

UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
NOVI SAD

Odsek/smer/usmerenje:
Računarstvo i automatika/Računarske nauke i informatika

SEMINARSKI RAD
iz predmeta
Sistemi za istraživanje i analizu podataka

Kandidat: **Teodor Vasić**

Broj indeksa: **E2 30/2020**

Tema rada: **Primena istraživanja i analize podataka u cilju
povećanja uspeha obrađivanja zahteva za kredit**

Mentor rada: **dr Aleksandar Kovačević**

Mesto i datum:
Novi Sad 2023.

SADRŽAJ

1. UVOD.....	5
2. OPIS PROBLEMA.....	8
2.1. Opis zadatka.....	8
3. SKUP PODATAKA - Eksplorativna analiza podataka.....	10
3.1. Opis atributa.....	10
3.2. Analiza raspodele ciljnog atributa	15
3.3. Problem nedostajućih vrednosti.....	16
3.4. Problem koreliranih atributa	16
3.5. Transformacija atributa.....	17
3.6. Istraživanje nominalnih atributa	21
3.7. Istraživanje numeričkih atributa	28
3.8. Analiza veza između više atributa i ciljnog atributa.....	30
4. FAZA MODELOVANJA I EVAULACIJE.....	33
4.1 Klasterovanje	33
4.1.1 Teorijske osnove	33
4.1.2 Primena klasterovanja.....	37
4.1.3 Proučavanje rezultata klasterovanja	39
Klaster 0	41
Klaster 1	41
Klaster 2	41
4.2 Analiza asocijativnih pravila	41
4.2.1 Teorijske osnove	41
4.2.2 Primena analize asocijativnih pravila	44
4.2.3 Proučavanje rezultata analize asocijativnih pravila.....	45
4.3 Regresija	49
4.3.1 Teorijske osnove	49
4.3.1.1 Jednostruka linearna regresija.....	49
4.3.1.2 Višestruka linearna regresija.....	54
4.3.2 Jednostruka regresija	54
4.3.2.1 Primena jednostrukih regresija.....	54
4.3.2.2 Proučavanje rezultata jednostrukih regresija	55
4.3.3 Višestruka regresija	59
4.3.3.1 Primena višestruke regresije	59
4.3.3.2 Proučavanje rezultata višestruke regresije.....	59

4.4	Klasifikacija	65
4.4.1	Stabla odlučivanja	66
4.4.1.1	Teorijske osnove	66
4.4.1.2	Primena i proučavanje rezultata metode stabla odlučivanja	72
4.4.2	K -najbližih suseda	76
4.4.2.1	Teorijske osnove	76
4.4.2.2	Primena i proučavanje rezultata metode k -najbližih suseda	79
4.4.3	Naivna <i>Bayes</i> -ova metoda.....	82
4.4.3.1	Teorijske osnove	82
4.4.3.2	Primena i proučavanje rezultata naivne <i>Bayes</i> -ove metode	84
4.4.4	Mašine potpornog vektora.....	85
4.4.4.1	Teorijske osnove	85
4.4.4.2	Primena i proučavanje rezultata metode maštine potpornog učenja	92
4.5	Evaluacija	94
5.	ZAKLJUČAK	98
	LITERATURA.....	99

1. UVOD

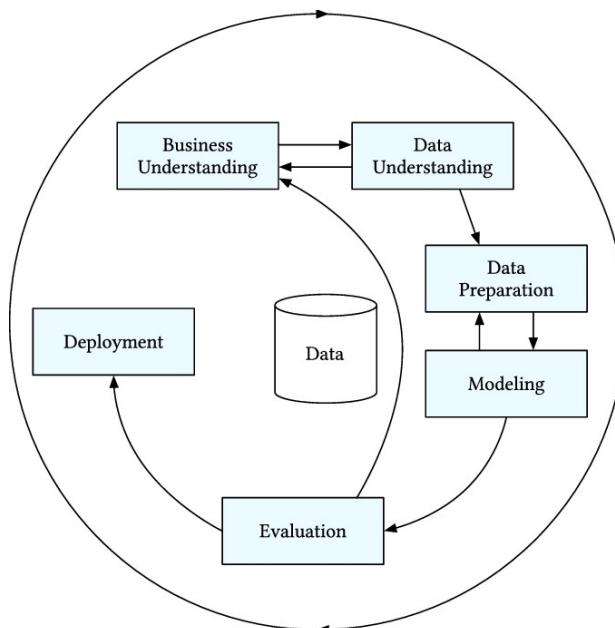
U današnjem digitalnom dobu, količina dostupnih podataka eksponencijalno raste. Od poslovnih transakcija i senzorskih merenja do društvenih mreža i mobilnih uređaja, podaci su postali ključni resurs za donošenje informiranih odluka i otkrivanje vrednih uvida. Međutim, sama količina podataka nije dovoljna za ostvarivanje njihove vrednosti. Potrebna je analiza podataka (engl. *Data mining*), proces koji nam pomaže razumeti, tumačiti i izvući smislene informacije iz velikih i kompleksnih skupova podataka.

Data mining, ili rudarenje podataka, je proces otkrivanja korisnih informacija, obrazaca i znanja iz velikih skupova podataka. To je grana analize podataka koja se bavi ekstrakcijom znanja iz sirovih podataka primjenom različitih tehnika, algoritama i statističkih metoda. Glavni cilj data mininga je otkrivanje skrivenih uzoraka, trendova i povezanosti u podacima koji mogu biti korisni za donošenje informiranih odluka. Ova tehnika ima široku primjenu u različitim područjima, uključujući poslovne, zdravstvene, financijske, marketinške i istraživačke domene.

Analiza podataka u ovom radu biće korišćena da se pokuša rešiti problem procene. Problem procene odnosi se na postupak procene vrednosti ili karakteristika neke varijable ili pojave na temelju raspoloživih podataka. Osnovna ideja je da se na temelju dostupnih informacija i statističkih metoda doneše procena ili predviđanje vrednosti ili ponašanja neke varijable koja nije direktno dostupna ili izmerena.

Ovaj rad će biti fokusiran na problem procene kojim budućim aplikantima treba biti odobren zahtev za kredit a kojima odbijen. Sam problem procene odobravanja kredita odnosi se na postupak procene rizika i sposobnosti samih aplikantata da otplaćuju kredit. Financijske institucije, kao što su banke, moraju doneti odluku o odobravanju ili odbijanju zahteva za kredit na osnovu dostupnih informacija o samom aplikantu koje imaju na raspolaganju.

Studija slučaja opisanog u ovom radu biće izvršena upotreboom procesa za primenu Istraživanja podataka nazvanog *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Po ovom procesu upotreba Istraživanja podataka sastoji se iz 6 faza prikazanih na slici 1.1.



Slika 1.1 *Cross-Industry Standard Process for Data Mining*

Ovaj rad koncipiran je tako, da nakon uvodnog dela, sledi poglavlje pod nazivom Opis problema. U ovom poglavlju je detaljno opisan problem procene rejtinga aplikanata za kredit jedne nemačke banke.

Treće poglavlje koje nosi naziv Skup podataka – sadrži konkretan opis i uvid u skup podataka koji će se korisiti prilikom pisanja seminariskog rada. Vršeno je istraživanje odnosa između atributa na osnovu čega su uočeni određeni šabloni među podacima. Dat je detaljan opis atributa koji opisuju klijenta banke sa svim mogućim vrednostima tih atributa. Nakon toga je prikazana analiza raspodele ciljnog atributa. Potom je predstavljen problem nedostajućih vrednosti, transformacija atributa i problem koreliranih atributa. Posle toga sledi istraživanje nominalnih odnosno numeričkih atributa. Na kraju poglavlja prikazana je analiza veza između više atributa i ciljnog atributa.

U četvrtom poglavlju, pod nazivom Formiranje modela, vršeno je formiranje modela koristeći sledeće tehnike: klasterovanje, asocijativna pravila, regresija (jednostruka, višestruka) i klasifikacija (stabla odlučivanja, k -najbližih suseda, naivna Bayes-ova metoda, maštine potpornog učenja). Osim teoretskih opisa svake tehnike pojedinačno, izvršena je i njihova primena na samom skupu podataka.

U sledećem poglavlju, pod nazivom Zaključak, dat je osvrt na postignute rezultate i uočene nedostatke.

2. OPIS PROBLEMA

U ovom radu su iskorišćene tehnike *data mining-a* kako bi se analizirala uspešnost dodeljivanja rejtinga aplikantima za kredit jedne nemačke banke.. Osnovni zadatak je bio da se formira funkcija ili pravilo za dodeljivanje rejtinga budućim aplikantima za kredit.

2.1. Opis zadatka

Dat je skup podataka o aplikantima za kredit. Svakom od aplikanata dodeljen je jedan od dva moguća rejtinga: "dobar za kredit" ili "loš za kredit". Rejting je dodeljen na osnovu podataka koji se sastoje od 20 atributa za svakog aplikanta (starost, kreditna istorija itd.) i podatke u vezi klijenta vezane za prethodne kredite(zaostatak sa plaćanjem rate, svrha i sl.).

Potrebno je formirati funkciju ili pravilo za dodeljivanje rejtinga budućim aplikantima(da li će banka da zaradi ili izgubi ako odobri kredit) kao i opisati svaki od skupova aplikanata: dobar i loš rejting i objasniti zašto nekome treba odobriti kredit, a nekome ne treba.

Ako banka ne doneše dobro odluku kojem aplikantu za kredit je rejting dobar, a kome aplikantu je loš, ona snosi određene novčane posledice u vidu gubitaka.

Glavni cilj banke jeste da poveća profit te banke, i da izbegne gubitke, tako što će najpričnije odrediti koji aplikanti imaju dobar a koji imaju loš rejting. Iz tog razloga modeli za klasifikaciju moraju da budu formirani sa tim ciljem. To dalje implicira da je u fazu evaluacije potrebno uvesti mere koje u obzir uzimaju zaradu banke

Biće razmotrena sva četiri moguća ishoda klasifikacije: *true positive, false positive, true negative i false negative*.

True negative (TN), model je predviđao da klijent ima loš rejting i to se pokazalo istinito. Očigledno da je da u ovom slučaju nema ni troška ni zarade za banku.

True positive (TP), model je predvideo da klijent ima dobar rejting te mu je i samim time zahtev za kredit odobren. Ako se uzme u računicu, naspram dostupnih podataka, da u proseku klijenti potražuju kredit u iznosu od 3270DM, sa vremenskom otplatom na 20 meseci gde u proseku banka daje kamatu od 3%, na osnovu računice $3270 * 0.03 * 20$ dolazimo do iznosa od 1962DM koliko banka zaradi u datom periodu.

False negative (FN), znači da je model predvideo da klijent ima dobar rejting za klijenta koji ima loš. Ovo je obično greška koja najviše košta. Posmatraće se najgori mogući slučaj, a to je da kredit uopšte ne bude vraćen tako da dolazimo do računice od profita banke + zahtevane sume = $1962 + 3270 = 5232$ DM.

False positive (FP), model je odredio da ne treba odobriti kredit klijentu koji ima dobar rejting. Očigledno je da je u ovom slučaju gubitak mogući profit, tačnije 1962DM.

		Odluka banke	
		Dobar	Loš
Stvarni rejting	Dobar	1962DM	-1962DM
	Loš	-5232DM	0DM

Tabela 2.1 Matrica troška

3. SKUP PODATAKA - Eksplorativna analiza podataka

Pojam eksplorativna analizu podataka (engl. Explorative Data Analysis – EDA) osmislio je John Turkey 1970. godine. Turkey je obično povezivao EDA sa detektivskim poslom. Uloga onoga ko analizira podatke jeste da „sluša“ podatke sve dok se ne pojavi verodostojna verzija podataka, iako takva verzija možda nije logički očekivana.

Eksplorativna analiza podataka je proces istraživanja i otkrivanja skrivenih obrazaca, trendova, povezanosti i značajnih informacija u skupu podataka. Cilj eksplorativne analize podataka je stvoriti intuitivno razumevanje podataka, generisati hipoteze i identifikovati potencijalne varijable ili faktore od interesa za dalju analizu. Ključni element eksplorativne analize podataka je vizualizacija podataka. Koristeći različite grafičke metode kao što su histogrami, grafikoni raspršenosti, dijagrami, možemo vizualno prikazati podelu, trendove i anomalije u podacima. Vizualizacija pomaže u otkrivanju nepravilnosti, izuzetaka ili potencijalnih vrednosti koje bi mogle biti zanimljive za dalje istraživanje.

3.1. Opis atributa

U ovom radu se obraduje skup podataka o aplikantima za kredit nemačke banke. Skup se sastoji od 1000 klijenata, gde je svakom od njih dodeljen jedan od dva moguća ishoda: „klijent je dobar za kredit“ i „klijent je loš za kredit“. Svaki klijent je opisan sa 20 atributa, 7 numeričkih i 13 kategoričkih. Naziv svakog atributa i njegovo značenje predstavljeno je u Tabeli 3.1.

R. br.	Naziv atributa	Značenje atributa
1	Stanje na računu	Trenutno stanje na računu u DM
2	Trajanje računa	Trajanje računa u mesecima
3	Istorija kredita	Da li je trenutno ili je u prošlosti klijent imao kredit
4	Svrha uzimanja	Razlog zbog kog klijent aplicira za kredit

5	Iznos kredita	Iznos kredita
6	Račun za štednju	Prosečan iznos na računu za štednju
7	Zaposlenost	Trenutan status zaposlenosti klijenta
8	Kamata	U procentima od tekućih prihoda
9	Pol i bračni status	Pol i bračni status
10	Žirant / Ko-aplikant	Da li ova aplikacija za kredit ima žiranta ili ko-aplikanta
11	Prebivalište	Koliko aplikant ima prebivalište u mestu u kome je banka mereno u godinama
12	Imovina	Da li aplikant posede imovinu
13	Godine starosti	Starost aplikanta merena u godinama
14	Plaćanje u ratama	Da li aplikant otplaćuje nešto u ratama i gde.
15	Stan/Kuća	Da li aplikant plaća stanarinu, posede stan/kuću ili živi besplatno
16	Postojeći krediti	Broj postojećih kredita u ovoj banci
17	Vrsta posla	Da li je aplikant aktivno zaposlen i stepen obrazovanja
18	Izdržavani	Broj lica koji zavise od prihoda aplikanta
19	Telefon	Telefon aplikanta
20	Radnik stranac	Da li je aplikant radnik iz druge države
21	Odobren kredit	Klasno obeležje koje označava da li aplikantu treba odobriti kredit ili ne

Tabela 3.1 Spisak atributa i njihovih značenja

U narednim tabelama biće opisan svaki kategorički atribut pojedinačno sa svojim mogućim vrednostima i njihovim opisima.

