Primjena tehnika strojnog učenja na otvorenim podacima

Pranjić, Anabela

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Anabela Pranjić

PRIMJENA TEHNIKA STROJNOG UČENJA NA OTVORENIM PODACIMA

ZAVRŠNI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Anabela Pranjić

Matični broj: 45869/17-R

Studij: Informacijski sustavi

PRIMJENA TEHNIKA STROJNOG UČENJA NA OTVORENIM PODACIMA

ZAVRŠNI RAD

Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, kolovoz 2020.

Anabela Pranjić

Izjava o izvornosti

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

Autorica potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Tema završnog rada je primjena tehnika strojnog učenja na otvorenim podacima. Odabran je skup podataka vezan uz prodaju nekretnina te se na navedenom skupu provodi deskriptivno i prediktivno modeliranje primjenom tehnika nadgledanog i nenadgledanog učenja. Cilj ovog rada je pronaći čimbenike koji utječu na cijenu nekretnine i u kojoj mjeri kako bi izradili dobre modele prema kojima će se moći postaviti njezina cijena. Nastojimo potvrditi povezanost cijene nekretnine sa kvadraturom unutarnjeg prostora, kvalitetom građevine i brojem prostorija. Prvi dio rada uključuje uvod u temu, teorijsku podlogu teme, opis problema i pregled rezultata sličnih istraživanja. U praktičnom dijelu izrađeni su deskriptivni modeli primjenom klasterske i faktorske analize te prediktivni modeli stabla odlučivanja i neuronske mreže. Izradom modela potvrdili smo jaku pozitivnu povezanost cijene nekretnine sa kvadraturom i kvalitetom gradnje te nešto slabiju pozitivnu povezanost cijene sa brojem prostorija. Pregledom istraživanja koji se bave sličnom tematikom, zaključili smo kako na cijenu utječu brojni faktori među kojima su i oni koje smo utvrdili kao čimbenike jakog utjecaja na cijenu nekretnine.

Ključne riječi: rudarenje podataka, strojno učenje, otvoreni podaci, nekretnine, prodaja

Sadržaj

1. Uvod

Napredak tehnologije te sve veća primjena računala u svim područjima ljudske djelatnosti, stvaranje baza podataka i povezivanje računala i drugih uređaja internetom uzrokuju rast količine podataka koja se svakodnevno stvara. [1] Prikupljanje, sortiranje i analiziranje velike količine podataka može biti poprilično zahtjevan posao te ga je nemoguće obaviti bez primjene računala. [1] Tehnike strojnog učenja uvelike olakšavaju posao analize i izdvajanja korisnih informacija iz velike količine podataka.

Tema završnog rada odnosi se na analizu skupa podataka o prodaji nekretnina na području King County države Washington, SAD u periodu od svibnja 2014. do svibnja 2015. godine. Podaci su preuzeti sa stranice *kaggle*. Temelj rada je analiza podataka, izrada deskriptivnih i prediktivnih modela, njihova evaluacija te usporedba rezultata sa drugim istraživanjima vezanim uz navedenu temu. Cilj je utvrditi koji faktori utječu na prodaju nekretnina te dati smjernice koje vode uspješnijoj prodaji. Za izradu modela korišten je alat BigML.

Rad je podijeljen na 7 glavnih poglavlja. U prvom poglavlju naglasak je na uvod u temu te njezino predstavljanje. Drugo poglavlje se odnosi na predstavljanje pojmova vezanih uz strojno učenje, tehnike strojnog učenja i primjene. Treće poglavlje je posvećeno nešto detaljnijem opisu problema, proučavanju trendova prodaje nekretnina, postavljanju hipoteze i usporedbi sa prethodnim istraživanjima kako bi se izdvojili slični zaključci. U četvrtom poglavlju su opisani podaci i postupak pripreme podataka uz opise atributa, tablice skupa podataka uz navedene tipove, vrijednosti, distribucije te utvrđivanje povezanosti atributa skupa podataka. Peto poglavlje se odnosi na modeliranje podataka – izradu deskriptivnih i prediktivnih modela. Deskriptivno modeliranje se sastoji od potpoglavlja vezanih uz klastersku i faktorsku analizu, a prediktivno od potpoglavlja primjene stabla odlučivanja i neuronske mreže. U šestom poglavlju dana je evaluacija modela podataka, njihova interpretacija, smjernice za poboljšanje i zaključci. U sedmom, zaključnom poglavlju, iznose se konačni rezultati rada i sumira istraživanje. Slijede poglavlja vezana uz popis literature, slika i tablica.

Kao što je navedeno, skup podataka korišten u radu je preuzet sa stranice kaggle.com pod nazivom „House Sales in King County, USA“ – skup podataka o nekretninama prodanih između svibnja 2014. i svibnja 2015.

2. Metodologija

Data mining, odnosno dubinska analiza podataka je primjena računarskih postupaka i alata koja ima veliku ulogu u analizi podataka te teži otkrivanju potencijalno korisnih informacija iz velikih skupova podataka. [1] Dubinska analiza podataka uključuje nekoliko faza – priprema podataka, analiza podataka, interpretacija rezultata. [1] Tijekom faza pripreme podataka i postupka analize nužno je sudjelovanje čovjeka ili još bolje eksperta određenog područja koji razumije značenje ulaznih podataka te može protumačiti dobivene rezultate. [1] Količina i kvaliteta ulaznih podataka ima velik utjecaj na kvalitetu rezultata dubinske analize. [1] Zbog toga je sam proces pripreme podataka iznimno važan te je vremenski najzahtjevniji dio dubinske analize podataka. [1]

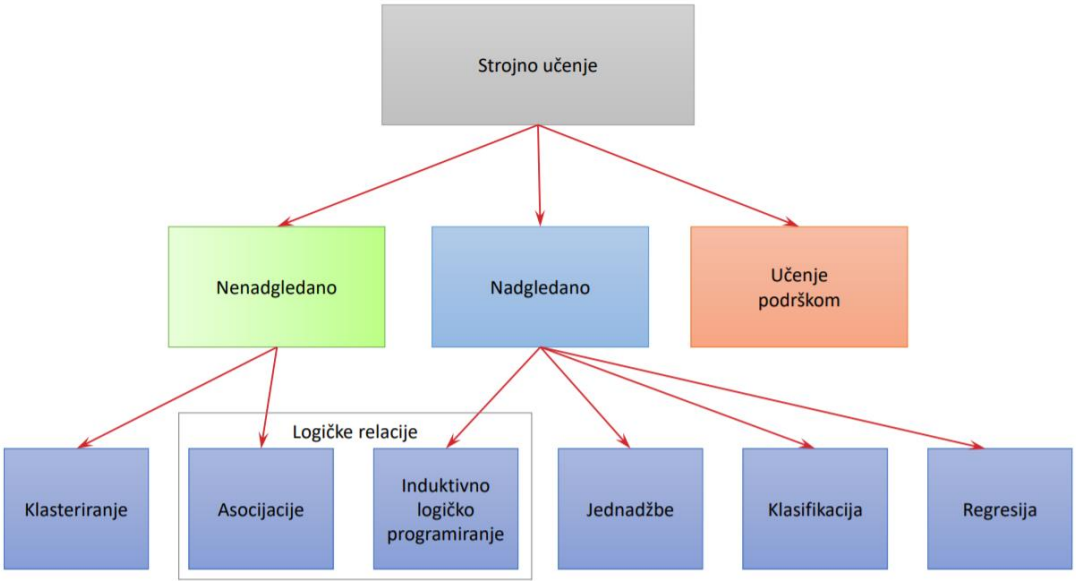
Strojno učenje (engl. *Machine learning*) jedna je od temeljnih tehnika dubinske analize podataka kojom se dubinska analiza bitno razlikuje od tradicionalne analize podataka. [1] Pripada području umjetne inteligencije te omogućava softverskim aplikacijama preciznije predviđanje rezultata bez eksplicitnog programiranja. [2] Ono omogućava analizu velikih količina podataka te daje brže i preciznije rezultate za identifikaciju mogućih profitabilnih mogućnosti ili opasnih rizika. [3] Kombinacija strojnog učenja i umjetne inteligencije sa kognitivnim tehnologijama može rezultirati učinkovitijom obradom velike količine podataka. [3] Algoritmi strojnog učenja koriste povijesne podatke za ulazne vrijednosti kako bi pomoću njih predvidjeli izlazne rezultate. [2] Rezultat strojnog učenja je najčešće model kojeg je moguće primijeniti sa ciljem poboljšanja sustava. [1] Uspješnost postupaka strojnog učenja mjeri se prema prediktivnoj točnosti konstruiranih modela nad onim podacima koji nisu korišteni u procesu učenja. [1] Nužan uvjet primjene rezultata strojnog učenja u postupku otkrivanja znanja je da čovjek na kraju te rezultate može razumjeti i interpretirati. [1]

Proces strojnog učenja započinje promatranjem i analiziranjem podataka, uočavanjem uzoraka, proučavanjem povijesnih podataka kako bi se donijele bolje odluke. [3] Primarni cilj je automatsko učenje računala bez ljudske intervencije. [3]

Tehnike strojnog učenja možemo podijeliti u nekoliko kategorija ovisno o primjeni algoritma i tipu podataka koji se nastoji predvidjeti. [2] Na slici 1 prikazana je moguća podjela tehnika strojnog učenja na 3 kategorije:

- nenadgledano učenje (eng. *Unsupervised learning*),
- nadgledano učenje (eng. *Supervised learning*) te
- učenje podrškom (eng. *Reinforcement learning*).

Postoje i ostala učenja: polu-nadzirano učenje, transduktivno učenje, relacijsko učenje, genetsko programiranje i dr. [4]



Slika 1: Taksonomija metoda strojnog učenja [5]

Nenadgledano učenje se koristi se kada su dani podaci bez ciljne vrijednosti (neoznačeni podaci). [3] Proučava kako sustav može definirati funkciju za opisivanje skrivene strukture iz neoznačenih podataka. [3] Sustav ne utvrđuje pravi izlaz već istražuje podatke te izvlači zaključke iz skupa podataka za opisivanje skrivenih struktura iz neobilježenih podataka. [3] Nenadgledano učenje primjenjuje se u raznim područjima poput: [4]

- marketing - grupiranje korisnika prema potrošačkim navikama
- biologija - grupiranje biljaka ili životinja prema sličnim značajkama
- rudarenje teksta (eng. *Text mining*) - grupiranje sličnih dokumenata
- pretraživanje informacija i grupiranje sličnih rezultata
- bioinformatika – grupiranje DNA-mikropolja
- obrada slike – sažimanje slike grupiranjem istobojnih piksela

Naziv nenadgledanog učenja je nastao zbog toga što se odnosi na vrstu učenja u kojem nema točnih odgovora te nema „učitelja“ koji bi nadgledao proces i pokazao ispravno znanje. [6] Zadaća algoritama nenadgledanog učenja je da pronađu zanimljive podatkovne strukture. [6] Cilj nenadgledanog učenja je grupirati podatke, otkriti strukturne pravilnosti među podacima te odrediti njihovo značenje. [4] Neki od algoritama nenadgledanog učenja su: [4]

- klaster analiza (eng. *Clustering*)

- otkrivanje anomalija (eng. *Outlier detection*)
- otkrivanje asocijacijskih pravila
- redukcija dimenzionalnosti podataka

Nadgledano učenje sadrži ulaznu varijablu (x) i izlaznu varijablu (y) te koristi algoritam za učenje kojim definira funkciju za izradu prediktivnih modela. [6] Cilj nadgledanog učenja je izraditi model za izradu predikcija na još neviđenim primjerima. [4] Naziv ovakvog učenja povezujemo s „učiteljem“ koji nadgleda proces učenja u kojem se koristi skup podataka za treniranje. [6] Proces učenja može završiti kada algoritam dosegne prihvatljivu razinu performansi i pouzdanosti. [6] Nadgledano učenje ima razne primjene poput: [4]

- predviđanje – na temelju ulaznih varijabli
- otkrivanje ekstremnih vrijednosti – iznimke
- ekstrakcija znanja – učenje modela koji se lako interpretiraju
- upravljanje – upravljački ulazi dobiveni kao izlaz regresije

Kod nadziranog učenja podaci su u obliku (ulaz, izlaz) = (x, y) te je potrebno pronaći preslikavanje y' = f(x). [7] Ako je izlazna varijabla diskretna ili nebrojčana vrijednost, radi se o klasifikaciji, a ako je kontinuirana ili brojčana vrijednost o regresiji. [7]

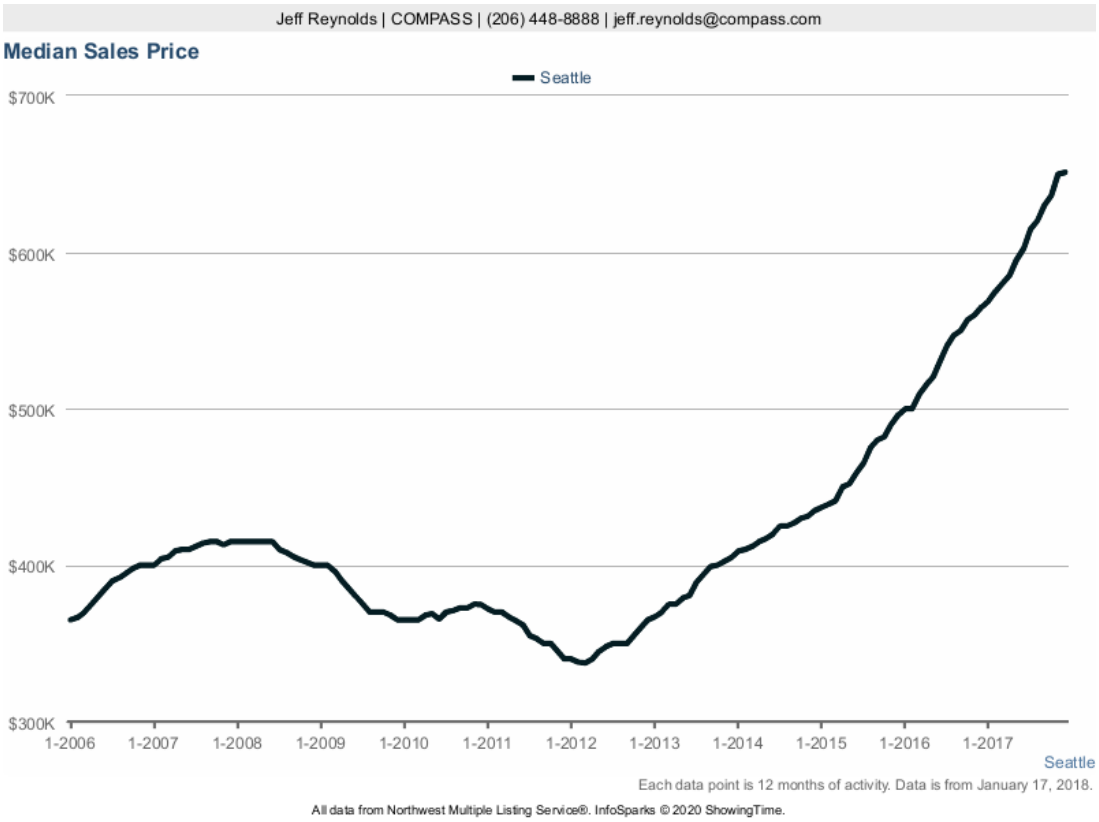
Učenje podrškom odnosi se na vrstu učenja koja je u interakciji s okolinom te proizvodi akcije i otkriva pogreške ili nagrade. [3] Ovakvo učenje omogućava računalima automatsko određivanje idealnog ponašanja u određenom kontekstu za maksimalizaciju performansi. [3] Zahtjeva povratne informacije za učenje onih akcija koje su dobre. To je poznato kao signal pojačanja. [3] Primjenjuje se u sljedećim područjima: [4]

