Rudarenje podataka u poduzetništvu

Mislav, Knez

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Mislav Knez

RUDARENJE PODATAKA U PODUZETNIŠTVU

ZAVRŠNI RAD

Varaždin, 2019.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Mislav Knez

Matični broj: 44164/15-R

Studij: Poslovni sustavi

RUDARENJE PODATAKA U PODUZETNIŠTVU

ZAVRŠNI RAD

Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, rujan 2019.

Mislav Knez

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

Ovaj rad fokusira se na rudarenje podataka, te metode rudarenja podataka koje nam služe za rudarenje podataka u poduzetništvu. U ovome radu objasniti će se što je to rudarenje podataka, koje metode rudarenja poznajemo, te koje metode su nam najkorisnije za primjenu u poduzetništvu. Također objasniti će se što su to GEM podaci, kako se oni prikupljaju te koja je njihova svrha. Na kraju rada biti će opisano par oglednih primjera kako bi se prikazalo kako se rudarenje podataka i GEM podaci koriste u poduzetništvu. U ovome radu kao izvori informacija korištene su knjige i internetski članci, te ostali dostupni internetski sadržaj.

Ključne riječi: rudarenje podataka, GEM podaci, poduzetništvo, stablo odlučivanja.

Sadržaj

1. Uvod

U današnjem svijetu, gdje se većina poslovanja odvija na webu, broj podataka koji su izrazito važni za poduzeće raste. U tim podacima poduzeće može otkriti nove i korisne, do sada nepoznate informacije za poduzeće. Kako bi poduzeće otkrilo te nove i korisne informacije koristi se rudarenjem podataka, to jest jednom od metoda rudarenja podataka. Dobivanjem tih informacija poduzeće može prognozirati ponašanje u budućem poslovanju.

U ovome radu u prvom poglavlju objasniti ću što je to rudarenje podataka, te koje metode su nam najpogodnije za primjenu u poduzetništvu. Također opisati ću i proces rudarenja podataka, te ću ukratko objasniti pojedine metode. U drugome poglavlju fokusirat ću se na GEM podatke. Objasniti ću što su to GEM podaci, kako ih dobivamo i za što nam služe. Nakon toga u trećem poglavlju odabrati ću par primjera iz poduzetništva u kojima su korištene neke metode rudarenja podataka te ću objasniti koje metode su korištene za koji slučaj, ulazne podatke i izlazne vrijednosti. Na kraju dolazi zaključak gdje ću sumirati bit cijelog ovoga rada.

2. Rudarenje podataka

Kroz niz godina u poduzećima dolazi do gomilanja velike količine poslovnih podataka. Ti podaci mogu skrivati do sada neke nepoznate, ali korisne informacije. Kako bi otkrili te podatke koristimo se nekom od metoda rudarenja podataka. Svrha tih podataka je da nam pomognu u donošenju poslovnih odluka na osnovu znanja koja smo otkrili u podacima.

2.1. Definicija rudarenja podataka

Rudarenje podataka (engl. *data mining*) možemo opisati kao netrivijalan postupak kojim pronalazimo nove, valjane, razumljive i potencijalno korisne oblike podataka. Kao oblik podataka misli se na neke nove otkrivene pravilnosti među podatkovnim varijablama. Rudarenje podataka većinom se provodi na velikim količinama podataka iz baza podataka radi otkrivanja novoga znanja i donošenja boljih poslovnih odluka. [1]

Rudarenje podataka također možemo definirati kao i proces pronalaženja zakonitosti u podacima. Ti podaci mogu biti organizirani u baze podataka, ali to mogu biti i neki drugi podaci kao što su: tekstualni podaci, nestrukturirani podaci koji su proizašli sa weba ili neki podaci koji su organizirani u vremenske serije. [2]

Pojam rudarenja podataka možemo gledati na dva načina, to jest šire i uže poimanje. Šire poimanje se odnosi na cjelokupni proces otkrivanja znanja iz raspoloživih podataka, a uže poimanje se odnosi na specifičnu fazu obrade podataka [3]. Cjelokupni proces rudarenja podataka uključuje sljedeće faze [3]:

- Određivanje cilja prema potrebama korisnika
- Priprema podataka(odnosi se i na izgradnju skladišta podataka)
- Rudarenje podataka
- Ocjena i upotreba otkrivenog znanja

Rudarenje podataka oslanja se na znanja iz informatike, matematike i statistike pri otkrivanju odnosa između podataka i generiranja novoga znanja [4]. Iako su znanja iz svih područja važna, možemo reći kako je napredak informatike i tehnologije najznačajniji za intenzivnije uvođenje rudarenja podataka u poduzeća [1].

U osnovi je to analitički proces koji rezultira novim znanjem [4]. Podaci koji se koriste za rudarenje podataka često su pohranjeni u skladišta podataka koja sadrže povijesne, nepromjenjive podatke koji su tolerantni na pogreške u transakcijama [3]. Također rudarenje možemo vršiti i na XML dokumentima, vezanim pojavama ili biološkim mrežama.

Možemo reći kako rudarenje podataka u današnje vrijeme postaje sve korisnije i kako ozbiljne organizacije ne mogu dobro poslovati, to jest donositi bitne odluke bez rudarenja podataka. Takva novo stečena znanja ne služe nam samo kako bi dobili poveznice između podataka nego i tome kako bi unaprijedili našu prodaju, marketing, poslovanje i druge bitne stvari za svako poduzeće. Rudarenje podataka u poduzeću većinom obavljaju poslovni analitičari koji se koriste nekim od programskih alata namijenjenim za rudarenje podataka, a za njegovo upravljanje nije potrebno poznavanje metoda rudarenja podataka. Treba navesti kako se rudarenje podataka može primijeniti i u drugim područjima kao što su: mehanika, medicina, genetika, farmacija, itd. [2]. Tako na primjer ako istražujemo neki novi lijek rudarenjem podataka možemo otkriti koje su to karakteristike ljudi kod kojih je određeni lijek djelovao, a koje kod kojih nije djelovao. Neovisno o području primjene upotreba pravilne i dobro iskorištene metode rudarenja podataka sposobne su nam otkriti zakonitosti iz velike mase podataka. Općenito gledajuću rudarenje podataka možemo primijeniti u svim područjima u kojima se raspolaže sa velikom masom podataka na osnovu kojih želimo otkriti određene pravilnosti, veze i zakonitosti. [2]

U poduzećima rudarenje podataka se najviše koristi u području marketinga. Marketing je sve više usmjeren kupcu, to jest zadržavanju postojećih kupaca, te stvaranju novih. Za to je zaslužno upravljanje odnosima s kupcima (engl. *Customer Relationship Managment – CRM*) [1]. Upravo na tom području rudarenje podataka je naročito učinkovito ako postoji dovoljna količina kvalitetnih podataka koje je poduzeće prikupilo o kupcima. Poduzeće na području marketinga rudarenje podataka najviše upotrebljava za [1]:

- Direktni marketing – ponude se šalju kupcima za koje postoji najveća vjerojatnost odziva
- Izradu profila kupaca - utvrđuje se uzorak ponašanja kupca da bi mu se kasnije poslala prilagođena ponuda
- Segmentaciju – utvrđivanje grupa kupaca s jednakim karakteristikama
- Istraživanje povezanosti prodaje različitih proizvoda – analiza kupovne košarice
- Stimulaciju kupovine drugih artikala istog poduzeća
- Zadržavanje kupaca

Osim u marketingu najveća upotreba rudarenja podataka je u bankarstvu i osiguranju. Kao posebna područja rudarenja podataka razvijaju se rudarenja teksta i Web rudarenja [1]. Rudarenje Weba je analiza podataka o posjeti pojedina Web stranici, te analiza puta kojim posjetitelj dolazi do portala s podacima. Takve nam analize omogućuju praćenje zanimanja kupca i personalizaciju prikaza Web stranica. Kod rudarenja tekstova analiziraju se tekstovi i unutar njih se traže pravila i uzorci. [1]

2.2. Zadaci rudarenja podataka

U ovome poglavlju nastojati ću objasniti zadatke rudarenja podataka. Postoji šest zadataka rudarenja podataka, a zadaci su [5]:

- Deskripcija
- Klasifikacija
- Estimacija (regresija)
- Predikcija
- Klasteriranje
- Asocijacija

2.2.1. Deskripcija

Istraživači i analitičari ponekad žele pronaći načine kako da opišu pravilnosti i trendove koji leže u podacima. To možemo objasniti na primjeru tako što recimo ljudi koji su dobili otkaz za vrijeme vladajuće stranke će manje vjerojatno podržati tu stranku tijekom idućih izbora. U većini slučaja ti ljudi će svoj glas dati nekoj drugoj stranci zbog želje za promjenom. To smo mogli zaključiti opisom trendova i pravilnosti za takav slučaj. Zbog toga modeli rudarenja podataka moraju biti što je više moguće transparentni, to jest rezultati modela moraju opisivati jasne pravilnosti koje se mogu intuitivno objasniti i interpretirati. Neke metode rudarenja podataka možemo razlikovati od drugih po tome koliko je njihova interpretacija transparentna. Kao primjer ljudima lako razumljive metode rudarenja podataka možemo navesti stabla odlučivanja. Ona su većinom intuitivna i lako razumljiva, dok s druge strane imamo metodu rudarenja podataka neuronske mreže koja je namijenjena za stručnjake u određenom području zbog nelinearnosti i kompleksnosti modela. [6]

2.2.2. Klasifikacija

U klasifikaciji varijablu čiju vrijednost predviđamo nazivamo ciljnom varijablom. Ciljna varijabla je kategorijska. Jedan od primjera takve varijable može biti mjesečni prihod, a njega možemo podijeliti u tri kategorije: niski, srednji i visoki. Pošto proces rudarenja podataka obrađuje veliku količinu podataka, svaki podatak mora sadržavati informaciju o ciljnoj varijabli i skup nekih ulaznih varijabli. Ulazne varijable mogu biti dob, spol i zanimanje. Na temelju ulaznih vrijednosti za postojeće zapise u bazi podataka analitičar pokušava odrediti prihode za one osobe koje još nisu zabilježene u bazi. [6]

U idućih nekoliko rečenica opisat ću kako bi izgledao proces za klasifikacijsku zadaću. Prvi korak bio bi pregledavanje skupa podataka koji sadržava ulazne i ciljne varijable. Na taj način algoritam uči koje kombinacije vrijednosti ulaznih varijabli su povezane sa ciljnim varijablama. Na kraju algoritam pregledava zapise koji nemaju zabilježene nikakve vrijednosti i dodjeljuju im određene klasifikacije. Podaci na kojima algoritam uči nazivaju se podaci za trening. [6]

Primjeri korištenja klasifikacijskih zadaća uključuju: određivanje je li neka transakcija kreditnom karticom ilegalna, dijagnosticiranje bolesti kod pacijenta, procjena određene hipoteke kao dobar ili loš rizik, određivanje je li neka oporuka krivotvorena, određivanje pokazuje li financijsko i osobno ponašanje na terorističku aktivnost. [6]

2.2.3. Estimacija (regresija)

Estimacija (regresija) slična je klasifikaciji, ali je razlika u tome što je ciljna varijabla numerička, a ne kategorička. Modeli su izgrađeni pomoću kompletnih zapisa, to jest onih koji sadržavaju ciljne varijable uz ulazne varijable. Zatim se procjenjuje vrijednost ciljne varijable na temelju vrijednosti ulaznih varijabla. Većina modela estimacije dolazi iz područja statističke analize, ali mogu se koristiti i neuronske mreže. Primjer zadaća estimacije uključuju: procjena potrošnje slučajno odabrane obitelji, procjena smanjenja brzine nogometaša nakon ozljede koljenja, procjena prosječne ocjene studenta na preddiplomskom studiju na temelju ocjena iz srednje škole. [6]