Tabela Stanje na računu

Stanje na računu	
A11	< 0 DM
A12	0 <= ... < 200 DM
A13	>= 200 DM ili redovna primanja u prethodnih godinu dana/a
A14	nema računa u ovoj banci

Tabela Istorija kredita

Istorija kredita	
A30	nema uzetih kredita ili svi krediti vraćeni na vreme (u bilo kojoj banci)
A31	svi krediti <u>u ovoj banci</u> vraćeni na vreme
A32	ima druge kredite ali su uplate do sada sve bile na vreme
A33	kašnjenje sa uplatama kod prošlih kredita
A34	kritičan račun

Tabela Svrha uzimanja

Svrha uzimanja	
A40	automobil (nov)
A41	automobil (polovan)
A42	nameštaj
A43	muzička tehnika/televizija
A44	bela tehnika
A45	opravke
A46	obrazovanje
A47	(odmor – ove vrednosti nema u skupu podataka?)
A48	prekvalifikovanje (<i>retraining</i>)

A49	biznis
A410	ostalo

Tabela Račun za štednju

Račun za štednju	
A61	< 100 DM
A62	100 <= ... < 500 DM
A63	500 <= ... < 1000 DM
A64	.. >= 1000 DM
A65	nema računa za štednju

Tabela Zaposlenost

Zaposlenost	
A71	nezaposlen
A72	1 godina
A73	1 <= ... < 4 godine
A74	4 <= ... < 7 godina
A75	.. >= 7 godina

Tabela Pol i bračni status

Pol i bračni status	
A91	muško, razveden
A92	žensko, uodata/razvedena
A93	muško, neoženjen
A94	muško, oženjen/udovac
A95	žensko, neudata

Tabela Žirant/Ko-aplikant

Žirant/Ko-aplikant	
A101	nema

A102	ko-aplikant
A103	žirant

Tabela Imovina

Imovina	
A121	Poseduje nekretninu
A122	Ulaganja u nekretnine / životno osiguranje
A123	Automobil, (ne onaj za koji traži kredit)
A124	nema

Tabela Plaćanja u ratama

Plaćanja u ratama	
A141	Banka
A142	Radnja
A143	Nema

Tabela Stan/Kuća

Imovina	
A151	Stanarina
A152	Poseduje
A153	Živi besplatno

Tabela Vrsta posla

Vrsta posla		
A171	nezposlen/nema	zvanično obrazovanje
A172	nema zvanično obrazovanje – ima	stalan posao
A173	zaposlen i ima	zvanično obrazovanje
A174	radi u menadžmentu/samo-zaposlen/visoko obrazovanje/oficir	

Tabela Telefon

Telefon	
A191	nema
A192	ima registrovan telefon pod svojim imenom

Tabela Radnik Stranac

Radnik stranac	
A201	Da
A202	Ne

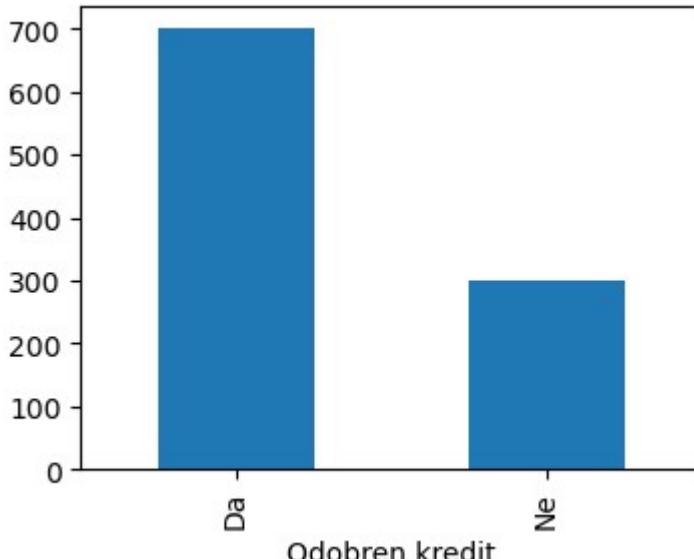
Tabela Odobren Kredit

Odobren kredit – ciljno obeležje	
1	Da
2	Ne

3.2. Analiza raspodele ciljnog atributa

Eksplorativna analiza se obavlja u više koraka. Prvi podrazumeva analizu raspodele ciljnog atributa. Sa slike 3.1 se vidi da se kod, od ukupno 1000

klijenta, čak 700 puta javlja vrednost „Da“, odnosno 300 puta vrednost „Ne“. Iz čega sledi da je odnos mogućih vrednosti u procentima 70% naspram 30% u korist „Da“ vrednosti.



Slika 3.1 Analiza raspodele ciljnog atributa Odobren kredit

3.3. Problem nedostajućih vrednosti

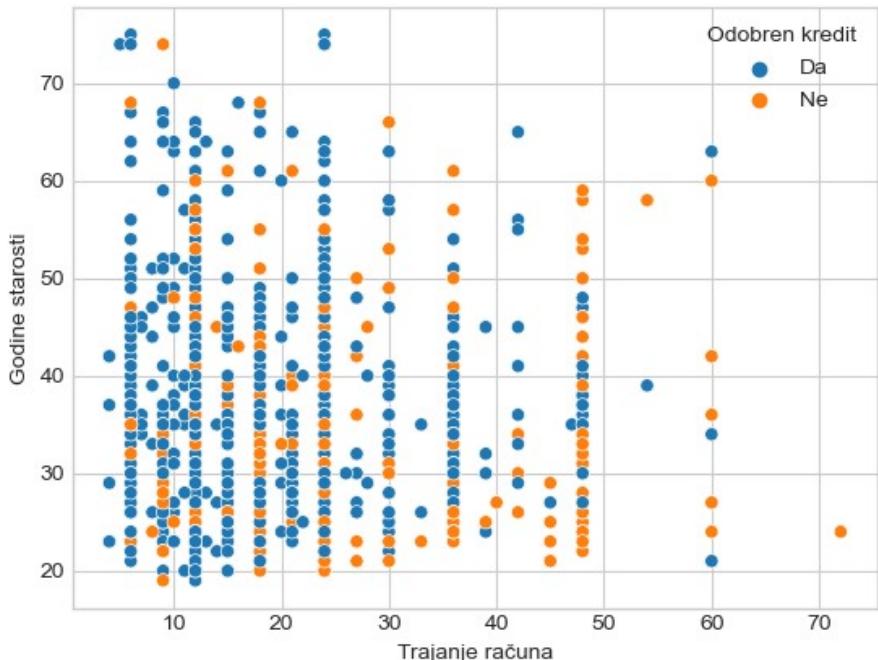
Proučavanjem datih podataka koristeći jednostavnu proveru: `data.isNull().values.any()` došlo se do zaključka da svi atributi u datom skupu podataka imaju vrednosti i sa te strane nije potrebna nikakva dodatna manipulacija sa te strane.

3.4. Problem koreliranih atributa

Problem koreliranih atributa je situacija u kojoj postoje visoke ili jake veze između različitih atributa u skupu podataka. Ova korelacija može biti pozitivna (kada se vrednosti dva atributa povećavaju zajedno) ili negativna (kada se vrednosti jednog atributa smanjuju dok se vrednosti drugog atributa povećavaju). Atributi u korelaciji predstavljaju potencijalni problem u istraživanju podataka iz dva razloga. Prvi je što tokom procesa

istraživanja dolazi do prenaglašenog značaja jednog atributa. Drugi je što pomenuti problem dovodi do formiranja nepouzdanih i nestabilnih modela.

Da bi se ispitalo postojanje problema koreliranih atributa na datom skupu podataka koristimo *scatter* funkciju *seaborn* modula u python-u. Po pozivanju funkcije sa datim podacima, prikazuje se dijagram za prikaz odnosa vrednosti atributa. Na slici 3.2 prikazan je odnos atributa *Godine starosti* - godine klijenta i *Trajanje računa* – koliko dugo klijent ima otvoren račun u banci. Dijagram jasno pokazuje da ne postoji korelacija između ova dva atributa.

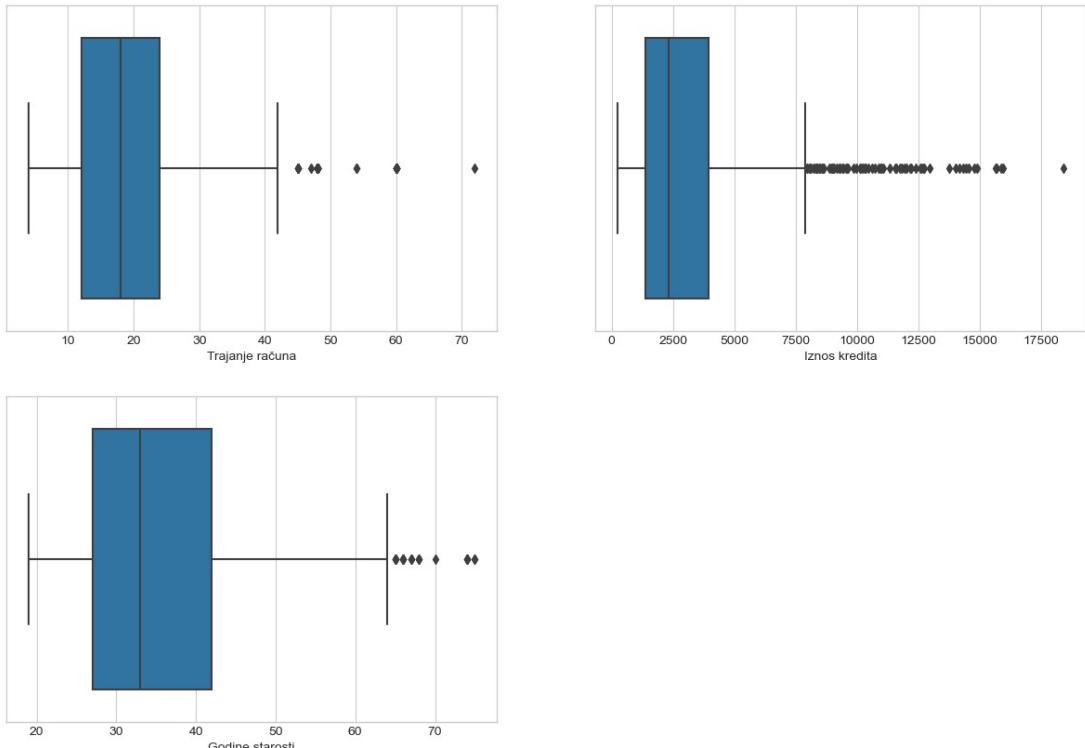


Slika 3.2 Odnos atributa *Godine starosti* i *Trajanje računa*

Nakon detaljne analize, gde su upoređivani odnosi svih kombinacija atributa, izведен je zaključak da u datom skupu podataka ne postoji problem korelacije atributa.

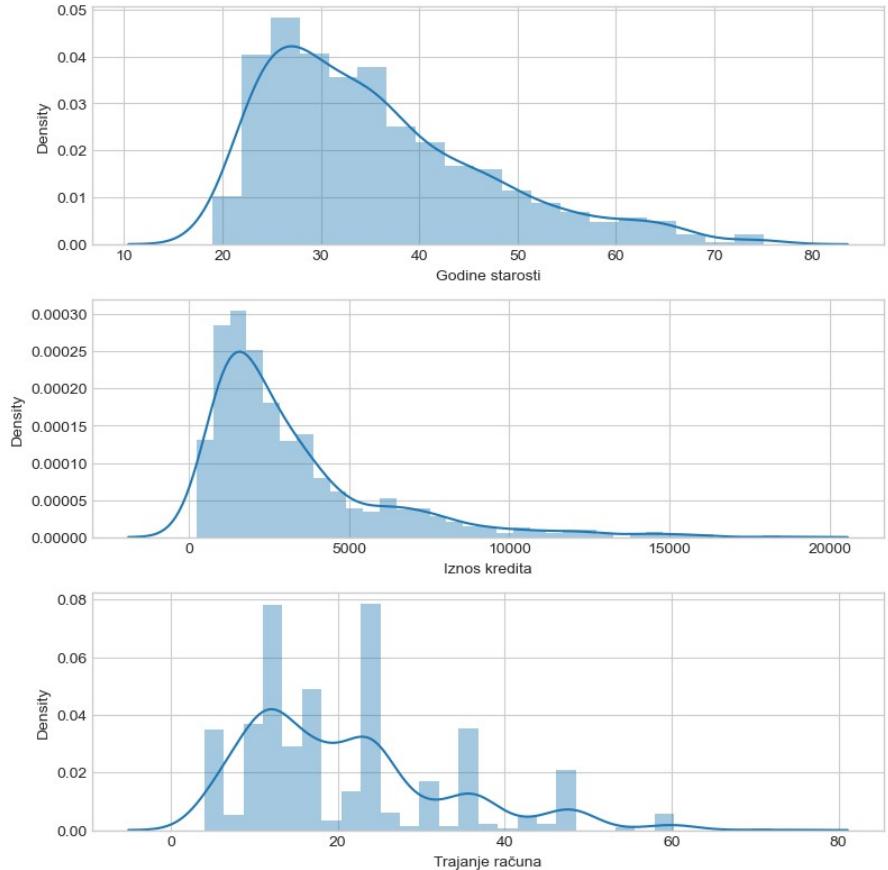
3.5. Transformacija atributa

Pre same transformacije atributa izvršićemo analizu istih, da bi došli do ideje šta tačno treba da uradimo da bi sama transformacija imala smisla i pomogla nam na putu do rešenja samog problema. Za početak ćemo izvršiti analizu numeričkih atributa. Koristeći funkciju *boxplot* biblioteke *seaborn* možemo viditi da su nam potrebne dodatne transformacije.