- robotika
- igranje igara
- autonomna navigacija

3. Opis problema

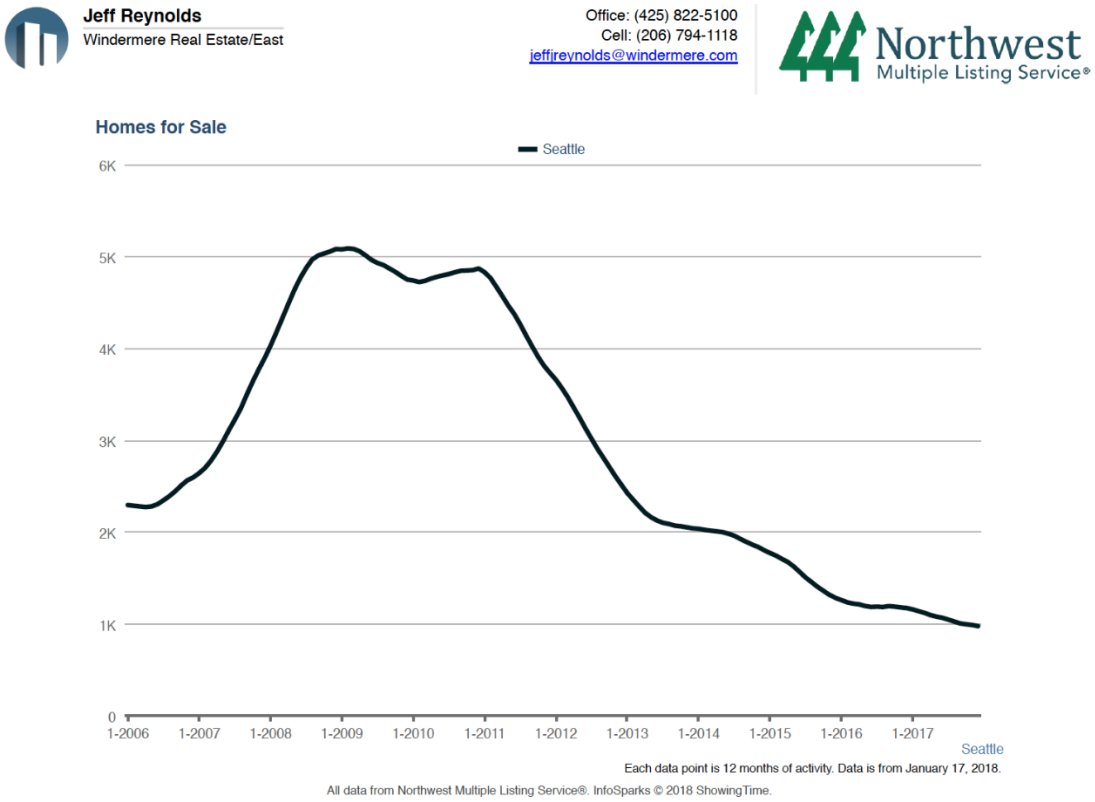
Podaci korišteni tijekom izrade rada odnose se na cijene nekretnina sa područja King County države Washington u Sjedinjenim Američkim Državama. Prema podacima iz 2020. godine, okrug King County broji otprilike 2.26 milijuna stanovnika što ga čini najnaseljenijim okrugom u državi Washington. [8] Grad Seattle je sjedište okruga te ujedno njegov najveći i najnaseljeniji grad. [9] Otprilike dvije trećine stanovništva okruga naseljava predgrađa grada Seattle. [9]

Zanimljivo je pratiti porast medijana cijene prodaje nekretnine te pad broja nekretnina dostupnih za prodaju tijekom perioda od 2006. do 2017. godine. Sljedeći dijagrami prikazuju ogromne razlike u cijenama nekretnina i broju dostupnih domova – jedanaestogodišnju povijest prodaje nekretnina na području grada Seattle.



Slika 2: Medijan cijene prodaje [10]

Slika 2 prikazuje medijan prodajne cijene nekretnine za svaku godinu od 2006. do 2017. godine. Uočava se nagli porast cijena od 2012. godine kada za nekretninu cijene od 337500$, vrijednost u 2017. godini iznosi 651000$. [10]



Slika 3: Broj nekretnina za prodaju [10]

Na slici 3 se može uočiti pad broja nekretnina raspoloživih za prodaju u periodu od 2013. godine do 2017. godine. Primjerice, 2009. godine bilo je oko 5085 raspoloživih nekretnina, a 2017. 973. [10] Porast broja stanovnika na području grada Seattle te veća potražnja za nekretninama uzrokovali su i porast cijena.

Skup podataka korišten u projektu je preuzet sa stranice kaggle.com te je naveden u popisu literature. Preuzeti podaci bilježe prodaje u periodu od svibnja 2014. do svibnja 2015. [11] Predviđanje cijene nekretnine je stvarni ekonomski problem kojem je cilj uskladiti iznos bankovnog kredita i stvarnu vrijednost nekretnine. [12] Korist od predviđanja cijene nekretnine imaju kupci i prodavači, a analiza podataka uključuje i definiranje ključnih čimbenika koji utječu na cijenu nekretnine.

Cijena nekretnine ovisi o brojnim faktorima: lokaciji, kvadraturi, broju soba, stanju nekretnine, godini gradnje, godini renovacije. Cilj projekta je odrediti ključne čimbenike koji utječu na cijenu i u kojoj mjeri te izraditi dobre modele koji daju pouzdane rezultate. Postavljene su sljedeće hipoteze: postoji pozitivna povezanost cijene nekretnine s kvadraturom unutarnjeg prostora; postoji pozitivna povezanost cijene nekretnine s kvalitetom gradnje; postoji pozitivna povezanost cijene nekretnine s brojem prostorija.

3.1. Pregled prethodnih istraživanja

Za razumijevanje pojma otvorenih podataka i ispitivanje kvalitete podataka, autori [13] su definirali sve aspekte otvorenih podataka te utvrdili čimbenike mjerenja njihove kvalitete. Autori su proširili definiciju otvorenih podataka, odnosno govore o aspektima otvorenih podataka, a ne atributima. Postoji nekoliko aspekata koji opisuju pojam otvorenih podataka poput: slobodnog preuzimanja, zanemarivog troška, javne dostupnosti, ponovne iskoristivosti, mogućnosti redistribucije, bez restrikcije od autorskih prava i patenata, strukturiranost za iskoristivost, zahtijevanje „otvorene“ licence, bez objave osobnih informacija, izrađenost tijekom poslovnih operacija, pripadnost poreznom obvezniku te dostupnost u velikim količinama. U radu su utvrđeni sljedeći mjerni podaci kvalitete: konzistentnost, potpunost, točnost, jedinstvenost, pristup i vidljivost, upotrebljivost i razumljivost, pravovremenost, vrijednost i detaljnost. Rad daje važan doprinos utvrđivanju definicije otvorenih podataka i uspostavljanju čimbenika kvalitete.

Skup podataka o cijenama nekretnina prodanih u periodu između 2006. i 2010. godine na području grada Ames države Iowa Sjedinjenih Američkih Država, poslužio je kao temelj izrade jednog istraživačkog rada na Stanfordu. Skup podataka se sastoji od 79 atributa (numeričkih i kategoričkih) i 1460 instanci. Autori su primijenili tehnike nadgledanog učenja: klasifikaciju i regresiju. U postupku klasifikacije, svrstali su nekretnine u sljedeće cjenovne rangove kako bi dobili uvid u njihovu distribuciju: [0, 100K), [100K, 150K), [150K, 200K), [200K, 250K), [250K, 300K), [300K, 350K), [350K, ∞). Za izradu prediktivnih modela, kao varijablu čija se vrijednost nastoji predvidjeti odabrana je cijena nekretnine. Izradili su nekoliko modela, a za problem klasifikacije najpreciznije rezultate dao je model SVC sa linearnom jezgrom (eng. *Linear kernel*). Nadalje, za problem regresije najboljim se pokazao model SVR sa gaussovom jezgrom (eng. *Gaussian kernel*). Završne analize u istraživanju pokazale su kako na cijenu nekretnine ponajviše utječu kvadratura, materijal krova i susjedstvo. [14]

Sljedeći istraživački rad ponovno se odnosi na predviđanje cijene nekretnine te kao skup podataka koristi podatke o prodaji 352 nekretnine na području okruga Petaling grada Kuala Lumpur države Malezije. Glavni atributi kojima se opisuju nekretnine odnose se na površinu zemljišta, kvadraturu, broj prostorija soba, broj kupaonica, godinu gradnje, stanje građevine, kvalitetu namještaja i lokaciju. Za predikciju cijene nekretnine, autori su izradili sljedeće modele: model neuronske mreže, *Adaptive Neuro-Fuzzy Inference Systems (ANFIS)* i *Fuzzy Least-Squares Regression (FLSR)*. Kao rezultat istraživanja, uočili su kako fizičke karakteristike građevine i lokacija imaju najveći utjecaj na cijenu nekretnine. [15]

Predviđanje cijene nekretnine izradom hednonskog modela (eng. *Hedonic model*) i modela neuronske mreže, autor je proveo nad podacima o prodaji nekretnina u gradu Christchurch, Novi Zeland. Neki od atributa kojima se opisuju nekretnine su: kvadratura, godina gradnje, vrsta građevine, broj prostorija, broj kupaonica, broj garaža, objekti na području nekretnine i lokacija. Većina navedenih atributa ima pozitivnu povezanost sa cijenom nekretnine, osim godine gradnje. Primjerice, nekretnina koja ima vrt ili veću kvadraturu je skuplja od one koja nema vrta ili ima manju kvadraturu. Godina gradnje ima negativnu povezanost sa cijenom iz razloga što su starije nekretnine jeftinije od onih nedavno izgrađenih. [16]

4. Opis i priprema podataka

U radu se analizira skup podataka vezan uz prodaju nekretnina. Ovo poglavlje je namijenjeno opisu podataka, postupku pripreme podataka, broju instanci skupa te utvrđivanju povezanosti atributa. Sljedeća tablica prikazuje popis atributa skupa podataka uz pripadne opise, tip podatka i vrstu distribucije.

Tablica 1: Popis atributa skupa podataka

Naziv atributa	**Opis**	**Tip podatka**	**Distribucija**
id	Jedinstvena identifikacijska oznaka za svaku prodanu nekretninu.	numerički	uniformna
date	Datum prodaje nekretnine.	kategorički	multimodalna
price	Prodajna cijena nekretnine.	numerički	eksponencijalna
bedrooms	Broj prostorija nekretnine.	numerički	unimodalna (nagnuta udesno)
bathrooms	Broj kupaonica nekretnine, gdje oznaka .5 označava prostorije sa toaletom bez tuša.	numerički	multimodalna
sqft_living	Kvadratura unutarnjeg prostora.	numerički	unimodalna (nagnuta udesno)
sqft_lot	Kvadratura zemljišta.	numerički	eksponencijalna
floors	Broj katova.	numerički	eksponencijalna
waterfront	Bilježi ima li nekretnina pogled na rivu (1) ili nema (0).	kategorički	unimodalna (nagnuta udesno)
view	Indeks 0-4 koji bilježi koliko dobar je bio pregled zemljišta.	numerički	unimodalna (nagnuta udesno)
condition	Indeks 1-5 koji označava stanje nekretnine. [17]	numerički	unimodalna (nagnuta ulijevo)

	1=Loše, zahtjeva renovaciju. 2=U redu, potrebno mnogo popravaka. 3=Prosječno, potrebno nekoliko manjih popravaka i dorađivanja. 4=Dobro, nije potrebno mnogo popravaka, ali nije ni sve novo. 5=Izvrsno, očuvani i održavani prostor i predmeti.		
grade	Indeks 1-13 koji označava kvalitetu gradnje. [17] 1-3= Nedostaju minimalni građevinski standardi. 4=Starija gradnja niske kvalitete. 5=Niski troškovi gradnje. Mala, jednostavna građevina. 6=Kvalitetni materijali i jednostavan dizajn. Niska ocjena. 7=Prosječna ocjena građevine i dizajna. 8=Bolji materijali za gradnju i oblikovanje interijera i eksterijera. Malo iznad prosjeka. 9=Bolja arhitektura, dizajn i kvaliteta. 10=Nekretnine ove kvalitete uglavnom imaju visokokvalitetne značajke: veća kvadratura, renovacije, bolji dizajn. 11=Nekretnine sa prilagođenim dizajnom, dodatnim sadržajima i luksuznijim mogućnostima. 12=Dizajn i gradnja po mjeri izvrsnih graditelja. Materijali najviše kvalitete.	numerički	normalna

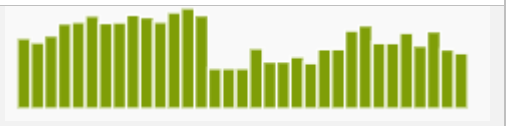
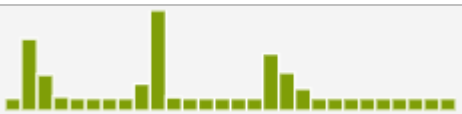
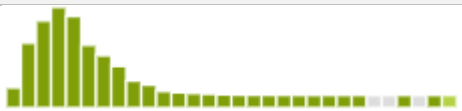
	13=Gradnja i dizajn po mjeri. Razina vile. Velika količina najkvalitetnijih materijala.		
sqft_above	Kvadratura unutarnjeg stambenog prostora iznad razine tla.	numerički	unimodalna (nagnuta udesno)
sqft_basement	Kvadratura unutarnjeg stambenog prostora ispod razine tla.	numerički	eksponencijalna
yr_built	Godina gradnje.	numerički	unimodalna (nagnuta ulijevo)
yr_renovated	Godina posljednje obnove.	numerički	unimodalna (nagnuta ulijevo)
zipcode	Poštanski broj.	numerički	multimodalna
lat	Zemljopisna širina.	numerički	unimodalna (nagnuta ulijevo)
long	Zemljopisna dužina.	numerički	eksponencijalna
sqft_living15	Prosječna kvadratura unutarnjeg stambenog prostora 15 najbližih kuća.	numerički	unimodalna (nagnuta udesno)
sqft_lot15	Prosječna kvadratura zemljišta 15 najbližih kuća.	numerički	eksponencijalna

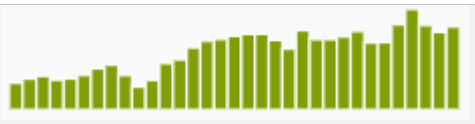
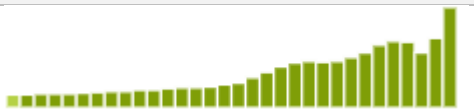
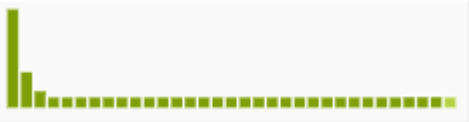
Izvor: [18]

Skup podataka sadrži 21 atribut i 21613 instanci. Atributi su uglavnom numeričkog tipa uz 2 kategorička – date i waterfront. Distribucije su različitog tipa, od eksponencijalne distribucije do multimodalne. Najčešća distribucija podataka je eksponencijalna. U skupu nema nedostajućih podataka što nam omogućava izradu boljeg, odnosno točnijeg modela.