2.2.4. Predikcija

Kod predikcije ciljna varijabla nam predstavlja buduću vrijednost. Pod pravilnim okolnostima bilo koja metoda koju smo koristili kod klasifikacije i estimacije možemo koristiti i kod predikcije. Možemo koristiti i tradicionalne metode kao što su procjena točke i intervala pouzdanosti, jednostavna linearna regresija i korelacija, višestruka regresija, ali možemo uključiti i metode rudarenja podataka kao što su neuronske mreže, stabla odlučivanja i k-najbliži susjed. Primjer predikcije je predviđanje cijena dionica u određenom vremenu u budućnosti. [6]

2.2.5. Klasteriranje

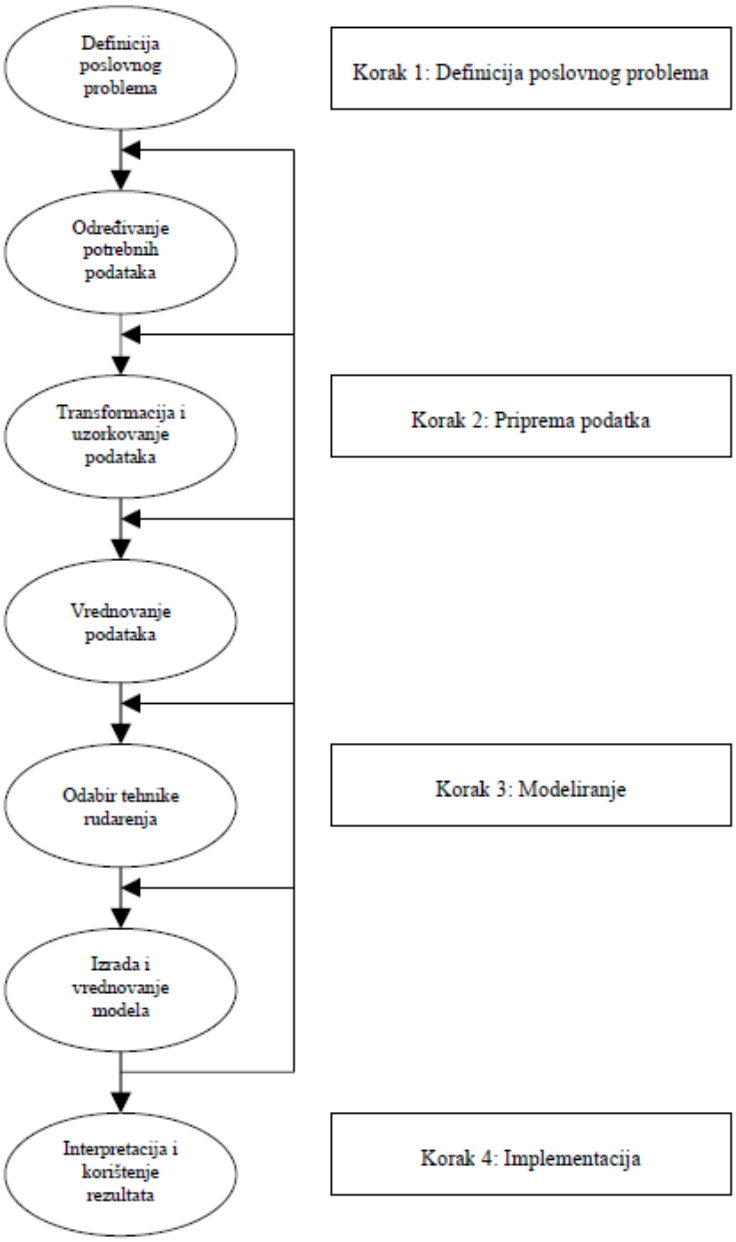
Odnosi se na grupiranje podataka u klastere. Klaster je skup podataka koji su međusobno slični, ali se razlikuju od skupa podataka u drugim klasterima. Kod klasteriranja ne postoji ciljna varijabla. Klasteriranje nam služi za segmentiranje cijelog skupa podataka u relativno homogene podgrupe ili klastere, s tim da sličnost podataka u klasterima bude maksimalna, a sličnost podataka jednog klastera s drugim minimalna. Često se koristi kao pred korak u procesu rudarenja podataka, a klasteri koji se dobiju koriste se kao ulazna varijabla za daljnje tehnike, na primjer neuronske mreže. Primjer klasteriranja je smanjenje broja atributa skupa podataka sa stotinama atributa. [6]

2.2.6. Asocijacija

Cilj asocijacije je otkrivanje koji atributi su međusobno povezani, odnosno određivanja pravila koja određuju odnos između dva ili više atributa. Većinom se koristi u poduzetništvu gdje je poznata pod nazivom analiza potrošačke košarice ili analiza afiniteta. Pravila asocijacije su oblika „Ako *uvjet*, tada *posljedica*“, zajedno s mjerom značaja i pouzdanosti vezanima za pravilo. Na primjer, određena trgovina ima podatke da je u subotu navečer od 300 kupaca 150 kupilo alkohol, a od njih 150, 100 ih je kupilo i cigarete. Asocijacijsko pravilo na temelju ovih podataka izgledalo bi ovako: „Ako *je kupac kupio alkohol*, tada *će kupiti i cigarete*“. Značaj ove tvrdnje je $\frac{150}{300} = 50\%$, a pouzdanost $\frac{100}{150} = 66.6\,\%$. Primjeri korištenja asocijacije mogu biti: istraživanje proporcije djece čiji roditelji im čitaju, a i sami su dobri čitači; otkrivanje koji artikli se prodaju zajedno; otkrivanje koji se artikli nikad ne prodaju zajedno. [6]

2.3. Proces rudarenja podataka

Rudarenje podataka je proces za koji ne postoji recept kako bi uvijek bio uspješan, to jest kako bi rezultirao pronalaženjem vrijednih informacija. Vjerojatnost uspjeha možemo povećati ako slijedimo korake procesa prikazane na Slika 1.



Slika 1. Proces rudarenja podataka [7]

U prvom koraku definira se poslovni problem, a drugi korak uključuje pripremu podataka koje uključuje njihovu pripremu, transformaciju, uzorkovanje i vrednovanje. Idući korak je modeliranje koje uključuje odabir modela rudarenja podataka i vrednovanje samoga modela. Zadnji korak je implementacija koja se sastoji od interpretacije dobivenih rezultata i njihove primjene. [7]

Proces rudarenja podataka je iterativan što nam omogućuje da se u svakom trenutku možemo vratiti na neki od prethodnih koraka. U nastavku ovoga poglavlja objasniti ću pojedine korake u procesu rudarenja podataka.

2.3.1. Definicija poslovnog problema

Prvi korak u procesu rudarenja podataka je definicija poslovnog problema. Kod definiranja poslovnog problema, poslovni problem moramo izraziti u obliku pitanja na koja ćemo na kraju procesa moći odgovoriti. Najbolji pristup definiranju problema je analiziranje područja u kojemu je rudarenje podataka već bilo uspješno, a nakon upoznavanja sa uspješnim primjenama možemo odabrati područje koje je najkritičnije za naše poduzeće. [7]

Ovaj korak služi nam i za određivanje koje osobe će raditi na projektu rudarenja podataka, a najčešće su to: specijalisti za rudarenje podataka, informatičari koji poznaju baze i skladišta podataka i stručnjaci koji su vezani za područje u kojemu vršimo rudarenje podataka. Također je važno da tim za rudarenje nadgleda netko od stručnih osoba iz menadžmenta koja ne mora biti direktno uključena u projekt, ali mora pomoći u rješavanju nekih ne planiranih problema. [7]

2.3.2. Priprema podataka

Priprema podataka, kao faza procesa rudarenja podataka, vremenski je najzahtjevnija i može obuhvaćati 60% do 90% vremena koje je potrebno za rudarenje podataka [8]. Podaci koje koristimo za rudarenje mogu biti pohranjeni u različitim oblicima, ali najčešće su to relacijske baze ili skladišta podataka. Specijalist za rudarenje, informatičar i stručnjak određuju koji će se podaci koristiti. Ovaj korak procesa sastoji se od sljedećih dijelova koji će biti objašnjeni u daljnjem tekstu [7]:

- Određivanje potrebnih podataka
- Transformacija podataka
- Uzorkovanje podataka
- Vrednovanje podataka

2.3.2.1. Određivanje potrebnih podataka

U ovome koraku određujemo podatke koje ćemo koristiti za izradu modela. Podaci koji se obično koriste za rudarenje pohranjeni su u obliku neke baze. Ta baza može biti baza klijenata, transakcija, narudžbi, itd. U ovome koraku određujemo još i koje ćemo varijable izbaciti iz analize i koja će nam varijabla biti ciljna. Na primjer, ako radimo analizu kreditnog rizika tada će nam ciljna varijabla biti ona koja opisuje je li klijent vratio kredit ili još uvijek duguje banci. Konačan rezultat ovog koraka je popis varijabli koji ćemo koristiti u izradi modela. [7]

2.3.2.2. Transformacija podataka

Ovaj korak služi nam za transformiranje varijabla iz baza u podatke koje možemo koristiti pri rudarenju podataka. Dobiveni podaci moraju biti u tabelarnom obliku gdje su stupci varijable, a redci opažanja. Svaki redak mora opisivati jedan podatak značajan za poduzeće. [7]

2.3.2.3. Uzorkovanje podataka

Korak uzorkovanja služi nam kako bi iz velikih količina podataka koje se nalaze u našoj bazi izdvojili manji dio podataka koji je potreban za naš model. Kod ovog koraka često se javlja pitanje: „Koliko je podataka dovoljno ?“. Podaci koji se izabiru za uzorak najčešće se izabiru slučajnim odabirom. Nakon što je izabran uzorak za izradu modela on se mora podijeliti na dva dijela i to tako da jedan dio podataka koristimo za izradu modela, a drugi dio podataka za testiranje modela. Ovakav pristup tipičan je za proces rudarenja podataka jer na taj način provjeravamo efikasnost modela na podacima koji nisu korišteni za izradu samoga modela. [7]

2.3.2.4. Vrednovanje podataka

Ovaj korak služi nam za analiziranje postojanja nekih netipičnih vrijednosti ili prljavih podataka. Netipične vrijednosti postoje u svim bazama podataka, a u bankarstvu u bazi podataka klijenata to bi mogao biti klijent sa izrazito visokim ili niskim primanjima. Možemo reći kako za prljave podatke vrijedi da su oni neke nepostojeće vrijednosti, nejasne definicije podataka ili neke netočne vrijednosti. Nepostojeće vrijednosti su one vrijednosti koje nemamo upisane kod klijenta u bazu podataka, na primjer spol klijenta. Nejasne definicije podataka javljaju se tijekom prebacivanja podataka iz jedne baze u drugu, a netočne vrijednosti su one vrijednosti koje nastaju tijekom pogrešnog unosa podataka u računalo. [7]

2.3.3. Modeliranje

Na početku modeliranja u procesu rudarenja podataka potrebno je napraviti analizu profila klijenta u kojoj se analiziraju odabrane karakteristike klijenta. Tijekom procesa rudarenja podatka obično se koriste metode: statistike, baze i skladišta podataka, umjetne inteligencije i vizualizacije [7]. Metode rudarenja podataka možemo podijeliti u tri kategorije [9]:

- Otkrivanje
- Klasifikacija
- Predviđanje

Metode otkrivanje odnose se na one metode koje traže pravilnosti u podacima bez prethodnog znanja o samim podacima. U te metode spadaju segmentiranje i asocijativna pravila. Metode klasifikacija varijable koriste za predviđanje određene kategorije, a u ovu metodu ubrajamo stabla odlučivanja, logit regresiju i neuronske mreže. Metode predviđanja koriste varijable kako bi predvidjele neku numeričku vrijednost. Koristi neuronske mreže, linearnu regresiju i metode vremenskih serija. Nakon odabira jedne od metoda koja nam koristi u našem procesu rudarenja podataka dobivene podatke potrebno je vrednovati, a za to nam služe podaci koje smo odabrali za testiranje metode. [7]

2.3.4. Implementacija

Ovaj korak odnosi se na interpretaciju samih rezultata procesa rudarenja podataka i na primjenu dobivenih rezultata. Ovdje je ključna uloga određenog stručnjaka za područje na kojem smo rudarili podatke zbog mogućnosti interpretacije podataka. Zbog lakše razumljivosti rezultata oni su najčešće prezentirani u obliku grafikona ili nekih pravila. Korištenje dobivenih rezultata ovisi o tome koliko su oni dobro predstavljeni. [7]

2.4. Metode rudarenja podataka

Kao znanstveni algoritmi većina metoda rudarenja podataka postoji već dugo godina. Njihova primjena u komercijalnom rudarenju počinje tek negdje u zadnjih 10 godina gdje dobivaju strateški značaj u poslovanju [3]. Već je prije navedeno kako su metode za rudarenje podataka preuzete iz različitih znanstvenih područja kao što su statistika, matematika, informatika. Izbor određene kombinacije tehnika koje ćemo primijeniti kod određene situacije ovisi nam o prirodi problema, dostupnosti podataka i vještinama stručnjaka [3].

Kod rudarenja podataka u većini slučajeva bavimo se izradom modela koji predstavlja algoritam. Taj algoritam povezuje ulaze i ciljne varijable ili izlaze. Za rudarenje podataka koristi se niz analitičkih metoda, a specifična faza obrade rudarenja podataka obično koristi jednu od sljedećih metoda [3]:

- Određivanje najbližeg susjeda
- Grupiranje
- Asocijativna pravila
- Stabla odlučivanja
- Neuronske mreže
- Genetski algoritmi

Gore nabrojane metode samo su neke od dostupnih metoda rudarenja podataka, to jest to su neke od najpopularnijih metoda. Još neke metode rudarenja podataka su [2]:

- Metoda potrošačke košarice
- Memorijsko temeljno razlučivanje
- Klasteriranje
- Bayesove mreže
- Neizrazita logika

U Tablica 1. možemo vidjeti učestalost korištenja pojedine metode rudarenja podataka prema odgovorima korisnika jednog od relevantnih internetskih izvora o otkrivanju znanja iz baza podataka koji se zove KDnuggets [3]. Prema Tablica 1. možemo vidjeti kako su najpopularnije metode za rudarenje podataka stabla odlučivanja, regresija i klasteriranje.

Metode rudarenja podataka učestalo korištene u posljednjih 12 mjeseci (N=203)	
Stabla odlučivanja/pravila (127)	62.6%
Regresija (104)	51.2%
Klasteriranje (102)	50.2%
Deskriptivna statistika (94)	46.3%
Vizualizacija (66)	32.5%
Asocijativna pravila (53)	26.1%
Analiza vremenskih serija (35)	17.2%
Neuralne mreže (35)	17.2%

Tablica 1. Metode rudarenja podataka prema učestalosti korištenja [10]

U današnje vrijeme postoje različite metode kojima se mogu analizirati prikupljeni podaci i otkrivati uzorci ponašanja. Sve te metode imaju isti cilj, a to je prikaz dosadašnjih kretanja podataka i izvođenje zaključka što ti podaci znače u nekom širom kontekstu [3]. Svaka od ovih metoda nastala je kao dugotrajni proces rada i istraživanja, te razvoja statističkih algoritama koji se mogu primjenjivati nad podacima.

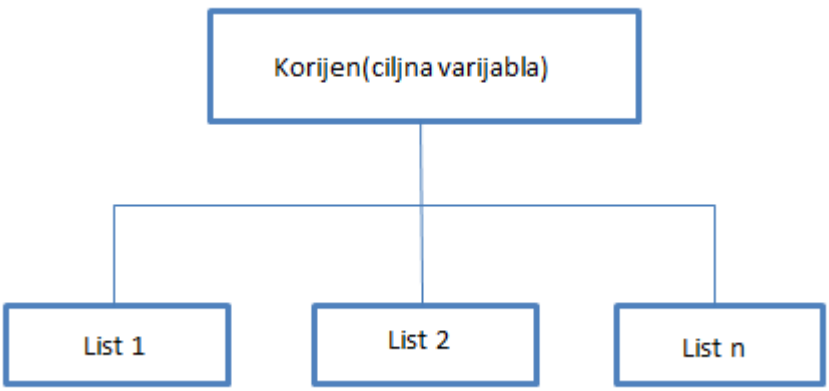
Uspješnost rudarenja podataka ovisi o odabranoj metodi i o kvaliteti samih podataka. Također potrebno je i jasno definirati problem koji istražujemo. Kod rudarenja podataka preporuča se upotreba više metoda na istom skupu podataka, kako bi se rezultati različitih metoda mogli međusobno usporediti. Time se postiže raznovrsnost, te više mogućih rješenja koja možemo primijeniti za određeni slučaj.

U nastavku ovoga poglavlja objasniti ću neke od metoda rudarenja podataka. Metode koje ću objasniti su: stabla odlučivanja, neuronske mreže, metoda potrošačke košarice, klasteriranje.

2.4.1. Stabla odlučivanja

Za razliku od tablice odlučivanja koja je statička, stabla odlučivanja možemo gledao kao dinamičku formu odlučivanja, a namjena im je klasificiranje atributa s obzirom na zadanu ciljnu varijablu [2].

Stablo odlučivanja može se formirati na temelju raznih algoritama, a neki od najpoznatijih su: ID3, c4.5, CHAID, CR&T i QUEST. Svim stablima odlučivanja neovisno o korištenom algoritmu zajedničko je to da ih je rezultate vrlo lako interpretirati u obliku odgovarajućih pravila. Struktura stabla odlučivanja vidljiva je na Slika 2. [2]



Slika 2. Struktura stabla odlučivanja [2]

Osnovni postupak konstrukcije stabla odlučivanja započinje tako da se definira jedinstveni korijen, odnosno ciljna varijabla koja reprezentira cijeli uzorak. Zatim ukoliko uzorci pripadaju istoj klasi čvor postaje list i označava se tom klasom. Ako to nije slučaj koristi se algoritam za selekciju atributa koji će najbolje razdijeliti uzorke na pod klase koje se nazivaju testnim atributom. Nakon toga stablo se dalje grana za svaku vrijednost testnog atributa. Ovi koraci ponavljaju se rekurzivno sve dok se ne dostigne određeni kriterij. [2]

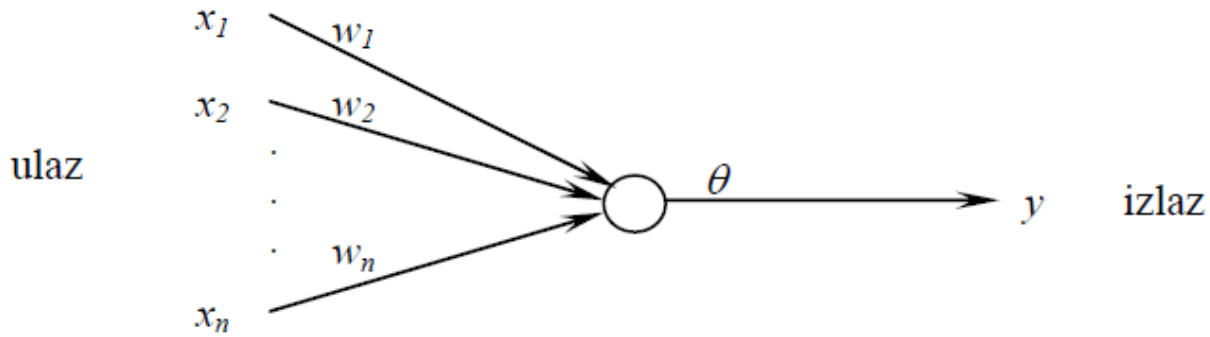
Stabla odlučivanja spadaju u klasifikacijske metode rudarenja podataka i uz njih se često koriste metode klasteriranja. Glavna primjena stabla odlučivanja je kod analize sklonosti potrošača kupnji nekog proizvoda ili usluge. Tako možemo otkriti koji potrošači kupuju koji proizvod, to jest kakve su karakteristike određenih potrošača koji kupuju neki proizvod. Ova metoda može koristiti i za segmentaciju tržišta pri čemu kao rezultat segmentacije dobijemo lako čitljiva pravila segmentacije. Na kraju treba naglasiti da prilikom formiranja stabla odlučivanja treba pratiti procjene pouzdanosti modela koje se temelje na uzorcima za treniranje i uzorcima za testiranje. [2]

2.4.2. Neuronske mreže

Neuronske mreže zamišljene su tako da djeluju slično ljudskom mozgu, a koriste se u analizi rizika i prognoziranju [1]. Kod ove metode rudarenje podataka počinje učenjem mreže pomoću podataka za koju je poznata vrijednost koju želimo prognozirati, a nakon toga naučeno znanje primjenjujemo na nekoj drugoj testnoj skupini podataka sve dok ne dobijemo zadovoljavajuće rezultate [1].

Metoda neuronskih mreža u posljednje vrijeme ima sve veću primjenu. Ove metode su u početku bile korištene za prepoznavanje uzoraka kao što su prepoznavanje rukopisa ili geometrijskih likova, a proces se temeljio na principu podudarnosti [11]. Ideja o neuronskim mrežama potječe iz neuropsihologije iz koje su iskorištena znanja o ponašanju živčane stanice. Može se reći kako ne postoji jedan univerzalni model neuronskih mreža koji bi bio primjenjiv za sve, nego se za svako područje mora raditi poseban model koji ima određene prednosti i mogućnosti u tome području [11].

Osnovni element neuronske mreže nazivamo neuron, a možemo ga opisati kao čvor grafa s ulaznim i izlaznim granama [12]. Neuron možemo vidjeti na Slika 3.



