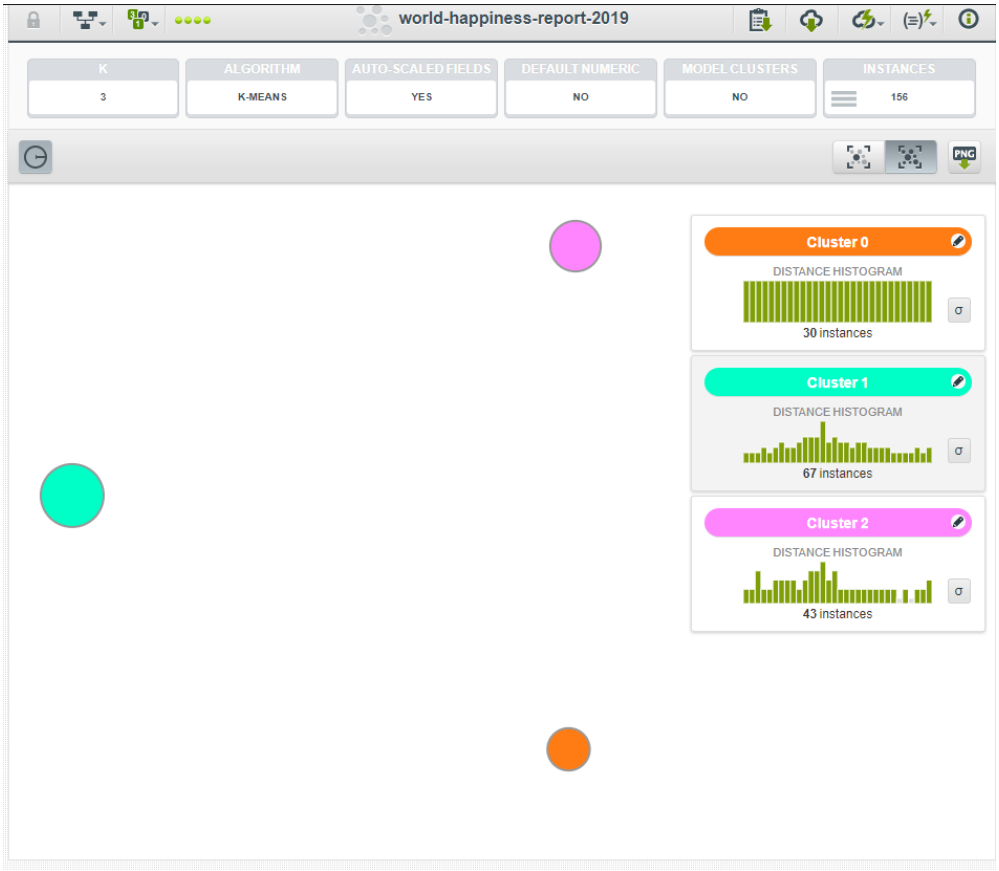
5.1.1. Klasterska analiza

Klasterska analiza ili klasterifikacija je postupak grupiranja elemenata u grupe odnosno klastere na način da su elementi unutar svakog klastera što je više moguće slični (udaljenosti unutar klastera se minimiziraju), a oni različitih klastera što različitiji (udaljenosti između klastera se maksimiziraju). [12] Klasterska analiza jedan je od glavnih zadataka rudarenja podataka te je česta tehnika statističke analize. [12] Koristi se u brojnim područjima, poput: financija, marketinga, analiziranja slika, bioinformatike, sažimanja podataka, računalne grafike i sl. [12]

Postavljaju se pitanja, koliko klastera odabrati, je li klasterska analiza dobra, koliko elemenata (instanci) treba sadržavati pojedini klaster, kako mjeriti je li jedan set klastera bolji od drugog i sl. Univerzalnog odgovora nema. Naime, provedbom klaster analize potrebno je odabrati različiti broj klastera i usporediti rezultate kako bi se utvrdila ispravnost modela. [13] Svaka mjera treba ispitati koncepte separacije i kohezije klastera. [13]

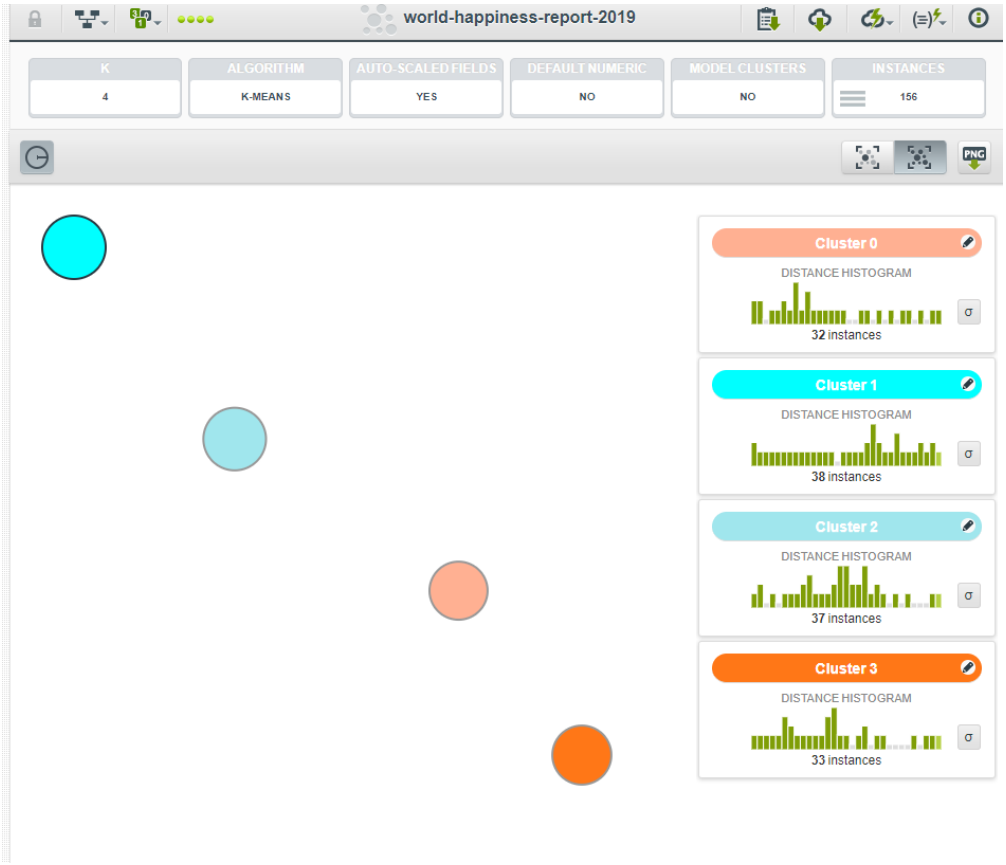
U alatu BigML provedena je klaster analiza na skupu podataka. Kako bi utvrdili koji je optimalan broj klastera koji će se pokazati kao najbolja grupacija elemenata, provedena je klasterska analiza sa 3, 4 i 5 klastera. Opisana je distribucija elemenata po klasterima, prikazani su elementi grupirani po klasterima te je priložen grafički prikaz klasterske analize za sva tri slučaja.

Rezultat klasterske analize je određeni broj klastera koji predstavljaju države grupirane po sličnim karakteristikama. Primjerice, države rangirane niže na ljestvici mjere sreće, pozitivnog utjecaja, društvene potpore te visokog ranga korupcije i negativnog utjecaja čine jednu grupu – jedan klaster.



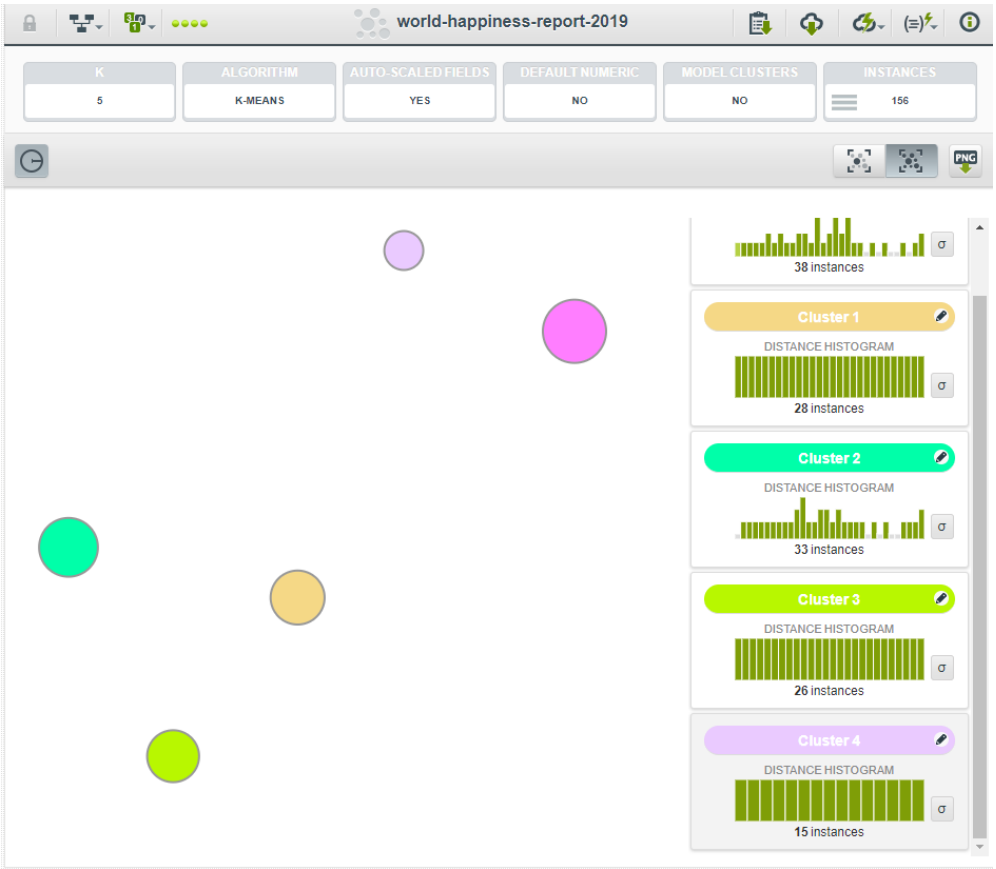
Slika 11: *Grafički prikaz klasterske analize sa 3 klastera (Izvor: samostalna izrada)*

Slika 11 prikazuje grafički prikaz klasterske analize sa 3 klastera. Distribucija instanci u klasteru 0 je najmanja, 30 instanci, u klasteru 1 je najveća, 67 instanci, a u klasteru 2 je 43. Uočavamo veću razliku u distribucijama elemenata po klasterima što nastojimo smanjiti. Stoga provodimo klastersku analizu sa 4 klastera kako bi pokušali približno izjednačiti broj elemenata po klasteru. Dobivenom klasterskom analizom ne uočavamo klastere koji strše, odnosno ne uočavamo klaster sa velikom razlikom u broju elemenata u odnosu na ostala dva klastera.