Tablica 2 prikazuje atribute skupa podataka iz alata BigML sa navedenim nazivom atributa, brojem nedostajućih i neispravnih podataka te histogramom distribucije. Vrste distribucija su određene u tablici 1 dok u tablici 2 grafički prikazujemo distribuciju.

Tablica 2: Popis atributa skupa podataka uz navedene nedostajuće i pogrešne podatke i prikaz distribucije

Naziv atributa	**Broj nedostajućih podataka**	**Broj pogrešnih podataka**	**Histogram**
id	0	0	
date	0	0	
price	0	0	
bedrooms	0	0	
bathrooms	0	0	
sqft_living	0	0	
sqft_lot	0	0	
floors	0	0	
waterfront	0	0	
view	0	0	
condition	0	0	

grade	0	0	
sqft_above	0	0	
sqft_basement	0	0	
yr_built	0	0	
yr_renovated	0	0	
zipcode	0	0	
lat	0	0	
long	0	0	
sqft_living15	0	0	
sqft_lot15	0	0	

Izvor: izrada pomoću alata BigML

Tablica 3: Minimalna, maksimalna i prosječna vrijednost za pojedini atribut unutar skupa podataka

Naziv atributa	**Minimalna vrijednost**	**Maksimalna vrijednost**	**Prosječna vrijednost**
id	1000102	9900000190	4580301520.86
price	75000	7700000	540088.14
bedrooms	0	3	3.37
bathrooms	0	775	84.7
sqft_living	290	13540	2079.90
sqft_lot	520	1651359	15106.97
floors	1	35	2.87
view	0	4	0.23
condition	1	5	3.41
grade	1	13	7.66
sqft_above	290	9410	1788.39
sqft_basement	0	4820	291.51
yr_built	1900	2015	1971.01
yr_renovated	0	2015	84.4
zipcode	98001	98199	98077.94
lat	474	477776	430655.33
long	-122519	-122	-110987.16
sqft_living15	399	6210	1986.55
sqft_lot15	651	871200	12768.46

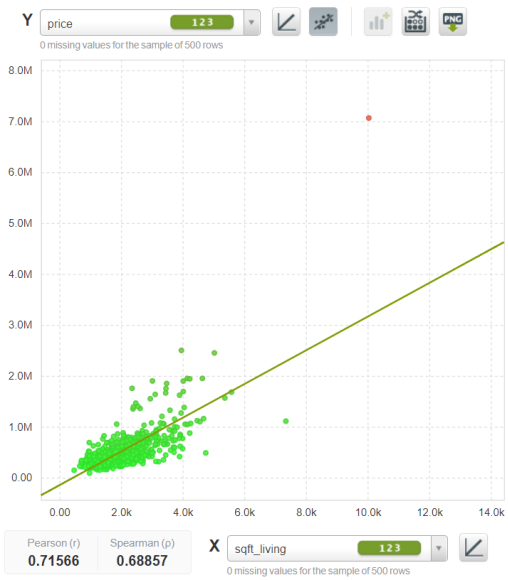
Izvor: izrada pomoću alata BigML

Tablica 3 prikazuje popis numeričkih atributa skupa podataka uz navedene minimalne, maksimalne i prosječne vrijednosti za svaki numerički atribut. Vrijednosti su izračunate pomoću alata BigML. Na popisu se ne nalaze kategorički atributi (date i waterfront) za koje ne alat ne može izračunati navedene vrijednosti. Atribut „date“ formata je „YYYYMMDDThhmmss“, a atribut „waterfront“ može imati vrijednost 0 ili 1. Kao što je navedeno u opisu problema, prodaje su zabilježene u periodu od svibnja 2014. do svibnja

2015. godine. Ako promotrimo distribuciju atributa „waterfront“ sa slike 3, možemo uočiti nesrazmjer u distribuciji podataka za vrijednosti 0 i 1. Čak 99.25% podataka (ukupno 21450 instanci) za vrijednost atributa „waterfront“ ima 0, odnosno 99.25% prodanih nekretnina nije imalo pogled na rivu. Može se zaključiti da atribut „waterfront“ ima mod false.

4.1. Korelacije

U nastavku poglavlja odredit će se povezanosti među varijablama te će se utvrditi one varijable koje su potencijali ulazni parametri za predviđanje ciljane varijable – cijena nekretnine. Korelacije objašnjavaju povezanosti varijabli te mogu pomoći utvrditi one varijable koje mogu utjecati na predviđanje vrijednosti izlazne varijable. [19] Korelacija je statistička metoda kojom se utvrđuje kako pojedina varijabla mijenja drugu varijablu sa kojom je povezana i/ili obratno. [19] Povezanost varijabli ne podrazumijeva uzročno-posljedičnu vezu, odnosno ako su dvije varijable povezane, ne mora nužno značiti da promjene jedne varijable uzrokuju promjene druge varijable. [19] Kako bi utvrdili povezanosti među varijablama, pomoću alata BigML moguće je kreirati dvodimenzionalne grafove na čijim osima (x i y) su navedene varijable čiji odnos promatramo (opcija *Scatterplot*). Prvi kvadrant koordinatnog sustava sadržava položaje podataka gdje svaka točka predstavlja jednu instancu skupa (nekretninu) sa vrijednostima svake varijable (cijenom, identifikacijskom oznakom, godinom gradnje…). Upravo prema položajima, odnosno uzorku kojeg kreiraju podaci, možemo grafički utvrditi povezanost. Kako u skupu podataka prevladavaju numerički atributi te je ciljana varijabla numeričkog tipa, za izračun korelacije proučavat će se i interpretirati vrijednosti Pearson i Spearman koeficijenata.



Slika 4: Korelacija varijabli sqft_living i price (Izvor: vlastita izrada)

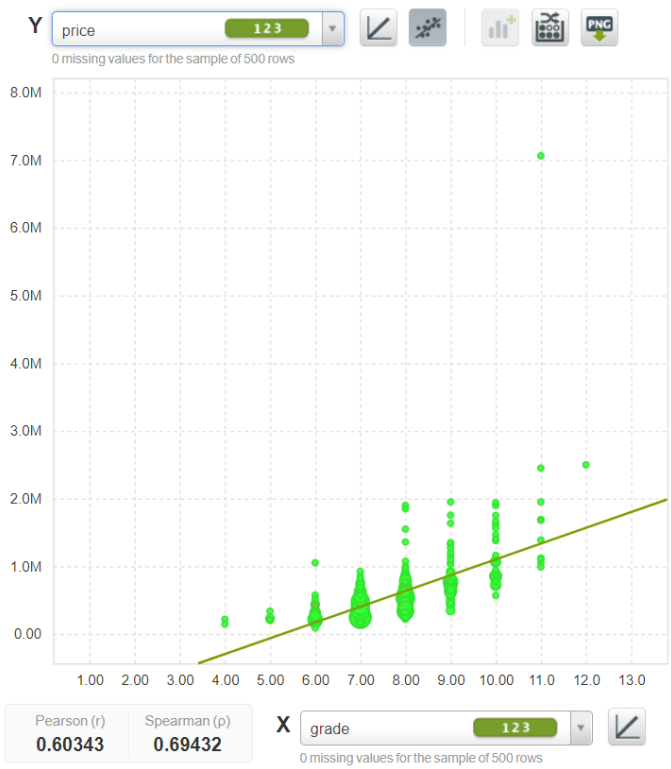
Slika 4 prikazuje odnos varijabli sqft_living i price. Kako se povećava kvadratura unutarnjeg prostora nekretnine, također se povećava i cijena nekretnine. Međutim, korelacija ove dvije varijable nije savršena. Primjerice, ako promatramo kvadraturu unutarnjeg prostora od 88.26 kvadratnih metara (950 kvadratne stope – *square feet*), postoje nekretnine sa manjom kvadraturom od 88.26m^2 i većom cijenom te nekretnine sa većom kvadraturom od 88.26m^2 i manjom cijenom. Dakle, iako korelacija nije savršena, postoji povezanost i tendencija povećanja cijene i kvadrature.

Pomoću alata BigML izračunat je Pearsonov koeficijent koji mjeri snagu i smjer linearne povezanosti dvije varijable. [20] Kreće se u rasponu od -1 do 1. [20] Ekstremne vrijednosti Pearsonovog koeficijenta, -1 i 1, označavaju savršenu linearnu povezanost dvije varijable gdje promjenu jedne varijable prati promjena druge. [20] Međutim, takvi slučajevi su rijetki u praksi. [20] Grafički prikaz se prepoznaje po položaju svih točaka na pravcu regresije koji raste (1) ili pada (-1). [20] Vrijednost 0 ne predstavlja linearni odnos te kako se jedna varijabla povećava, ne postoji tendencija da se druga varijabla povećava ili smanjuje. [20] Vrijednost Pearsonovog koeficijenta između 0 i -1 ili 0 i 1 označava da postoji odnos između varijabli. [20] Taj odnos je slabiji što je apsolutna vrijednost bliža 0, odnosno tada je povezanost među varijablama manja. [20] Suprotno tome, ako je apsolutna vrijednost Pearsonovog koeficijenta bliža 1, povezanost među varijablama je veća. [20] Takav odnos se grafički prikazuje kao odudaranje položaja točaka od pravca regresije što je apsolutna vrijednost Pearsonovog koeficijenta bliža 0, odnosno približavanje točaka pravcu regresije što je apsolutna vrijednost Pearsonovog koeficijenta bliža 1. [20] Pozitivna vrijednost Pearsonovog koeficijenta označava da kada se vrijednost jedne varijable poveća, vrijednost druge varijable također se povećava. [20] Takav odnos se grafički prikazuje kao nagib prema većim vrijednostima prvog kvadranta koordinatnog sustava. [20] Negativan Pearsonov koeficijent predstavlja slučajeve kada se vrijednost jedne varijable povećava, a vrijednost druge varijable ima tendenciju pada. [20] Takav odnos se grafički prikazuje kao nagib prema manjim vrijednostima y osi. [20]

Sada kada su objašnjene moguće vrijednosti Pearsonovog koeficijenta, možemo objasniti njegovu vrijednost sa slike 3. Pearsonov koeficijent za varijable price i sqft_living iznosi 0.71566 što je pozitivan broj i po apsolutnoj vrijednosti je bliži jedinici. Iz navedenog zaključujemo kako postoji srednje jaka povezanost između cijene i kvadrature te da u većini slučajeva povećanje kvadrature nekretnine uzrokuje i porast njezine cijene.

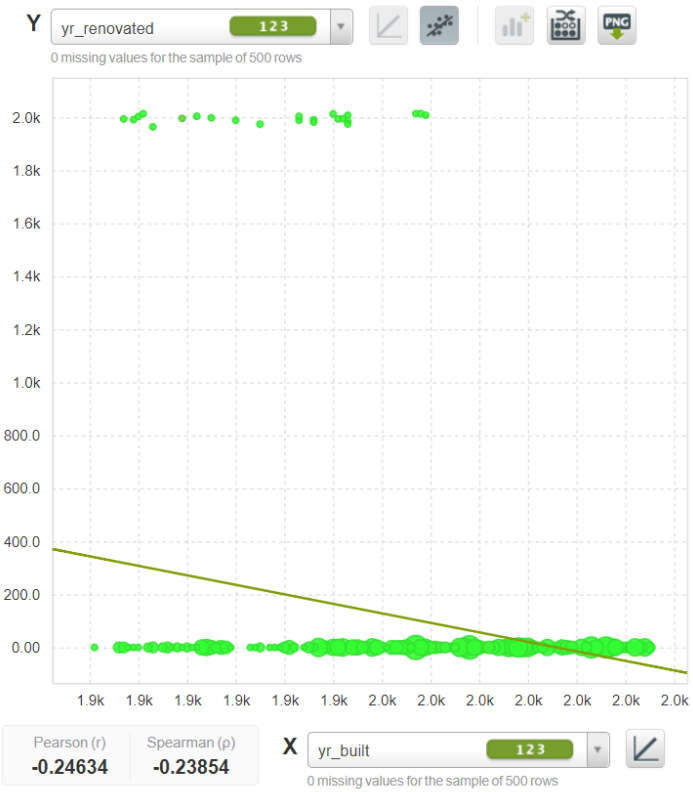
Korelacija varijabli nije potpuno linearna, odnosno točke nisu grupirane oko pravca regresije. U tom slučaju dobro je promotriti Spearmanov koeficijent. U statistici, Spearmanov koeficijent procjenjuje koliko dobro se odnos dvije varijable može opisati korištenjem monotone funkcije. [21] Monotonu funkciju koja opisuje odnos između dvije varijable definira nekoliko svojstava: kako se vrijednost jedne varijable povećava, tako se povećava i vrijednost druge varijable ili kako se vrijednost jedne varijable povećava, tako vrijednost druge varijable

opada. [21] Ali, ne konstantnom brzinom, dok je u linearnom odnosu stopa povećanja / smanjenja konstantna. [21] Vrijednost Spearmanovog koeficijenta sa slike 4 iznosi 0.68857 što ukazuje na monotoni rast funkcije. Nadalje, pozitivna vrijednost Spearmanovog koeficijenta koja je po apsolutnoj vrijednosti bliža 1 ukazuje na odnos između varijabli u kojem porast vrijednosti varijable sqft_living (povećanje kvadrature nekretnine) nikada neće uzrokovati pad varijable price (cijene nekretnine). Vrijednost Spearmanovog koeficijenta od 0.68857 ukazuje na jaku vezu između varijabli.



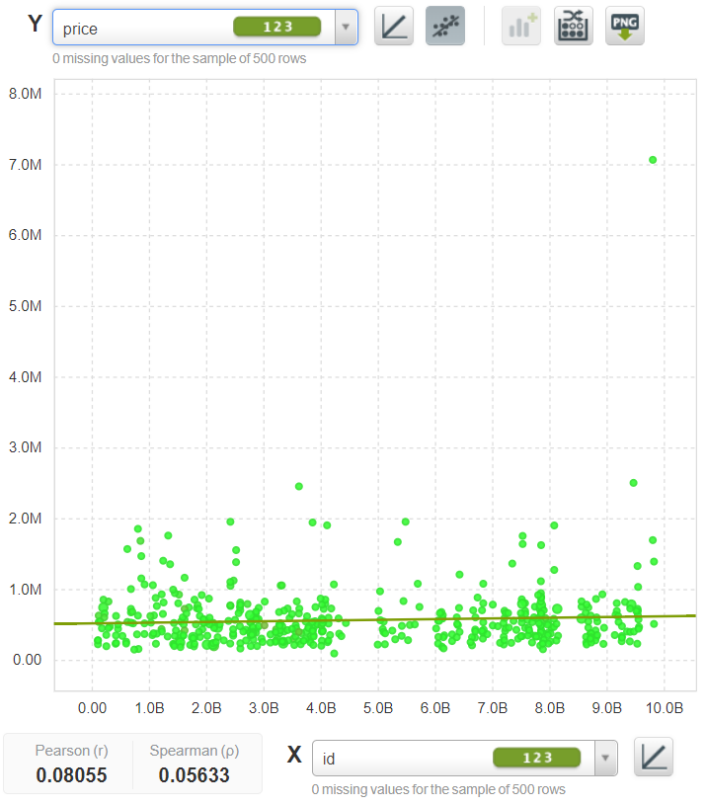
Slika 5: Korelacija varijabli grade i price (Izvor: vlastita izrada)

Slika 5 sadrži grafički prikaz korelacije varijabli grade (indeks kvalitete gradnje) i price (cijena nekretnine). Poput primjera korelacije sa slike 4, vrijednosti Pearsonovog i Spearmanovog koeficijenta su pozitivne vrijednosti bliže jedinici po apsolutnoj vrijednosti. Vrijednost Pearsonovog koeficijenta iznosi 0.60343, a Spearmanovog koeficijenta 0.69432. Uočavamo pozitivnu vezu između varijabli iz vrijednosti koeficijenata te promatranjem grafičkog prikaza. Možemo utvrditi kako su varijable povezane te da povećanje indeksa kvalitete gradnje može uzrokovati povećanje cijene, odnosno kvalitetnije građene i dizajnirane nekretnine imaju veću cijenu od nekretnina nižih kvaliteta gradnje.