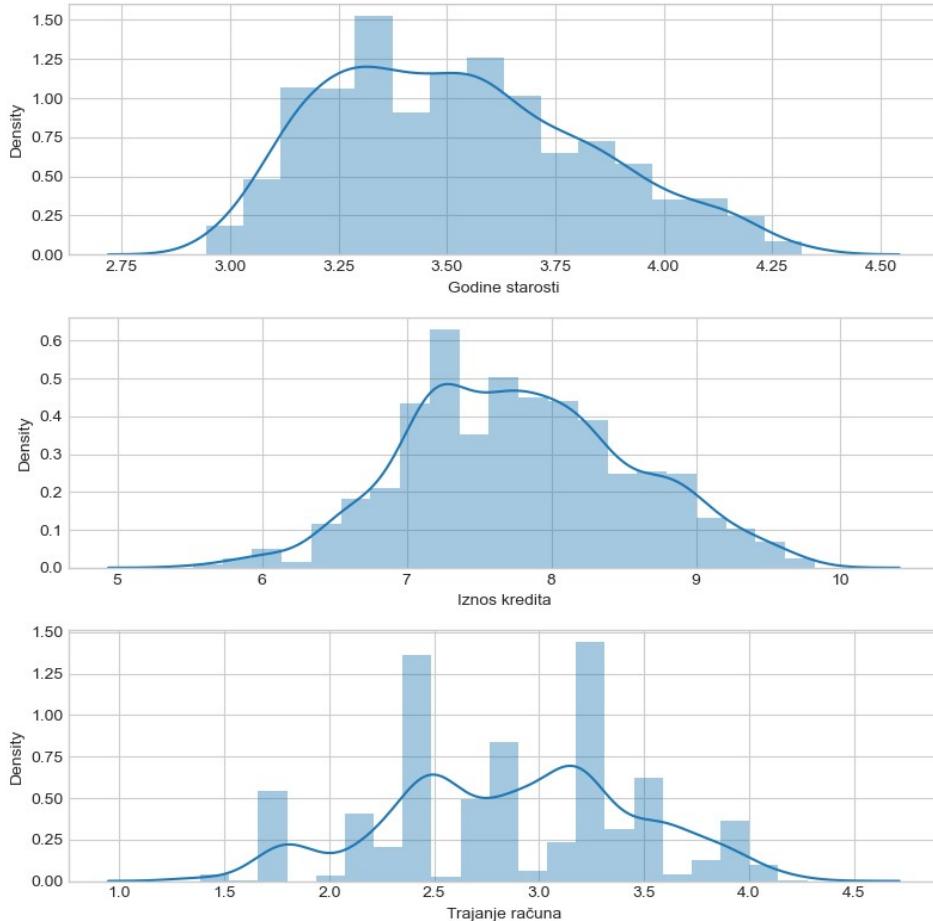
Slika 3.3 Pregled odstupanja numeričkih atributa

Sa slike 3.3 vidimo da postoje određena odstupanja nekih od numeričkih atributa. Da bi videli histograme atributa *Trajanje računa*, *Godine starosti* i *Iznos kredita*, kao i tačnu raspodelu naspram srednje vrednosti atributa koristićemo funkciju *distplot* biblioteke *seaborn* i rezultat možemo videti na slici 3.4. Na rezultatu možemo da vidimo da je raspodela vrednosti neravnomerna i da su uglavnom vrednosti sa leve strane srednje vrednosti.



Slika 3.4 Raspodela vrednosti numeričkih atributa

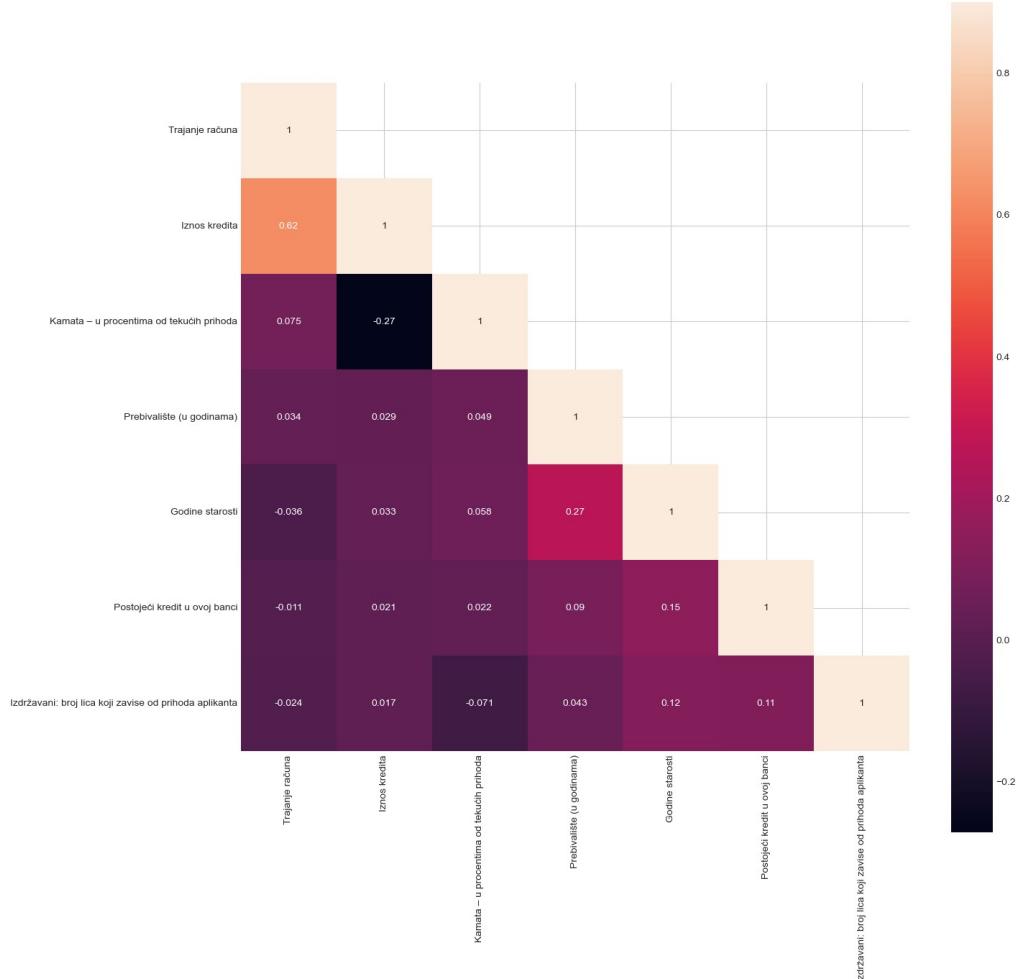
Ako posmatramo histogram atributa, prikazan na slici 3.5, posle logaritamske transformacije koju smo izveli koristeći funkciju *log* iz biblioteke *numpy* vidimo da su podaci sada mnogo ravnomernije distribuirani u odnosu na srednju vrednost.



Slika 3.5 Histogram nakon logoritamske transformacije atributa

Veoma važan deo faze pripreme i razumevanja podataka, pored transformacija atributa, jeste i istraživanje odnosa atributa klase i ostalih atributa. Tokom ovog istraživanja mogu se saznati potencijalne veze i šabloni koji mogu biti od velike koristi tokom faze modelovanja. Idealno bi bilo kada bi istražili vezu svakog od atributa sa atributom klase. Ovaj posao je vremenski veoma zahtevan, pošto obuhavajući skupovi uglavnom imaju veliki broj atributa. Iz tog razloga potrebno je istraživanje ograničiti na neki podskup atributa za koji bi nekako trebalo da inicijalno predvidimo da će imati značajne veze sa klasnim atributom.

Dobar način da se postigne prethodni cilj je upotreba korelacije tj. mere koeficijent korelacije. Da bi uvideli koeficijente korelacije pogledaćemo toplotnu mapu (eng. *Heat map*) na slici 3.6. Ovde možemo da vidimo da najveći koeficijent imaju *Trajanje računa* i *Iznos kredita*.



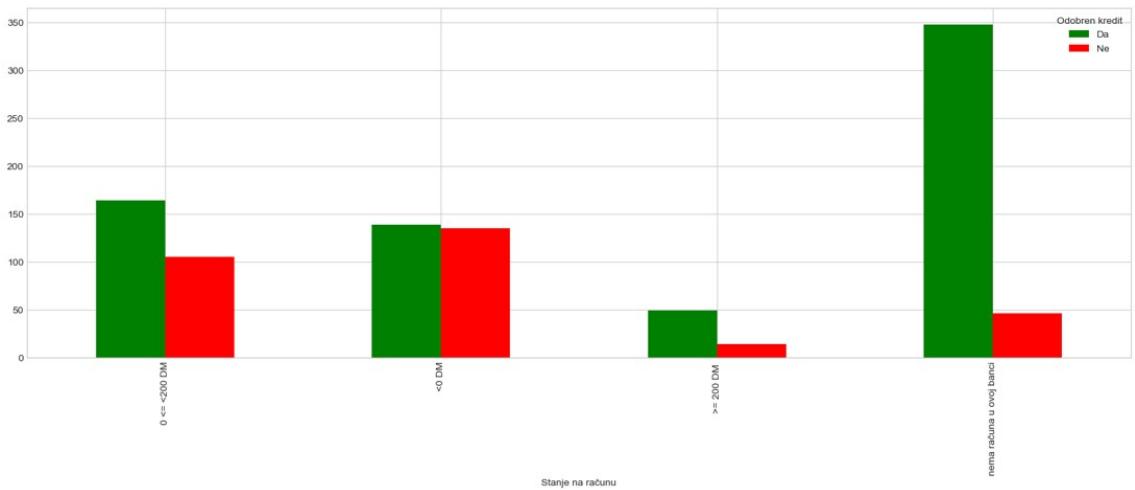
Slika 3.6 Tabela atributa sa absolutnom vrednošću većom od 0.1.

3.6. Istraživanje nominalnih atributa

Krajnji zadatak ovog rada jeste da se odredi da li je budući aplikant za kredit odgovarajući ili treba da bude odbijen. Iz tog razloga, u fazi eksplorativne analize podataka, vrši se posmatranje veza ostalih atributa sa

ciljnim atributom koji je indikator da li je nekome odobren kredit ili nije. Za ovaj proces bira se *Histogram* dijagram.

Za ilustraciju primera bira se atribut *stanje na računu* – da li je podrazumevano da klijent ima novac na računu, kako bi se predstavila veza između tog atributa i ciljnog atributa. Sa slike 3.7 se vidi da je najveći procenat onih klijenata kojima je odobren kredit, koji nemaju tekući račun u ovoj banci. Da bi se ova pretpostavka potvrdila kreira se tabela sa uporednim prikazom ciljanog atributa i *stanje na računu*. Tabela je prikazana na slici 3.8.



Slika 3.7 Histogram veze ciljanog atributa i atributa stanje na računu

	Odobren kredit	Da	Ne
Stanje na računu			
0 <= <200 DM	164	105	
<0 DM	139	135	
>= 200 DM	49	14	
nema računa u ovoj banci	348	46	

Slika 3.8 Uporedni prikaz ciljanog atributa i atributa stanje na računu

Na osnovu podataka prikazanih na slici 3.8, možemo doći do nama interesantnih vrednosti. Prva je procenat onih ljudi kojima je odobren kredit i nemaju tekući račun u ovoj banci. Ta vrednost se računa po

formuli $348/(348+46)$ i iznosi 88,32%. Druga je procenat onih kojima je odobren kredit a na tekućem računu imaju više od 200 DM i njih dobijamo putem formule $49/(49+14)$ i iznosi 77,78%. Treća vrednost je vrednost klijenata kojima je odobren kredit a na tekućem računu imaju manje od 200 DM i do njih dolazimo preko formule $164/(164+105)$ i iznosi 60,97% i četvrta grupa klijenata kojima je odobren kredit a nemaju sredstva na tekućem računu iznosi $139/(139+135)$ i iznosi 50,71%.

Na sledećim slikama prikazane su tabele uporednog prikaza ciljanog atributa sa atributima *istorija kredita, svrha uzimanja, račun za štednju, zaposlenost, kamata – u procentima od tekućih prihoda, pol i bračni status, žirant i ko-aplikant, prebivalište, imovina, plaćanje u ratama, stan/kuća, postojeći kredit, vrsta posla i izdržavani broj lica* respektivno. Analogno principu računanja vrednosti u prethodnom primeru vrši se računanje tih vrednosti i za ostale atribute, sa izuzetkom odnosa ciljanog atributa i atributa *telfon, radnik stranac*.

Odobren kredit	Da	Ne
Istorija kredita		
drugi krediti plaćani na vreme	361	169
kašnjenje sa uplatama	60	28
kritičan račun	243	50
nema uzetih kredita	15	25
svi krediti u ovoj banci vraćeni na vreme	21	28

Slika 3.9 Uporedni prikaz atributa *istorija kredita* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.9, vidi se da je procenat onih kojima je odobren kredit a istorija kredita im je sledeća:

- drugi krediti plaćani na vreme je 68,11%,
- kašnjenje sa uplatama je 68,18%,
- kritičan račun je 82,94%,
- nema uzetih kredita je 37,50%,
- svi krediti vraćeni na vreme je 42,86%

Dakle, izvodi se zaključak da ne postoji neko vidljivo pravilo koje se ističe naspram koga je nekoj grupi ljudi češće odobren kredit od ostalih grupa.

Odobren kredit	Da	Ne
Svrha uzimanja		
automobil (nov)	145	89
automobil (polovan)	86	17
bela tehnika	8	4
biznis	63	34
muzička tehnika/televizija	218	62
nameštaj	123	58
obrazovanje	28	22
ostalo	7	5
popravka	14	8
prekvalifikovanje	8	1

Slika 3.10 Uporedni prikaz atributa *svrha uzimanja* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.10, vidi se da je procenat onih kojima je odobren kredit a koji su uzeli kredit za:

- automobil (nov) je 61,97%,
- automobil (polovan) je 83,50%,
- belu tehniku je 66,67%,
- biznis je 64,95%,
- muzička tehnika/televizija je 77,86%,
- nameštaj je 67,96%,
- obrazovanje je 56,00%,
- ostalo je 58,33%,
- popravku je 63,64%,
- prekvalifikovanje je 88,89%.