Slika 12: *Grafički prikaz klasterske analize sa 4 klastera (Izvor: samostalna izrada)*

Na slici 12 se vidi grafički prikaz sa 4 klastera gdje klaster 0 ima distribuciju od 32 instance što je najmanja distribucija, klaster 1 ima distribuciju od 38 što je najveća distribucija, klaster 2 ima distribuciju od 37, a klaster 3 distribuciju od 33. Sva četiri klastera imaju približno jednaku distribuciju elemenata što je upravo ono što nastojimo postići. Ne uočavamo klastere koji strše. Pokušat ćemo provesti ponovnu klastersku analizu, ovoga puta sa 5 klastera kako bi dobili još manju razliku u distribucijama elemenata po klasterima. Također, iz slike prikaza 4 klastera uočavamo kako su sva 4 klastera podjednako udaljena te sadrže približno jednak broj instanci.



Slika 13: *Grafički prikaz klasterske analize sa 5 klastera (Izvor: samostalna izrada)*

Grafički prikaz klasterske analize sa 5 klastera prikazan je na slici 13. Distribucije elemenata po klasterima su sljedeće: klaster 0 ima distribuciju od 38 instanci što je najveća distribucija, klaster 1 ima distribuciju od 28, klaster 2 ima distribuciju od 33, klaster 3 ima distribuciju od 26 instanci, a klaster 4 distribuciju od 15 što je najmanja distribucija. Nema klastera koji strše.

Klasterskom analizom sa 5 klastera nismo uspjeli smanjiti razlike u distribuciji po klasterima, zapravo smo razliku u distribuciji povećali. Stoga se odlučujemo za 4 klastera koji će činiti finalnu grupaciju elemenata po sličnim karakteristikama.

Temeljem klasterske analize sa 3, 4 i 5 klastera i njihovih distribucija, možemo zaključiti kako je za ovaj skup podataka optimalan broj klastera 4 jer navedeni prikaz daje najmanje razlike u distribucijama instanci po klasterima. Sljedeće stranice predstavljaju detaljniji uvid elementa klastera koji služe za smislenu grupaciju država ovoga skupa podataka.



Slika 14: *Elementi klastera (Izvor: samostalna izrada)*

Slika 14 daje uvid u elemente klastera, distribucije po klasterima i podatke o vrijednostima (aritmetička sredina, standardna devijacija i sl.).

Klaster 0 ima distribuciju od 32 instance. Najmanju prosječnu vrijednost u klasteru ima atribut „Ladder“, a najveću atribut „Positive affect“.

Klaster 1, kao što je navedeno, ima najveću distribuciju instanci, 38. Najmanju prosječnu vrijednost u klasteru ima atribut „Healthy life expectancy“, a najveću atribut „Corruption“.

Klaster 2 ima distribuciju od 37 instanci. Najmanju prosječnu vrijednost ima atribut „Corruption“, a najveću atribut „Social support“.

Klaster 3 ima distribuciju od 33 instance. Najmanju prosječnu vrijednost u klasteru ima „Positive affect“, a najveću atribut „SD of Ladder“.

Rezultati klasterske analize predstavljaju skupine država grupirane po sličnim karakteristikama. Klaster 0 predstavlja države koje imaju visok rang mjere sreće, pozitivnog utjecaja, društvene podrške, slobode, darežljivosti i GDP per capita. Odnosno, takve države su „najsretnije“. Klaster 1 predstavlja države sa nešto nižim rangom mjere sreće, pozitivnog utjecaja i GDP per capita. Takve države imaju manju mjeru sreće nego države grupirane u klasteru 0. Klaster 2 čine države sa najnižim rangom sreće, pozitivnim utjecajem, rangom slobode i GDP per capita. Takve države su najmanje „sretne“. Klaster 3 čine države koje su manje „sretne“ nego države klastera 0 i 1 što zaključujemo po elementima mjere sreće, pozitivnog utjecaja, društvene potpore, slobode i GDP per capita.

5.1.2. Faktorska analiza

Faktorska analiza je statistička metoda koja opisuje korelacijske odnose između nekoliko varijabli u smislu nekoliko osnovnih i neupadljivih slučajnih komponenti koje ćemo nazvati faktorima. [14] Primjerice, moguće je da varijacije u nekoliko promatranih varijabli uglavnom odražavaju varijacije u nekoliko nevidljivih (osnovnih) varijabli. [14] Faktorska analiza traži takve zajedničke varijacije kao odgovor na neprimijećene, skrivene varijable. [14] Promatrane varijable su modelirane kao linearne kombinacije potencijalnih faktora, a sama faktorska analiza ima za cilj pronaći neovisne, skrivene varijable. [14] Faktorska analiza povezana je sa analizom glavnih komponenti (PCA). [15] Analiza glavnih komponenti je tehnika koja se koristi kako bi se naglasile varijacije i istaknule povezanosti u skupu podataka. [15] Često se kositi i kako bi se skup podataka prikazao na razumljiv način za istraživanje i vizualizaciju. [15]

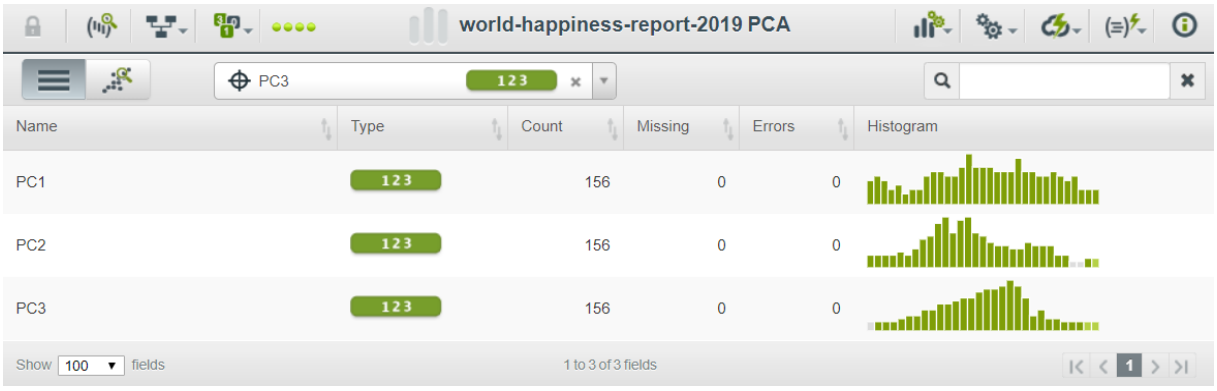
Analiza glavnih komponenti se koristi za transformaciju skupa podataka kako bi se reducirala dimenzionalnost. [16] Uglavnom se primjenjuje u područjima poput bioinformatike, analize portfolia, procesiranja signala, kvantitativnih financija koje sadržavaju vrlo velik broj varijabli što rezultira smanjenjem optimalnosti performansa strojnog učenja. [17]

Faktorska analiza se koristi kako bi skup podataka sa mnogo atributa transformirali u skup sa manjim brojem atributa. [18] Zbog toga se često naziva i tehnikom reduciranja dimenzije. [18] Možemo reducirati dimenziju podatka u jednu ili više „super-varijablu“ koja ima težinske atribute i tako predstavlja jedan atribut umjesto njih više. [18] Tehnika koja se najčešće koristi za reduciranje dimenzije podataka je Analiza glavnih komponenti (PCA). [18] U ovom poglavlju će se provesti redukcija podataka koristeći pristup ekstrakcije podataka. Alat BigML pruža opciju PCA konfiguracije koja kreira određeni broj komponenata.