Slika 6: Korelacija varijabli yr_built i yr_renovated (Izvor: vlastita izrada)

Vrijednosti Pearsonovog i Spearmanovog koeficijenta za varijable yr_built (godina gradnje nekretnine) i yr_renovated (posljednja godina renovacije nekretnine) su negativne i po apsolutnoj vrijednosti bliže 0, odnosno predstavljaju slabu povezanost varijabli. Slika 6 prikazuje grafički prikaz korelacije navedenih varijabli. Slaba povezanost se uočava kao udaljeniji položaj točaka od pravca regresije. Nadalje, pravac regresije pada što označava da porast vrijednosti jedne varijable uzrokuje smanjenje vrijednosti druge. Dakle, ako je godina gradnje nekretnine starija (manja vrijednost), posljednja godina renovacije je veća (nedavno obavljena renovacija) i obratno. Međutim, slaba povezanost među varijablama ukazuje da godina gradnje ne utječe u velikoj mjeri na godinu renovacije i obratno.



Slika 7: Korelacija varijabli price i id (Izvor: vlastita izrada)

Slika 7 prikazuje odnos varijabli price i id, odnosno odnos cijene nekretnine i pripadne identifikacijske oznake. Prije promatranja grafičkog prikaza i vrijednosti Pearsonovog i Spearmanovog koeficijenta, možemo pretpostaviti kako identifikacijska oznaka nekretnine ne bi trebala imati veliki utjecaj na njezinu cijenu zbog toga što predstavlja samo jedinstvenu oznaku po kojoj se pojedina nekretnina može identificirati. Međutim, identifikacijske oznake se mogu dodjeljivati prema kategorijama nekretnina i sl. U objašnjenjima atributa skupa podataka nije specificiran način dodjeljivanja identifikacijske oznake nekretninama stoga ne možemo zaključiti da postoji ili ne postoji veza između navedene dvije varijable. Promotrimo li Pearsonov i Spearmanov koeficijent, vidimo da su njihove vrijednosti pozitivne te po apsolutnoj vrijednosti bliže 0 što sugerira slabu vezu između dvije varijable te da identifikacijska oznaka ne utječe u velikoj mjeri na cijenu nekretnine i obratno.

5. Modeliranje podataka

Ovo poglavlje odnosi se na usporedbu deskriptivnog i prediktivnog modeliranja koji će detaljnije biti opisani u sljedećim potpoglavljima. Uz njih, opisat će se i tehnike pomoću kojih će se izraditi deskriptivni i prediktivni modeli uz pomoć alata BigML.

Glavne razlike između deskriptivnog i prediktivnog modeliranja odnose se na sadržaj kojeg opisuju, procese koje uključuju, preciznost rezultata i vrsti pristupa. [22] Tablica 4 prikazuje usporedbu deskriptivnog i prediktivnog modeliranja uz primjere korištenja.

Tablica 4: Usporedba deskriptivnog i prediktivnog modeliranja

Deskriptivno modeliranje	**Prediktivno modeliranje**
Opisuje što se dogodilo u prošlosti analizirajući pohranjene podatke.	Opisuje što bi se moglo dogoditi analizirajući povijesne podatke.
Uključuje procese prikupljanja podataka i rudarenja podataka.	Uključuje statistiku i tehnike predviđanja.
Deskriptivno modeliranje definiramo kao proces pronalaska korisnih informacija prilikom analize velike količine podataka.	Prediktivno modeliranje definiramo kao proces koji uključuje predviđanja koja su veoma korisna organizacijama prilikom donošenja odluka.
Primjeri deskriptivnog modeliranja uključuju izvješća o prodaji, izvješća o prihodima, izvješća o performansama i sl.	Primjeri prediktivnog modeliranja uključuju analize kreditnih rezultata, izvješća o predviđanjima i sl.
Pruža pouzdane rezultate u izvješćima.	Rezultati prediktivnog modeliranja nisu pouzdani, ne mogu sa sigurnošću tvrditi što će dogoditi, ali mogu sugerirati što bi se moglo dogoditi.
Tip pristupa deskriptivnog modeliranja je reaktivan te podrazumijeva rješavanje problema nakon što su nastupili.	Tip pristupa prediktivnog modeliranja je proaktivan i podrazumijeva otklanjanje problema prije nego što nastupe. Odnosno, sprečavanje njihovog nastanka.

Izvor: [22]

5.1. Deskriptivni modeli

Deskriptivna analiza je statistička metoda koja se koristi za pregled i pretraživanje povijesnih podataka kako bi se uočili uzorci među podacima. [23] Može pomoći organizacijama razumjeti prošle događaje kako bi na temelju njih donijele odluke. [22] Skupljanje podataka (eng. *Data aggregation*) i rudarenje podataka (eng. *Data mining*) dvije su tehnike koje se često koriste u deskriptivnoj analizi za pronalazak povijesnih podataka. [23] Cilj deskriptivne analize je sažeti i pretvoriti podatke u smislene informacije za kreiranje izvještaja. [24] Nadalje, pruža detaljniju analizu podataka kako bi se dobili odgovori o onome što se dogodilo i što se upravo događa. [24] U radu su odabrane dvije tehnike deskriptivnog modeliranja podataka: klasterska i faktorska analiza koje su detaljnije obrađene.

5.1.1. Klasterska analiza

Klasterska analiza se odnosi na algoritme koji grupiraju slične instance u grupe nazvane *klasteri*. [25] Krajnji rezultat klasterske analize je grupa klastera u kojoj je svaki klaster različit od ostalih, a instance unutar klastera su slične jedne drugima. [25] Klasterska analiza je računalno težak problem – za stvarne probleme, računala nisu u mogućnosti pronaći sve moguće načine na koje se instance mogu grupirati u klastere. [25] Do sada je razvijeno nekoliko tisuća algoritama koji pokušavaju pronaći približno točna rješenja problema. [25] Neki od njih su: hijerarhijski algoritmi (eng. *Hierarchical clustering*), algoritam k-srednjih vrijednosti (eng. *K-means cluster analysis*), analiza latentne klase (eng. *Latent class analysis*). [25]

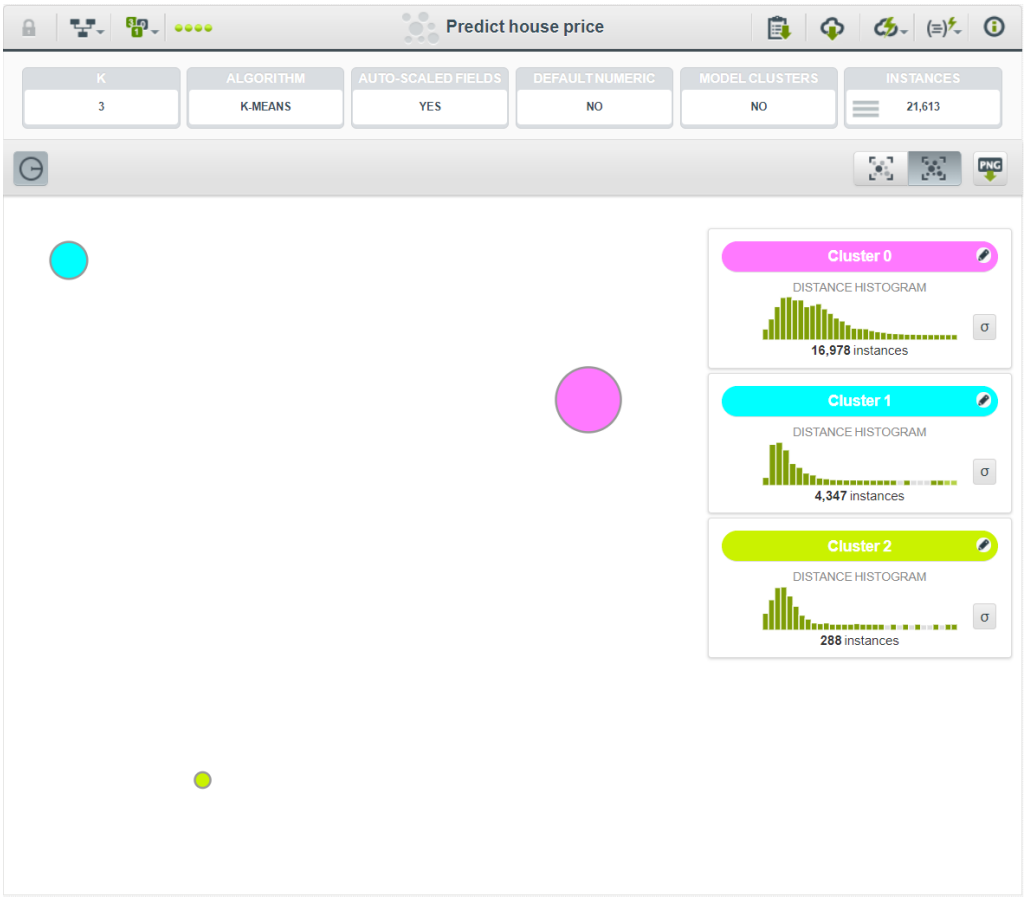
Tipične primjene klasterske analize: [26]

- Samostalni alat koji daje uvid u distribuciju podataka
- Korak predprocesiranja drugim algoritmima u inteligentnim sustavima

Primjene u realnim domenama: [26]

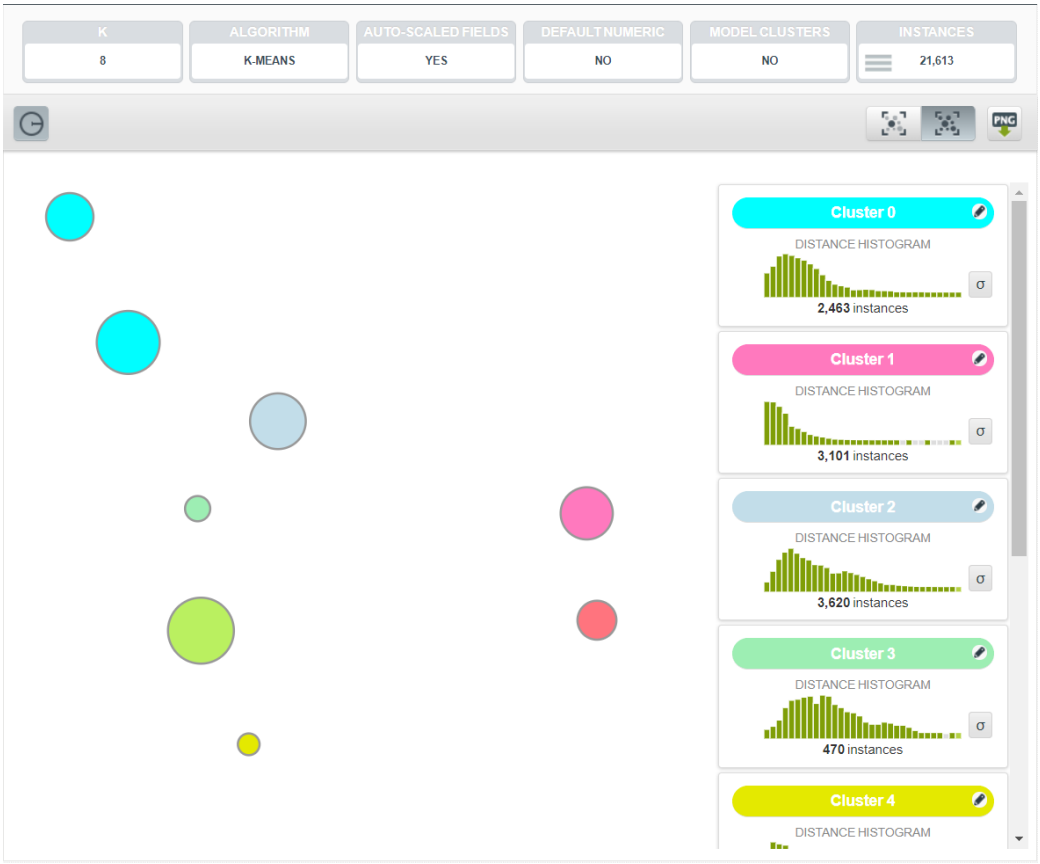
- Marketing: otkrivanje različitih skupina u bazama podataka korisnika i upotreba tog znanja u razvoju ciljanih marketinških programa
- Kompresija/segmentacija slike: grupiranje koherentnih piksela
- Bankarska/internet sigurnost: otkrivanje prijevara i neželjene pošte (spam)
- Pretraživanje informacija: Google search, vijesti temeljene na temi
- Biologija: taksonomija živih stvari poput obitelji, roda i vrsta

Nakon kratkog opisa metode, slijedi provedba klasterske analize u alatu BigML kako bi utvrdili optimalan broj klastera, opisali distribuciju instanci u klasterima te grafički prikazali rezultate.