Ovi podaci vode do zaključka da je kredit najviše odobravan za grupu klijenata koji su naveli kao svrhu prekvalifikaciju, a najviše ljudi se prijavilo za kredit za svrhu kupovine muzičke tehnike ili televizije.

Odobren kredit	Da	Ne
Račun za štednju		
100 <= <500 DM	69	34
500 <= < 1000 DM	52	11
<100 DM	386	217
>= 1000 DM	42	6
nema račun za štednju	151	32

Slika 3.11 Uporedni prikaz atributa *račun za štednju* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.11, vidi se da je procenat onih klijenata kojima je odobren kredit a koji imaju račun za štednju sa:

- $100 <= <500$ je 66,99%,
- $500 <= <1000$ je 82,54%,
- < 1000 je 64,01%,
- $>= 1000$ je 87,50%,
- nema račun za štednju 82,51%.

Izvodi se zaključak da se grupi klijenata koji na računu za štednju imaju više od 1000 DM najčešće odobrava kredit, a slede ih grupe koje nemaju račun za štednju ili na istom imaju jednalo ili više od 500 DM a manje od 1000 DM.

Odobren kredit	Da	Ne
Zaposlenost		
$1 <= < 4$ godina	235	104
$4 <= < 7$ godina	135	39
$<= 1$ godina	102	70
$>= 7$ godina	189	64
nezaposlen	39	23

Slika 3.12 Uporedni prikaz atributa *zaposlenost* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.12, vidi se da je procenat onih kojima je odobren kredit a zaposleni su na:

- $1 <= < 4$ godine je 69,32%,
- $4 <= < 7$ godine je 77,59%,
- < 1 godine je 59,30%,
- $>= 7$ godina je 74,70%,
- nezaposlen je 62,90%.

Iz ovoga dolazimo do zaključka da nema značajnih odstupanja prilikom odobravanja kredita gore navedenim grupama sa time da je najčešće odobren kredit klijentima koji rade četiri ili više godina, a manje ili tačno 7.

Odobren kredit	Da	Ne
Kamata – u procentima od tekućih prihoda		
1	102	34
2	169	62
3	112	45
4	317	159

Slika 3.13 Uporedni prikaz atributa *kamata* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slikama 3.13, vidi se da je procenat onih kojima je odobren kredit sa kamatama:

- 1 je 75,00%,
- 2 je 73,16%,
- 3 je 71,34%,
- 4 je 66,60%.

Odobren kredit	Da	Ne
Pol i bračni status		
muško, neoženjen	402	146
muško, oženjen/udovac	67	25
muško, razveden	30	20
žensko, udata/razvedena	201	109

Slika 3.14 Uporedni prikaz atributa *pol i bračni status* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.14, vidi se da je procenat onih kojima je kredit odobren a koji su:

- neoženjeni muškarci je 73,36%,
- oženjen ili udovac muškarac je 72,83%,
- razveden muškarac je 60,00%,
- nisu stambeni lični kredit je 64,84%.

Izvodi se zaključak da je kredit najčešće odobren neoženjenim muškarcima.

Odobren kredit	Da	Ne
Žirant Ko-aplikant		
ko-aplikant	23	18
nema	635	272
žirant	42	10

Slika 3.15 Uporedni prikaz atributa *žirant ko-aplikant* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.15, vidi se da je procenat onih kojima je kredit odobren a koji imaju:

- ko-aplikanta 56,10%,
- nemaju žiranta je 70,01%,
- imaju žiranta je 80,77%.

Dolazimo do zaključka da se najčešće odobrava kredit klijentima koji imaju žiranta.

Odobren kredit	Da	Ne
Žirant Ko-aplikant		
ko-aplikant	23	18
nema	635	272
žirant	42	10

Slika 3.16 Uporedni prikaz atributa *žirant ko-aplikant* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.15, vidi se da je procenat onih kojima je kredit odobren a koji imaju:

- ko-aplikanta 56,10%,
- nemaju žiranta je 70,01%,
- imaju žiranta je 80,77%.

Dolazimo do zaključka da se najčešće odobrava kredit klijentima koji imaju žiranta.

Odobren kredit	Da	Ne
Vrsta posla		
nema zvanično obrazovanje: ima stalan posao	144	56
nezposlen/nema zvanično obrazovanje	15	7
radi u menadžmentu/samo-zaposlen/visoko obrazovanje	97	51
zaposlen i ima zvanično obrazovanje	444	186

Slika 3.17 Uporedni prikaz atributa *vrsta posla* i *odobren kredit*

Na osnovu podataka iz tabele, prikazane na slici 3.15, vidi se da je procenat onih kojima je kredit odobren a koji:

- nemaju zvanično obrazovanje: ima stalan posao je 72,00%,
- nezaposlen/nema zvanično obrazovanje je 68,18%,
- radi u menadžmentu / samo-zaposlen / visoko obrazovanje je 65,54%,
- zaposlen i ima zvanično obrazovanje 70,48%.

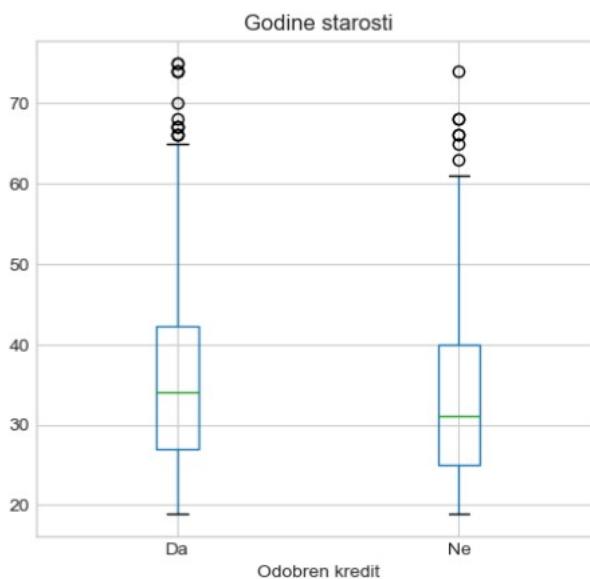
Dolazimo do zaključka da klijenti koji su zaposleni sa zvaničnim obrazovanjem najčešće apliciraju za kredit i čak 70% njih dobije pozitivan odgovor od banke.

3.7. Istraživanje numeričkih atributa

U ovoj fazi eksplorativne analize vrši se posmatranje veza između numeričkih atributa i ciljnog atributa *odobren kredit*. Za ovaj korak se bira *bar plots/Boxplots* vrsta dijagrama, koja omogućava da se tačno vizuelno odredi da li su numerički atribut i ciljni atribut međusobno u vezi i da li je numerički atribut od značaja za dalju analizu. Za iscrtavanje ovog dijagrama koristi se funkcija *boxplot* iz *Matplotlib* biblioteke.

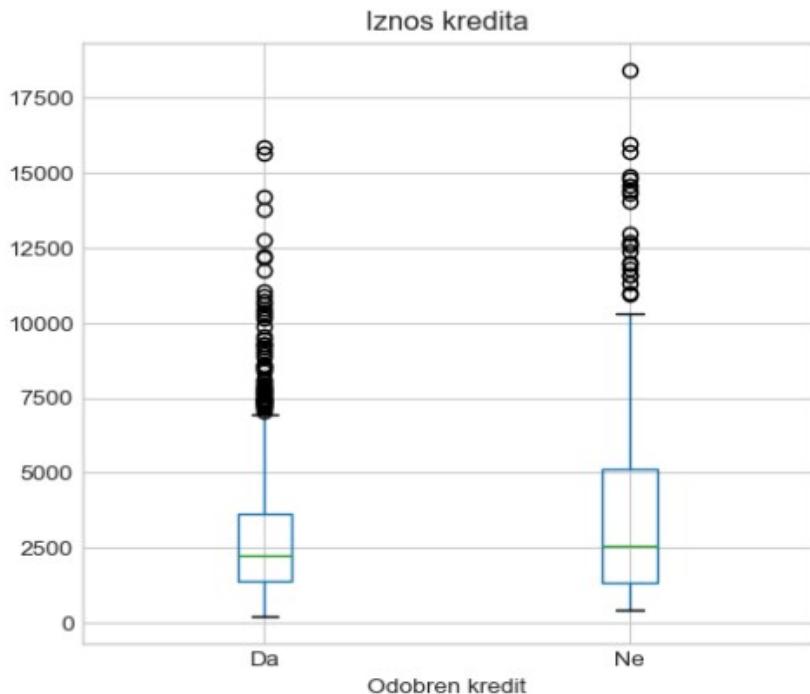
Ako raspored podataka na grafikonu izgleda slično, kutije su u istoj liniji, to znači da kontinuirana numerička promenljiva nema efekta na ciljni atribut, dakle varijable nisu u korelaciji jedna sa drugom.

Na primer ako se pogleda grafikon na slici 3.18 gde vidimo odnos između *godine starosti* i *odobren kredit* kutije su u sličnoj liniji. To znači da klijenti kojima je kredit odbijen i klijenti kojima je kredit odobren imaju isti broj godina starosti. Dakle, ne može da se razlikuje odobrenje i odbijanje na osnovu starosti podnosioca zahteva. Dakle iz vizuelnog pregleda možemo doneti zaključak da ove dve varijable nisu u korelaciji.



Slika 3.18 Boxplot dijagram vrednosti atributa *godine starosti* i *odobren kredit*

Sa druge strane ako pogledamo na dijagram na slici 3.19 možemo videti da za odnos *iznos kredita i odobren kredit* da su prikazane suprotne karakteristike, tako da pomoću vizuelne inspekcije, možemo reći da su *iznos kredita i ciljana varijabla* međusobno u vezi, i isto važi za odnos između *trajanje računa i ciljnoe variabile*.



Slika 3.19 Boxplot dijagram vrednosti atributa *iznos kredita i odobren kredit*

Da bi se potvrdila pretpostavka donešena vizuelnim putem koristi se ANOVA test pomoću funkcije *f_oneway* iz biblioteke *scipy.stats*. Kada se primeni ANOVA test, ako je dobijeni rezultat veći od 0,05 znači da varijablu nećemo uzeti u obzir. Na slici 3.20 može se videti rezultat ANOVA testa za gore navedene varijable.

```
Trajanje računa is correlated with Odobren kredit | P-Value: 6.488049877187189e-12
Iznos kredita is correlated with Odobren kredit | P-Value: 8.797572373533373e-07
Godine starosti is correlated with Odobren kredit | P-Value: 0.003925339398278295
```

Slika 3.20 Rezultat ANOVA testa

Uzimajući u obzir rezultat ANOVA testa dolazimo do sledećeg zaključka koji je prikazan u tabeli 3.3.

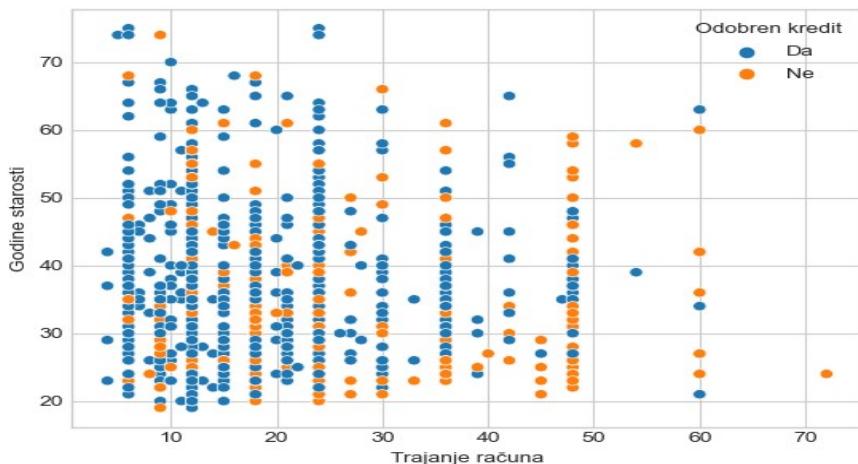
Atribut	Značajna veza sa ciljanim atributom(da/ne)
<i>godine starosti</i>	DA
<i>iznos kredita</i>	DA
<i>trajanje računa</i>	DA

Tabela 3.3 Veze numeričkih atributa sa cilnjim atributom

3.8. Analiza veza između više atributa i ciljnog atributa

Iako možda ne postoji direktna veza između jednog atributa i ciljnog atributa moguće je da taj atribut u kombinaciji sa drugim atributima pruža očigledne veze sa atributom. To znači da je potrebno koristiti tehnike prikaza podataka koje mogu da nam prikažu odnose više od jednog atributa sa cilnjim atributom.

Za ilustraciju primera biraju se atributi *godine starosti* – broj godina starosti klijenta koji je aplicirao za kredit i *trajanje računa* – trajanje računa u mesecima, kako bi se predstavila veza između ovih atributa i ciljnog atributa *odobren kredit*.