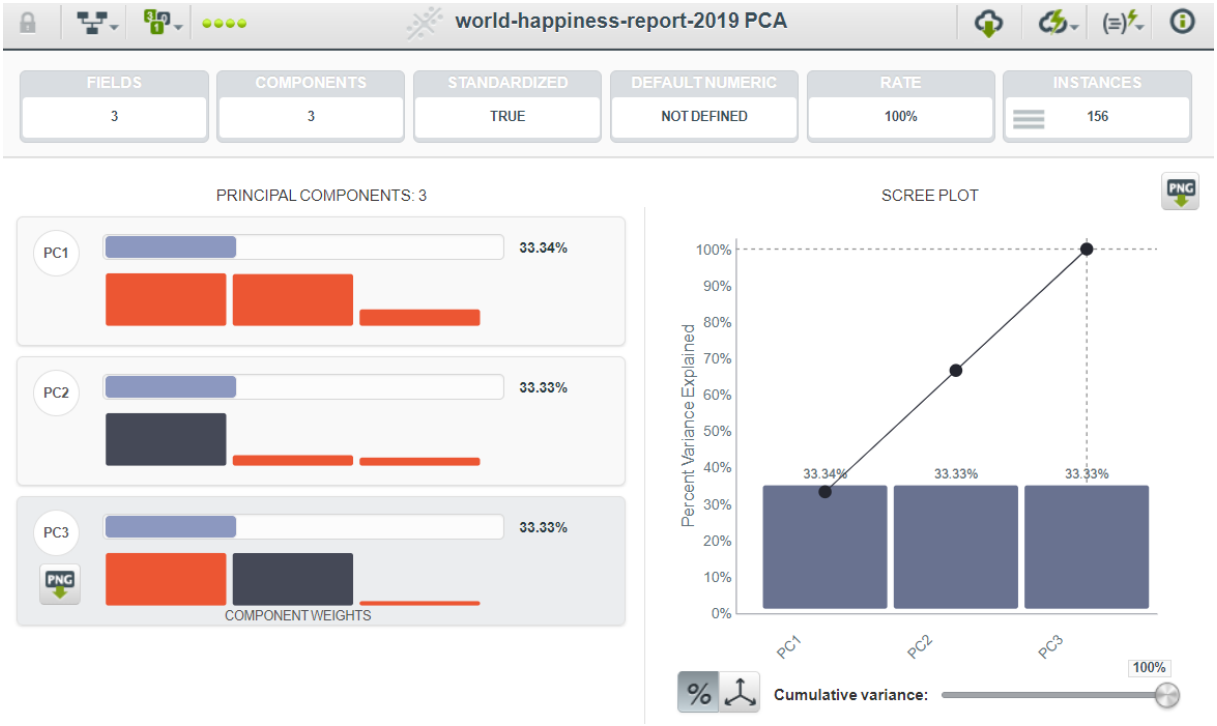


Slika 15: *Tehnika PCA (Izvor: samostalna izrada)*

Slika 15 prikazuje grafički prikaz svojstvenih vrijednosti dobivenih PCA konfiguracijom. Iz grafičkog prikaza svojstvenih vrijednosti uočava se manji pregib između prve i druge komponente. Stoga vrijednosti kumulativne varijance postavljamo na vrijednost 42.44% što odgovara zbroju varijanci prve i druge komponente. Rezultat kreiranja skupa podataka je dobivene tri komponente.



Slika 16: *Skup podataka dobiven tehnikom PCA (Izvor: samostalna izrada)*



Slika 17: *Rezultat tehnike PCA (Izvor: samostalna izrada)*

Svaka komponenta PC1, PC2 i PC3 se sastoje od atributa određenih težina. Težinski atributi prve komponente su Ladder, Social support, odnosno mjera zadovoljstva životom i društvena potpora. Težinski atributi druge komponente su Positive affect i Freedom, dakle pozitivni utjecaj i sloboda. Treća komponenta za težinske atribute ima Country i Corruption, odnosno naziv države i mjeru korupcije. Komponente PC1, PC2 i PC3 čine „super-varijable“ koje se sastoje od više atributa i tako predstavljaju skup podataka reduciranih dimenzija kojima je lakše rukovati.

5.2. Prediktivni modeli

Glavna zadaća prediktivnog modeliranja podataka je kreiranje rezultata kojima se previđaju budući ishodi. [10] Djeluje na način tako da se najprije analiziraju postojeći podaci te popune oni nedostajući najboljim pogađanjima. [11] Važno je znati da niti jedan statistički algoritam ne može sa stopostotnom sigurnošću predvidjeti budućnost zbog toga što se prediktivna analiza temelji na vjerojatnostima. Primjere primjene prediktivnih modela možemo pronaći u raznim granama. Recimo, predviđanje koje proizvode će kupci kupiti, predviđanje prognoze, predviđanje rasta i pada dionica, predviđanje razvoja dijabetesa kod osobe i sl. [11] Pruža odgovore na pitanja: što bi se moglo dogoditi i zašto bi se nešto moglo dogoditi. [10] U završnom radu su definirane i razrađene dvije tehnike prediktivnog modeliranja podataka: stablo odlučivanja i neuronska mreža.

5.2.1. Stablo odlučivanja

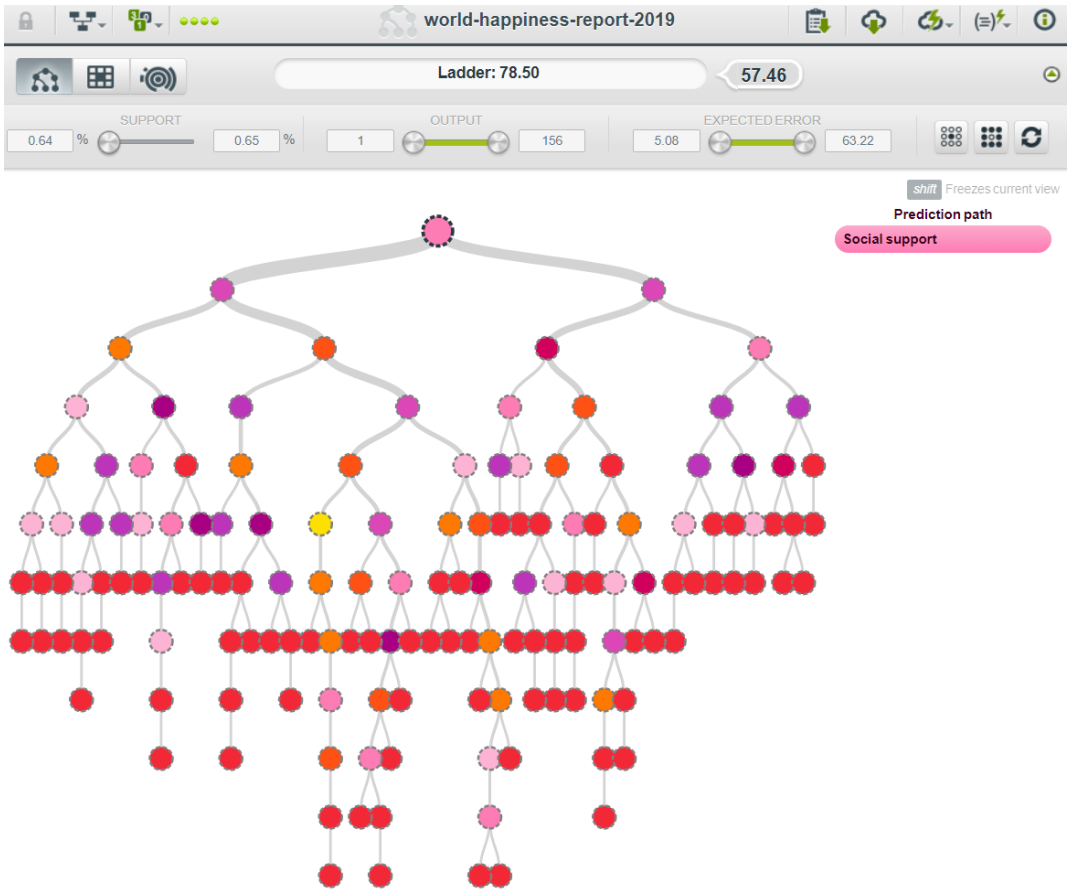
Stablo odlučivanja je primjer nadgledanog strojnog učenja i prediktivna metoda gdje se skup podataka kontinuirano dijeli prema definiranim parametrima. [19] Ulazni podatak, kao i izlazni podatak (podatak koji se predviđa) definiran je u postavkama koji se podešavaju prilikom kreiranja stabla odlučivanja. U završnom radu izlazni podatak koji se želi predvidjeti je „*Ladder*“, odnosno mjera sreće za pojedinu državu. Stablo odlučivanja je definirano čvorovima i listovima, gdje listovi predstavljaju odluke i konačne rezultate, a čvorovi mjesta gdje se podaci razdvajaju ovisno o odluci. [19] Jedna od važnijih prednosti stabla odlučivanja je svojstvo pridruživanja specifične vrijednosti problemu, odluci i rezultatu pojedine odluke, što rezultira reduciranjem dvosmislenosti u odlučivanju. [20] Svaki mogući scenarij odluke je prikazan grananjem, čvorovima i listovima što olakšava pregled i razumijevanje skupa podataka. [20]

Na sljedećim stranicama prikazani su rezultati stabla odlučivanja, pouzdanost predviđanja, navedeni su atributi koji imaju velik utjecaj na pojedinu odluku, broj instanci u granama te distribucija predviđanja.

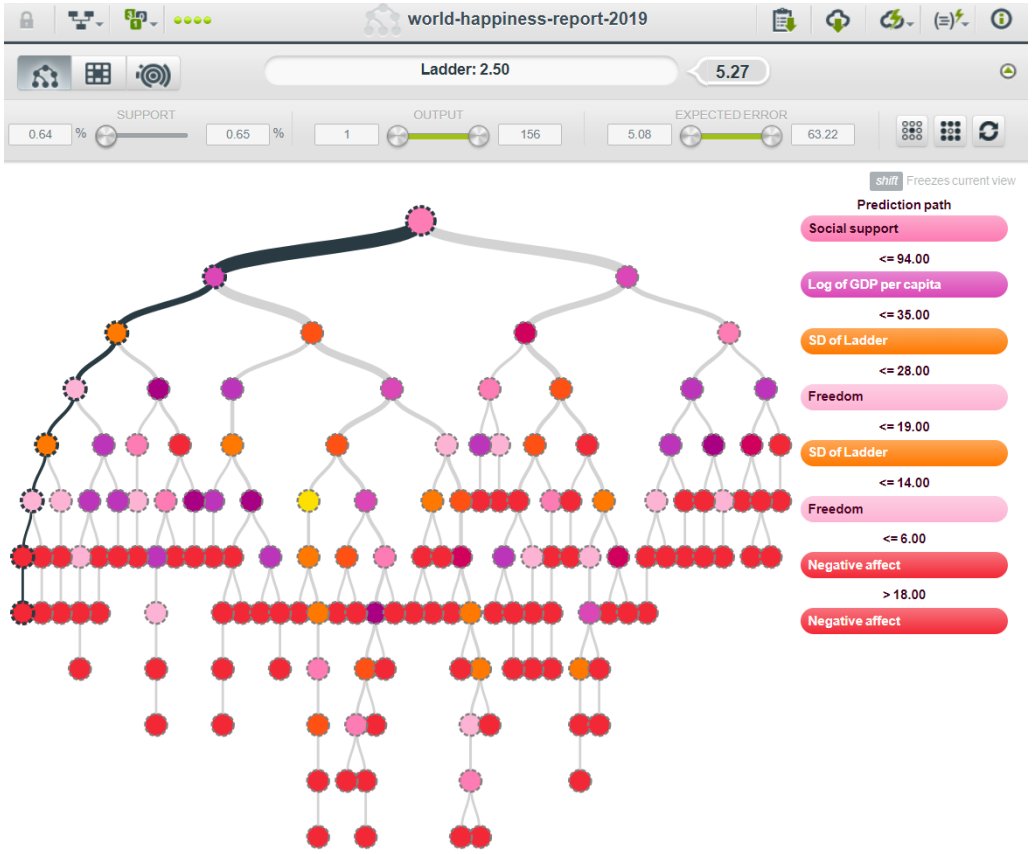
Alat pruža tri načina kreiranja stabla odlučivanja sa različitim metodama obrezivanja (*pruning methods*). Isprobane su metode poput Smart pruning, Active statistical pruning i No

statistical pruning te je zaključeno kako metoda Active statistical pruning daje najveću pouzdanost – 57,46%.