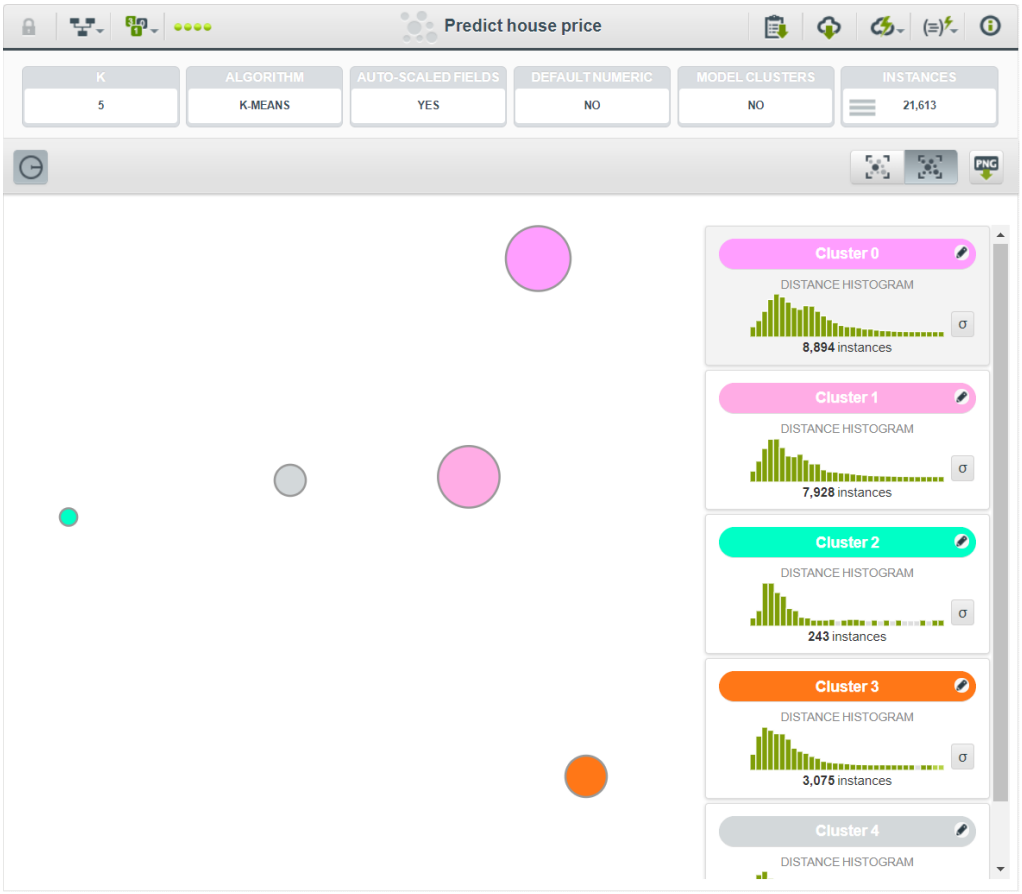
Slika 8: Klasterska analiza sa 3 klastera (Izvor: vlastita izrada)

Slika 8 odnosi se na grafički prikaz klasterske analize sa 3 klastera sa različitim distribucijama instanci. Uočavamo kako je distribucija instanci u klasteru 0 najveća (16978 instanci), a u klasteru 2 najmanja (288 instanci). Distribucija instanci u klasteru 1 iznosi 4347. Možemo zaključiti da klaster 0 strši, odnosno ima bitno veću distribuciju instanci u odnosu od klastera 1 i 2. Između klastera postoje bitne razlike u distribucijama što nastojimo smanjiti, stoga ćemo provesti analizu sa 8 klastera i usporediti rezultate.



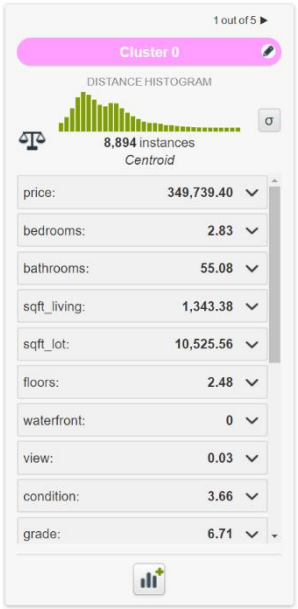
Slika 9: Klasterska analiza sa 8 klastera (Izvor: vlastita izrada)

Klasterska analiza sa 8 klastera prikazana je na slici 9 te uočavamo klastere sa različitim distribucijama. Klasteri su smješteni relativno blizu jedni drugima te promatranjem vrijednosti pojedinih varijabli unutar klastera poput cijene, kvadrature i dr, možemo uočiti sličnosti. Prema nekakvoj općoj definiciji, cilj grupiranja instanci u klastere je da instance unutar klastera budu slične jedna drugoj i različite od instanci drugih klastera. [26] Odnosno, da se udaljenosti unutar klastera minimiziraju, a između klastera maksimiziraju. [26] Ovakvim grupiranjem možda jesmo dobili nešto manju razliku u distribucijama instanci, no također smo dobili slične grupacije. Primjerice, promotrimo karakteristike klastera 0 i 5. Distribucija instanci unutar klastera 0 iznosi 2463 instanci, a unutar klastera 5 1520. Nekakav prosjek cijena nekretnina klastera 0 iznosi 449749.16$, prosječna kvadratura je 1612.89 kvadratnih stopa, broj kupaonica je 2-3, prosječni indeks stanja nekretnina je 3.06 itd. Navedene karakteristike instanci klastera 5 iznose: prosječna cijena nekretnina - 434670.62$, prosječna kvadratura - 1686.88 kvadratnih stopa, broj kupaonica je 3, prosječni indeks stanja nekretnina je 3.46. Možemo zaključiti kako postoje određene sličnosti među instancama navedenih klastera te kako bi se zapravo takve instance mogle grupirati u jedan klaster. Osim navedenih klastera, uočene su određene sličnosti među instancama i u drugim klasterima. Stoga nastojimo iznova provesti klastersku analizu, ovaj put sa 5 klastera.

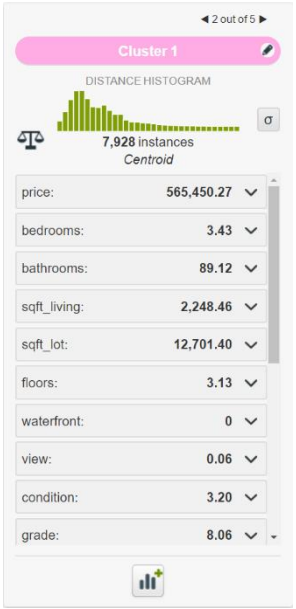


Slika 10: Klasterska analiza sa 5 klastera (Izvor: vlastita izrada)

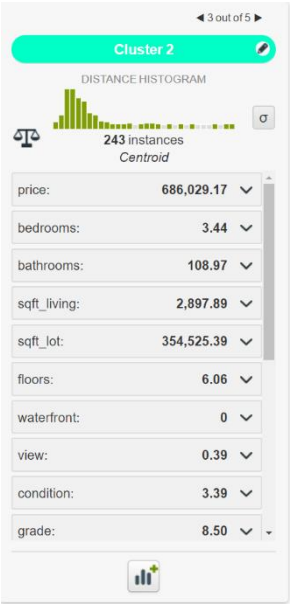
Slika 10 prikazuje grupaciju instanci u 5 klastera. Distribucije instanci unutar klastera su sljedeće: najveću distribuciju ima klaster 0 u iznosu od 8894 instanci, najmanju distribuciju ima klaster 2 sa 243 instance, klaster 1 sadrži 7928 instanci, klaster 3 3075 instanci, a klaster 4 1473 instanci. Distribucije instanci među unutar klastera su nešto međusobno sličnije nego što je to kod primjera grupacije sa 8 ili 3 klastera, s tim da klaster 2 iskače po malom broju instanci (243). Promotrimo li grafički prikaz, uočavamo kako su klasteri nešto bliže položeni nego u primjeru grupacije sa 3 klastera što ukazuje na postojanje određenih sličnosti među instancama klastera. Međutim, grupacija instanci u 3 klastera daje dosta velike razlike u distribucijama među klasterima, dok grupacija u 5 klastera bitno smanjuje te razlike. Za optimalan broj klastera možemo uzeti broj 5 koji daje najmanje razlike u distribucijama od provedenih analiza. Povećanjem broja klastera dolazi do smanjenja udaljenosti između klastera čemu ne težimo. Isto tako odabirom manjeg broja klastera povećava se razlika u distribuciji instanci unutar klastera. Slijedi detaljniji uvid osobina klastera za smisleniju grupaciju instanci skupa podataka.



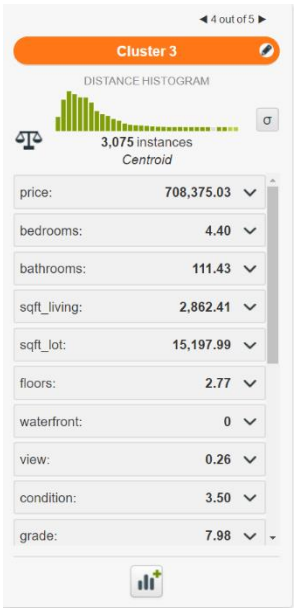
Minimum:	0.19
Mean:	0.50
Median:	0.46
Maximum:	1.44
Std dev:	0.17
Sum:	4,465.12
Sum sq:	2,513.48
Variance:	0.03



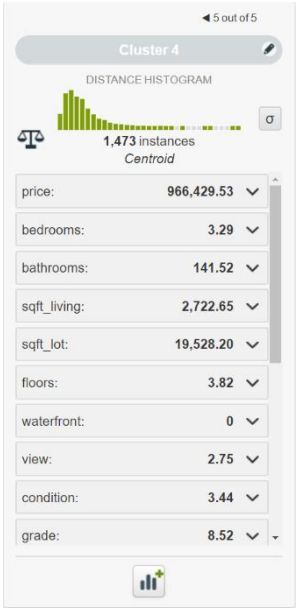
Minimum:	0.17
Mean:	0.50
Median:	0.44
Maximum:	1.74
Std dev:	0.21
Sum:	3,940.18
Sum sq:	2,298.48
Variance:	0.04



Minimum:	0.39
Mean:	1.17
Median:	1.01
Maximum:	5.37
Std dev:	0.61
Sum:	283.75
Sum sq:	421.21
Variance:	0.37



Minimum:	0.25
Mean:	0.62
Median:	0.57
Maximum:	5.10
Std dev:	0.25
Sum:	1,908.20
Sum sq:	1,377.02
Variance:	0.06



Minimum:	0.34
Mean:	0.81
Median:	0.73
Maximum:	4.13
Std dev:	0.34
Sum:	1,189.04
Sum sq:	1,126.84
Variance:	0.11

Slika 11: Opis klastera (Izvor: vlastita izrada)

Slika 11 daje opis 5 klastera, histogram distribucije, broj instanci po klasteru, minimalne i maksimalne vrijednosti, aritmetičke sredine, medijane i sl. Svaki klaster ima nekoliko tisuća instanci, osim klastera 2 koji ima 243 instance. Promatranjem histograma, uočava se unimodalna (nagnuta udesno) distribucija za svaki klaster.

Klasterskom analizom instance su podijeljene u 5 grupa (klastera) po sličnim karakteristikama. Klaster 4 predstavlja grupu nekretnina sa najvišom prosječnom prodajnom cijenom (prosjek: 966429.53$). Klaster 0 predstavlja nekretnine sa najnižom prosječnom prodajnom cijenom (prosjek: 349739.40$). Klasteri 1, 2 i 3 predstavljaju nekretnine sa prosječnom prodajnom cijenom koja je u sredini – između najniže i najviše prodajne cijene. Prosječna prodajna cijena nekretnina za klaster 1 iznosi 565450.27$, za klaster 2 iznosi 686029.17$ i za klaster 3 iznosi 708375.03$. Dakle, od klastera 1, 2 pa do klastera 3 prosječna cijena se povećava za otprilike 100000$. Promotrimo li kvadraturu nekretnina, klaster 0 koji predstavlja grupu nekretnina sa najnižom prosječnom prodajnom cijenom, ima i najnižu prosječnu kvadraturu nekretnina u iznosu od 1343.38 kvadratnih stopa. S druge strane, klaster 2 sadrži prosječnu kvadraturu nekretnina od 2897.89 kvadratnih stopa.

Ovakvom grupacijom nekretnine možemo podijeliti na skupe, jeftine i umjerene cijene. Cijenu nekretnina prate i ostala obilježja poput povećanja kvadrature, ocjene, stanja nekretnina, broj prostorija, postojanje pogleda na rivu i sl.

5.1.2. Faktorska analiza

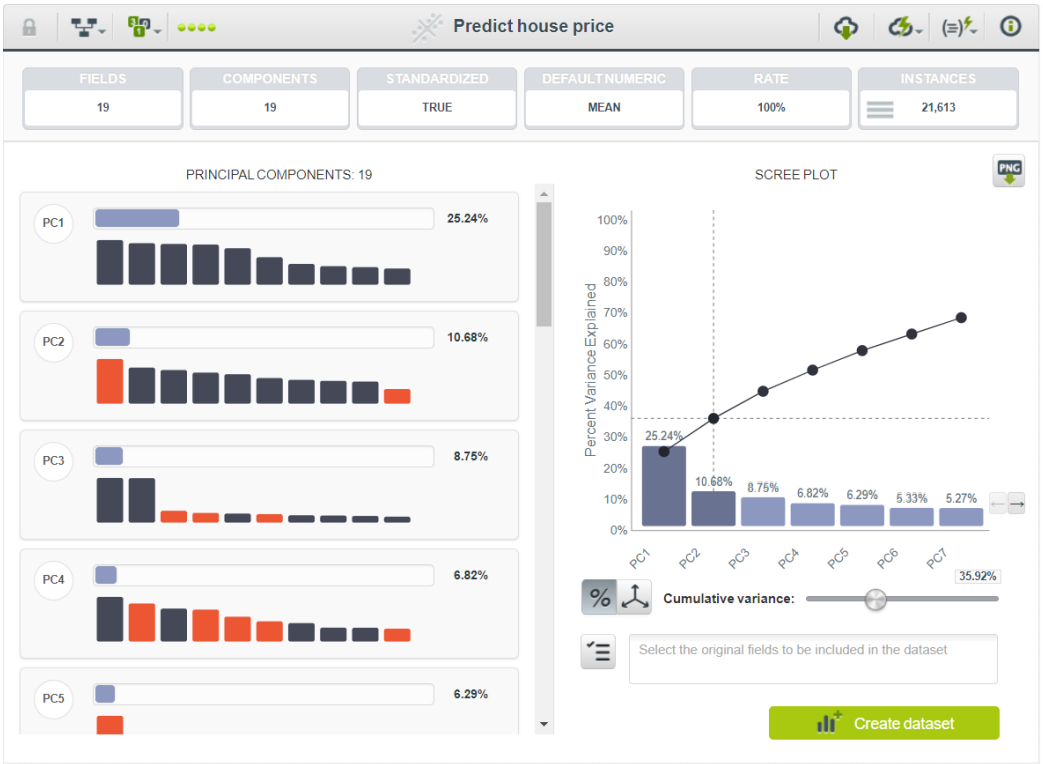
Faktorska analiza je tehnika koja se koristi za reduciranje velikog broja varijabli u manje faktore. [27] Ovom tehnikom se izdvaja najveća zajednička varijanca varijabli i stavlja u zajednički rezultat. [27] Faktorska analiza je dio općeg linearnog modela (eng. *General linear model* - GLM) te postavlja nekoliko pretpostavki: postoji linearni odnos, nema multikolinearnosti, uključuje relevantne varijable u analizu te postoji prava korelacija između varijabli i faktora. [27] Glavna značajka faktorske analize je postojanje zajedničkih uzoraka među varijablama. [28]

Postoji nekoliko metoda za izdvajanje faktora iz skupa podataka poput: [27]

- Analiza glavnih komponenata (eng. *Principal component analysis*)
- Analiza zajedničkih faktora (eng. *Common factor analysis*)
- Faktoriranje slika (eng. *Image factoring*)
- Metoda maksimalne vjerojatnosti (eng. *Maximum likelihood method*)

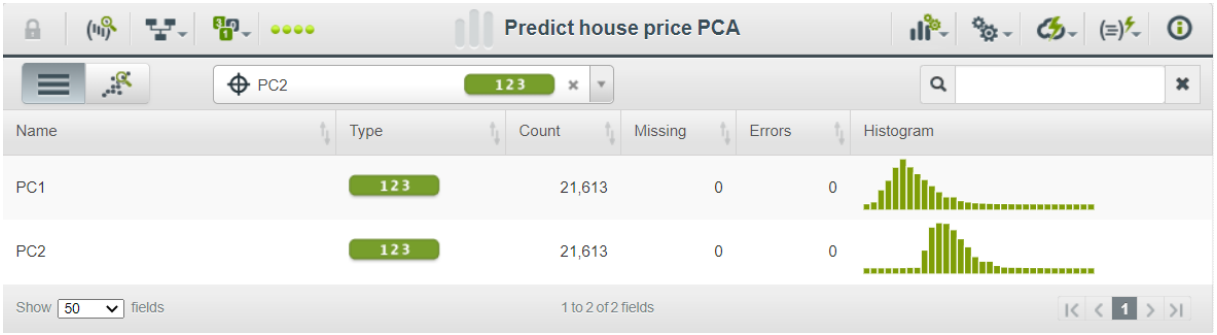
Analiza glavnih komponenata (PCA) je tehnika reduciranja dimenzija velikih skupova podataka koja povećava interpretativnost i minimizira gubitak informacija. [29] Kreira nove nekorelirane varijable koje maksimiziraju varijancu. [29] Također je jedna od najstarijih i najčešće korištenih metoda za reduciranje dimenzija varijabli. [29]

Provest ćemo redukciju podataka pomoću PCA konfiguracije koju nudi BigML, promotriti broj dobivenih komponenata i atribute koji ih određuju.