Slika 3. Neuron [12]

Sa Slika 3. možemo vidjeti kako ulazni dio neurona čini vektor ulaznih vrijednosti $(x_1, x_2, \dots x_n)$, a izlazni je jedna realna vrijednost y. Ulazne jedinice kombiniraju ulaze $(w_1, w_2, \dots w_n)$ u jedinstveni rezultat koji se preusmjerava u funkciju transformacije. Funkcija transformacije zatim kalkulira izlaznu vrijednost i najčešće poprima vrijednost između 0 i 1. Kombinacijska i transferna funkcija čine zajedno aktivacijsku funkciju neurona. Kod kombinacijske funkcije najčešće se primjenjuje funkcija sume s težinskim koeficijentima, a za funkciju transformacije najčešće se rabe funkcije sigmoida, te linearna i hiperbolična tangentna funkcija. Najčešće rabljenja sigmoidna funkcija: $S(x) = \frac{1}{1+ e^{-x}}$. U ovoj formuli x je rezultat kombinacijske funkcije koja je sumarna funkcija produkta težinskih koeficijenata i ulaznih parametara. [11]

Kako bi neuronske mreže što bolje odgovarale određenom zadatku one moraju učiti. Temeljni princip učenja neuronskih mreža su veze između eksperimentalnih uzoraka, a razlikujemo: autoasocijativno učenje i heteroasocijativno učenje. Kod autoasocijativnog učenja uzorci se pridružuju sebi samima, a u heteroasocijativnom učenju dva različita tipa uzorka se pridružuju jedan drugom [2]. Postupke učenja ovisno o tome koriste li vanjske signale dijelimo na [2]:

- Nadzirano učenje
- Nenadzirano učenje
- Pridruženo pojačavajuće učenje

Kod nadziranog učenja metodama povratne veze uklanjaju se razlike između željenog izlaza i stvarnog izlaza. Kod nenadziranog učenja naglasak je na samoorganizaciji, to jest klasteriranju. Kod nenadziranog učenja ne postoji povratna veza. [2]

Pridružujuće pojačavajuće učenje ima povratnu vezu koja indicira kada se aktualni i željeni uzorci podudaraju. U modelu širenja pogrešaka u neuronskoj mreži indikator uspješnosti je sveukupna pogreška. To je ukupna vrijednost svih razlika i trebala bi biti null. Na osnovu ove vrijednosti varijable u procesu učenja neuronske mreže odlučuje treba li prekinuti ili nastaviti proces učenja. Temeljna ideja ovih sustava je predviđanje budućih događaja, no mogu imati i širi spektar primjene. [2]

Kao što je već rečeno ne postoji univerzalni model neuronskih mreža primjenjiv u svim slučajevima. Svi do sada razvijeni modeli koriste se u određenom, za to namijenjenom, području. Jedan od prvih modela u praksi bio je McCulloh i Pittov model. Njihov princip se temelji na delta pravilu, a sam model koristi se u prepoznavanju uzoraka rukopisa i izgleda objekata. Nedostatak ovoga modela je nemogućnost uočavanja parnosti, povezanosti i simetrije što može biti problem kod rješavanja određenih problema. [2]

Idući model koji se može primijeniti u praksi bio je Hopfieldov model koji se temelji na autoasocijativnosti. Ovaj model namijenjen je za rješavanje numeričkih problema i ima visok stupanj primjenjivosti. Hopfieldov model nedostatke, kao što su problem lokalnih minimuma i maksimuma, rješava nadogradnjom Boltzmanovog načela preuzetog iz područja fizike. Kohonenove samoorganizirajuće mape su model neuronskih mreža koji se temelji na nenadziranom učenju. Ovaj model ima veliki potencijal primjene u području poslovne inteligencije. [2]

Sam tijek rješavanja problema biti će opisan u ovom odlomku. Nakon otkrivanja problema donosi se odluka o njegovom rješavanju, odnosno je li problem prigodan za rješavanje metodom neuronskih mreža. Ukoliko se može riješiti primjenom neuronskih mreža odabire se određeni model i određuje broj skrivenih slojeva i ulazno, izlaznih neurona. Nakon izbora skupa podataka počinje etapa inicijalizacije i učenja mreže. Po završetku učenja mreža se testira, a ako su rezultati zadovoljavajući mreža se primjenjuje u praksi. Ukoliko rezultati nisu zadovoljavajući vraćamo se na neki od prethodnih koraka i ponavljamo proces. [2]

Neuronske mreže snažan su alat, ali ih karakterizira skromna mogućnost interpretacije. Zbog toga se u većini slučajeva metode neuronskih mreža kombiniraju sa ostalim koncepcijama i metodama koje mogu pomoći pri interpretaciji rezultata. Neuronske mreže najviše se koriste kod prognoziranja trendova i predviđanja na temelju povijesnih podataka. Isto tako uspješno se mogu koristiti i u robotici i za prepoznavanje uzoraka kod alarmnih sustava. One mogu biti samo jedna karika u procesu rudarenja podataka gdje se mogu koristiti za segmentiranje ili predikciju, a na dobivene rezultate možemo primijeniti ostale metode rudarenja podataka. [2]

2.4.3. Metoda potrošačke košarice

Metoda potrošačke košarice najveću primjenu ima u manjim poduzećima kao što su maloprodajni centri, a podaci koji se koriste tijekom obrade su podaci s računa koji se izdaju na POS aparatima. Najčešće se ti podaci nalaze u dvije relacijske tablice od kojih prva služi za registraciju broja računa, vremena, datuma i sata; a druga je povezana sa relacijskim ključem. U njoj se nalaze podaci o šiframa proizvoda i količini kupljenih proizvoda. Analizom potrošačke košarice otkrivamo neka skrivena pravila koja se odnose na prodaju robe. [11]

Kao temeljnu definiciju metode potrošačke košarice možemo reći da je to metoda koja se svodi na otkrivanje asocijativnih pravila koja prikazuju koji se parovi artikala i s kojom vjerojatnošću kupuju zajedno. Matematički to možemo opisati kao asocijativno pravilo: $x \rightarrow y, gdje\ je\ x \cap y = 0.$ Mjera podrške skupa artikala X izražava se kao omjer transakcija koje sadrže skup transakcija u odnosu na ukupni broj transakcija. [2]

Pouzdanost se može definirati kao postotak transakcija. Ako te transakcije sadrže artikl X onda sadrže i artikl Y. Pouzdanost možemo iskazati preko sljedeće formule: $(Pouzadnost(X \rightarrow Y) = podrška(X \cup Y)/podrška(X).$ [2]

Metoda potrošačke košarice zanimljiva je po tome što pruža određenu slobodu pri utvrđivanju pravila s obzirom na vjerojatnosti koje se pojavljuju u transakcijama, a koje su povezane s pojavama parova proizvoda. Dizajner metode potrošačke košarice mora odrediti donju granicu vjerojatnosti koja utječe na formiranje pravila. [2]

Najčešće korištena metoda u procesu analize potrošačke košarice je a prior algoritam. Njegov glavni nedostatak je njegova kompleksnost i osjetljivost na umnožavanje elemenata analize. Umnožavanje elemenata analize dovodi do povećanja broja kombinacija, ali postoje i određenje tehnike kojima se to povećanje može smanjiti ali ne i u potpunosti riješit. [2]

Tehnike koje nam služe za smanjivanje elemenata prilikom korištenja a prior algoritma su: metoda formiranja prividnih varijabli i metoda grupiranja skupa proizvoda na temelju zajedničkih karakteristika. Ukoliko se odlučimo za smanjivanje elemenata a prior algoritma pojavljuje se nedostatak preciznosti analize [2]. Osnova funkcija a prior algoritma može se opisati u dva koraka [2]:

- Pronalaženje frekventnih artikala ili skupova artikala
- Generiranje asocijativnih pravila na temelju frekventnih artikala ili skupova artikala

Efikasnost a prior algoritma može se povećati redukcijom broja transakcija ili particioniranjem baze. Redukcija broja transakcija teži smanjivanju broja kombinacija, a metoda particioniranja baze dijeli bazu u nekoliko osnovnih particija. Nakon podjela u svakoj particiji računa frekvenciju pojavnosti skupa artikla u svakoj particiji i odabire najfrekventnije skupove. [2]

2.4.4. Klasteriranje

Klasteriranje možemo opisati kao princip koji nam služi za algoritamsko pronalaženje skupine podataka koji imaju zajednička svojstva. Navodi se kao jedna od glavnih metoda rudarenja podataka i kao nezaobilazna metoda pretprocesiranja podataka [13]. Algoritmi za klasteriranje pokušavaju naći sličnosti unutar neke skupine podataka, po nekom određenom atributu. Sličnost između podataka unutar neke skupine najčešće se računa primjenom funkcije udaljenosti kao što je Euklidska ili Manhattan udaljenost [2]. Postoji čitav niz algoritama za metodu klasteriranja, ali među njima možemo izdvojiti najpopularnija dva koja će biti objašnjena u daljnjem tekstu [2]:

- k-means algoritam
- aglomerativni hijerarhijski algoritam

2.4.4.1. K-means klasteriranje

Osnova ove metode svodi se na dijeljenje populacije na k segmenata gdje svaki segment ima n sličnih elemenata. Sličnost tih elemenata algoritama računa na temelju neke od funkcija udaljenosti [2]. Kod K-means klasteriranja za računanje udaljenosti u većini slučajeva koristi se funkcija udaljenosti Euklidska distanca. Isto tako za računanje udaljenosti koriste se i samoorganizirajuće mape koje spadaju u metodu neuronskih mreža. Ovaj algoritam koristi se za pretprocesiranje podataka kod rudarenja podataka i u kasnijim analizama. [13]

Ovu metodu možemo objasniti na sljedeći način. Na početku se izabire proizvoljno k segmenata koji predstavljaju klastere. Nakon toga odabire se središte za svaki od k segmenata. Zatim slijede ponavljajuće radnje, a to su: pridruživanje pomoću funkcije udaljenosti elemente populacije najbližim klasterima; računanje nove vrijednosti središta klastera za svaki klaster pojedinačno kao prosječnu vrijednost svih elemenata unutar klastera. Ovi koraci se ponavljaju sve dok se mijenjaju vrijednosti središta klastera. [2]

Ovaj algoritam razvio je J. B. MacQueen 1976. godine. U njegovoj originalnoj verziji za početne vrijednosti klastera uzima se prvih k vrijednosti iz populacije, gdje k predstavlja željeni broj klastera. Koristi se i metoda gdje se odabire slučajni broj k elemenata koji postaju centralne vrijednosti. Također postoje i verzije algoritma odabira početnih vrijednosti koje bi trebale optimizirati proces klasteriranja, a u tu skupinu možemo uvrstiti postupak sortiranja osnovne populacije. Kao glavni nedostatak ove metode navodi se to što analitičar izabire broj klastera, te je u praksi potrebno proći nekoliko iteracija procesa klasteriranja kako bi se dobio zadovoljavajući broj reprezentativnih klastera. [2]

2.4.4.2. Hijerarhijsko klasteriranje

Ova vrsta klasteriranja počinje grupiranjem objekata u stablo klastera. Na Hijerarhijsko klasteriranje možemo gledati kao na aglomerativno ili divizijsko hijerarhijsko klasteriranje, ovisno o tome krećemo li se od vrha prema dnu ili od dna prema vrhu. Glavni nedostatak ovog algoritma je to što ponavljanje procesa klasifikacije nije moguće na istoj razini stabla nakon što je izvršeno dijeljenje populacije u klastere na toj razini. [2]

Aglomerativno hijerarhijsko klasteriranje je klasteriranje metodologijom od dna prema vrhu svrstavanjem svakog pojedinačnog elementa u njegov posebni klaster. Nakon toga stvaraju se novi klasteri koji povezuju temeljne klastere u sve veće grupe sve dok svi elementi na kraju ne formiraju zajednički klaster ili dok se ne ostvari uvjet prekida daljnjeg klasteriranja. [2]

Divizijsko hijerarhijsko klasteriranje od aglomerativnog se razliku jedino po tome što je smjer klasteriranja od vrha prema dnu. Kod ovog algoritma temeljni klasteri, koji sadrže sve elemente populacije, se dijele na manje klastere sve dok svaki od elemenata ne formira vlastiti klaster ili dok se ne ostvari uvjet prekida daljnjeg klasteriranja. [2]

3. Global Entrepreneurship Monitor (GEM)

Global Entrepreneurship Monitor (GEM) je najvažnija svjetska studija o poduzetništvu. Na početku ovoga poglavlja objasniti ću kako je nastala GEM inicijativa, a nakon toga što su to GEM podaci, kako se prikupljaju i za što se koriste.