U korijenskom čvoru stabla pouzdanost je 57,46%, a atribut „Social support“ ima ulogu putanje predviđanja (prediction path) u važnosti od 60,82%. U prvom lijevom čvoru pouzdanost iznosi 43,58%. Lijeva grana stabla sadrži 94 instance, odnosno 60,26% ukupnih instanci. Atributi „Log of GDP per capita“ i „Social support“ čine putanju predviđanja gdje je važnost atributa „Log of GDP per capita“ 18,92%, a atributa „Social support“ 60,82%. U prvom desnom čvoru pouzdanost je 33,52%. Desna grana stabla sadrži 61 instanci što čini 39,10% ukupnih instanci. Putanju predviđanja čine atributi „Social support“, „Log of GDP per capita“. Važnost atributa „Social support“ je 60,82%, atributa „Log of GDP per capita“ je 18,92%.

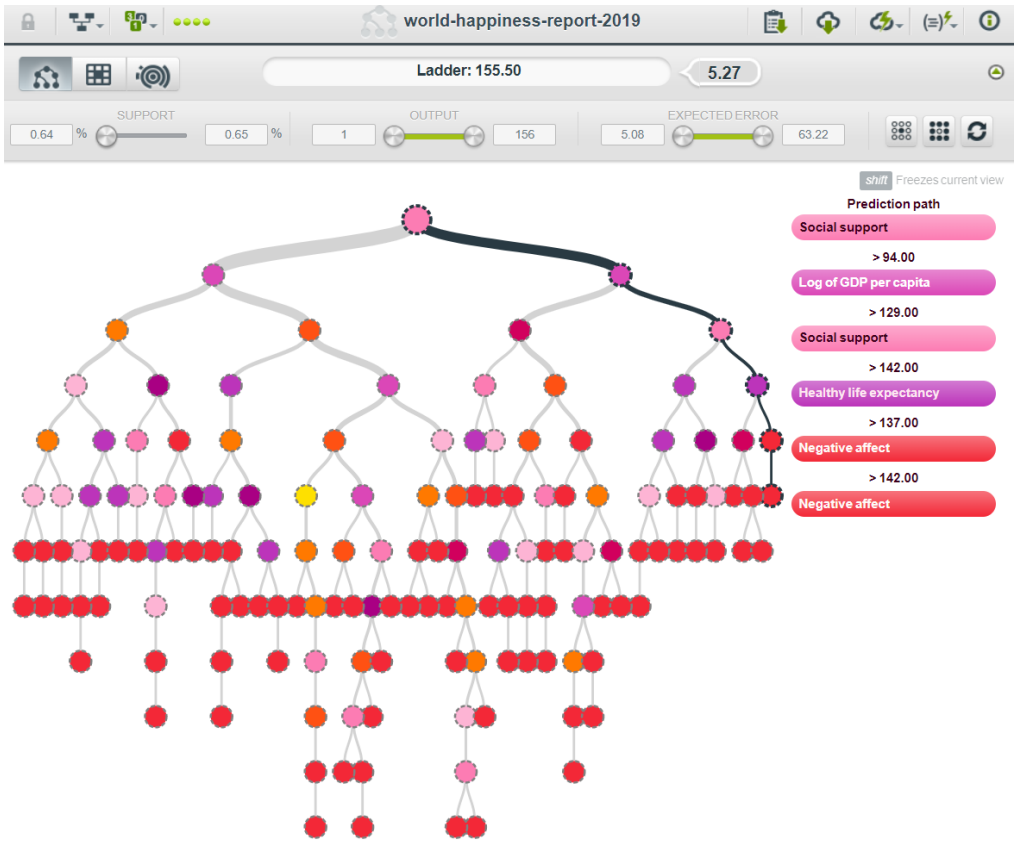


Slika 18: *Stablo odlučivanja (Izvor: samostalna izrada)*



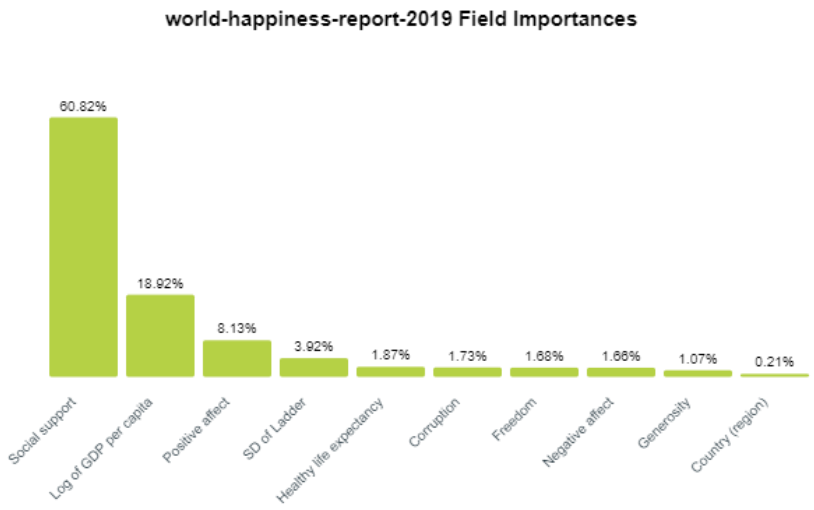
Slika 19: *Lijeva grana stabla sa opisom lijevog lista (Izvor: samostalna izrada)*

Krajnje lijevi list daje za rezultat mjeru sreće u vrijednosti 2,5 što predstavlja države visoko rangirane po mjeri sreće, odnosno čiji su stanovnici „najsretniji“. Na rezultat odluke utjecali su sljedeći atributi: rang društvene podrške je manji ili jednak 94, rang GDP per capita je manji ili jednak 35, rang standardne devijacije mjere sreće je manji ili jednak 28, rang slobode je manji ili jednak 19, rang negativnog utjecaja je veći od 18. Pouzdanost takvog predviđanja je 5,27.



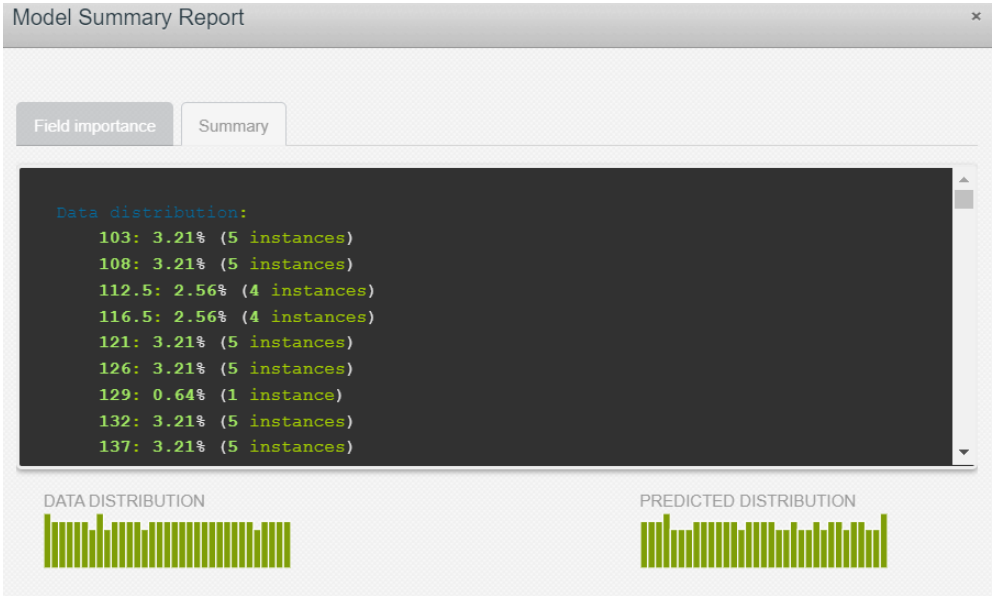
Slika 20: *Desna grana stabla sa opisom desnog lista (Izvor: samostalna izrada)*

Krajnje desni list za rezultat daje rang mjere sreće 155,5. Države ranga mjere sreće 155,5 su one koje se nalaze pri dnu ljestvice i obično su njihovi stanovnici najmanje „sretni“. Na rezultat ove odluke utjecali su sljedeći atributi: rang društvene podrške je veći od 94, rang GDP per capita je veći od 129, rang očekivanja trajanja života je veći od 137, rang negativnog utjecaja je veći od 142. Pouzdanost ovakvog predviđanja je 5,27.