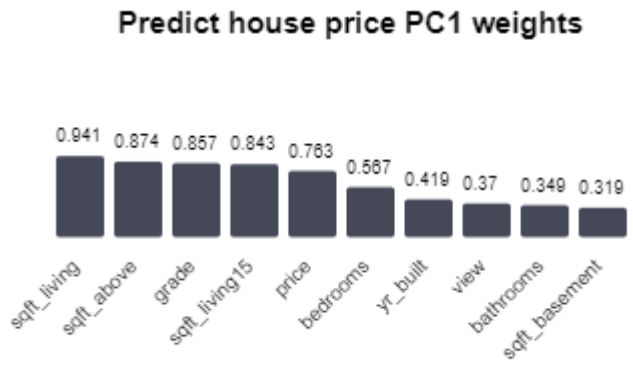
Slika 12: Analiza glavnih komponenata (Izvor: vlastita izrada)

Slika 12 se odnosi na grafički prikaz svojstvenih vrijednosti dobivenih analizom glavnih komponenata. Za vrijednost kumulativne varijance postavljena je vrijednost 35.92% koja je suma varijanci prve i druge komponente. Odabirom opcije *Create dataset* dobivamo skup podataka od 2 komponente kao što je prikazano na slici 13. Distribucija instanci prve komponente je unimodalna (nagnuta udesno), a druge komponente je normalna.



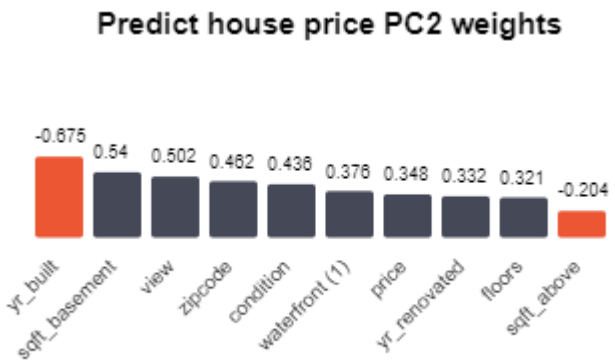
Slika 13: Skup podataka dobiven analizom glavnih komponenata (Izvor: vlastita izrada)

Slika 14 sadrži pregled varijabli prve komponente određenih težina. Težinske varijable prve komponente su: kvadratura nekretnine, kvadratura unutarnjeg stambenog prostora iznad razine tla, indeks kvalitete gradnje te prosječna kvadratura unutarnjeg stambenog prostora najbližih 15 susjeda. Navedene varijable koje imaju najveću težinu unutar prve komponente postavljaju veći utjecaj kvadrature i kvalitete gradnje na cijenu nekretnine.



Slika 14: Varijable prve komponente (Izvor: samostalna izrada)

Varijable druge komponente sa pripadnim težinama promatramo sa slike 15. Težinske varijable druge komponente su: godina gradnje, kvadratura unutarnjeg stambenog prostora ispod razine tla, indeks pregleda zemljišta. Navedena komponenta bilježi cijenu nekretnine prema navedenim varijablama sa najvećim težinama.



Slika 15: Varijable druge komponente (Izvor: samostalna izrada)

5.2. Prediktivni modeli

Prediktivna analiza pruža informacije organizacijama o tome što bi se moglo dogoditi – predviđa moguće ishode temeljene na trenutnim podacima. [22] Uključuje primjene nadgledanog učenja korištenog za predviđanje ciljane vrijednosti ili varijable. [24] Obavlja zadatke poput prediktivnog modeliranja, predviđanja, simulacija i upozorenja koji uključuju odgovore na pitanja poput: što bi se moglo dogoditi, kakav je ishod ako se ovi trendovi nastave, koje se akcije trebaju poduzeti. [24] U radu su detaljnije obrađene dvije tehnike prediktivnog modeliranja: stablo odlučivanja i neuronska mreža.

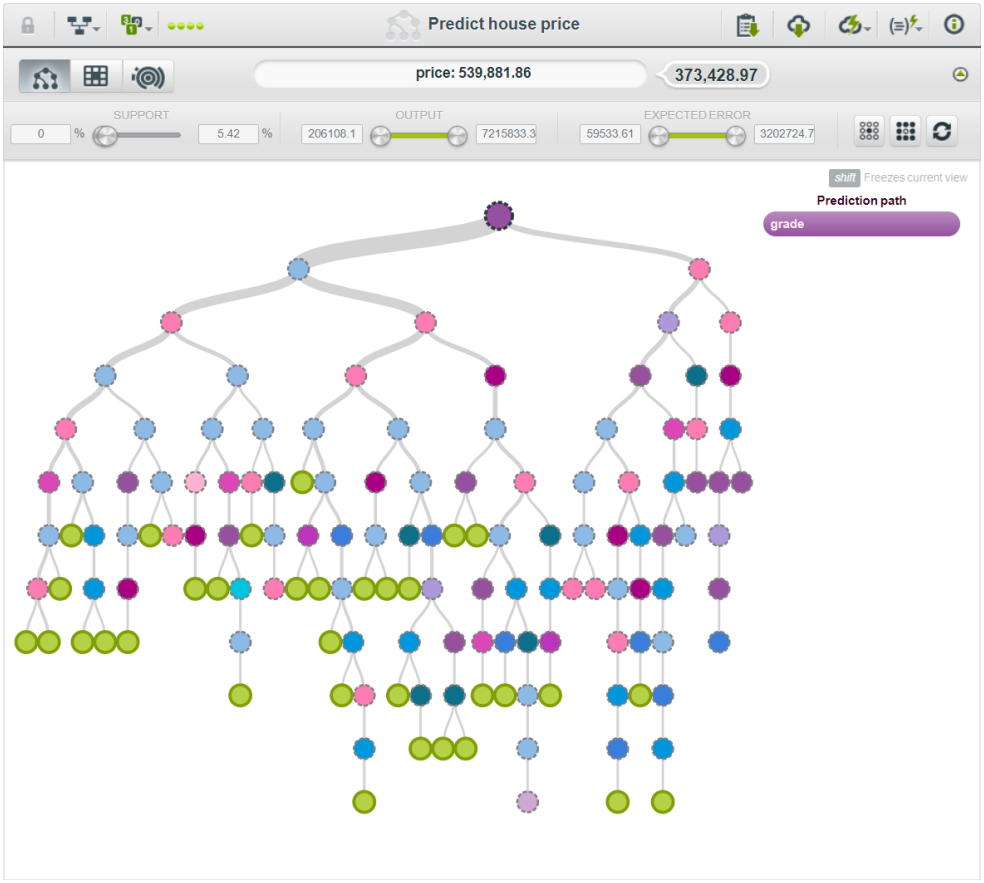
5.2.1. Stablo odlučivanja

Stablo odlučivanja je vrsta nadgledanog učenja koji sadrži predefiniranu ciljanu varijablu (eng. *Predefined target variable*) i najčešće se koristi u klasifikacijskim problemima, no može se koristiti za kategoričke i kontinuirane ulazne i izlazne varijable. [30] Stablo odlučivanja je grafički prikaz modela odluka i njihovih mogućih posljedica te uključuje ishode slučajnih događaja, troškove resursa i korisnost. [30] Jedan je od načina prikaza algoritma koji sadrži samo uvjetne kontrolne izjave. [30] Strukturom nalikuje na dijagram tijeka u kojem svaki unutarnji čvor predstavlja „test“ na atributu (primjerice, ima li osoba dijabetes ili nema), svaka grana ishod testa, a svaki čvor lista oznaku klase odnosno donesenu odluku. [30] Putevi od korijena do lista predstavljaju pravila klasifikacije. [30] Stablo odlučivanja daje prediktivan model kojeg obilježava velika točnost, stabilnost i lakoća tumačenja. [30] Može se koristiti za predviđanje vrijednosti ciljane kontinuirane varijable. No, modeli regresije i neuronske mreže su prikladniji za takve predikcije. [30] Zbog jednostavnosti, stabla odlučivanja se koriste u raznim područjima poput financija, inženjerstva, zdravstva, obrazovanja i sl. [30] Glavne prednosti stabla odlučivanja uključuju jednostavnost razumijevanja, rad sa numeričkim i kategoričkim varijablama te ga odlikuje korisnost u fazi istraživanja podataka. [30] Neki od nedostataka uključuju povećanje kompleksnosti računanja ukoliko postoji više klasa, prilikom rada sa kontinuiranim numeričkim varijablama dolazi do gubitka informacija tijekom kategoriziranja varijabli te ima manju pouzdanost predikcije za razliku od drugih algoritama strojnog učenja. [30]

Slijedi prikaz i tumačenje rezultata stabla odlučivanja izrađenog u alatu BigML. Moguće je izraditi različita stabla pomoću nekoliko metoda obrezivanja (eng. *Pruning methods*) koje se odnose na reduciranje veličine stabla na način da uklanjaju dijelove stabla koji imaju zanemarivu važnost na klasificiranje instanci. Isprobane su metode Smart pruning, Active

statistical pruning i No statistical pruning. Isprobane su navedene metode obrezivanja i nisu uočene bitne razlike, odnosno očekivana greška se nije smanjila.

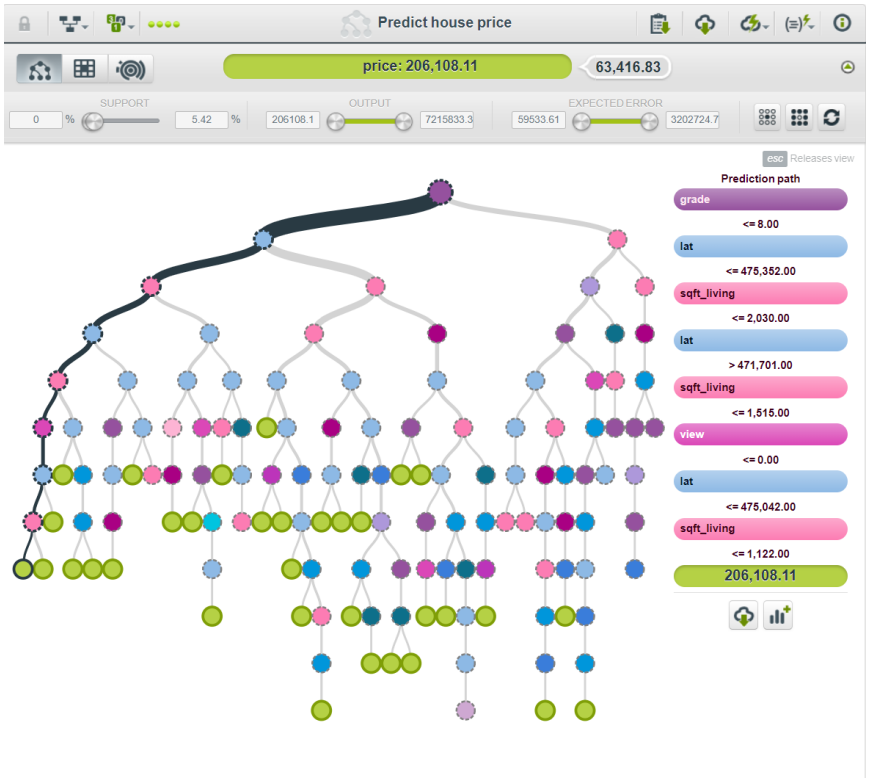
Slika 16 prikazuje dobiveno stablo odlučivanja dobivenog metodom Smart pruning i 512 čvorova. Za objective field, odnosno varijablu na temelju koje radimo predikciju smo izabrali cijenu nekretnine te promatramo koji faktori i u kojoj mjeri utječu na nju.



Slika 16: Stablo odlučivanja izrađeno metodom Smart pruning (Izvor: vlastita izrada)

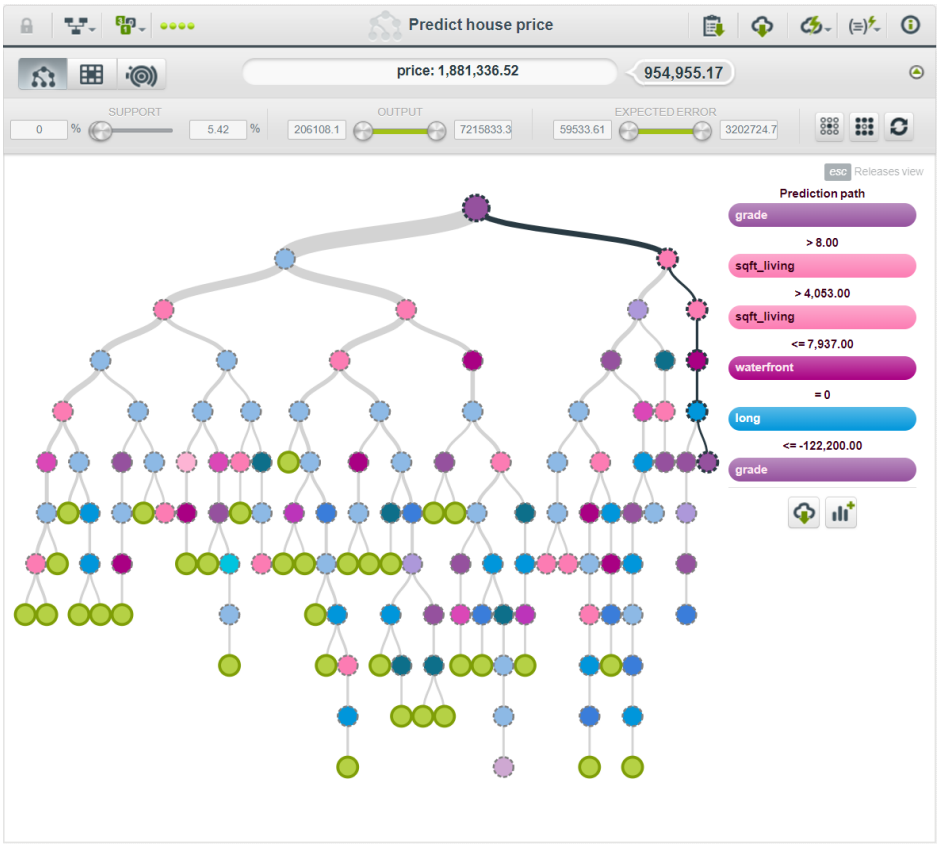
Nije bilo moguće izračunati pouzdanost u korijenskom čvoru, no navedena je greška u iznosu od 373428.97$ za trenutnu cijenu od 539881.86$. Atribut grade, odnosno indeks kvalitete gradnje ima glavnu ulogu putanje predviđanja (eng. *Prediction path*) u važnosti od 43.27%.