3.1. Povijest i definicija GEM-a

Osnovali su je 1997. godine Michael Hay iz London Business School i Bill Bygrave sa Babson College kao prototipno istraživanje. Pod vodstvom Paul Reynoldsa 1999. istraživanju se priključuje deset država od kojih je sedam bilo članica G7 ekonomije (Kanada, Francuska, Njemačka, Italija, Japan, Ujedinjeno Kraljevstvo i Sjedinjenje Američke Države), a ostale tri su bile Danska, Finska i Izrael. Pod njegovim vodstvom projekt je do 2003. godine narastao do 31 člana. Kako bi mogli upravljati interesima država članica projekta, 2004. godine osniva se Global Entrepreneurship Research Association (GERA) koja služi kao nadzorno tijelo GEM-a. GERA je neprofitna organizacija kojom upravljaju predstavnici država članica, dvije osnivačke institucije i sponzorske institucije. [13]

Kroz ogroman i središnje koordiniran, međunarodno proveden način skupljanja podataka GEM nam omogućuje pristup visokokvalitetnim informacijama, izvješćima i pričama koje nam omogućuju bolje razumijevanje poduzetništva. Također to je i stalno rastuća zajednica članova koji imaju korist od tih podataka. GEM je pouzdan izvor podataka za neke od najvećih međunarodnih organizacija kao što su Ujedinjeni narodi, Svjetski ekonomski forum, Svjetske banke i Organizacije za ekonomsku suradnju i razvoj. Oni nude prilagođene skupine podataka, posebna izvješća i stručno mišljenje. [14]

U svakoj ekonomiji GEM promatra dva elementa, a to su [14]:

- Poduzetničko ponašanje i stavovi pojedinaca
- Nacionalni kontekst i kako to utječe na poduzetništvo

Informacije koje se dobiju promatranjem pažljivo analiziraju GEM istraživači koji nakon analize dobivaju saznanja o kvaliteti okruženja za poduzetništvo. Cilj GEM istraživanja, zbog kojega je i pokrenut, bio je otkriti zašto su neke države više poduzetničke od drugih. [14]

GEM je jedinstven jer [15]:

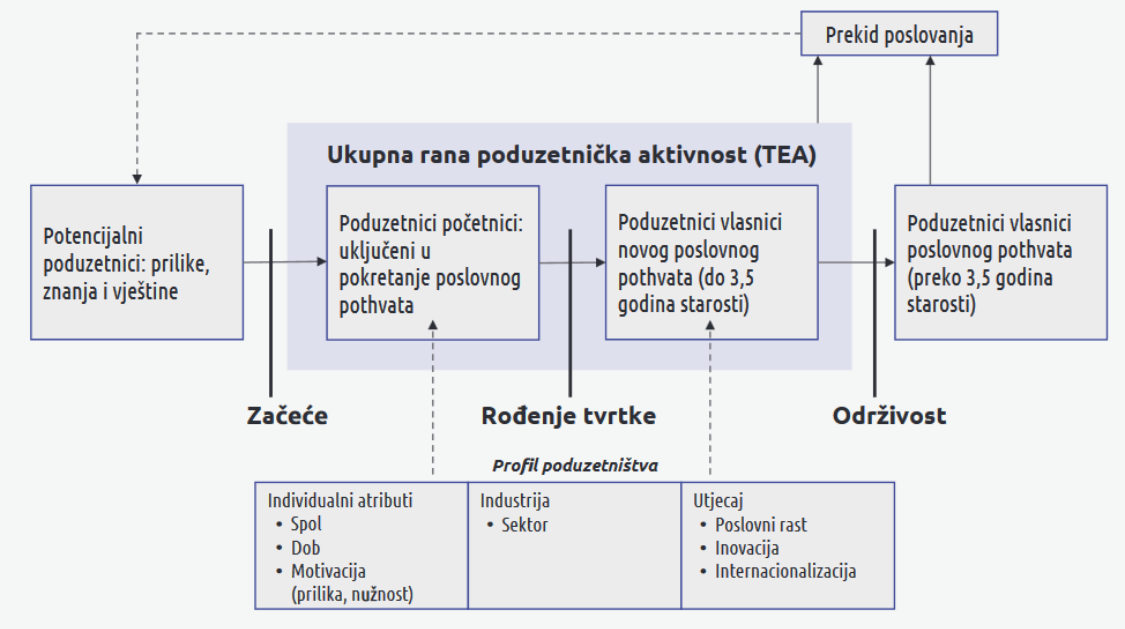
- Prikuplja primarne podatke o poduzetništvu
- Fokusira se na individualnog poduzetnika
- Ima isti pristup u cijelome svijetu
- Prepoznaje poduzetništvo kao proces
- Ima povijesne baze podataka koje su izuzetno opsežne
- Dom je preko 500 stručnjaka koji istražuju poduzetništvo
- U stanju je pratiti neformalne poduzetničke aktivnosti koje službene statistike ne bilježe

Glavni način kojim se GEM služi za prikupljanje podataka sastoji se od dva komplementarna alata: Ispitivanje populacije odraslih (APS) i Ispitivanje nacionalnih stručnjaka (NES). Oba načina prikupljanja biti će objašnjena u daljnjim poglavljima. Samo prikupljanje podataka je središnje koordinirano, a svaka država u kojoj se prikupljaju podaci prije samog prikupljanja je podvrgnuta nizu provjera. Također i nakon prikupljanja podataka i sami podaci se provjeravaju prije nego se objave. Takav proces prikupljanja osigurava da su GEM podaci najviše kvalitete. U GEM-u svaku državu predstavlja „Nacionalni tim“ koji vodi neka istraživačka ili akademska institucija odgovorna za prikupljanje podataka u toj državi. [16]

3.2. GEM podaci

Kao tri glavne karakteristike GEM istraživanja možemo navesti dostupnost baze prikupljenih podataka, specifičnu organizacijsku strukturu, te procjene na godišnjoj razini i izvješća o poduzetničkoj aktivnosti. Kao što sam već prije naveo kod GEM istraživanja su nam najkorisnija dva komplementarna alata Ispitivanje populacije odraslih (eng. Adult Population Surveys – *APS)* i Ispitivanje nacionalnih stručnjaka (eng. National Expert Survey – *NES)*. Kao rezultat Ispitivanja nacionalnih stručnjaka u bazu se zapisuju podaci koji nam govore o poduzetničkoj prirodi u državama Globalne poduzetničke mreže. U bazama podataka koje su nastale kao rezultat Ispitivanja populacije odraslih upisuju se podaci o poduzetničkim aktivnostima u zemljama kroz njihovu usporedbu. [17]

Kod prikupljanja GEM podataka GEM prati poduzetnički proces i mjeri različite komponente tog procesa. Pod komponente misli se na različite etape prilikom poduzetničkog poduhvata, na primjer od percepcije društvenih vrijednosti fokusiranih na poduzetništvo, percepcije o prilikama, poduzetničkih namjera, pa sve do poduzetničkog poduhvata u nastajanju i u ranoj fazi djelovanja. Svaki taj indikator poduzetničke aktivnosti može se analizirati s obzirom na određena obilježja (individualni atributi, industrija, utjecaj). Komponente poduzetničkog procesa i GEM operativne definicije možemo vidjeti na Slika 4. [18]



Slika 4. Poduzetnički proces i GEM operativne definicije [18]

Isto tako važno je i napomenuti kako GEM istraživanja razlikuju više kategorija poduzetničke aktivnosti. Kategorije koje razlikuju GEM istraživanja su: Početnici, Novi poduzetnici, Poduzetnici, Zaposlenici, Poduzetnici uključeni u ranu poduzetničku aktivnost. Prikaz tih kategorija i njihova obilježja možemo vidjeti na Slika 5. [18]

Poduzetnička aktivnost odrasle populacije, 18-64 godina starosti	TEA indeks, u % od odrasle populacije		„odrasla" poduzeća, u % od odrasle populacije	Poduzetnička aktivnost zaposlenika, u % od odrasle populacije*
	Početnici Sam ili sa drugima pokušava pokrenuti vlastiti posao, samozapošljavanje	**Novi poduzetnici** Vlasnik poduzeća/obrta, od 3 do 42 mjeseca starog	**Poduzetnici** Vlasnik poduzeća/obrta, starijeg od 42 mjeseca	**Zaposlenici** Zaposlenik razvija novi proizvod/uslugu za poduzeće u kojem radi, u protekle tri godine i sada je uključen u takve aktivnosti
	Poduzetnici uključeni u ranu poduzetničku aktivnost (iz uočene prilike vs. iz nužde)			

Slika 5. Kategorije poduzetničke aktivnosti [18]

Hrvatska se pridružila GEM istraživanjima 2002. godine. U Hrvatskoj ispitivanje populacije odraslih provodi Agencija za istraživanje tržišta IPSOS zajedno s GEM istraživačkim timom Hrvatske i globalnim GEM koordinacijskim timom. Ispitivanje nacionalnih stručnjaka provodi Centar za politiku razvoja malih i srednjih poduzeća i poduzetništvo CEPOR i GEM istraživački tim [18]. U daljnjem tekstu objasniti ću detaljnije što je to Ispitivanje populacije odraslih, a što Ispitivanje nacionalnih stručnjaka.

3.2.1. Ispitivanje populacije odraslih

Ispitivanje populacije odraslih jedinstveni je instrument koji koristi GEM za mjerenje razine i prirode poduzetničke aktivnosti u državama širom svijeta. Nacionalni timovi upravljaju ovim istraživanjem čiji se uzorak sastoji od najmanje 2000 ispitanika. Samo ispitivanje provodi se svake godine u isto vrijeme koje je uglavnom između travnja i lipnja. Ispitivanje obavlja neovisni dobavljač ankete kojeg je odabrao ekonomski GEM tim neke države. Ispitivač predaje prikupljene GEM podatke, koje zatim pregleda GEM središnji tim po nekoliko određenih kriterija. Sirovi podaci šalju se izravno GEM timu na pregled, provjeru kvalitete i kako bi se izradili izračuni prije nego se sami podaci objave javnosti. [19]

Cilj ovog ispitivanja je istražiti ulogu pojedinca u životnom ciklusu poduzetničkog procesa. Fokus istraživanja nije samo na općenite karakteristike poslovanja, nego i na ljudsku motivaciju za pokretanjem vlastitog poduzeća, radnje koje poduzima prilikom pokretanja poduzeća, kao i stavove pojedinca vezane uz poduzetništvo [19]. Do ispitanika u ovome istraživanju dolazi se jednim od sljedećih načina: kontaktom preko telefona ili mobitela; intervjuom oči u oči. Odabir jednog od ovih načina ovisio je o tome koliko je način komunikacije bio razvijen u određenim državama [17].