Slika 21: *Važnosti varijabli (Izvor: samostalna izrada)*

Slika 21 prikazuje popis varijabli koje su važne za predviđanje. Uočavamo kako je najvažnija varijabla „Social support“ s vrijednosti važnosti od 60,82%. Uz nju veliku važnost ima varijabla „Log of GDP per capita“ s vrijednosti važnosti od 18,92%.



Slika 22: *Grafovi distribucija (Izvor: samostalna izrada)*

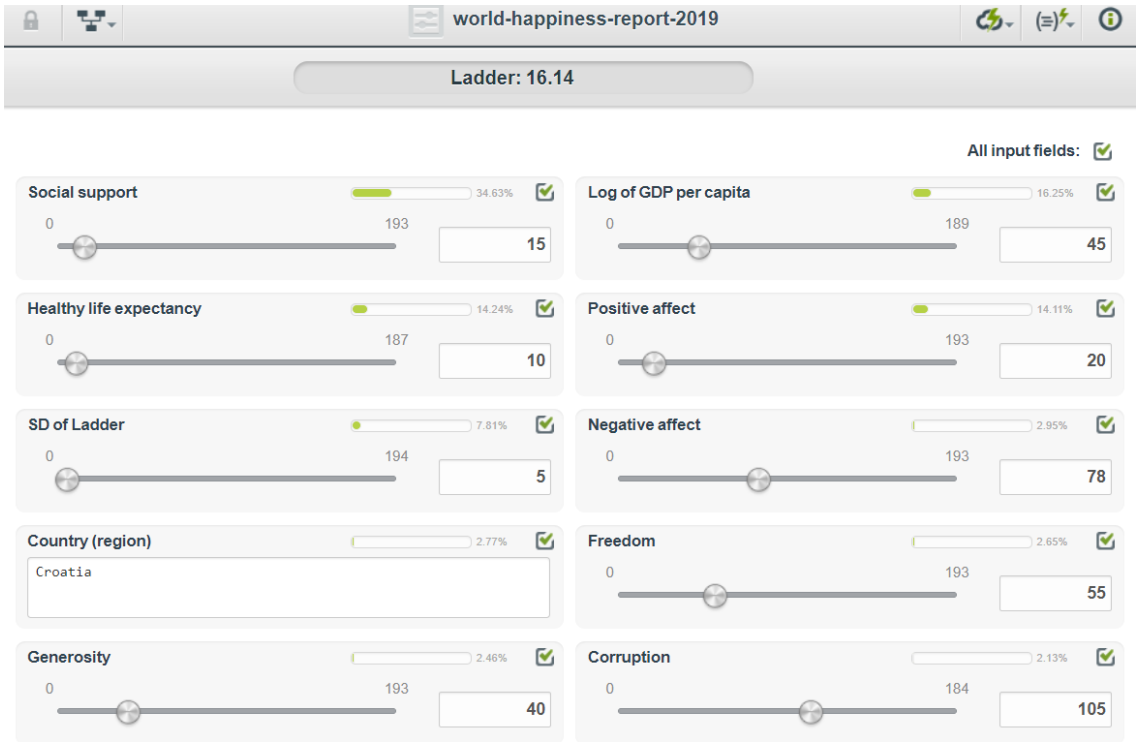
Prema slici 22 uočavamo da su grafovi trenutačne distribucije podataka i predviđene distribucije podataka približno jednaki. Time se zaključuje da je pouzdanost podataka visoka te predviđanje dobro.

5.2.2. Neuronska mreža

Neuronska mreža je tehnika strojnog učenja modelirana po uzoru na ljudski mozak. [21] Glavna svrha tehnike je izdvajanje skrivenih uzoraka unutar podataka što je osobito važno primjerice kod fotografija, videa ili govora. [21] Neuronska mreža se sastoji od sljedećih komponenti: input layer – potencijalni deskriptivni faktori koji mogu pomoći u predikciji; hidden layer – broj slojeva koje definira korisnik sa specifičnim brojem neurona u svakom sloju; output layer – utječe na ono što se nastoji predvidjeti; weights – svaki neuron u nekom sloju je potencijalno povezan sa svakim susjednim neuronom, gdje težine stvaraju važnost tih poveznica. [21]

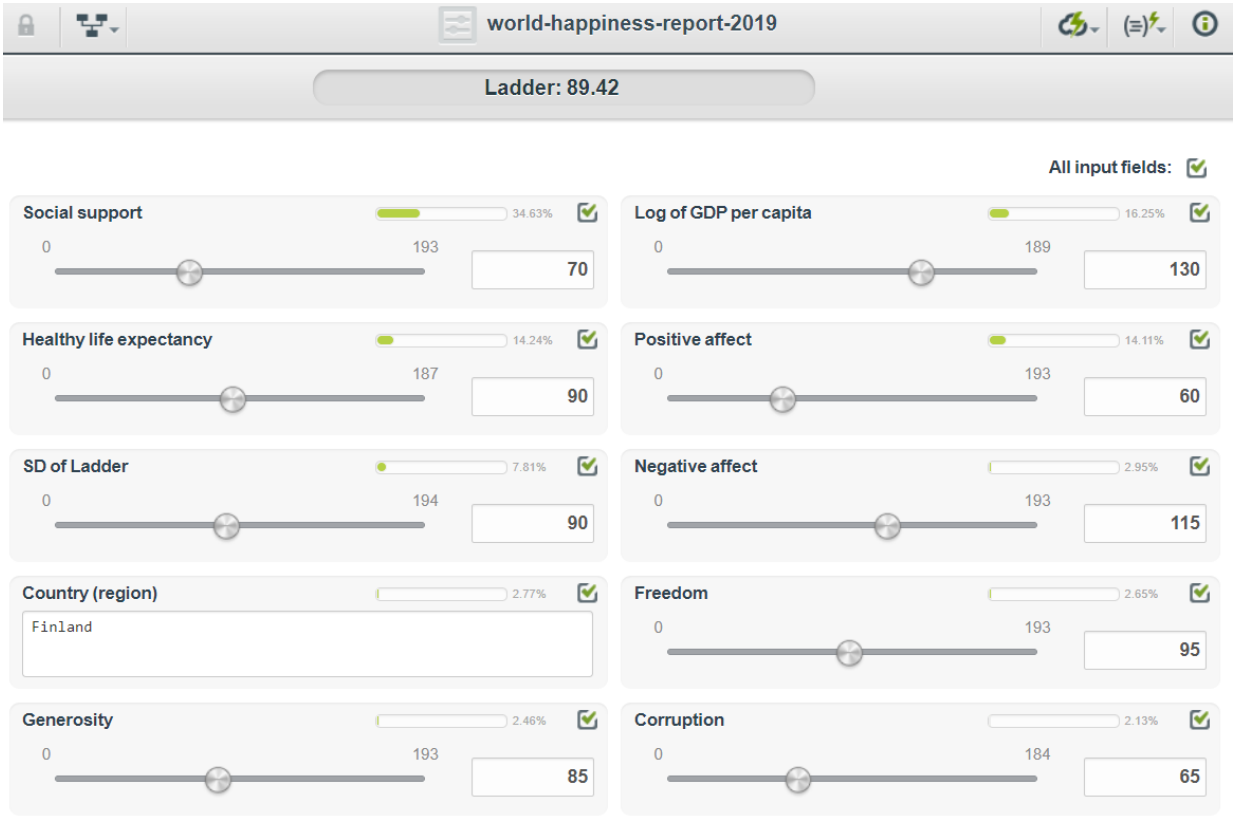
Za zavisnu varijablu prediktivnog modela primjenom neuronskih mreža izabran je atribut „Ladder“, odnosno mjera sreće. Prediktivni model kao rezultat daje rang mjere sreće za pojedinu državu temeljem ranga ostalih atributa koji u većoj ili manjoj mjeri utječu na krajnji rezultat predviđanja.

Isprobavanjem različitog broja skrivenih neurona (10,11,12 i 13), uočavamo kako prediktivni model sa 12 skrivenih neurona daje najveću pouzdanost. Na sljedećim stranicama prikazani su i opisani primjeri provedbe predviđanja podešavanjem vrijednosti atributa kako bi predvidjeli mjeru sreće za pojedinu državu.



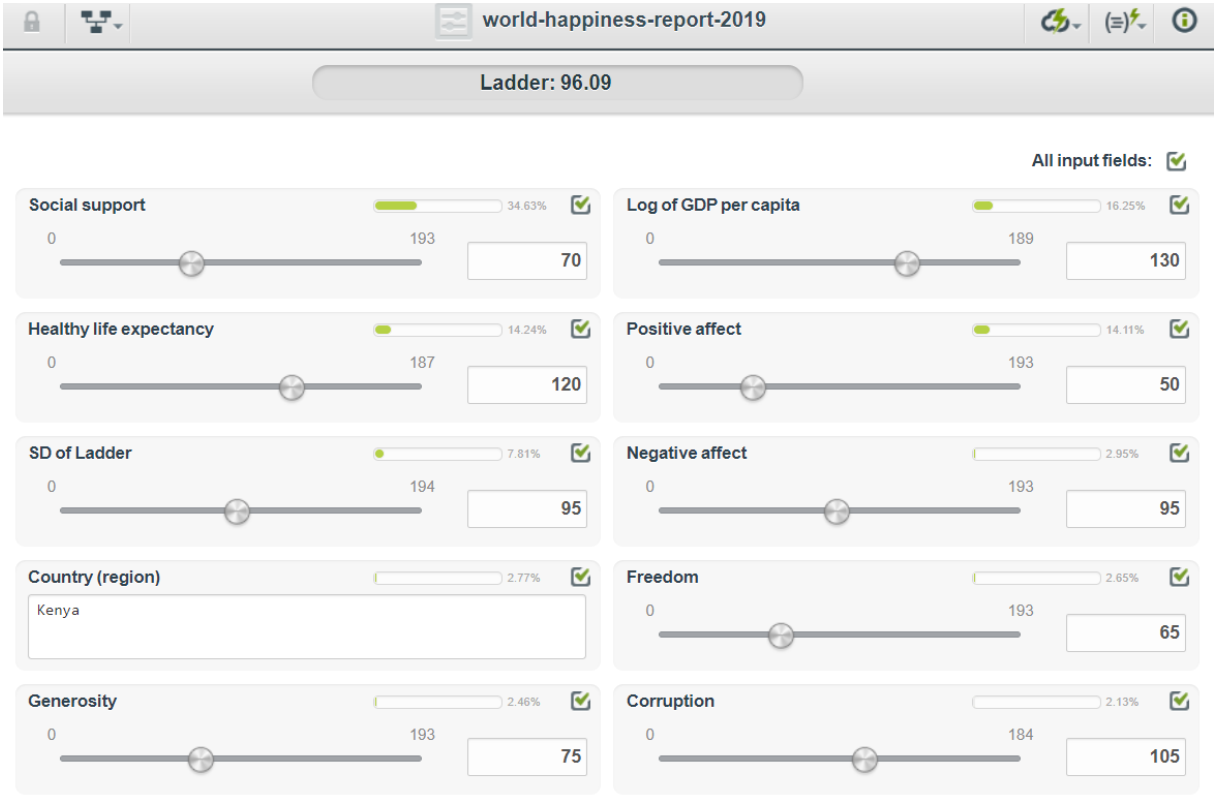
Slika 23: *Prvi primjer predviđanja (Izvor: samostalna izrada)*