Slike 17 i 18 prikazuju rezultate predviđanja za krajnji lijevi čvor, odnosno krajnji desni uz očekivanu grešku i atribute koji utječu na putanju predviđanja.



Slika 17: Lijeva grana stabla sa opisom lijevog lista (Izvor: vlastita izrada)

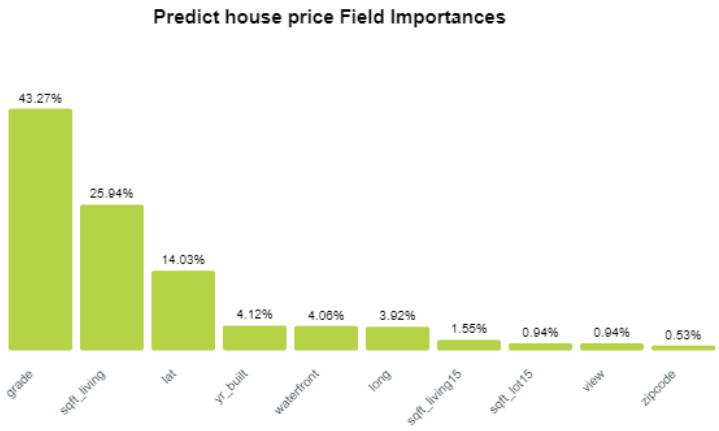
Lijeva grana stabla sadrži 17362 instance, odnosno 80.33% ukupnih instanci. Očekivana greška prvog čvora je otprilike 199663.92$. U prvom lijevom čvoru atributi grade (indeks gradnje) i lat (zemljopisna širina) čine putanju predviđanja u važnosti od 43.27%, odnosno 14.03%. Krajnji lijevi čvor sadrži 642 instance, odnosno 2.97% ukupnih instanci u skupu. Predviđena cijena nekretnina tog čvora iznosi 206108.11$ sa greškom od 63416.83$. Na rezultat odluke su utjecali sljedeći atributi: indeks gradnje, zemljopisna širina, kvadratura i indeks pregleda nekretnine.



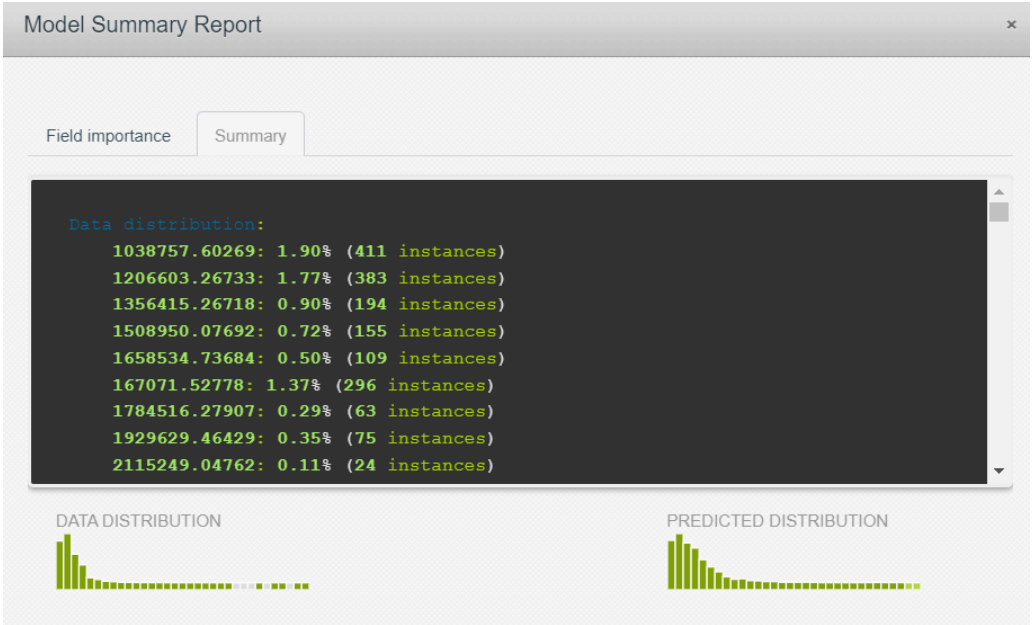
Slika 18: Desna grana stabla sa opisom desnog lista (Izvor: vlastita izrada)

Desna grana stabla sadrži 4251 instance, dakle preostalih 19.67% ukupnih instanci. Dakle prvi lijevi čvor predvodi brojem instanci. Prethodno je navedeno kako na cijenu nekretnina prvog lijevog čvora ponajviše utječu indeks gradnje i zemljopisna širina, dok na cijenu nekretnina prvog desnog čvora u najvećoj mjeri utječu atributi indeks gradnje i kvadratura. Predviđena cijena prvog desnog čvora iznosi 959962.41$ uz očekivanu grešku od 581542.03$. Dakle, skuplje nekretnine su tijekom prve podjele raspoređene u desnoj grani, dok su one jeftinije raspoređene u lijevu granu. Krajnji desni čvor (desni list) sadrži 224 instance, tj 1.04% ukupnih instanci sa predviđenom cijenom od 1881336.52$ i očekivanom greškom od 954955.17$. Na donesenu odluku utjecali su sljedeći atributi: indeks gradnje, kvadratura, pogled na rivu, zemljopisna širina.

Slika 19 odnosi se na grafički prikaz varijabli koje imaju određenu važnost u predviđanju. Varijabla grade (indeks gradnje) ima najveću važnost od 43.27%. Sljedeće po važnosti su varijable kvadrature (važnost: 25.94%) i zemljopisne širine (važnost: 14.03%).



Slika 19: Važnost varijabli u predviđanju (Izvor: vlastita izrada)



Slika 20: Izvještaj modela i grafovi distribucije (Izvor: vlastita izrada)

Grafovi trenutačne distribucije podataka i predviđene distribucije podataka sa slike 20 su jednakog tipa – unimodalna distribucija (nagnuta udesno) te time daju visoku pouzdanost predviđanja cijene nekretnine uz određena odstupanja.

5.2.2. Neuronska mreža

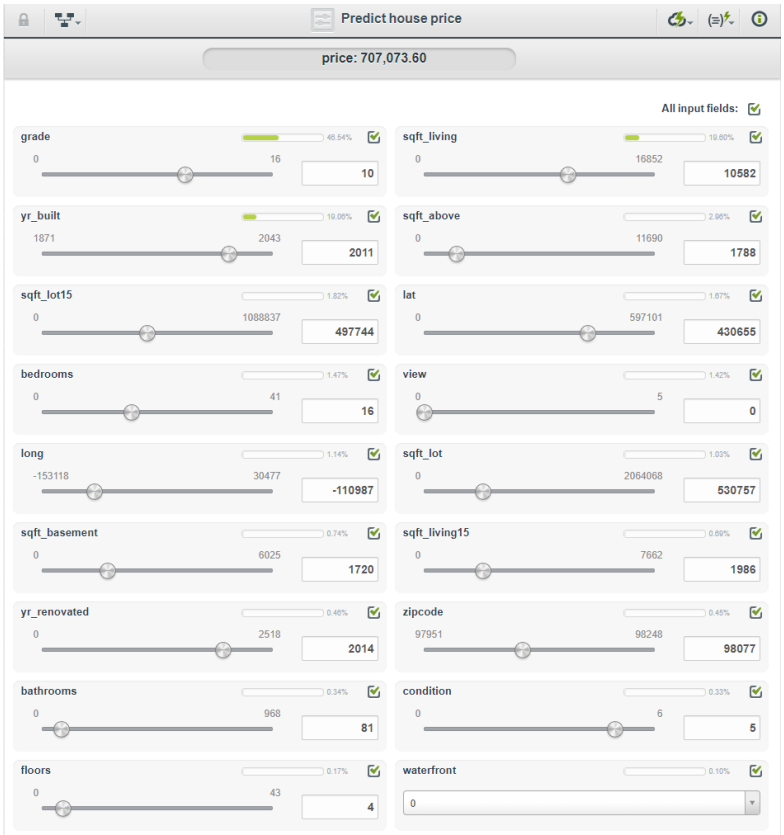
Neuronska mreža predstavlja niz algoritama koji nastoje prepoznati temeljne odnose u skupu podataka i oponašaju način rada ljudskog mozga. [31] Može se prilagoditi promjenama ulaznih podataka te na taj način generirati najbolji mogući rezultat bez potrebe za redizajniranjem izlaznih kriterija. [31] Jednostavna neuronska mreža sastoji se od 3 sloja: ulazni sloj, skriveni sloj i izlazni sloj. [31] Primjenjuje se u razne svrhe poput: financija, detekcija prevara, procjene rizika i sl. [31]

Kao i za izradu stabla odlučivanja, atribut price, odnosno cijena nekretnine je odabrana za zavisnu varijablu koju nastojimo odrediti izradom modela te nastojimo utvrditi koji atributi u kojoj mjeri utječu na rezultat predviđanja.

Za izradu modela neuronske mreže potrebno je podesiti broj skrivenih neurona koje dobivamo kao aritmetičku sredinu neurona na ulazu i izlazu te taj iznos variramo dodavajući ili oduzimajući brojeve od tri do pet. U našem slučaju, variramo iznose od 24 do 26 neurona. Najpreciznije rezultate je davao model sa 24 skrivena neurona. Slijede primjeri provođenja predviđanja i uočavanja predviđene cijene nekretnine, prikaz atributa koji utječu na predikciju te mjere u kojima utječu.

Predviđena cijena nekretnine sa slike 21 iznosi 707073.60$. Atributima poput indeksa kvalitete gradnje, kvadrature, godine gradnje, broja prostorija i sl. dodijeljene su veće vrijednosti te se stoga i očekivala visoka cijena nekretnine.

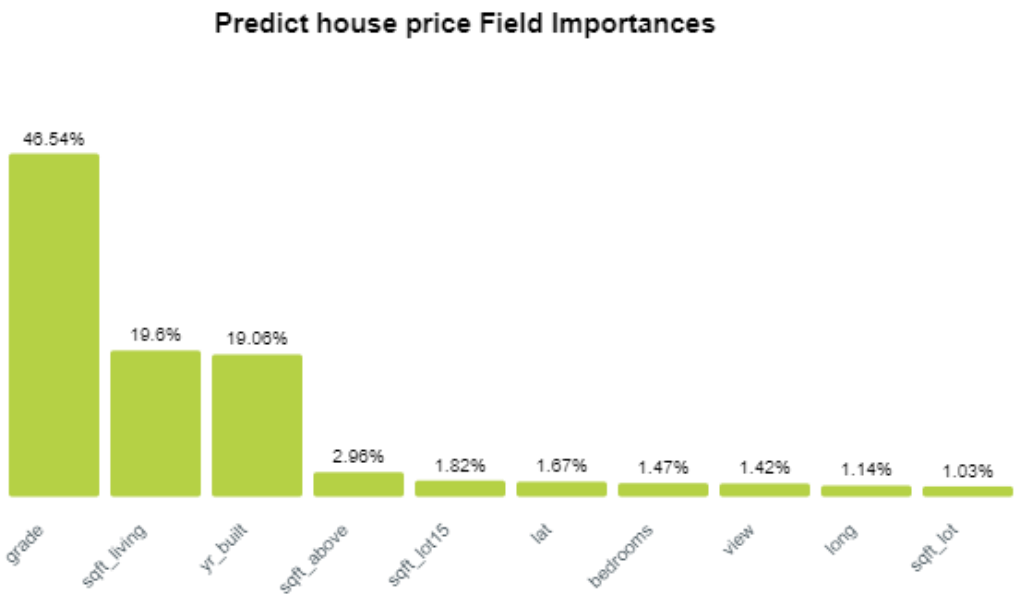
Sljedeće predviđanje cijene vidljivo je na slici 22 te cijena nekretnine iznosi 12144.33$. Za ovaj primjer podesili smo atribute na daleko niže vrijednosti i time zadali lošiju kvalitetu nekretnine, mali broj prostorija i katova, malu kvadraturu, godinu gradnje 1967. te smo isključili atribut vezan uz godinu renovacije. Za očekivati je bilo da bi model predvidio daleko nižu cijenu nekretnine što smo upravo i postigli.



Slika 21: Prvi primjer predviđanja cijene nekretnine (Izvor: samostalna izrada)

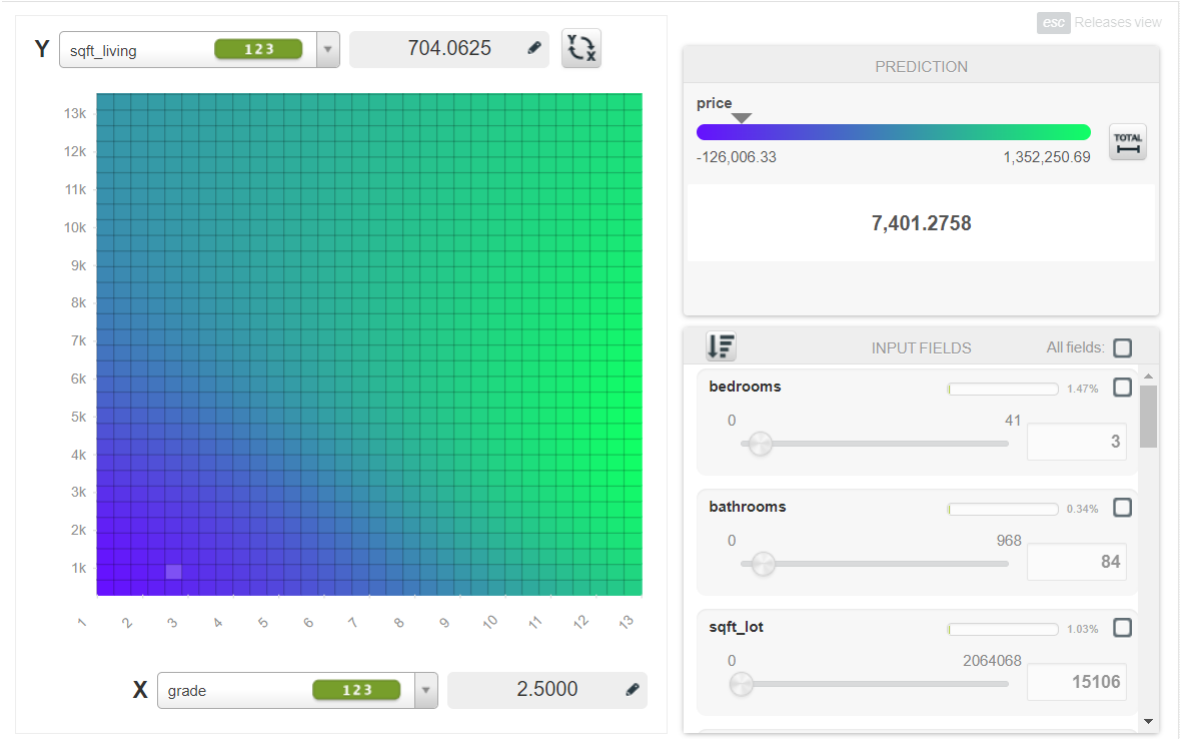
Slika 22: Drugi primjer predviđanja cijene nekretnine (Izvor: samostalna izrada)

Promotrimo atribute koji uvelike utječu na predikciju. Slika 23. sadrži prikaz atributa i pripadne važnosti. Možemo uočiti kako atribut grade (indeks kvalitete gradnje) daleko više utječe na cijenu nekretnine u odnosu na ostale atribute. Važnost tog atributa iznosi 46.54%. Sljedeći atribut koji se ističe po važnosti je kvadratura nekretnine u važnosti od 19.6% te godina gradnje sa istom važnosti. Promatranjem atributa i njihovih važnosti možemo razumjeti kako je cijena nekretnine iz drugog primjera (slika 22.) bitno opala smanjenjem indeksa kvalitete građevine te smanjenjem kvadrature.