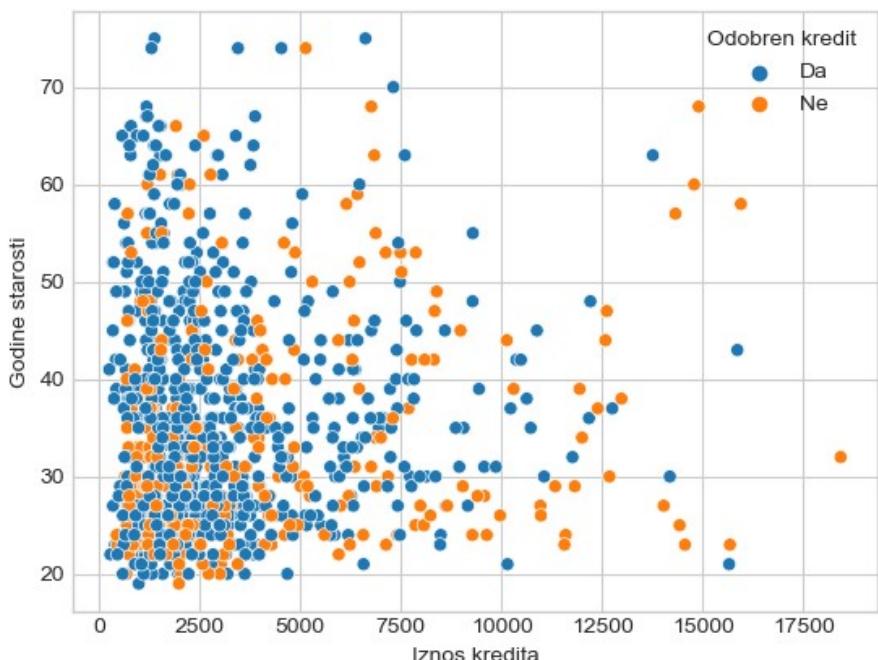


Slika 3.21 Analiza veza između atributa *godine starosti* i *trajanje računa* i *odobren kredit*

Ako se posmatra gornji levi ugao grafikona, prikazanog na slici 3.21, vidimo da su tu klijenti koji su stariji od 60 godina i koji imaju račun u banci manje od 30 meseci. Sa slike se jasno vidi da je većini ovih klijenata zahtev za kredit odobren.

Isto tako možemo da vidimo da je većini klijenata, nebitno koliko godina starosti imaju, a koji imaju račun u banci manje od 30 godina, zahtev za kredit odobren. Ali kako se odmiče broj meseci broj zahteva za kredit biva odbijen.

U nastavku se biraju atributi *godine starosti* – broj godina klijenta i *iznos kredita* – iznos koji klijent potražuje, kako bi se predstavila veza između ovih numeričkih atributa i ciljnog atributa *odobren kredit*.



Slika 3.22 Analiza veza između atributa *age* i *duration* i *Y*

Sa grafika, prikazanog na slici 3.16, vidi se da se klijentima koji potražuju kredit na manji novčani iznos, i koji su stariji od 25 godina, mnogo češće biva odobren kredit nego klijentima koji su mlađi od 30 godina koji potražuju isti novčani iznos.

Dakle, dolazi se do zaključka da se klijentima koji su stariji i koji imaju račun u banci manje od 30 meseci, a koji potražuju kredit na manji iznos novca u većini slučaja bude odobren kredit. Što dovodi do logičnog zaključka da klijentima koji su stariji i potražuju kredit na manje iznose bude u više slučaja odobren nego klijentima koji pripadaju mlađoj starosnoj grupi i koji potražuju kredite na veće iznose od 7500DM.

4. FAZA MODELOVANJA I EVAULACIJE

U prethodne dve faze (faza shvatanja konkretnog problema i faza razumevanja i pripreme podataka) izvršeno je proučavanje skupa podataka i odlučivanje koji atributi treba da budu zadržani a koje treba da eliminišemo. Osim toga, takođe je izvršena transformacija atributa kod kojih je postojala potreba za tim.

Sada dolaze na red dve ključne faze:

1. modelovanje tj. formiranje modela koji će automatski predvideti kome uputiti poziv
2. evaluacija modela koja će nam reći da li uopšte ima smisla primenjivati naše modele iz prethodne faze

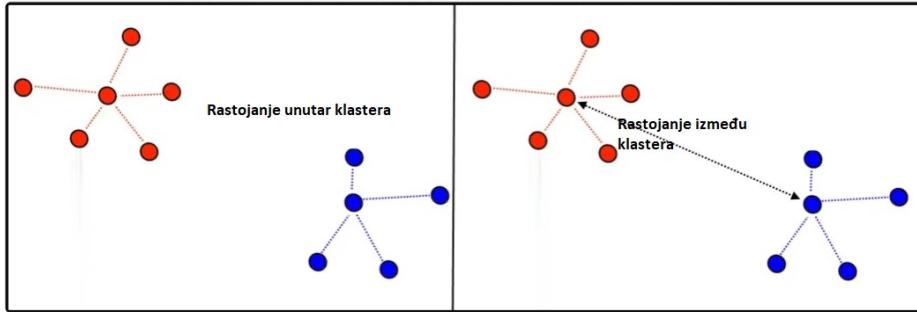
4.1 Klasterovanje

4.1.1 Teorijske osnove

U prvom koraku procesa modelovanja, započet ćemo sa klasterovanjem skupa podataka sa svrhom otkrivanja obrazaca i utvrđivanja upotrebljivosti pridruživanja klasteru kao osobine u konstrukciji modela.

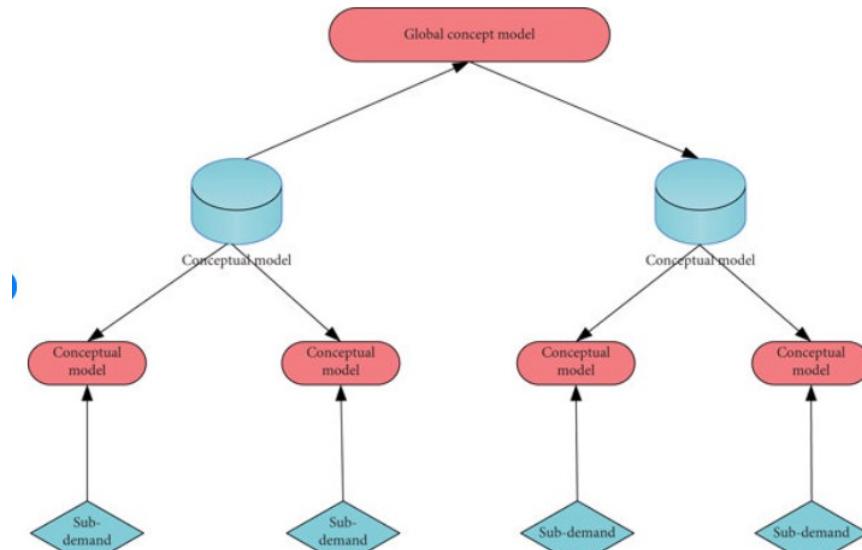
Klasterovanje podataka može se smatrati jednim od ključnih izazova nenadgledanog učenja (engl. *unsupervised learning*). Suština klasterovanja leži u otkrivanju obrazaca ili struktura u neobeleženim podacima. Konceptualno, klasterovanje podrazumeva organizaciju elemenata u kolekcije sa sličnostima među članovima iste kolekcije, ali i razlikama u poređenju sa članovima drugih kolekcija. Ovako formirane grupe se nazivaju klasterima, a predstavljaju skupove objekata koji dele određene sličnosti među sobom, dok se istovremeno razlikuju od objekata u drugim klasterima.

Postoje dve vrste klasterovanja. Prvi je klasterovanje zasnovano na rastojanju (engl. *distance-based clustering*) , gde se dva ili više objekata svrstavaju u isti klaster ukoliko su međusobno "bliži" prema unapred definisanoj meri udaljenosti. Klasterovanje zasnovano na udaljenosti je prikazano na slici 4.1.



Slika 4.1 Klasterovanje zasnovano na udaljenosti

Drugi tip klasterovanja je konceptualno klasterovanje, što je ilustrovano na slici 4.2. Ova forma klasterovanja se bazira na ideji da će dva ili više objekata biti smatrana delom istog klastera ukoliko dele isti koncept ili princip. Pored samih klastera, ovaj pristup klasterisanju takođe generiše opise za svaki klaster.

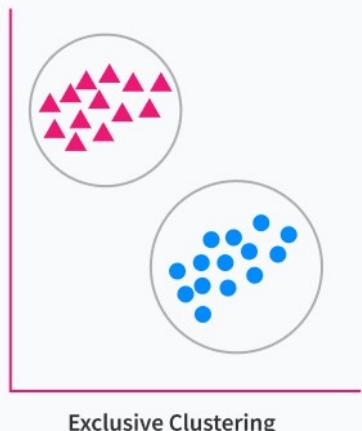


Slika 4.2 Konceptualno klasterovanje

Algoritmi za klasterovanje mogu se klasifikovati u tri grupe.

Ekskluzivno (partitivno) klasterovanje (engl. *Exclusive Clustering*)

Ovo algoritam za klasterovanje obezbeđuje da svaki element može da bude u najviše jednom klasteru. Na slici 4.3 je prikazan primer gde svaki objekat je član samo jednog klastera.

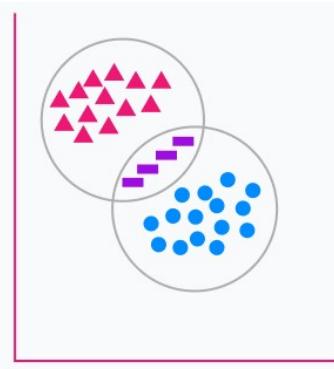


Exclusive Clustering

Slika 4.3 Ekskluzivno (partitivno) klasterovanje

Klasterovanje sa preklapanjem (engl. *Overlapping Clustering*)

Ova vrsta klasterovanja, koja je prikazana na slici 4.4, omogućava da se elementi strukturiraju unutar dva ili više klastera. Dok klasterovanje sa delimičnim preklapanjem dozvoljava da se, na primer, jedna osoba svrsta i kao zaposlena i kao student, striktno grupisanje bi zahtevalo od te osobe da izabere jednu ulogu.

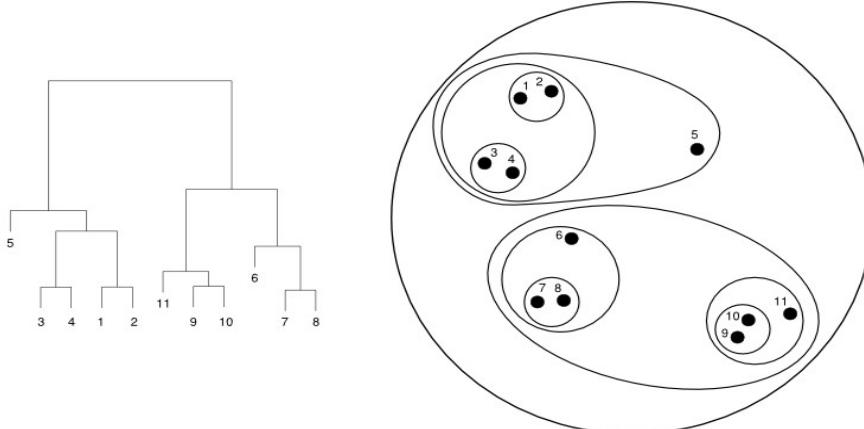


Overlapping Clustering

Slika 4.4 Klasterovanje sa preklapanjem

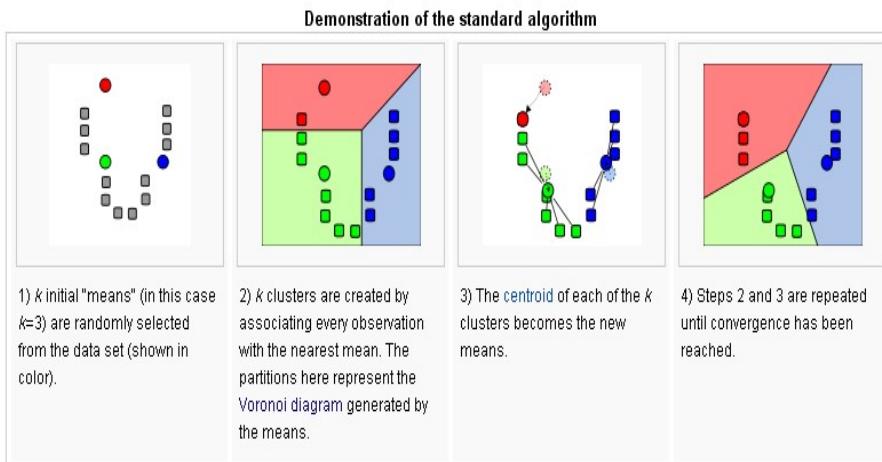
Hijerarhijsko klasterovanje (engl. *Hierarchical Clustering*)

Ovo klasterovanje je bazirano na osobini ugnježdavanja klastera. Dakle, postoje klasteri koji se nalaze u okviru većeg klastera kao što je i prikazano na slici 4.5.



Slika 4.5 Hijerarhijsko klasterovanje

U ovom radu korišćeno je ekskluzivno (partitivno) klasterovanje i to K-sredina (engl. *K-means*) algoritam, prikazan na slici 4.6.



Slika 4.6 K-sredina algoritam

K-sredina algoritam predstavlja jedan od najjednostavnijih algoritama u domenu nенадгледаног учења, чија сврха је решавање изазова

klasterovanja. Ova procedura prati jednostavan i nezahtevan pristup za razvrstavanje datog skupa podataka u unapred određeni broj klastera (tačnije, K klastera).