Podešavanjem postavki prikazanih na slici 23 kao rezultat predviđanja dobiva se rang mjere sreće u vrijednosti od 16,14 za Hrvatsku. Prema vrijednostima podešenim na slici, predviđa se rangiranje na otprilike 16.mjesto u odnosu na ostale države svijeta po sreći i zadovoljstvu stanovnika. Originalni podaci o rangovima atributa iz istraživanja iz 2019. godine Hrvatsku svrstavaju na 75. mjesto prema rangu mjere sreće. Pomicanjem rangova atributa društvene podrške, očekivanog trajanja života, GDP per capita i pozitivnog utjecaja prema gore, predviđa se pomicanje prema gore na ljestvici mjere sreće.



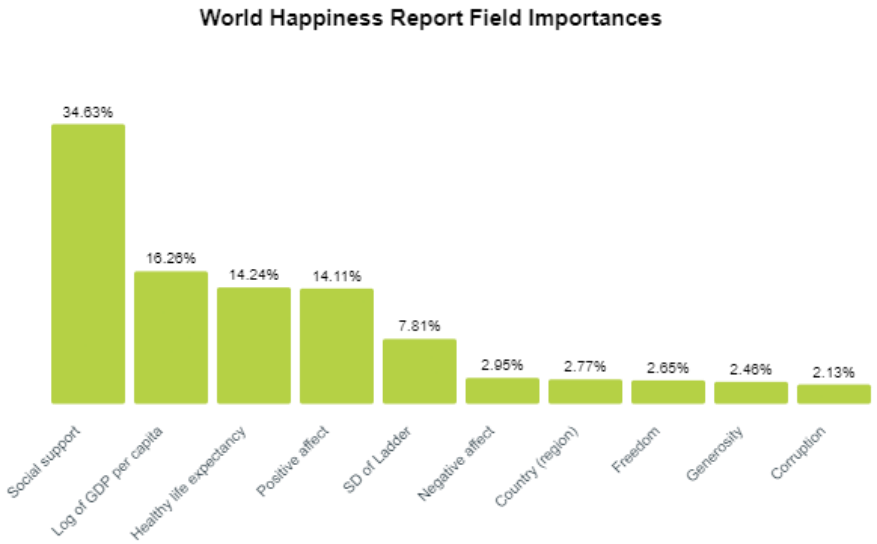
Slika 24: *Drugi primjer predviđanja (Izvor: samostalna izrada)*

Na slici 24 prikazan je rezultat predviđanja mjere sreće za Finsku prema podešenim postavkama vidljivih na slici. U izvornom skupu podataka, Finska je rangirana na 1.mjestu prema mjeri sreće, no ako bismo izmijenili podatke i postavili ih na nešto drukčije vrijednosti, mjera sreće bi se spustila na 89. mjesto u odnosu na druge države svijeta.



Slika 25: *Treći primjer predviđanja (Izvor: samostalna izrada)*

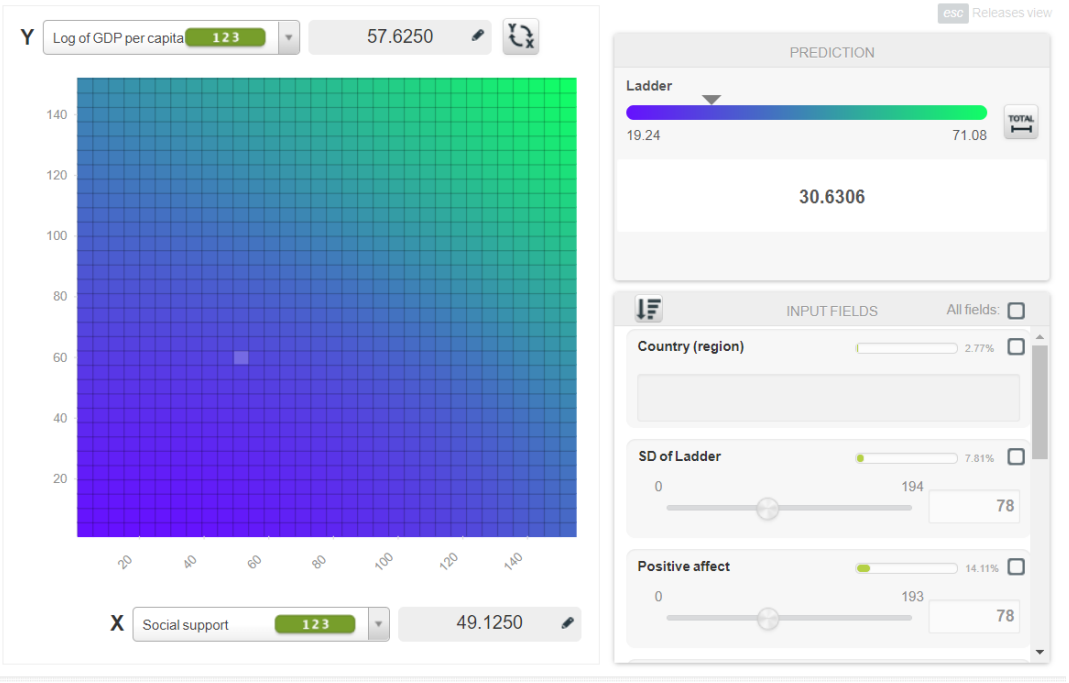
Slika 25 prikazuje predviđanje za Keniju. Prema izvornim podacima, Kenija je rangirana na 121. mjesto po mjeri sreće, a rezultat predviđanja prema gore podešenim postavkama svrstava je na 96. mjesto u odnosu na druge države.



Slika 26: *Deepnet Summary Report (Izvor: samostalna izrada)*

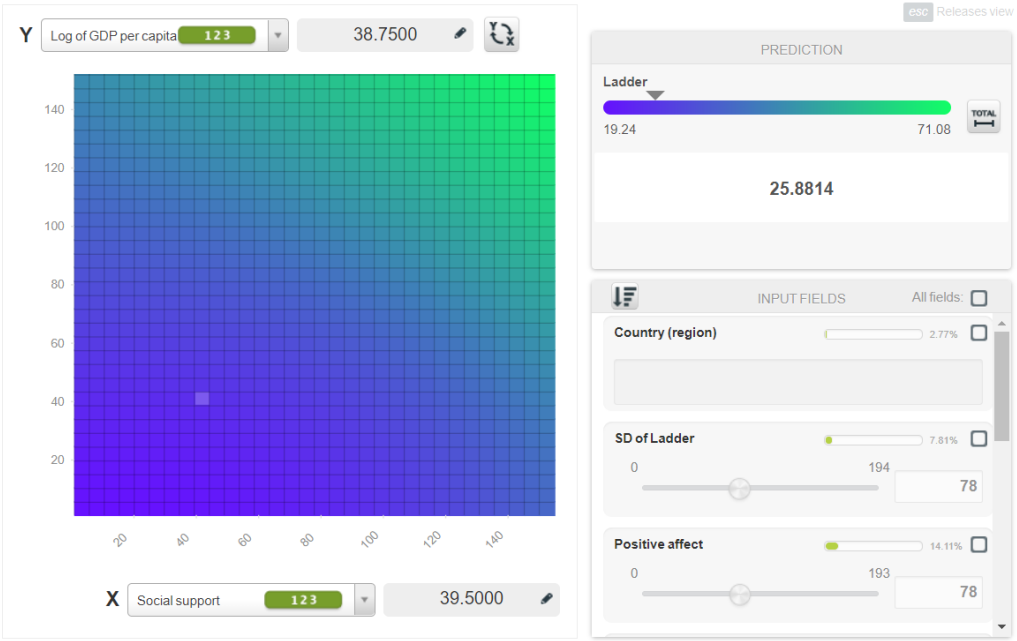
Slika 26 prikazuje varijable na temelju kojih najbolje možemo predvidjeti vrijednost izlazne varijable. Uočavamo kako se po važnosti ističe varijabla „Social support“, odnosno društvena potpora (važnost 34,63%) što ima smisla jer na mjeru sreće kod ljudi mnogo znači utjecaj i potpora društva i okoline u kojoj žive.

Provedeno je i predviđanje usporedbom dvije varijable te pozicioniranje na grafu kako bi se uočio rang mjere sreće za pojedinu državu.