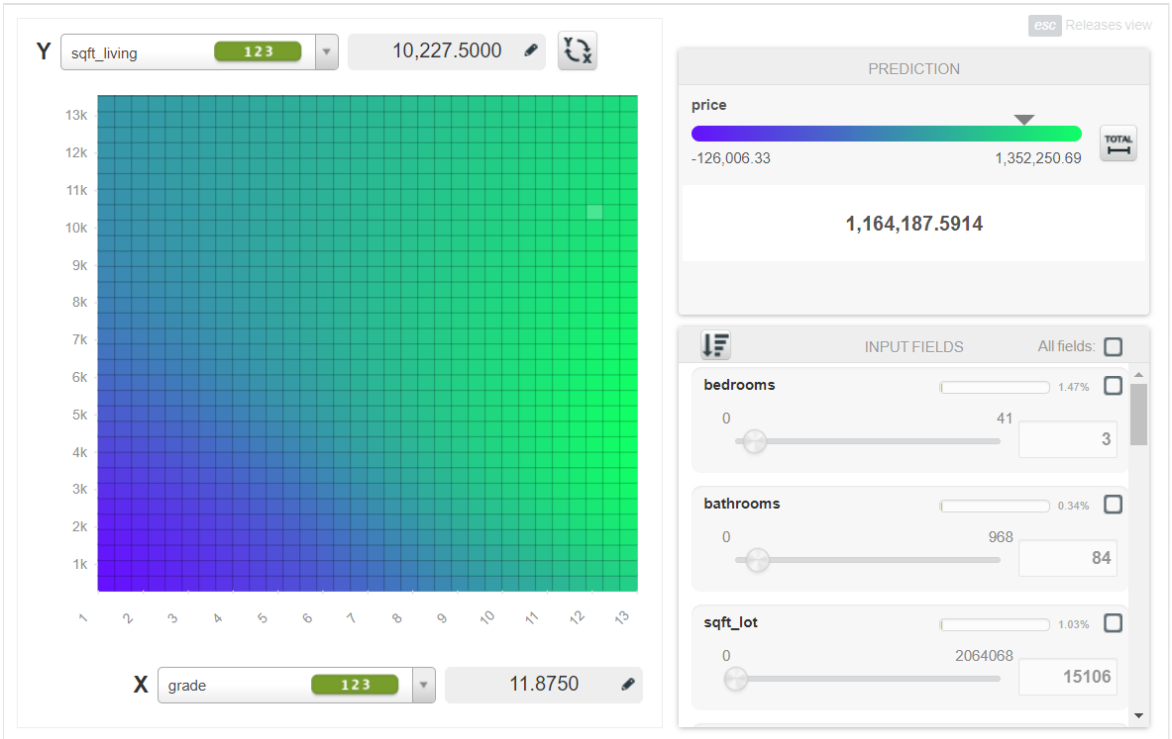


Slika 23: Izvještaj modela neuronske mreže (Izvor: vlastita izrada)

Slike 24 i 25 prikazuju predviđanje cijene nekretnine zadavanjem vrijednosti atributa grade (indeksa kvalitete gradnje) i sqft_living (kvadrature). Promatran je odnos navedena dva atributa iz razloga što najviše utječu na rezultat prema slici 23. Kao što smo vidjeli u ranijim primjerima, zadavanjem male kvadrature i niskog indeksa kvalitete gradnje dobili smo poprilično nisku cijenu nekretnine u iznosu od 7401.2758$ (slika 24). Povećanjem vrijednosti navedenih atributa za očekivati je daleko višu cijenu nekretnine što smo upravo i dobili. Prema slici 25, predviđena cijena nekretnine iznosi 1164187.5914$ za daleko veću vrijednost kvadrature i indeksa kvalitete gradnje nego što smo zadali u prvom primjeru (slika 25).



Slika 24: Prvi primjer predviđanja cijene nekretnine pomoću atributa grade i sqft_living (Izvor: vlastita izrada)



Slika 2510: Drugi primjer predviđanja cijene nekretnine pomoću atributa grade i sqft_living (Izvor: vlastita izrada)

6. Interpretacija i evaluacija modela

Provedenim deskriptivnim i prediktivnim modeliranjem utvrdili smo čimbenike koji utječu na cijenu nekretnine. Usporedimo modele stabla odlučivanja i neuronske mreže koji ukazuju na sljedeće čimbenike: indeks kvalitete gradnje i kvadratura su se pokazali kao čimbenici koji imaju najveći utjecaj na formiranje cijene prema oba modela. Nadalje, prema modelu stabla odlučivanja kao treći čimbenik po važnosti uključuje zemljopisnu širinu, a godina gradnje se ističe kod modela neuronske mreže. Podešavanjem vrijednosti navedenih atributa uočavaju se značajne promjene cijene te time potvrđujemo postavljene hipoteze kako postoji pozitivna povezanost između cijene nekretnine sa kvadraturom unutarnjeg prostora i kvalitetom gradnje. Također je i potvrđena pozitivna povezanost cijene nekretnine sa brojem prostorija, ali u nešto slabijem odnosu. Prediktivni modeli su pokazali kako broj prostorija ima dosta manji utjecaj na cijenu nego što smo očekivali.

Klasterska analiza rezultirala je grupiranjem instanci u klastere odnosno grupe čije su pripadne instance slične po različitim obilježjima. Usporedbom različitog broja klastera, utvrđeno je kako grupiranje unutar 5 klastera daje najoptimalnije rezultate, odnosno na taj način kreira kategorije nekretnina koje možemo odrediti kao skupe i jeftine (dvije krajnosti) te tri kategorije koje ubrajaju nekretnine umjerenih cijena. Unutar tih grupa (klastera) osim slične cijene, nekretnine imaju slične vrijednosti ostalih obilježja poput kvadrature, lokacije, godine gradnje i sl.

Sada kada smo utvrdili čimbenike koji najviše utječu na cijenu nekretnine možemo dati smjernice za njezino pravilno određivanje. Prilikom formiranja cijene nekretnine promatramo razne faktore koji su već navedeni i opisani te nekretnine svrstavamo u cjenovni rang. Dakle veća cijena se dodjeljuje nekretninama sa većim kvadraturama za razliku od prosječne cijene, nekretninama izgrađenim od kvalitetnih materijala, sa privlačnim dizajnom interijera i eksterijera te onima koje su nedavno renovirane ili novije gradnje.

7. Zaključak

Cilj rada bio je provesti tehnike strojnog učenja i pomoću njih izraditi deskriptivne i prediktivne modele nad skupom podataka o prodaji nekretnina, utvrditi čimbenike koji utječu na cijenu te dati smjernice za preciznije određivanje cijene nekretnine. Provedene su tehnike nadgledanog i nenadgledanog učenja koje uključuju klastersku analizu, faktorsku analizu, stablo odlučivanja i neuronsku mrežu. Utvrdili smo kako na cijenu nekretnine ponajviše utječu kvaliteta gradnje i kvadratura. Prediktivni modeli su pokazali dobre rezultate. Rezultati su ponajviše dobri zbog velikog broja instanci skupa podataka i nepostojanja nedostajućih instanci. Međutim, predviđanja nisu sto posto pouzdana i uvijek može doći do velikih odstupanja te je potrebno uzimati i nekakve druge faktore u obzir prilikom određivanja cijene poput opadanja cijene nekretnine kroz vrijeme i drugih ekonomskih faktora. Rezultati ovog istraživanja mogu pomoći u razumijevanju tih faktora i proučavanju variranja cijene obzirom na njihovo podešavanje te promatranje odstupanja predviđene cijene od one realne.

Za izradu modela koristio se alat BigML (opsežna online platforma za strojno učenje) kao što je naglašeno u prethodnim poglavljima. Alat omogućava vrlo jednostavno korištenje i obradu podataka te niz mogućnosti analize i manipulacije podataka.

Kao što je navedeno, izradom i analizom modela smo potvrdili hipoteze postavljene na početku izrade rada – potvrdili smo pozitivnu povezanost cijene nekretnine sa kvalitetom gradnje i kvadraturom. No, očekivali smo kako će i neki drugi faktori poput pogleda na rivu ili broja prostorija imati velik utjecaj na cijenu što rezultati nisu pokazali.

Na kraju, uspoređivanjem dobivenih rezultata sa sličnim istraživanjima koji su se bavili predviđanjem cijena nekretnina na području grada Ames (Iowa, SAD), na području okruga Petaling (Kuala Lumpur, Malezija) i grada Christchurch (Novi Zeland) možemo zaključiti kako na cijenu nekretnine utječu brojni faktori među kojima su i oni koje smo u ovom radu utvrdili kao čimbenike velikog utjecaja – kvadratura i kvaliteta gradnje.

Popis literature

[1] Institut Ruđer Bošković, „Otkrivanje znanja dubinskom analizom podataka, Priručnik za istraživače i studente“. [Na internetu]. Dostupno na: http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf . [Pristupljeno: 14-kolovoz-2020].

[2] M.Rouse, „Machine Learning“, 2016. [Na internetu]. Dostupno na: https://searchenterpriseai.techtarget.com/definition/machine-learning-ML . [Pristupljeno: 14-kolovoz-2020].

[3] Expert System, „What is Machine Learning? A definition“, 2020. [Na internetu]. Dostupno na: https://expertsystem.com/machine-learning-definition/ . [Pristupljeno: 15-kolovoz-2020].

[4] D.Marković, Strojno učenje, 2018. Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za fiziku. Osijek.

[5] B.Kliček, D.Oreški, Otkrivanje znanja u podacima, 2015. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Varaždin.

[6] J.Brownlee, „Supervised and Unsupervised Machine Learning Algorithms“, 2016. [Na internetu]. Dostupno na: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ . [Pristupljeno: 16-kolovoz-2020].

[7] B. D. Bašić, J. Šnajder, Uvod u strojno učenje, 2016. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva. Zagreb.

[8] King County, „Demographic Trends of King County“, 2020. [Na internetu]. Dostupno na: https://www.kingcounty.gov/independent/forecasting/King%20County%20Economic%20Indicators/Demographics.aspx . [Pristupljeno: 18-srpanj-2020].

[9] T.M. Reddy, „Analysis and Prediction of House Sales in King County, USA“, 2018. [Na internetu]. Dostupno na: https://rpubs.com/Mani/report . [Pristupljeno: 18-srpanj-2020].

[10] J. Reynolds, „10 Year Chart Of The Seattle Real Estate Market Is Mind Blowing, Up 93% Since The Bottom“, 2018. [Na internetu]. Dostupno na: http://www.urbancondospaces.com/10-year-chart-seattle-real-estate-is-mind-blowing/ . [Pristupljeno: 19-srpanj-2020].

[11] Skup podataka: House Sales in King County, USA, 2016. [Na internetu]. Dostupno na: https://www.kaggle.com/harlfoxem/housesalesprediction . [Pristupljeno: 18-srpanj-2020].

[12] I. Bershadskiy, „Predicting House Price Using Regression Algorithm for Machine Learning“, 2020. [Na internetu]. Dostupno na: https://yalantis.com/blog/predictive-algorithm-for-house-price/ . [Pristupljeno: 18-srpanj-2020].

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

[14] H. Yu, J. Wu, „Real Estate Price Prediction with Regression and Classification“, 2016. [Na internetu]. Dostupno na: http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf . [Pristupljeno: 27-kolovoz-2020].

[15] A. G. Sarip, M. B. Hafez, M. N. Daud, „Application Of Fuzzy Regression Model For Real Estate Price Prediction“, 2016. [Na internetu]. Dostupno na: https://ejournal.um.edu.my/index.php/MJCS/article/view/6889 . [Pristupljeno: 27-kolovoz-2020].

[16] V. Limsombunchai, „House Price Prediction: Hedonic Price Model vs. Artificial Neural Network“, 2004. [Na internetu]. Dostupno na: http://researcharchive.lincoln.ac.nz/bitstream/handle/10182/5198/House_%20price_%20prediction.pdf?sequence=1&isAllowed=y . [Pristupljeno: 27-kolovoz-2020].

[17] King County, „Residential Glossary of Terms“, 2017. [Na internetu]. Dostupno na: https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r . [Pristupljeno: 23-srpanj-2020].

[18] J.P. Murillo, „King County Home Sales: Analysis and the limitations of a multiple regression model“, 2016. [Na internetu]. Dostupno na: https://rstudio-pubs-static.s3.amazonaws.com/155304_cc51f448116744069664b35e7762999f.html . [Pristupljeno: 23-srpanj-2020].

[19] Analytics Vidhya, „What Is Correlation in Machine Learning?“, 2020. [Na internetu]. Dostupno na: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47 . [Pristupljeno: 9-kolovoz-2020].

[20] J.Frost, Introduction to Statistics: An Intuitive Guide, ebook, 2019.

[21] J.Ramzai, „Clearly explained: Pearson V/S Spearman Correlation Coefficient“, 2020. [Na internetu]. Dostupno na: https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8 . [Pristupljeno: 13-kolovoz-2020].

[22] Educba, „Predictive Analytics vs Descriptive Analytics“. [Na internetu]. Dostupno na: https://www.educba.com/predictive-analytics-vs-descriptive-analytics/ . [Pristupljeno: 18-kolovoz-2020].

[23] Valamis, „What is Descriptive Analytics?“. [Na internetu]. Dostupno na: https://www.valamis.com/hub/descriptive-analytics . [Pristupljeno: 18-kolovoz-2020].

[24] 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: 19-kolovoz-2020].

[25] T. Bock, „What is Cluster Analysis?“, 2018. [Na internetu]. Dostupno na: https://www.displayr.com/what-is-cluster-analysis/ . [Pristupljeno: 23-kolovoz-2020].

[26] D. Oreški, Klaster analiza, materijal s laboratorijskih vježbi iz kolegija Otkrivanje znanja u podacima ak. god. 2019./2020. Sveučilište u Zagrebu, Fakultet organizacije i informatike. Varaždin.

[27] Statistics Solutions, „Factor Analysis“. [Na internetu]. Dostupno na: https://www.statisticssolutions.com/factor-analysis-sem-factor-analysis/ . [Pristupljeno: 24-kolovoz-2020].

[28] M. Rahn, „Factor Analysis: A Short Introduction, Part 1“, 2012. [Na internetu]. Dostupno na: https://www.theanalysisfactor.com/factor-analysis-1-introduction/ . [Pristupljeno: 24-kolovoz-2020].

[29] I.T. Jolliffe, J. Cadima, „Principal component analysis: a review and recent developments“, 2016. [Na internetu]. Dostupno na: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 . [Pristupljeno: 24-kolovoz-2020].

[30] R.S. Brid, „Decision Trees — A simple way to visualize a decision“, 2018. [Na internetu]. Dostupno na: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb . [Pristupljeno: 25-kolovoz-2020].

[31] J. Chen, „Neural Network“, 2020. [Na internetu]. Dostupno na: https://www.investopedia.com/terms/n/neuralnetwork.asp . [Pristupljeno: 26-kolovoz-2020].

Popis slika

Popis tablica