Ispitivanje populacije odraslih sastoji se od dvije glavne strukturne komponente: moduli i blokovi. Na početku istraživanje se sastoji od više različitih modula. Moduli su grupe pitanja koja se razlikuju po tome što su obavezna ili neobavezna i na temelju teme pitanja. Ti moduli se raspodjeljuju unutar upitnika u različitim blokovima. Blokovi se mogu razlikovati prema vrsti ispitanika kojima su pitanja namijenjena. Na primjer možemo reći da je modul koji sadrži pitanja vezana o poslovnim mrežama uključen u jedan blok pitanja koji je namijenjen vlasnicima tvrtki, a drugi blok je namijenjen početnicima. [19]

Temeljni modul ispitivanja populacije odraslih odnosi se na skup obaveznih pitanja koja se ne mijenjaju iz godine u godinu i ta pitanja moraju biti prisutna u svim anketama koje vrše GEM nacionalni timovi. Ta pitanja su raštrkana po blokovima a podaci dobiveni iz njih koriste se za dobivanje glavnih pokazatelja GEM-a kao što je stopa poduzetničke aktivnosti u ranoj fazi (TEA) o kojoj ću pisati kasnije. [19]

Kod ispitivanja populacije odraslih postoji i tako zvani „Special Topic“ modul pitanja koji se mijenja svake godine. Ta se pitanja usredotočuju na različite aspekte poduzetništva koji nisu pokriveni u temeljnom modulu ispitivanja populacije odraslih. Pitanja ovog tipa mogu biti uključena u jedan ili više blokova ovisno o vrsti ispitanika. Iako je ovaj modul obavezan za sve GEM nacionalne timove najčešće se u njemu postavljaju neobavezna pitanja. Isto tako, ukoliko želi, GEM nacionalni tim može uključiti skup neobaveznih pitanja kako bi proširili anketu. [19]

3.2.2. Ispitivanje nacionalnih stručnjaka

Od svojega osnutka GEM je predložio da se dinamika poduzetništva može povezati s uvjetima koji poboljšavaju ili otežavaju stvaranje novih poduzeća. U GEM metodologiji ti uvjeti se nazivaju Poduzetnički okvirni uvjeti (eng. Entrepreneurial Framework Conditions – *EFC*). Oni su jedna od najvažnijih sastavnica bilo kojeg poduzetničkog ekosustava. Mogu se smatrati jednim od najvažnijih razloga razumijevanja stvaranja i rasta poduzeća. Ovisno o stanju ovi uvjeti utječu na postojanje poduzetničke mogućnosti, sposobnosti i sklonosti što određuje dinamiku poslovanja. Očekivano je kako različite regije i gospodarstva imaju različitu kvalitetu EFC-a koji izravno utječe na ulazne i izlazne rezultate poduzetničkih aktivnosti. Ispitivanje nacionalnih stručnjaka služi nam za procjenu stanja EFC-a. [20]

Ispitivanje nacionalnih stručnjaka dio je standardnog GEM istraživanja i ono služi za procjenu EFC-a i ostalih tema koje se vežu uz poduzetništvo. Namjera ovog istraživanja je pribaviti stavove dodatnih stručnjaka o temama kao što su: podrška u ženskom poduzetništvu, poslovnom poticanju visokog rasta, nekim pitanjima vezanima za trenutni GEM ciklus istraživanja. Ispitivanje nacionalnih stručnjaka pokrenuto je zbog nedostatka usklađenih mjerila koja bi se mogla koristiti kao specifični EFC indeksi. Ovo ispitivanje je jedini izvor međunarodno usporedivih podataka koji se posebno bave čimbenicima koji povećavaju ili smanjuju poslovanje novih ili rastućih tvrtki. Može se reći kako je ovo ispitivanje slično drugim istraživanjima koja obuhvaćaju stručne prosudbe za procjenu specifičnih nacionalnih uvjeta, ali ovo se ispitivanje najviše razlikuje po tome što se usredotočuje samo na EFC. [20]

Ispitivanje nacionalnih stručnjaka koristi upitnik za prikupljanje mišljenja stručnjaka o različitim temama od koje je svaka dizajnirana tako da dotakne određenu dimenziju specifičnog EFC-a. Teme uključene u ispitivanje nacionalni stručnjaka su [20]:

- Poduzetničke financije – dostupnost financijskih sredstava
- Vladina politika – u kojoj mjeri politika podržava poduzetništvo
- Vladin program za poduzetništvo – prisutnost i kvaliteta programa koji pomažu malim i srednjim poduzetnicima
- Poduzetničko obrazovanje – do koje je mjere obrazovanje i osposobljavanje odgovorno za upravljanje malim i srednjim poduzećima
- Prijenos istraživanja i razvoja – u kojoj će mjeri nacionalna istraživanja i razvoj dovesti do novih poduzetničkih prilika u malim i srednjim poduzećima
- Komercijalna i pravna struktura – prisutnost pravnih, komercijalnih, računovodstvenih institucija koja promiču mala i srednja poduzeća
- Uredba o unosu – čine je dinamika tržišta i otvorenost tržišta
- Fizička infrastruktura – jednostavnost pristupa fizičkim resursima
- Kulturne i društvene norme – u kojoj mjeri društvo prihvaća nove poslovne metode koje mogu povećati prihode

Ispitivanje nacionalnih stručnjaka bira stručnjake koji će vršiti istraživanje na temelju ugleda i iskustva u određenom području. Svake godine bira se najmanje 36 stručnjaka za svako GEM gospodarstvo koji se osobno ispituju ili anketiraju. Nakon ispitivanja svi podaci, nacionalni i regionalni se središnje usklađuju. Proces usklađivanja podataka uključuje njihovu provjeru kvalitete i izračunavanje varijabli za određenu temu namijenjenu za mjerenje određenog aspekta EFC-a. Korištenjem ovog ispitivanja svaki odabrani stručnjak u svakoj državi dodjeljuje pojedinačne ocjene vrijednosti i tako omogućava međunarodne usporedbe sa ostalim stručnjacima. [20]

Ispitivanje nacionalnih stručnjaka koristi Likertovu ljestvicu kod anketiranja. Kod Likertove ljestvice, u ovom slučaju, 1 znači da stručnjak ponuđenu izjavu označava kao potpuno lažnu, a 5, 7 ili 9 (ovisno o verziji ljestvice) označuju da je izjava potpuno istinita. Također kod anketiranja stručnjaka traži se njihovo mišljenje o nekim, prema njima, najvažnijim institucionalnim uspjesima i ograničenjima za poticanje poduzetništva u njihovoj državi. Isto tako stručnjaci mogu davati i preporuke, a na kraju se bilježe informacije o anketiranom stručnjaku. [20]

Od 2015. godine Likertova ljestvica koja se koristi kod ispitivanja nacionalnih stručnjaka proširena je na 9 bodova. Postoje još ljestvice od 5 i 7 bodova. Ljestvica od 5 bodova koristi se od 2000. godine pa sve do danas. Ljestvica od 7 bodova služi nam u istraživačke svrhe jer omogućuje kombiniranje podataka GEM stručnjaka sa podacima stručnjaka Globalnog izvješća o konkurentnosti (eng. Global Competitiveness Report – *GCR*). Zadnja na redu je ljestvica od 9 bodova koja daje detaljniju sliku poduzetničkog okvira, te nam zbog toga omogućuje primjenu sofisticiranijih statističkih metoda koje zahtijevaju širenje podataka. [20]

3.2.3. Indeksi

Na temelju konceptualnih okvira GEM-a i prikupljenih podataka putem ispitivanja izračunavaju se brojni pokazatelji koji se uključuju u globalna i nacionalna izvješća. Tri osnovna GEM indeksa koja pružaju dobar uvid u stanje poduzetništva su [21]:

- Ukupna poduzetnička aktivnost u ranoj fazi (eng. Total Early-Stage Entrepreneurial Activity – *TEA*)
- Poduzetnička aktivnost zaposlenika (eng. Entrepreneurial Employee Activity - *EEA*)
- Društvena poduzetnička aktivnost (eng. Social Entrepreneurial Activity - *SEA*)

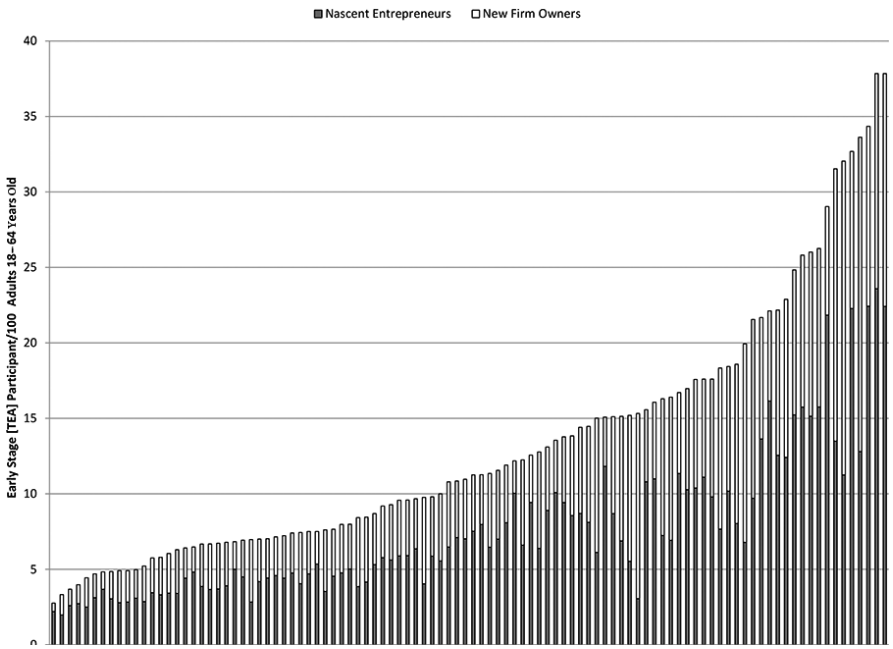
TEA indeks je pojam koji obuhvaća pojedince starosti između 18 i 64 godina koji tek počinju poduzetnički pothvat ili one koji se već neko vrijeme, najčešće do 42 mjeseca, bave poduzetništvom. Ovaj indeks može se dodatno poboljšati pružanjem informacija povezanih inkluzivnošću (spol, dob), utjecajem (rast, poslovanja, inovacije, internacionalizacije) i industrijom (sektori). TEA indeks biti će više objašnjen u daljnjem tekstu. EEA indeks je pojam koji nam govori o stopi uključenosti zaposlenika u poduzetničke aktivnosti. U te poduzetničke aktivnosti spadaju: razvoj, lansiranje nove robe ili usluge, pokretanje nove poslovnice, nova osnivanja. SEA indeks pokazuje nam stopu pojedinaca koji se bave poduzetničkim aktivnostima s društvenim ciljem [21]. U novije vrijeme pojavljuje se jedan novi indeks naziva Indeks poduzetničkog duha (eng. GEM's Entreprenurship Spirit Indeks - *GESI*). Sam indeks sastoji se od tri DA/NE pitanja. Ta tri pitanja sadrže teme percepcije prilika, poduzetničke samoučinkovitosti i svijesti, a za njih se koriste podaci dobiveni putem ispitivanja populacije odraslih [22].