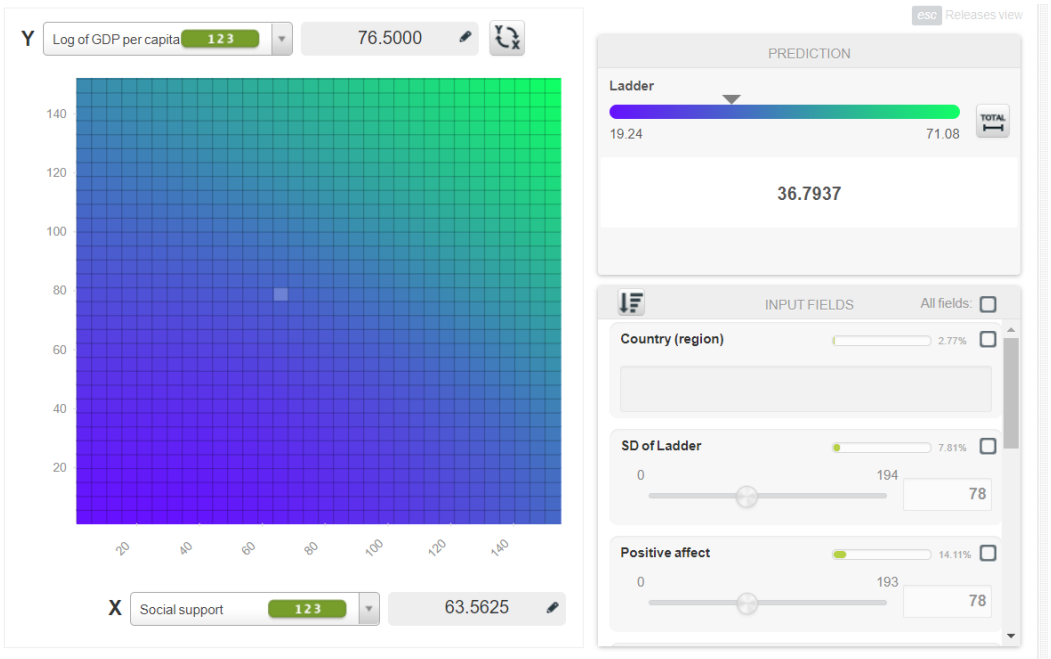
Slika 27: *Prvi primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada)*

Slika 27 prikazuje ovisnost numeričkih varijabli „Social support“ – vrijednost 49,1250 i „Log of GDP per capita“ – vrijednost 57,6250. Za navedene podatke predviđa se rang mjere sreće 30,6306. Ovakav način predviđanja omogućava podešavanje vrijednosti atributa društvene potpore i ranga GDP per capita kretanjem koordinatnim sustavom te mijenjanje vrijednosti x i y osi.



Slika 28: *Drugi primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada)*

Slika 28 prikazuje ovisnost numeričkih varijabli „Social support“ – vrijednost 39,5000 i „Log of GDP per capita“ – vrijednost 38,7500. Za navedene podatke predviđa se rang mjere sreće 25,8814.



Slika 29: *Treći primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada)*

Slika 29 prikazuje ovisnost numeričkih varijabli „Social support“ – vrijednost 63,5625 i „Log of GDP per capita“ – vrijednost 76,5000. Za navedene podatke predviđa se rang mjere sreće 36,7937.

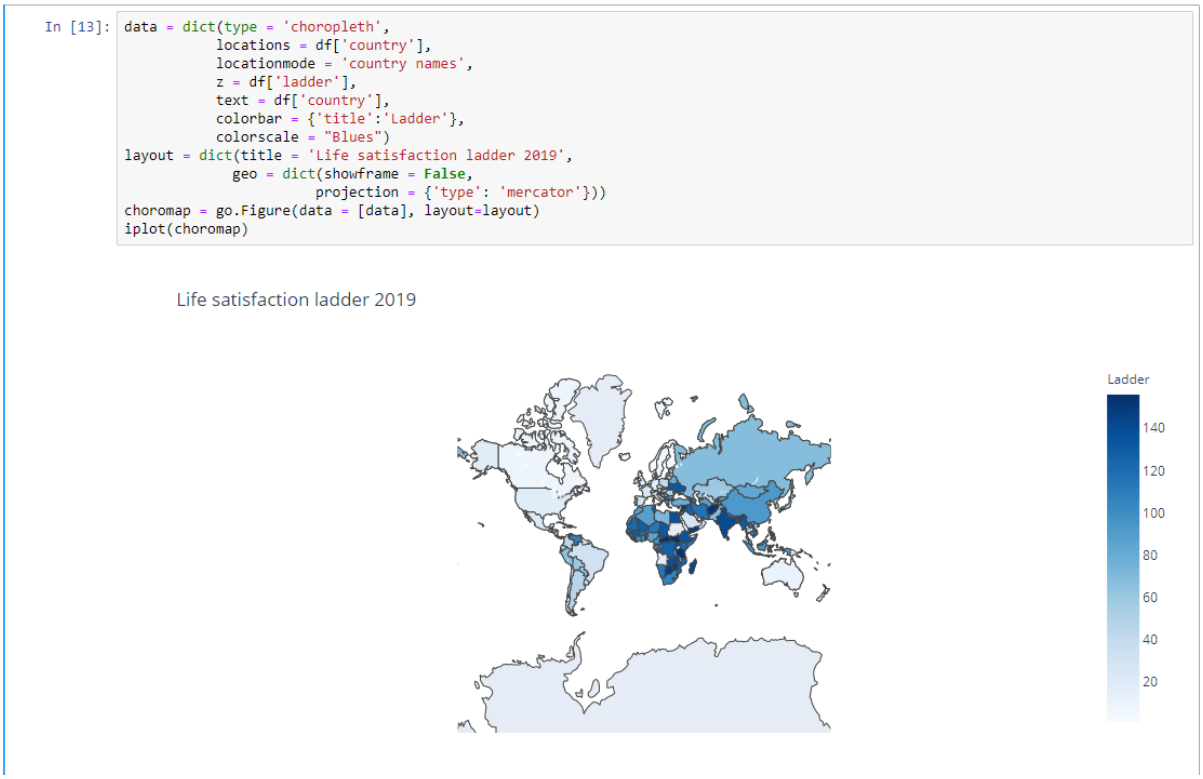
6. Interpretacija i evaluacija modela

Deskriptivnim i prediktivnim modeliranjem skupa podataka došli smo do rezultata o tome koji faktori u kojoj mjeri utječu na sreću i zadovoljstvo stanovnika države. Prediktivni modeli stabla odlučivanja i neuronske mreže dali su za rezultate sljedeće: atributi koji u najvećoj mjeri utječu na predviđanje i mjeru sreće su društvena potpora i GDP per capita (BDP po stanovniku). Oba atributa ispitanici smatraju bitnima za svoju percepciju sreće, stoga države koje imaju visok rang na ljestvici za oba atributa se smatraju „najsretnijima“.

Klasterskom analizom grupirane su države po sličnom rangu mjere sreće. Provedbom analize sa nekoliko broja klastera, pokazalo se kako je klasterska analiza sa 4 klastera najoptimalnija, odnosno daje najpreciznije rezultate. Jedan klaster svrstava „najsretnije“ države svijeta, jedan najmanje „sretne“ države svijeta, a ostala dva nešto manje „sretne“ države. Primjerice, promatranjem izvornog skupa podataka, stanovnici prvih 30-tak država se percipiraju najsretnijima. S druge strane, stanovnici zadnjih 30-tak država se percipiraju se najmanje sretnima, a ostale države koje čine središnji dio skupa podataka svrstavamo u ostala dva klastera kao nešto manje sretne države.

Ovakvom grupacijom država dobivamo uvid da one razvijenije države, sa višim GDP per capita i društvenom potporom imaju bolju mjeru sreće što potvrđujemo i primjenom prediktivnih modela. Stoga, kako bi pojedine države poboljšale percepciju sreće svojih stanovnika, trebale bi provesti razne metode koje bi rezultirale povećanjem GDP per capita primjerice otvaranjem novih radnih mjesta, ulaganjem u djelatnosti od kojih država može profitirati, izvoziti u druge države, poboljšati obrazovanje, poboljšavanje infrastrukture, smanjenjem populacije u zemljama u kojima je to nužno i sl. Također, osiguravanjem boljeg života razvija se i bolja društvena potpora što rezultira sretnijim i zadovoljnijim životom.

U programskom jeziku Python, izrađena je vizualizacija mjere sreće prema državama. Grafički prikaz predstavlja 156 država svijeta iz skupa podataka gdje tamnija nijansa plave boje predstavlja najmanje „sretne“ države, a svijetlije plava „najsretnije“ države. Za države obojene bijelom bojom nema podataka. Sljedeća slika prikazuje rezultat. Inspiracija za programsko rješenje je link prethodnog istraživanja naveden u literaturi pod brojem 22.



Slika 30: *Programsko rješenje (Izvor: samostalna izrada)*

Na početku programa uključuju se potrebne biblioteke za računanje i crtanje grafičkih elemenata poput numpy, pandas, matplotlib, seaborn, plotly. Zatim se u program učita skup podataka koji se nastoji analizirati, navedu stupci i broj instanci po stupcima. Nadalje, skup instanci dijelimo prema kontinentima: Azija, Europa, Sjeverna i Južna Amerika, Australija, Afrika što nam omogućava svrstavanje država u odgovarajuće kontinente. Posljednji detalji odnosili su se na definiranje boja, granica i teksta. [22]

7. Zaključak

U ovom se radu analizira javno dostupan skup podataka „*World Happiness Report*“ koji predstavlja godišnje izdanje Ujedinjenih naroda i sadrži rangove mjera sreće temeljenih na raznim faktorima koji utječu na sreću i percepciji sreće stanovnika pojedine države.

Provedbom nekoliko analiza na skupu podataka došlo se do određivanja faktora koji u određenoj mjeri utječu na mjeru sreće u svijetu. Zaključeno je da društvena podrška u najvećoj mjeri utječe na mjeru sreće. Slijedi je BDP po stanovniku kao jedan od faktora koji nije nužno činitelj kvalitete i zadovoljstva života, no ima veću ulogu u njezinom poboljšanju. Novac ne osigurava sreću i kvalitetu života, no zasigurno je poboljšava.