Osnovna zamisao je da se utvrdi pozicija k centroida, po jedan za svaki klaster. Potrebno je temeljno promisliti o mestima gde bi ove tačke trebalo da se smeste, jer različiti izbori pozicija pružaju raznolike rezultate. Preporučuje se da se ove tačke razmestite što je moguće udaljenije jedna od druge, kako bi se postigli optimalni ishodi. Nakon ovog koraka, svaki element iz skupa podataka se veže za najbližu tačku koja predstavlja centar. Kada se sve veze izvrše, završava se prva faza grupisanja, pružajući grubu sliku o grupama. Dalje, neophodno je da se ponovno izračuna pozicija tačaka koje predstavljaju centre svake grupe. Ovo se postiže pronalaženjem središta svih tačaka unutar određene grupe. Na ovaj način se dobijaju novi centri oko kojih se formiraju nove grupe. Ceo proces se ponavlja dok god se centri ne prestanu pomerati. Merilo koje se koristi za procenu uspešnosti K-srednjeg grupisanja je zbir kvadrata grešaka (engl. *Sum of Squared Error - SSE*):

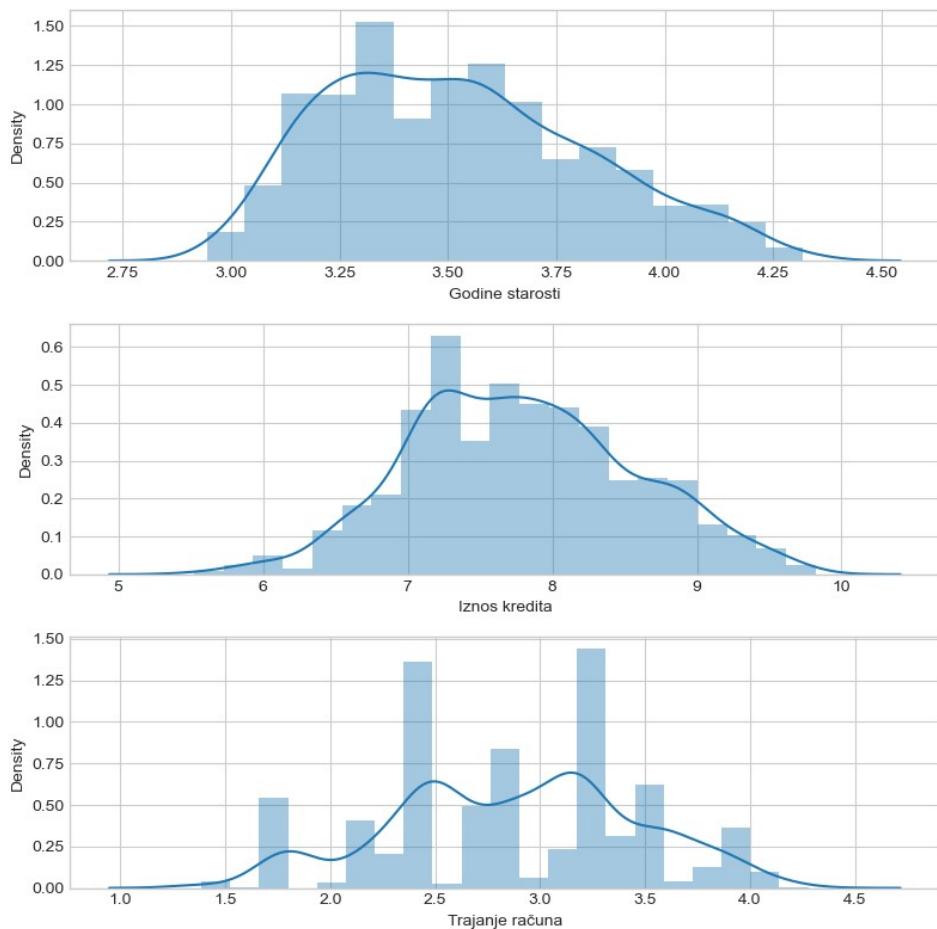
$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x),$$

gde je x podatak (tačka) iz klastera C_i a m_i je reprezentativna tačka za klaster C_i . Na osnovu ovog podatka, od dva data klastera možemo odabratи onaj sa manjom greškom. Jednostavan način za smanjenje SSE je povećanje broja klastera K , međutim treba imati na umu da dobro klasterovanje sa manjim K može da ima nižu SSE nego loše klasterovanje sa većim K .

4.1.2 Primena klasterovanja

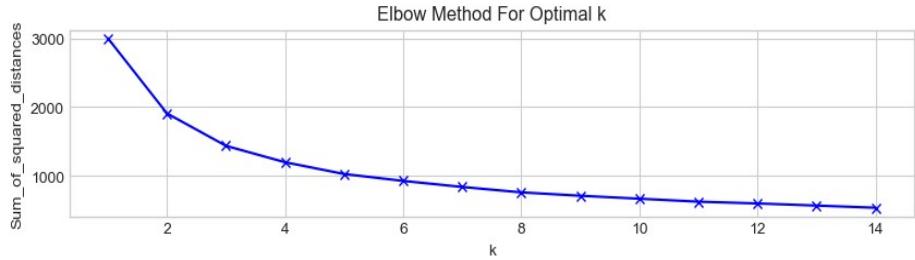
Da bi se izvrsilo klasterovanje, za početak se bira skup odgovarajućih atributa koji bi dali najbolje rezultate. Za najbolje rezultate za dati skup podataka biramo atributе *godine starosi*, *trajanje kredita* i *iznos kredita*. Zatim se treba izvršiti kontrola ovih atributa da bi se utvrdilo da li su potrebni dodatni koraci za njihovu obradu da bi se dobili bolji rezultati klasterovanja.

Kao što se može primetiti iz poglavlja 3.5 transformacija atributa, nad novim skupom podataka primenom logaritamske transformacije koju smo izveli koristeći funkciju *log* iz biblioteke *numpy* na slici 4.7 vidimo da su podaci sada mnogo ravnomernije distribuirani u odnosu na srednju vrednost i samim tim omogućavaju bolje rezultate klasterovanja.



Slika 4.7 Vrednosti atributa posle logaritamske transformacije

Za dobijanje klastera koristi se algoritam K-sredina. Kako je ranije navedeno da bi algoritam bio uspešan treba da se odredi optimalni broj centrioda. Kako bismo došli do optimalnog broja centrioda koristimo *Elbow Method*, ali pre toga treba da odradimo skaliranje i centriranje varijabli, to je preduslov za pravilno korišćenje algoritma K-sredina. Za tu svrhu se koristi *StandardScaler* iz *sklearn* biblioteke. Potom pravimo grafikon gore pomenute metode (slika 4.8).

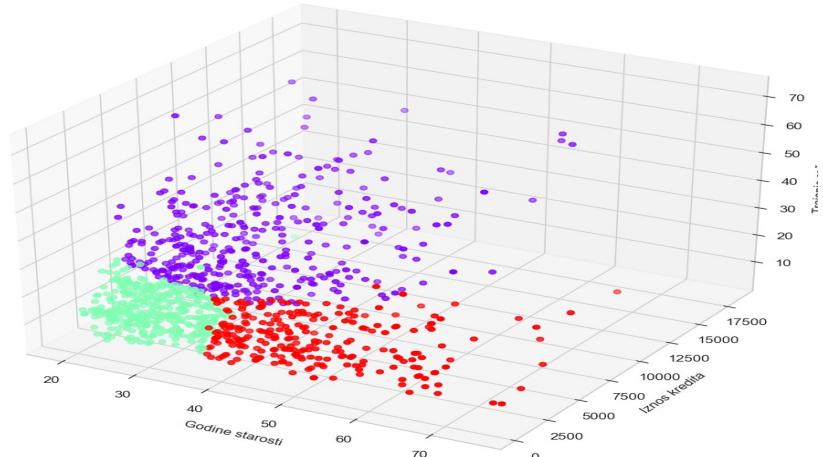


Slika 4.8 Grafik Elbow metode

Kako može da se vidi na slici za optimalni broj centrioda se bira broj 3. Kako je odabran optimalni broj centrioda pomoću funkcije *Kmeans* iz biblioteke *sklearn.cluster*. Pozivom ove funkcije dobijaju se klasteri.

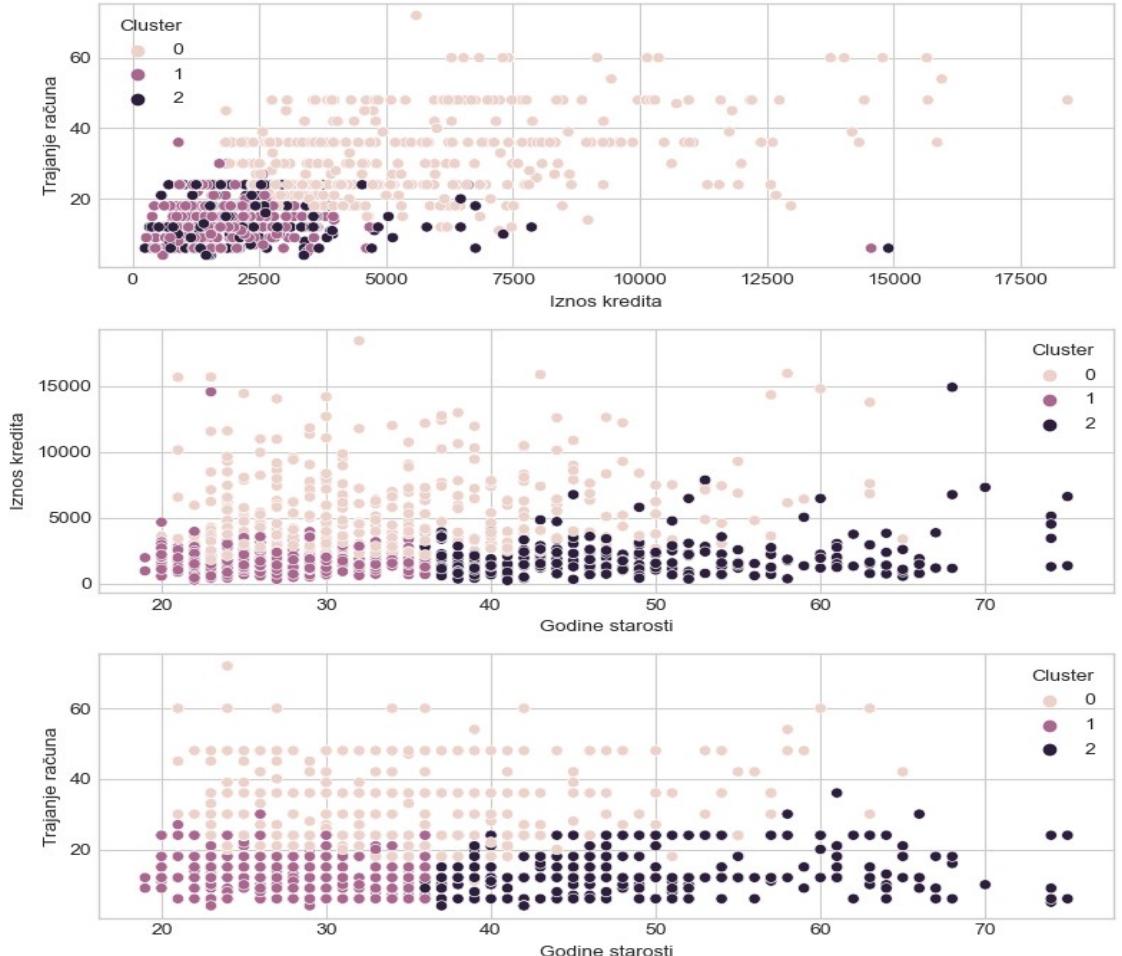
4.1.3 Proučavanje rezultata klasterovanja

U procesu klasterovanja primjenjen je algoritam K-sredina sa 3 centroida. Primećeno je da je ovo najoptimalniji broj za K, jer bilo koji broj manji ili veći od 3 formira izuzetno nepravilne klastere. Na slici 4.9 se može videti raspodela dobijenih klastera u 3d prostoru.



Slika 4.9 Klasteri dobijeni algoritmom K-sredina sa 3 centroida

Na slici 4.10 se može videti kako u dva slučaja postoji očigledno razdvajanje između klastera dok u jedanom slučaju i ne toliko očigledno.



Slika 4.10 razdvajanje između klastera

Klasteri dobijeni primenom algoritma K-sredina su prikazani na slici 4.11.

Cluster	Iznos kredita	Godine starosti	Trajanje računa
0	1970.5	48.6	13.9
1	5698.9	34.2	32.3
2	1746.4	27.7	14.4

Slika 4.11 Klasteri dobijeni algoritmom K-sredina

Klaster 0

Na slici 4.11 se vidi da klasteru 0 pripadaju klijenti koji potražuju niži srednji novčani iznos, za otplatu kredita se opredeljuju da vremenski period bude kraći, i ovoj grupi pripadaju klijenti koji su po godinama stariji građani.

Klaster 1

Na slici 4.11 se vidi da klasteru 1 pripadaju klijenti koji potražuju viši srednji novčani iznos, za otplatu kredita se opredeljuju da vremenski period bude duži, što je zbog velikog iznosa i logično, i ovoj grupi pripadaju klijenti koji su po godinama građani srednjeg doba.

Klaster 2

Na slici 4.11 za klaster 2 se vidi da njemu pripadaju klijenti koji potražuju niži novčani iznos, za otplatu kredita se opredeljuju na kraći vremenski period i možemo reći da ovoj grupi klijenata pripadaju osobe koje su mlađe od prethodne dve grupe.

4.2 Analiza asocijativnih pravila

4.2.1 Teorijske osnove

Asocijativni obrasci predstavljaju proces identifikacije elemenata koji se zajedno javljaju u određenim situacijama. Drugim rečima, asocijativni obrasci predstavljaju tehniku za pronalaženje veza u pojavljivanju pojedinačnih elemenata. Oni pružaju informacije o tome koliko često se ti događaji pojavljuju zajedno. Ovaj metod se često primenjuje u analizi podataka koji prate razne transakcije, kao što su prodaja, nabavke i slično.

<i>TID</i>	<i>Elementi</i>
1	Hleb, Mleko
2	Hleb, Pelena, Pivo, Jaje
3	Mleko, Pelena, Pivo, Sok
4	Hleb, Mleko, Pelena, Pivo
5	Hleb, Mleko, Pelena, Sok

Tabela 4.1 Transakcije Market-Basket

Prilikom ilustracije načina funkcionisanja asocijativnih pravila često se koristi primer analize transakcija u prodavnica. Analiza potrošačke korpe, tabela 1., nizom transakcija – pojedinačni računi prodavnice, otkrivaju skrivena pravila koja se tiču prodaje proizvoda. Cilj metode je otkriti ta pravila. Dakle, uočiti zakonitost ako kupac kupuje proizvod X tada će购置 i proizvod Y , uz određenu verovatnost temeljenu na istorijskim podacima – posmatranje pojavljivanja proizvoda kroz duži vremenski period. Pomenuta zakonitost se može prikazati u sledećem obliku:

$$X \rightarrow Y.$$

X i Y su skupovi artikala (*itemset*) koji su neprazni disjunktni tj. nemaju zajedničkih elemenata. X se naziva prepostavka (*antecedent*), Y posledica ili zaključak (*consequent*) a implikacija znači istovremeno događanje ali ne i uzročnost. Uzročnost zahteva istorijske podatke o odnosu posmatranih atributa.