Pošto je minimalna veličina uzorka kod ispitivanja 2000 ispitanika GEM je uveo TEA indeks kako bi se što više smanjila mogućnost pogreške. Svim ispitanicima koji se vode kao početnici ili vlasnici novih poduzeća dodjeljuje se broj 1, a svima ostalima 0. U Tablica 2. možemo vidjeti podatke koji se odnose za razdoblje od 2000. do 2014. godine. Ovi podaci temelje se na odgovorima preko 2 milijuna ispitanika iz 104 države i uključuju osobe od 18 do 64 godina. U Tablica 2. možemo vidjeti kako u globalnom prosjeku poduzetnici početnici čine 8,1 na 100 odraslih, vlasnici novih poduzeća 5,9 na 100 odraslih, a globalni prosjek TEA indeksa iznosi 13,6. [17]

Broj na 100 odraslih između 18 i 64 godina (Globalni prosjek)	
Poduzetnici početnici	8.1
Vlasnici novih poduzeća	5.9
TEA indeks	13.6
Vlasnici ostvarenih poduzeća	9.3

Tablica 2. Globalne mjere sudjelovanja u stvaranju poslovanja 2000. - 2014. [17]

Podaci pokazuju kako postoje određeni faktori koji utječu na razinu TEA indeksa. Također postoje i određenje činjenice koje su otkrivene, a vezane su za istraživanje, na primjer postoje varijacije između država u rasprostranjenosti stvaranja novih poduzeća, veza između ekonomskog rasta i stvaranja poslova, te kako razvijenije države stvaraju manje poslova. Razliku TEA indeksa između poduzetnika početnika i vlasnika novih poduzeća možemo vidjeti na Slika 6. [17]



Slika 6. Razlika TEA indeksa između poduzetnika početnika i vlasnika novih poduzeća [17]

Uspoređujući 2013. i 2014. godinu uočeno je postojanje visokog stupnja stabilnosti koja sve više postaje zajednička karakteristika promatranih država. Smatralo se kako zemlje s vrlo niskim ili vrlo visokim BDP-om imaju veću vjerojatnost stvaranja poslova. Ta tvrdnja dovodi se u pitanje jer se pokazalo da ne postoji „U“ oblik kada se radi o razvijenosti države i stvaranju posla i zaključeno je da siromašnije države imaju mnogo veći stupanj stvaranja poslova. [17]

4. Pregledni primjeri

U ovome poglavlju pisati ću o nekoliko primjera vezanih uz poduzetništvo koji su koristili neku od metoda rudarenja podataka u svrhu željenog istraživanja. Za svaki primjer ukratko ću navesti što se istraživalo, koji podaci su korišteni tijekom istraživanja, kojom metodom rudarenja podataka je izvršeno istraživanje i kakvi su bili rezultati istraživanja.

4.1. Klasifikacija poduzetničkih namjera studenata

Kako bi studentima omogućili što bolju obrazovnu podlogu koja će podržati njihove poduzetničke namjere potrebno je u početku takve poduzetničke namjere i prepoznati. Cilj ovoga istraživanja bio je razviti model koji će klasificirati studente prema poduzetničkim namjerama koristeći metode neuronskih mreža, rudarenja podataka i strojeva za podršku vektora. Istraživanje je provedeno na hrvatskom sveučilištu, a za uzorak su uzeti studenti prve godine studija. [23]

Veliki broj istraživača usredotočen je na faktore koji određuju izbor karijere učenika i na faktore koji objašnjavaju njihovu neusklađenost stavova i namjera. Istraživači Luethje i Franke su napravili model koji je za uzorak imao studente strojarstva. Tim modelom pokazali su da je poduzetnički stav studenata usko povezan s namjerom da se pokrene novi pothvat, ali i da osobnost studenata nije izravno povezana s poduzetničkim namjerama. Zatim, Krueger je testirao dva najpopularnija modela namjere (Ajzenova teorija planiranog ponašanja i Shaperov model poduzetničkog događaja) i pronašao snažnu statističku podršku za oba modela. U ovom istraživanju ta dva modela se kombiniraju kod izbora ulaznih varijabli. U prethodnim istraživanjima većinom su korištene metode višestruke regresije i strukturnog modeliranja, dok se u ovom koriste metode strojnog učenja: neuronske mreže, stabla odlučivanja i strojevi za podršku vektora. [23]

U ovome istraživanju koriste se tri metode rudarenja podataka: neuronske mreže, stabla odlučivanja i strojevi za podršku vektora. Pošto sam u ovome radu objasnio neuronske mreže i stabla odlučivanja, u ovome odlomku pokušati ću ukratko objasniti strojeve za podršku vektora. To je metoda strojnog učenja koja se koristi za nelinearno preslikavanje ulaznih vektora u prostor s velikim dimenzijama. Ova metoda proizvodi binarni klasifikator. Osnovno načelo ove metode je da traži optimalni rezultat koji udovoljava zahtjevu za klasifikaciju. Nakon toga koristi algoritam za izradu granica uz istovremeno osiguravanje točnosti ispravne klasifikacije. [23]

Uzorak podataka koji se koristio za ovo istraživanje sastoji se od 237 redovnih studenata poslovne administracije koji su pohađali prvu godinu studija na Sveučilištu J.J. Strossmayera u Osijeku, Hrvatska. Samo istraživanje provedeno je u srpnju 2010. godine. Kod izbora ulaznih varijabli u ovome istraživanju autori su se vodili prema dva svjetski priznata modela, a to su: Ajzenova teorija planiranog ponašanja i Shaperov model poduzetničkog događaja. Osim podataka o studentu korišteni su i neki drugi prediktori koji su grupirani u tri skupine: očekivani poduzetnički ishod, društvene norme i uvjerenja o poduzetništvu i poduzetničkoj karijeri. U modelu 1 korišteno je 46 ulaznih varijabli, dok je u modelu 2 korišteno njih 17. Četiri ulazne varijable bile su kategoričke, dok se ostatak ulaznih varijabli ocjenjivao ocjenama od 1 do 5. U svrhu treniranja modela, skup podataka podijeljen je na podatke za treniranje i podatke za testiranje. Izlazni sloj svih modela sastojao se od dvije vrijednosti, vrijednost 1 za postojanje namjera bavljenja poduzetništvom i 0 za nepostojanjem poduzetničke namjere. Kako bi osigurali jednaku vjerojatnost za obje izlazne vrijednosti koristi se slojevito uzorkovanje. Ono nam koristi kako bi sačuvali jednaku raspodjelu studenata s pozitivnim i negativnim namjerama u uzorku za treniranje. Struktura uzoraka u ovome istraživanju prikazana je u Tablica 3. [23]

Uzorak	**1 (postojanje poduzetničkih namjera)**	**0 (ne postojanje poduzetničkih namjera**	**Ukupno**	**%**
Treniranje	69	69	138	58.23
Testiranje	82	17	99	41.77
Ukupno	151	86	237	100.00

Tablica 3. Struktura uzoraka [23]

Cilj ovoga istraživanja bio je analizirati uspješnost tri metode strojnog učenja u modeliranju poduzetničkih namjera studenata. Sve tri korištene metode upotrijebljene su na istom uzorku podataka. Rezultatima ovog istraživanja pokazano je kako je model s najvećom preciznošću u klasifikaciji model neuronskih mreža, iako se njegova brzina pogodaka rezultata ne razlikuje previše od ostala dva modela. Rezultati pokazuju kako model neuronskih mreža pokazuje najveću osjetljivost, dok je model stabla odlučivanja najspecifičniji u prepoznavanju pozitivnih poduzetničkih namjera. Model stabla odlučivanja mogao bi se i koristiti za potvrđivanje rezultata jer je uspješniji u prepoznavanju stvarnih pozitivnih učenika. Rezultati ovog istraživanja ograničeni su samo za promatrani skup podataka, te se na njima trebaju provesti daljnja testiranja kako bi se mogli generalizirati. [23]

4.2. Primjena rudarenja podataka u poduzetničkoj analizi

Kreativno poduzetništvo smatra se jednim od najvažnijih faktora gospodarskog razvoja, pogotovo u društvu koje se temelji na znanju. Kod procesa poduzetničkog razvoja temeljnu ulogu imaju sveučilišta i poduzetnički ekosustav. U ovome istraživanju korišteno je rudarenje podataka kako bi se poboljšalo odlučivanje kod poduzetničkog menadžmenta, a podaci korišteni u istraživanju su podaci CCEEmprende projekta. Ovo istraživanje koristi metode rudarenja podataka: pravila asocijacije, stabla odlučivanja i logičku regresiju. Glavni cilj ovog istraživanja je predviđanje uspjeha i identificiranje najvažnijih čimbenika koji su povezani s uspjehom ili neuspjehom projekata u poduzetništvu. [24]

U ovome istraživanju koriste se nadzirane metode rudarenja podataka. Pod nadzirane metode rudarenja misli se na klasifikacijske metode, a ovom slučaju koriste se metode: pravila asocijacije, stabla odlučivanja i logička regresija. Pošto sam već objasnio stabla odlučivanja i pravila asocijacije u daljnjemu tekstu ukratko ću objasniti metodu logičke regresije. Metoda logičke regresije koristi se u situacijama u kojima je potrebno predvidjeti prisutnost ili odsutnost nekih karakteristika ili ishoda na temelju vrijednosti varijabli predviđanja. Metoda logičke regresije slična je metodi linearne regresije, ali je uvjetovana modelima gdje je ovisna varijabla binarna. Koeficijenti logičke regresije mogu se koristiti za procjenu neparnih omjera za svaku neovisnu varijablu u modelu. Logička regresija pomaže u izradi veza više opcionalne regresije između ovisne i nekih neovisnih varijabli. [24]

Korištenjem rudarenja podataka i statističkih metoda u ovome istraživanju prikazani su najvažniji čimbenici koji se odnose na uspjeh poduzetničkih projekata na sveučilištu. Podaci koji se koriste u ovom istraživanju prikupljeni su na Ekonomskom fakultetu (FCCEEA) Univerziteta Republike (spa. *Facultad de Ciencias Economicas y de Administracion (FCCEEA) of the Universidad de la Republica*) programom CCEEmprende. Podaci su prikupljeni putem ankete o svim polaznicima programa CCEEmprende Entrepreneurs Program. Obuhvaćao je 63 poduzetnika od kojih je 41% žena i 59% muškaraca, a sudjelovali su u programu u razdoblju od 2007. do 2010. godine. [24]

Samo istraživanje obuhvaća sljedeće podatke: sociodemografske karakteristike poduzetnika, razlog pokretanja poduzeća, različite vrste podrške (obiteljska, moralna, ekonomska), razlozi za polaženjem CCEmprende programa, financiranje projekta i pokazatelj uspjeha. [24]

Metodom stabla odlučivanja u ovome istraživanju pokazalo se kako je najveća značajka kod otvaranja novoga poduzeća mogućnost financiranja i kako većina poduzeća ne može opstati bez toga. U rezultatima istraživanja korištenjem metode asocijacije došlo je do zaključka kako su mogućnost financiranja i neovisna radna situacija najvažniji sporedni slučajevi kod pokretanja poduzeća i kako ćemo, ukoliko to zadovoljavamo, vjerojatno otvoriti poduzeće. S druge strane ukoliko nemamo novaca za financiranje poduzeća i ukoliko imamo ovisnu radnu situaciju istraživanje je pokazalo da ćemo teško otvoriti poduzeće. Model logičke regresije u ovome istraživanju govori nam da je kod pokretanja poduzeća najvažnije imati izvor financija i već postojeća zaposlenost kao poduzetnik. [24]

Prema dobivenim rezultatima možemo zaključiti kako rezultati ovog istraživanja izdvajaju dva čimbenika o kojima ovisi uspješnost poduzetničkog projekta, a ta dva čimbenika su: poduzetnikovo financiranje i poduzetnikov prethodni status zaposlenja. Ovo istraživanje otvorilo je mogućnosti za buduće radove kao što su razmatranje drugih načina uspješnosti projekta, primjene ostalih metoda rudarenja podataka, te proširenje rada na druge skupine poduzetnika i poduzeća. [24]