U teorijskom djelu rada opisana su 2 prethodna istraživanja slična ovom radu te se opisuje proces rudarenja podataka CRISP-DM. Također je objašnjeno deskriptivno te prediktivno modeliranje i dan je opis metoda koje će se koristiti.

Izradom rada potvrđena su očekivanja kako će na mjeru sreće najviše utjecati društvena potpora i BDP po stanovniku, no očekivan je veći utjecaj percepcije korupcije koja nije imala velik utjecaj. Rudarenje podataka pokazalo se kao izvanredna metoda za pronalazak informacija do kojih bi inače bilo veoma teško ili nemoguće doći. Primjena rudarenja podataka morala bi se početi koristiti što je više moguće jer je potencijal iznimno velik, a pitanja koja rješava, poput pitanja kako bi se mogla poboljšati sreća ljudi u svijetu, su ozbiljna i teška, a ovim metodama se doista olakšavaju u velikoj mjeri.

Alat koji se koristio za izrade deskriptivnih i prediktivnih modela je BigML, vrlo jednostavan i moćan alat za strojno učenje koji pruža niz opcija i načina analize podataka. Primjeri i slike grafičkih prikaza izvezene su iz alata uz dodatna objašnjenja. Također, izradom programskog rješenja bolje smo predočili države i odgovarajuću mjeru sreće prema podacima iz istraživanja.

U šestom poglavlju dane su i smjernice u svrhu poboljšanja percepcije sreće stanovnika poput otvaranja novih radnih mjesta, ulaganja u djelatnosti od kojih država može profitirati, izvoza u druge države, poboljšanja obrazovanja i infrastrukture, smanjenja populacije u zemljama u kojima je to nužno i dr. Osiguravanjem kvalitetnijeg života razvija se i bolja društvena potpora što rezultira sretnijim i zadovoljnijim životom.

8. Popis literature

[1] Microstrategy, „Data Mining Explained“. [Na internetu]. Dostupno na: https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained , [Pristupljeno: 13-kolovoz-2020].

[2] PyData, „Meta S. Brown: CRISP-DM: The dominant process for Data Mining | PyData London 2015“, 2015. [Na internetu]. Dostupno na: https://www.youtube.com/watch?v=civLio11SjQ, [Pristupljeno: 23-kolovoz-2020].

[3] Prof. dr. sc. Božidar Kliček, Doc.dr.sc. Dijana Oreški, materijal s predavanja iz kolegija Otkrivanje znanja u podacima ak. god. 2019./2020., preuzeto s e-learning sustava Moodle [Pristupljeno: 23-veljača-2020].

[4] Smart Vision Europe, „CRISP-DM“, 2015. [Na internetu]. Dostupno na: http://crisp-dm.eu/ , [Pristupljeno: 23-kolovoz-2020].

[5] Skup podataka: World Happiness Report 2019. [Na internetu]. Dostupno na: https://www.kaggle.com/PromptCloudHQ/world-happiness-report-2019, [Pristupljeno: 3-siječanj-2020].

[6] D. Oreški, I. Pihir, I. Kedmenec, „THE ASSOCIATION BETWEEN THE NATIONAL INTELLECTUAL CAPITAL COMPONENTS AND THE QUALITY OF LIFE“, 2016.

[7] D.Oreški, I. Kedmenec, „NATIONAL INTELLECTUAL CAPITAL OF EU-15 COUNTRIES FROM 1995 TO 2011“, 2015.

[8] European Commission, „What is open data?“. [Na internetu]. Dostupno na: https://www.europeandataportal.eu/elearning/en/module1/#/id/co-01 , [Pristupljeno: 14-rujan-2020].

[9] D. Oreški, B. Kliček, B. Šlibar, „Aspects of open data and illustrative quality metrics: literature review“, 2018.

[10] TechDifferences, „Difference Between Descriptive and Predictive Data Mining“, 2019. [Na internetu]. Dostupno na: https://techdifferences.com/difference-between-descriptive-and-predictive-data-mining.html, [Pristupljeno: 4-siječanj-2020].

[11] D. Bachar, „Descriptive, Predictive and Prescriptive Analytics Explained“, 2020. [Na internetu]. Dostupno na: https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/, [Pristupljeno: 4-veljača-2020].

[12] Poslovna učinkovitost, „Klasterizacija k-means algoritmom u Excel-u“ , 2017. [Na internetu]. Dostupno na:

https://www.poslovnaucinkovitost.eu/kolumne/poslovanje/klasterizacija-k-means-algoritmom-u-excelu, [Pristupljeno: 4-siječanj-2020].

[13] Oreški D., Klaster analiza, materijal s laboratorijskih vježbi iz kolegija Otkrivanje znanja u podacima ak. god. 2019./2020., preuzeto s e-learning sustava Moodle [Pristupljeno: 4-siječanj-2020].

[14] J. Andersen, „Factor Analysis 101“, 2019. [Na internetu]. Dostupno na: https://towardsdatascience.com/factor-analysis-101-31710b7cadff, [Pristupljeno: 11-siječanj-2020].

[15] V. Powell, „Principal Component Analysis“, 2016. [Na internetu]. Dostupno na: http://setosa.io/ev/principal-component-analysis/, [Pristupljeno: 11-siječanj-2020].

[16] BigML, „Principal Component Analysis (PCA): Dimensionality Reduction!“, 2018. [Na internetu]. Dostupno na: https://blog.bigml.com/2018/12/05/principal-component-analysis-pca-dimensionality-reduction/, [Pristupljeno: 11-siječanj-2020].

[17] BigML. „Introduction to Principal Component Analysis: Dimensionality Reduction Made Easy“, 2018. [Na internetu]. Dostupno na: https://blog.bigml.com/2018/12/07/introduction-to-prinicipal-component-analysis-dimensionality-reduction-made-easy/, [Pristupljeno: 11-siječanj-2020].

[18] Qualtrics, „What is factor analysis and how does it simplify research findings?“ [Na internetu]. Dostupno na: https://www.qualtrics.com/experience-management/research/factor-analysis/, [Pristupljeno: 11-siječanj-2020].

[19] M. Kulkarni, „Decision Trees for Classification: A Machine Learning Algorithm“, 2017. [Na internetu]. Dostupno na: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html, [Pristupljeno: 12-siječanj-2020].

[20] Bright Hub, „A Review of Decision Tree Analysis Advantages“, 2019. [Na internetu]. Dostupno na: https://www.brighthubpm.com/project-planning/106000-advantages-of-decision-tree-analysis/, [Pristupljeno: 12-siječanj-2020].

[21] D. Kellett, „Making Data Science Accessible – Neural Networks“ [Na internetu]. Dostupno na: https://www.kdnuggets.com/2016/08/making-data-science-accessible-neural-networks.html, [Pristupljeno: 12-siječanj-2020].

[22] Natevegh, „World Happiness difference between east and west“, [Na internetu]. Dostupno na: https://www.kaggle.com/natevegh/world-happiness-difference-between-east-west?fbclid=IwAR0owEbRDiisAzKJgovQzZBTpDW6HgFMLUPEt3N93qfYfp0Q8gLS4q2tew, [Pristupljeno: 13-siječanj-2020].

9. Popis tablica

Tablica 1: *Popis podataka* 22
Tablica 2: *Usporedba deskriptivnih i prediktivnih metoda rudarenja podataka* 26

10. Popis slika

Slika 1: CRISP-DM dijagram [3] 4
Slika 2: Razumijevanja poslovanja [4] 6
Slika 3: Razumijevanje podataka [4] 8
Slika 4: Priprema podataka [4] 10
Slika 5: Modeliranje [4] 12
Slika 6: Evaluacija [4] 14
Slika 7: Diseminacija [4] 16
Slika 8: Skup podataka (Izvor: vlastita izrada) 23
Slika 9: Korelacija između atributa „Ladder“ i „Healthy life expectancy“ (Izvor: vlastita izrada) 24
Slika 10: Korelacija između atributa „Corruption“ i „Healthy life expectancy“ (Izvor: vlastita izrada) 25
Slika 11: Grafički prikaz klasterske analize sa 3 klastera (Izvor: samostalna izrada) 28
Slika 12: Grafički prikaz klasterske analize sa 4 klastera (Izvor: samostalna izrada) 29
Slika 13: Grafički prikaz klasterske analize sa 5 klastera (Izvor: samostalna izrada) 30
Slika 14: Elementi klastera (Izvor: samostalna izrada) 31
Slika 15: Tehnika PCA (Izvor: samostalna izrada) 34
Slika 16: Skup podataka dobiven tehnikom PCA (Izvor: samostalna izrada) 34
Slika 17: Rezultat tehnike PCA (Izvor: samostalna izrada) 35
Slika 18: Stablo odlučivanja (Izvor: samostalna izrada) 37
Slika 19: Lijeva grana stabla sa opisom lijevog lista (Izvor: samostalna izrada) 38
Slika 20: Desna grana stabla sa opisom desnog lista (Izvor: samostalna izrada) 39
Slika 21: Važnosti varijabli (Izvor: samostalna izrada) 40
Slika 22: Grafovi distribucija (Izvor: samostalna izrada) 40
Slika 23: Prvi primjer predviđanja (Izvor: samostalna izrada) 42
Slika 24: Drugi primjer predviđanja (Izvor: samostalna izrada) 43
Slika 25: Treći primjer predviđanja (Izvor: samostalna izrada) 44
Slika 26: Deepnet Summary Report (Izvor: samostalna izrada) 44
Slika 27: Prvi primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada) 45
Slika 28: Drugi primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada) 46
Slika 29: Treći primjer predviđanja pomoću dva atributa (Izvor: samostalna izrada) 46
Slika 30: Programsko rješenje (Izvor: samostalna izrada) 48