Bitni pojmovi asocijativne analize su dati u sledećoj tabeli.

POJAM	FORMULA	OPIS
Skup artikala (<i>Itemset</i>)	X, Y	Skup jednog ili više artikala. {Mleko, Hleb, Pelena}
Frekventni skup artikala (Frequent Itemset)	$s \geq \text{minsup}$	Kolekcija čija potpora veća ili jednaka pragu <i>minsup</i>
Potporni broj (σ)	$\sigma(\{X\})$	Frekvencija pojavljivanja skupa artikala u transakciji. $\sigma(\{\text{Mleko, Pelena, Pivo}\}) = 2$

Potpore (s)	$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$, gde je $\sigma(X \cup Y)$ ukupan broj transakcija koje sadrže X i Y , a N ukupan broj transakcija.	Deo transakcija koje sadrže skup artikala Za implikaciju oblika {Mleko, Pelena} → {Pivo} poverenje bi se računalo prema sledećoj formuli: $s = \frac{\sigma(\text{Mleko, Pelena, Pivo})}{ \text{T} } = \frac{2}{5} = 0.4$
Poverenje (c)	$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$, gde je $\sigma(X \cup Y)$ ukupan broj transakcija koje sadrže X i Y , a $\sigma(X)$ ukupan broj transakcija koje sadrže X .	Meri učestalost pojavljivanja elemenata Y u transakcijama koje sadrže X . Za implikaciju oblika {Mleko, Pelena} → {Pivo} poverenje bi se računalo prema sledećoj formuli: $c = \frac{\sigma(\text{Mleko, Pelena, Pivo})}{\sigma(\text{Mleko, Pelena})} = \frac{2}{3} = 0.67$

Tabela 4.2 Bitni pojmovi asocijativne analize

Pravilo $X \rightarrow Y$ u transakcionom skupu T sadrži **podršku s** (eng. *Support*), gde je s procenat transakcija u transakcionom skupu T , koje sadrže $X \cup Y$ ili i X i Y . Dalje, ovo postaje verovatnoća $P(X \cup Y)$.

Poverenje c (eng. *Confidence*) pravila $X \rightarrow Y$ u skupu transakcija T je procenat transakcija koje ako sadrže A , sadrže i B . Ovo predstavlja uslovnu verovatnocu $P(X|Y)$.

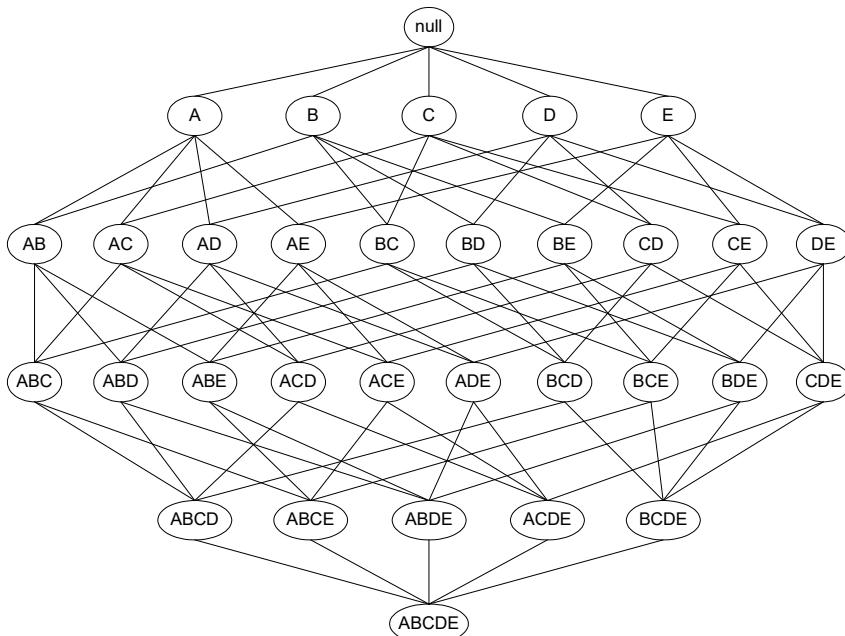
Pronađena pravila smatraju se interesantnim ukoliko zadovoljavaju minimalnu zadatu vrednost podrške i minimalnu zadatu vrednost poverenja. Drugim rečima, potpora mora biti veća ili jednaka *minsup* praga i poverenje mora biti veće ili jednako *minconf* praga. Dakle, da bi izveli asocijativna pravila potrebno je pre svega formirati listu svih asocijativnih pravila, zatim izračunati potporu i poverenje za svako pravilo i konačno odstraniti pravila koja ne zadovoljavaju pragove *minsup* i *minconf*. Ovakav pristup generisanja pravila, gde se proveravaju pomenute mere za sva moguća pravila, je praktično neizvodljiv. Zbog toga se izvođenje asocijativnih pravila deli na dva dela:

1. Generisanje frekventnog skupa artikala, prikazano na slici 4.12 – generisati sve frekventne skupove artikala za koje je potpora veća ili jednaka od $minsup$. Generisanje frekventnog skupa artikala je još uvek računski skupo. Za d elemenata postoji 2^d mogućih kandidatskih skupova a ukupan broj mogućih asocijativnih pravila se izračunava prema formuli:

$$\begin{aligned} R &= \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right] \\ &= 3^d - 2^{d+1} + 1 \end{aligned}$$

Što znači da, ukoliko je $d=6$ tada je ukupan broj frekventnih skupova $2^6=64$ a ukupan broj asocijativnih pravila 602.4

2. Generisanje pravila – generisati pravila visokog poverenja iz svakog frekventnog skupa artikala, gde je svako pravilo binarno particionisanje frekventnog skupa artikala.



Slika 4.12 Generisanje frekventnog skupa

4.2.2 Primena analize asocijativnih pravila

Za postupak izvođenja asocijativnih koristi se *Apriori* algoritam iz *mlxtend* biblioteke. *Apriori* algoritam radi sa kategoričkim podacima. Potrebno je da podaci budu odgovarajuće obrađeni. Numeričke atribute je potrebno pretvoriti u kategoričke binove koristeći funkciju *cut* iz *Pandas* biblioteke. Ova predobrada je važna kako bi se definisali značajni elementi za rudarenje asocijativnih pravila.

Zatim je potrebno kodirati sve kategoričke atribute u binarne vrednosti (0 ili 1). Za tu svrhu se koristi funkcija *get_dummies* iz *Pandas* biblioteke *value*. Kada je skup podataka obrađen i pripremljen, nad njim se upotrebljava *Apriori* algoritam tako što mu se pored skupa podataka prosleđuje i prag minimalne podrške. Prag minimalne podrške se prilagođava putem *min_support* parametra.

Nakon primene gore pomenutih funkcija i posle dobijenih čestih skupova elemenata, za generisanje asocijativnih pravila koristi se funkcija *association_rules* tako što se proslede česti skupovi elemenata i prosledi se prag minimalnog poverenja putem *min_threshold* parametra. Poziv ka ovoj funkciji generiše asocijativna pravila. Proces analize asocijativnih pravila je prikazan na slici 4.13.

U rezultatu *association_rules* su pored potpore i pouzdanosti dostupne još neke mere. Najznačajnije dve su *lift* i *conviction*.

Lift pravila tumači se kao odnos potpore dobijene na osnovu podataka i očekivane potpore ako X i Y ne zavise jedan od drugog:

1. $lift=1$ znači da su X i Y nezavisni tj. pravilo je beznačajno.
2. $lift>1$ znači pozitivnu vezu tj. ako znamo da je neko kupio X znamo da će verovatno kupiti i Y .
3. $lift<1$ znači negativnu vezu tj. ako znamo da je neko kupio X znamo da verovatno neće kupiti Y , što je takođe interesantno.

Conviction pravila tumači se kao odnos očekivane frekvencije da se X javlja bez Y (tj. da će pravilo biti pogrešno) i frekvencije da se X javlja bez Y dobijene na osnovu podataka. Recimo ako pravilo ima *conviction* 1.5, to pravilo bi bilo 1.5 puta (50%) češće pogrešno da su X i Y nezavisni, što znači da je pravilo interesantno.

4.2.3 Proučavanje rezultata analize asocijativnih pravila

Posmatraćemo samo pravila koja u zaključku imaju ciljni atribut *odobren kredit* koji nam govori da se klijent pretplatio na oročenu štednju. Na slici 4.13 prikazana su neka izvedena asocijativna pravila.

U tabelama 4.3, 4.4, 4.5 i 4.6 su opisana neka interesantna pravila.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
15	(Pol i bračni status _muško, neoženjen)	(Odobren kredit)	0.548	0.7	0.402	0.733577	1.047967	0.0184	1.126027	0.101264
17	(Žirant Ko-aplikant_nema)	(Odobren kredit)	0.907	0.7	0.635	0.700110	1.000158	0.0001	1.000368	0.001693
19	(Plaćanje u ratama_nema)	(Odobren kredit)	0.814	0.7	0.590	0.724816	1.035451	0.0202	1.090179	0.184071
21	(Stan Kuća_poseduje)	(Odobren kredit)	0.713	0.7	0.527	0.739130	1.055901	0.0279	1.150000	0.184464
22	(Vista posla_zaposlen i ima zvanično obrazovanje)	(Odobren kredit)	0.680	0.7	0.444	0.704762	1.006803	0.0030	1.016129	0.018262
24	(Iznos kredita po kategorijama_mali iznos)	(Odobren kredit)	0.914	0.7	0.658	0.719912	1.028446	0.0182	1.071094	0.321623
63	(Radnik stranac_Plaćanje u ratama_nema)	(Odobren kredit)	0.782	0.7	0.560	0.716113	1.023018	0.0126	1.056757	0.103211
67	(Stan Kuća_poseduje, Radnik stranac)	(Odobren kredit)	0.685	0.7	0.502	0.732847	1.046924	0.0225	1.122951	0.142288
73	(Iznos kredita po kategorijama_mali iznos, Radnik stranac)	(Odobren kredit)	0.880	0.7	0.626	0.711364	1.016234	0.0100	1.039370	0.133120
164	(Plaćanje u ratama_nema, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.742	0.7	0.538	0.725067	1.035811	0.0186	1.091176	0.134002
168	(Stan Kuća_poseduje, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.647	0.7	0.475	0.734158	1.048797	0.0221	1.128488	0.131803
170	(Žirant Ko-aplikant_nema, Vista posla_zaposlen i ima zvanično obrazovanje)	(Odobren kredit)	0.570	0.7	0.404	0.708772	1.012531	0.0050	1.030120	0.028782
173	(Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.830	0.7	0.597	0.719277	1.027539	0.0160	1.068670	0.157651
177	(Stan Kuća_poseduje, Plaćanje u ratama_nema)	(Odobren kredit)	0.576	0.7	0.443	0.769097	1.098710	0.0398	1.299248	0.211891
180	(Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema)	(Odobren kredit)	0.749	0.7	0.558	0.744993	1.064276	0.0337	1.176440	0.240615
184	(Stan Kuća_poseduje, Iznos kredita po kategorijama_mali iznos)	(Odobren kredit)	0.664	0.7	0.500	0.753012	1.075731	0.0352	1.214634	0.209524
188	(Iznos kredita po kategorijama_mali iznos, Vista posla_zaposlen i ima zvanično obrazovanje)	(Odobren kredit)	0.587	0.7	0.422	0.718910	1.027014	0.0111	1.067273	0.063688
258	(Radnik stranac_Plaćanje u ratama_nema, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.718	0.7	0.516	0.718663	1.026661	0.0134	1.066337	0.092089
266	(Stan Kuća_poseduje, Žirant Ko-aplikant_nema, Radnik stranac)	(Odobren kredit)	0.625	0.7	0.456	0.729600	1.042286	0.0185	1.109467	0.108187
273	(Radnik stranac_Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.806	0.7	0.574	0.712159	1.017370	0.0098	1.042241	0.088006
281	(Stan Kuća_poseduje, Plaćanje u ratama_nema, Radnik stranac)	(Odobren kredit)	0.550	0.7	0.419	0.761818	1.088312	0.0340	1.259542	0.180324
288	(Radnik stranac_Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema)	(Odobren kredit)	0.719	0.7	0.529	0.737744	1.051063	0.0257	1.135263	0.172891
297	(Stan Kuća_poseduje, Iznos kredita po kategorijama_mali iznos, Radnik stranac)	(Odobren kredit)	0.639	0.7	0.476	0.744914	1.064163	0.0287	1.176074	0.167020
304	(Iznos kredita po kategorijama_mali iznos, Radnik stranac_Vista posla_zaposlen i ima zvanično obrazovanje)	(Odobren kredit)	0.569	0.7	0.406	0.713533	1.019332	0.0077	1.047239	0.044004
439	(Stan Kuća_poseduje, Plaćanje u ratama_nema, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.524	0.7	0.400	0.763359	1.090513	0.0332	1.267742	0.174370
444	(Plaćanje u ratama_nema, Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.683	0.7	0.509	0.745242	1.064631	0.0309	1.177586	0.191506
452	(Stan Kuća_poseduje, Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.602	0.7	0.449	0.745847	1.065496	0.0276	1.180392	0.154447
458	(Stan Kuća_poseduje, Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema)	(Odobren kredit)	0.540	0.7	0.423	0.783333	1.119048	0.0450	1.384615	0.231267
490	(Radnik stranac_Plaćanje u ratama_nema, Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema)	(Odobren kredit)	0.661	0.7	0.488	0.738275	1.054679	0.0253	1.146243	0.152933
507	(Radnik stranac_Iznos kredita po kategorijama_mali iznos, Žirant Ko-aplikant_nema_Stan Kuća_poseduje)	(Odobren kredit)	0.583	0.7	0.431	0.739280	1.056114	0.0229	1.150658	0.127415
519	(Radnik stranac_Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema, Stan Kuća_poseduje)	(Odobren kredit)	0.516	0.7	0.400	0.775194	1.107420	0.0388	1.334483	0.200413

Slika 4.13 Izvedena asocijativna pravila

Support i *confidence* imaju relativno iste vrednosti za sva pravila. Mere koje pokazuju da li pravilo ima pozitivnu ili negativnu vezu i da li je važno, su *lift* i *conviction*.