4.3. Upotreba rudarenja podataka u menadžmentu

U današnje vrijeme kada postoji puno konkurentnih poduzeća i organizacija, promjene koje se događaju u poslovnome okruženju čine strateški menadžment učinkovitim oblikom menadžmenta u poduzećima. Strateški menadžment je suvremena strategija i ona zahtjeva postavljanje određenih strategija i planova, te njihovo izvršavanje kako bi poduzeća mogla postizati ciljeve. Rudarenje podataka pokazalo se jako dobrom metodom kod planiranja aktivnosti strateškog menadžmenta u poslovnom procesu. U ovome istraživanju prikazane su discipline strateškog menadžmenta i metode rudarenja podataka koje se upotrebljavaju kod strateškog menadžmenta. [25]

U ovome istraživanju kod rudarenja podataka korištena su pravila asocijacije, a algoritmi pravila asocijacije koji su se koristili su: apriori, prediktivni apriori i tertius algoritam. Metodu rudarenja podataka pravila asocijacije objasnio sam ranije u radu, ali nisam ulazio detaljno u algoritme pravila asocijacije stoga će oni biti objašnjeni u daljnjem tekstu. Apriori algoritam je jedan od najpopularnijih algoritama koji se koristi kod pravila asocijacije. Ovaj algoritam pronalazi skupove podataka čija potpora nije manja od minimalnog praga potpore. Glavna svrha ovoga algoritma temelji se na pretraživanju po razinama k skupova podataka od kojih se dobivaju k + 1 skupovi podataka. Jedna od važnijih karakteristika ovog algoritma je ta da ukoliko neki skup podataka nije frekventan, nisu frekventni ni njegovi nadskupovi. [25]

Prediktivni apriori algoritam temelji se na apriori algoritmu. Ovaj algoritam pretražuje najbolja n pravila koja imaju veze s ispravljenom vrijednošću pouzdanosti, a koje se temelje na potpori. Cilj ovog algoritma je maksimalno povećati očekivanu točnost pravila asocijacije. Kod rangiranja rezultata prediktivni apriori algoritam u obzir uzima pouzdanost, potporu i prediktivne mjere točnosti. Tertius algoritam je algoritam pravila asocijacije koji traži rečenice s najvećom vrijednosti poklapanja. Ovaj algoritam se koristi logikom prvoga preda, a skeniranje same baze ovisi o broju rečenica u pravilima. Zbog toga izvršavanje ovog algoritma je relativno dugo. [25]

Skup podataka koji je korišten u ovome istraživanju dobiven je na Sveučilištu Celala Bayara. U tome skupu podataka nalazilo se 692 studenta preddiplomskih studija smjera Poslovne administracije. Od ukupnog broja svih studenata koji su sudjelovali u istraživanju bilo je 349 muškaraca i 343 žene. Glavni atributi koji su korišteni u skupu podataka su: klasa, dob, spol, nacionalnost, mjesto rođenja, obiteljsko i bračno stanje, posljednji stupanj obrazovanja majke i oca, broj braće i sestara, vrsta studija, primanja i status roditelja. Kako bi se na ovom skupu podataka mogla primijeniti pravila asocijacije potrebno je obraditi skup pomoću konverzijskog filtra koji pretvara podatke numeričkog tipa u nominalni oblik. Kako bi se pronašla jaka asocijativna pravila svi studenti podijeljeni su u dvije skupine ovisno o ocjenama studenata (skupina uspješnih i skupina neuspješnih studenata). [25]

U rezultatima istraživanja u kojem je korišten apriori algoritam najbolje je utvrđeno pravilo koje govori da su studenti u dobi od 19 godina koji su redovno upisali studijski program na prvoj godini preddiplomskog studija. Kod rezultata prediktivnog apriori algoritma utvrđene su poveznice između sljedeća četiri uvjeta: klasa, status uspješnosti, status primanja i posljednji stupanj obrazovanja majke. Rezultati tertius algoritma nisu tako čvrsti kao kod apriori i prediktivnog apriori algoritma. Sva pravila koja su dobivena tertius algoritmom upućuju na to da studenti s niskim obiteljskim primanjima i s majkama koje imaju niži stupanj obrazovanja, također imaju i očeve nižeg stupnja obrazovanja. [25]

Potreba rudarenja podatak u strateškom menadžmentu je neizbježna. U ovome istraživanju prikazana je jedna od metoda rudarenja podataka u obrazovanju na podacima o studentima koji su prikupljeni sa informacijskog sustava Sveučilišta Celala Bayara. Rezultati ovoga istraživanja pokazuju kako postoji jaka veza između obrazovanja roditelja, obiteljskih primanja i uspješnosti studiranja. Prema tim rezultatima strateški menadžment bi trebao isplanirati da već na prvoj godini studija organizira savjetovanja za studente za koje se očekuje da će biti neuspješni. [25]

5. Zaključak

U ovome radu u tri glavna poglavlja opisao sam rudarenje podataka, Global Entrepreneurship Monitor i nekoliko preglednih primjera. U prvome poglavlju objasnio sam rudarenje podataka, njegovu definiciju, koje su mu zadaće, proces samog rudarenja podataka i neke od metoda. Nakon toga u poglavlju Global Entrepreneurship Monitor opisao sam kako je nastala GEM inicijativa i njenu povijest, GEM podatke i kako se oni prikupljaju i koriste, te indekse koje GEM koristi. U zadnjem poglavlju analizirao sam par primjera koji su koristili rudarenje podataka u poduzetništvu i područjima vezanima za poduzetništvo.

Mogu zaključiti kako je rudarenje podatak sve više prisutno u poduzetništvu zbog velikog gomilanja podataka koje se stvara u poduzećima. U tom velikom skupu podataka sama poduzeća mogu naći neke vrlo korisne informacije za samo poslovanje i samim time poduzeća u tim podacima vide veliku korist. Potrebno je navesti kako poduzeća koja ne rudare vlastite podatke i na taj način vrše analizu zaostaju za konkurencijom koja uspješno iskorištava znanja stečena iz rudarenja podataka. Isto tako bitno je da podaci koje poduzeće analizira budu kvalitetni, to jest korisni za poduzeće. Također za kvalitetno rudarenje podataka potrebno je izabrati pravilnu metodu rudarenja podataka za određeni skup podataka. Tu u fokus dolaze stručnjaci za rudarenje podataka koji su sve više traženi. Ukoliko poduzeće pravilno rudari dostupne podatke i uspješno ih analizira i primjenjuje u poslovanju tada ono može očekivati pozitivne rezultate razvoja i poslovanja.

Popis literature

[1] M. I. Štemberger, J. Jaklič, M. Varga, *Informacijska tehnologija u poslovanju*, Zagreb: Element, 2004.

[2] Ž. Panian, G. Klepac, *Poslovna inteligencija*, Zagreb: Masmedia, 2003.

[3] Ž. Garača, M. Jadrić, *Rudarenje podataka: različiti aspekti informacijskog društva*, Split: Ekonomski fakultet, 2011.

[4] M. Berry, G. Linoff, *Data mining tehniques, For marketing, sales, and customer relationship managment*, Indiana: Wiley Publishing Inc, 2004.

[5] D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley-Interscience, 2004.

[6] M. Policki, *Proces data mininga nad podacima o prodaji tekstila*, Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike, 2009. Preuzeto 25. srpnja 2019. sa: https://urn.nsk.hr/urn:nbn:hr:211:086283

[7] M. Pejić Bach, „Rudarenje podataka u bankarstvu“, *Zbornik Ekonomskog fakulteta u Zagrebu,* Vol. 3, No. 1, str. 181 – 193, 2005. Preuzeto 23. srpnja 2019. sa: https://hrcak.srce.hr/26220

[8] D. Pyle, *Data Preparation for Data Mining*, San Franciso: Morgan Kaufmann, 1999.

[9] M. J. A. Berry, G.S. Linoff, *Mastering Data Mining*, Chichester: Wiley, 2000.

[10] KDnuggets, *Data Mining / Analytic Methods,* 2007., Preuzeto 30. srpnja 2019. sa: https://www.kdnuggets.com/polls/2007/data_mining_methods.htm

[11] G. Klepac, *Primjena inteligentnih računalnih metoda u managementu,* Zagreb: Sinergija-nakladništvo d.o.o., 2001.

[12] F. Ujević, „*Postupci analize podataka u izgradnji profila korisnika usluga*“. Magistarski rad. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, 2004. Preuzeto 30. srpnja 2019. sa: http://www.maturskiradovi.net/forum/attachment.php?aid=1442

[13] Global Entrepreneurship Research i London Business School, *ORIGINS,* Preuzeto 08. kolovoza 2019. sa: https://gemconsortium.org/wiki/1140

[14] Global Entrepreneurship Research i London Business School, *WHAT IS GEM?,* Preuzeto 08. kolovoza 2019. sa: https://gemconsortium.org/about/news

[15] Global Entrepreneurship Research i London Business School, *WHY IS GEM UNIQUE?,* Preuzeto 08. kolovoza 2019. sa: https://gemconsortium.org/about/news

[16] Global Entrepreneurship Research i London Business School, *HOW DOES GEM WORK?,* Preuzeto 08. kolovoza 2019. sa: https://gemconsortium.org/about/news

[17] P. D. Reynolds, „*Global Entrepreneurship Monitor (GEM) Program: Development, Focus, and Impact“*, Prosinac 2017., Preuzeto 08. kolovoza 2019. sa: https://oxfordre.com/business/view/10.1093/acrefore/9780190224851.001.0001/acrefore-9780190224851-e-156

[18] S. Singer, N. Šarlija, S. Pfeifer, S. O. Peterka, „*Što čini Hrvatsku (ne)poduzetničkom zemljom?*“, Zagreb: CEPOR, 2018. Preuzeto 08. kolovoza 2019. sa: http://www.cepor.hr/wp-content/uploads/2018/05/GEM-2017-za-web-FINAL.pdf

[19] Global Entrepreneurship Research i London Business School, *WHAT IS THE ADULT POPULATION SURVEY (APS)?,* Preuzeto 12. kolovoza 2019. sa: https://www.gemconsortium.org/wiki/1141

[20] Global Entrepreneurship Research i London Business School, *WHAT IS THE NATIONAL EXPERT SURVEY (NES)?,* Preuzeto 12. kolovoza 2019. sa: https://www.gemconsortium.org/wiki/1142

[21] S. Singer, J. E. Amorós, D. M. Arreola, „*GLOBAL ENTREPRENEURSHIP MONITOR 2014 GLOBAL REPORT*“, 2015. Preuzeto 14. kolovoza 2019. sa: https://www.gemconsortium.org/report

[22] I. Naletina, „Rudarenje podataka u poduzetništvu“, Završni rad. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin 2018., Preuzeto 14. kolovoza 2019. sa: https://urn.nsk.hr/urn:nbn:hr:211:310012

[23] M. Zekić-Sušac, S. Pfeifer, I. Đurđević, „*Classification of entrepreneurial intentions by neural networks, decision trees and support vector machines*“, Croatian Operational Research Review Vol. 1, 2010. Preuzeto 20. kolovoza 2019. sa: https://hrcak.srce.hr/93438

[24] E. Hochsztain, A. Tasistro, M. Messina, „*Data Mining applications in entrepreneurship analysis*“, International Workshop on Data Mining with Industrial Applications, str. 25 – 29, 2015.

[25] A. Onan, V. Bal, B. Y. Bayam, „*The Use of Data Mining for Strategic Management: A Case Study on Mining Association Rules in Student Information System*“, Croatian Journal of Education, Vol.18; No.1, str. 41 – 70, 2016. Preuzeto 22. kolovoza 2019. sa: https://hrcak.srce.hr/155471

Popis slika

Popis tablica