Pravilo koje je prikazano u tabeli 4.3, nam govori da ono ima slabu pozitivnu vezu jer je vrednost *lift*-a 1,019 što je više od 1. Ovo nas navodi na zaključak da ako znamo da je klijent podigao mali iznos kredita, da je radnik stranac i da je zaposlen sa zvaničnim obrazovanjem da će mu biti odobren kredit. Vrednost *conviction*-a je 1,047 što znači da bi ovo pravilo bilo 4,7% više pogrešno da odobren kredit ne zavisi od malog iznosa

kredita, da je radnik stranac, zaposlenosti i činjenice da klijent ima zvanično obrazovanje. Dakle, pravilo prikazano u tabeli 4.3 jeste značajno.

Pravilo {Iznos kredita po kategorijama_mali iznos, Radnik stranac, Vrsta posla_zaposlen i ima zvanično obrazovanje} → {odobren kredit}	
Support	0,406
Confidence	0,714
Lift	1,019
Conviction	1,047

Tabela 4.3 Pravilo $\{(Iznos\ kredita\ po\ kategorijama_mali\ iznos,\ Radnik\ stranac,\ Vrsta\ posla_zaposlen\ i\ ima\ zvanično\ obrazovanje)\} \rightarrow \{odobren\ kredit\}$

Analizom pravila, prikazanog u tabeli 4.4, vidimo da ono ima vrednost 1,098 za lift. Znači da ako klijent posede stan ili kuću i ako nema nikakva plaćanja u ratama velika je verovatnoća da će mu biti odobren kredit. Conviction za ovo pravilo je 1,299 što znači da bi pravilo bilo 29,9% više pogrešno da odobrenje kredita ne zavisi od toga da li klijent posede nekretninu i da li plaća nešto na rate. Rezultati analize ovog pravila nas dovode do zaključka da je ono veoma interesantno.

Pravilo {Stan/Kuća_poseduje, Plaćanje u ratama_nema} → {odobren kredit}	
Support	0,443
Confidence	0,769
Lift	1,098
Conviction	1,299

Tabela 4.4 Pravilo $\{ Stan/Kuća_poseduje,\ Plaćanje\ u\ ratama_nema\} \rightarrow \{odobren\ kredit\}$

U tabeli 4.5 prikazano je pravilo $\{Stan/Kuća_poseduje,\ Iznos\ kredita\ po\ kategorijama_mali\ iznos,\ Plaćanje\ u\ ratama_nema\} \rightarrow \{odobren\ kredit\}$. Vrednost lift-a ($1,119 > 1$) nam govori da ako klijent posede stan ili kuću i ako nema nikakva plaćanja u ratama i iznos traženog kredita je mala, velika je verovatnoća da će mu biti odobren kredit. Vrednost mere conviction je 1,384 što znači da bi pravilo bilo 38,4% više pogrešno da odobrenje kredita ne zavisi od toga da li klijent posede nekretninu i da li plaća nešto na rate i iznos traženog kredita je mali. Rezultati analize ovog pravila nas dovode do zaključka da je ono veoma interesantno

Pravilo {Stan/Kuća_poseduje, Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema} → {odobren kredit}	
Support	0,423
Confidence	0,783
Lift	1,119
Conviction	1,384

Tabela 4.5 Pravilo {Stan/Kuća_poseduje, Iznos kredita po kategorijama_mali iznos, Plaćanje u ratama_nema} → {odobren kredit}

Analizom pravila, prikazanog u tabeli 4.6, vidimo da ono ima vrednost 1,049 za *lift*. Znači da ako se podrazumeva da klijent poseduje stan ili kuću i ukoliko nema žiranta ili ko-aplikanta velika je verovatnoća da će mu zahtev za kredit biti odobren. *Conviction* za ovo pravilo je 1,128 što znači da bi pravilo bilo 12,8% više pogrešno da odobrenje kredita ne zavisi od činjenice da klijent poseduje stan ili kuću i od činjenice da li klijent ima žiranta ili ko-aplikanta. Rezultati analize ovog pravila nas dovode do zaključka da je ono veoma interesantno.

Pravilo {Stan/Kuća_poseduje, Žirant Ko-aplikant_nema} → {odobren kredit}	
Support	0,475
Confidence	0,734
Lift	1,049
Conviction	1,128

Tabela 4.6 Pravilo {Stan/Kuća_poseduje, Žirant Ko-aplikant_nema} → {odobren kredit}

Da bi se uočena značajna pravila iskoristila pri odobravanju kreditnog zahteva, potrebno je da se za svako od pravila napravi po jedan novi atribut koji će se uključiti u skup podataka. Na slici 4.14 prikazan je proces formiranja novih atributa na osnovu izvedenih asocijativnih pravila.

```
# kreiranje novih kolona na osnovu asocijativnih pravila
def apply_rules(row):
    for idx, rule in rules.iterrows():
        if set(rule['antecedents']).issubset(row) and set(rule['consequents']).issubset(row):
            return 'Rule ' + str(idx)

data['association_rule'] = data.apply(apply_rules, axis=1)
```

Slika 4.14 Proces formiranja novih atributa na osnovu izvedenih asocijativnih pravila

4.3 Regresija

4.3.1 Teorijske osnove

Istraživanje povezanosti među različitim varijablama predstavlja ključni izazov u svakom naučnom polju. Ponekad uspevamo da identifikujemo preciznu funkcionalnu vezu između varijabli, ali često se to ne dešava. U situacijama kada to nije moguće, koristimo statističke metode kako bismo kvantifikovali prosečne promene u jednoj varijabli koje su izazvane promenama u drugoj ili više varijabli. Regresijska analiza se upravo bavi ovim konceptima. Cilj regresijske analize je da otkrije i meri ovakve veze. Drugim rečima, osnovni cilj primene regresijske analize jeste da na osnovu poznate variable predvidimo vrednost nepoznate variable, koristeći matematičku jednačinu koja opisuje njihovu međusobnu zavisnost.

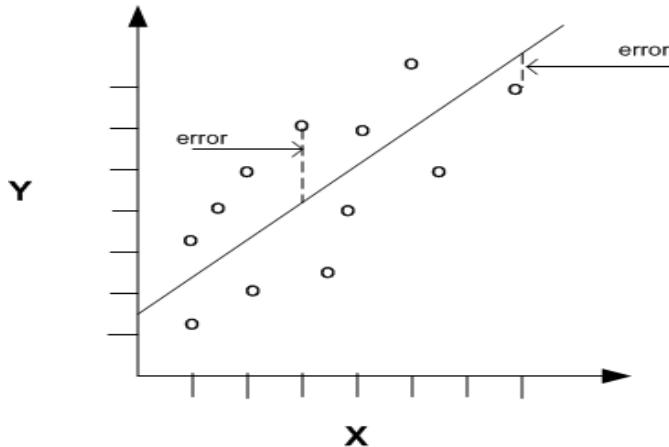
Regresiona analiza može biti jednostruka ili višestruka. U oba slučaja imamo jednu zavisnu promenljivu koja se računa na osnovu jedne promenljive, ukoliko se radi o jednostrukoj regresiji, ili dve ili više njih, ukoliko se radi o višestrukoj regresiji.

4.3.1.1 Jednostruka linearna regresija

Kao što je prethodno i pomenuto u jednostrukoj regresiji, prikazanoj na slici 4.29, figurišu jedna nezavisna i jedna zavisna promenljiva. Jednačina prave (regresiona jednačina) predstavlja linearnu aproksimaciju zavisnosti promenljivih x i y i sledećeg je oblika:

$$\hat{y} = b_0 + b_1 x .$$

\hat{y} je procenjena vrednost zavisne promenljive, a b_0 i b_1 predstavljaju y -isečak regresione prave i nagib prave respektivno. B_0 i b_1 nazivaju se koeficijenti regresione prave.



Slika 4.29 Jednostruka linearna regresija

Međutim, nekada postoji razlika između stvarne (*actual*) vrednosti i procenjene (*estimated*). Ta razlika tj. $\hat{y} - y$ naziva se predikciona greška ili ostatak (*residual*). Metoda najmanjih kvadrata minimizuje zbir kvadrata grešaka. Suština je u tome da se kroz grupu tačaka može provući više pravih linija, a najbolja je ona kod koje je zbir kvadrata odstupanja tačaka od regresione prave najmanji mogući.

Koeficijenti regresione jednačine izračunavaju se prema sledećim izrazima:

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i)(\sum y_i)] / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Koeficijent determinizacije

U domenu regresione analize, postoji niz različitih varijacija, čija važnost utiče na relevancu regresionog modela u svrhu prognoze. Ova raznolikost se manifestuje kroz raznovrsne kvantitativne vrednosti kvadrata odstupanja. Sve ove bitne numeričke informacije su sažete u tabeli 4.7 radi lakšeg pregleda.

NAZIV	FORMULA	OPIS
Suma kvadrata predikcionih grešaka <i>(sum squares error)</i>	$SSE = \sum (\hat{y} - y)^2$	Mera greške u predikciji koja proizilazi iz upotrebe regresione jednačine.
Ukupna suma kvadrata <i>(sum squares total)</i>	$SST = \sum_{i=1}^n (y - \bar{y})^2$	Mera varijabilnosti zavisne promenljive y oko svoje srednje vrednosti \bar{y} ne uzimajući u obzir nezavisnu promenljivu (x).
Suma kvadrata regresije <i>(sum squares regression)</i>	$SSR = \sum_{i=1}^n (\hat{y} - \bar{y})^2$	Mera poboljšanja predikcije kada se koristi x promenljiva u odnosu na slučaj u kome se ignoriše.

Tabela 4.7 Sume kvadrata odstupanja

Koeficijent determinacije meri koliko se regresiona jednačina tj. dobijeni linearni odnos x i y uklapa u uzorak podataka i definisan je sa:

$$r^2 = \frac{SSR}{SST}$$

r^2 interpretira se kao deo varijabilnosti koji je objašnjen pomoću regresije. Vrednost koeficijenta determinacije nalazi se u intervalu $[0,1]$, a prilikom primene regresije treba da se teži upravo maksimalnoj vrednosti 1.

Standardna greška procene

Mera tačnosti procene dobijene pomoću regresione jednačine naziva se standardna greška procene (s). Računa se prema jednačini:

$$s = \sqrt{\frac{SSE}{n - m - 1}},$$

gde je n broj slogova u datom skupu podataka, m je broj nezavisnih promenljivih (u slučaju jednostrukе regresije važi $m=1$) a SSE predstavlja sumu kvadrata predikcionih grešaka. Regresioni model bi trebalo da teži što većoj preciznosti tj. što manjim vrednostima s .

Koeficijent korelacije

Koeficijent korelacije (r) predstavlja indikator „snage“ odnosa dve promenljive a računa se prema sledećoj formuli:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

s_x i s_y su standardne devijacije za vrednosti promenljivih x i y . Računaju se na osnovu izraza:

$$s_x = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}} \text{ i } s_y = \sqrt{\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n}}$$

Ukoliko je vrednost koeficijenta korelacije blizu 1 to znači da su varijable pozitivno korelirane i da porastom vrednosti varijable x raste i vrednost varijable y . U slučaju kada je vrednost koeficijenta korelacije blizu -1 to znači da su varijable negativno korelirane. Tada je porast vrednosti x varijable povezan sa smanjenjem vrednosti y varijable. Ostale vrednosti koeficijenta korelacije vode do zaključka da varijable nisu u korelaciji.

ANOVA (Analysis of Variance) tabela

Ova tabela prikazuje statistike značajne za regresiju. U tabeli 4. 8 pobrojani su elementi statistika sa objašnjnjima.

ELEMENT	OPIS
Regression Statistics	
Multiple R	Koeficijent korelacije
R^2	Koeficijent determinacije određenosti
Standard Error	Standardna greška procene
Adjusted R^2	Popravljen R^2
Observations	Broj uzoraka
ANOVA	
Regression	Regresija
Residual	Predikciona greška
SS (Sum of Squares)	Suma grešaka
MS (Mean Square)	Srednja suma grešaka
Significance of F	P-vrednost za F test značajnosti

T - Statistics	
<i>Coefficients</i>	Prikazuje b_0 i b_1 koeficijente regresione jednačine.
<i>P-value</i>	Prikazuje p vrednosti za pojedinačne t -testove za nezavisne promenljive.
<i>T-test</i>	Služi za testiranje hipoteze da ne postoji veza između date nezavisne promenljive i zavisne promenljive. Hipoteza se odbacuje ako važi $p < \beta$.
β	Prag koji bira korisnik. Obično se uzima vrednost 0.05. Znači ako je $p < \beta$ zaključujemo da postoji veza između x i y tj. da je x korisna za regresiju.
<i>Lower 95%</i>	Interval poverenja za vrednosti koeficijenata. Ako se vrednost 0 ne nalazi u ovom intervalu možemo da zaključimo da je promenljiva korisna za regresiju.
<i>Upper 95%</i>	

Tabela 4.8 Statistike regresije

ANOVA tabela, prikazana na slici 4.30, prikazuje varijaciju suma grešaka, gde F statistika služi za testiranje hipoteze da ne postoji zavisnost između svih nezavisnih i zavisne promenljive. Vrednost koja je označena je p vrednost, pa ako je $p < \alpha$ prepostavka da ne postoji zavisnost se odbacuje, gde je $\alpha = 1 -$ interval poverenja. Dakle, ako je $\alpha=0.05$ onda smo 95% sigurni da zavisnost postoji, sa napomenom da parametar α bira korisnik i zavisi od oblasti, obično je 0.05.

Source of Variation	Sum of Squares	df	Mean Square	F
Regression	SSR	m	$MSR = \frac{SSR}{m}$	$F = \frac{MSR}{MSE}$
Error (or residual)	SSE	$n - m - 1$	$MSE = \frac{SSE}{n - m - 1}$	
Total	$SST = SSR + SSE$	$n - 1$		

Slika 4.30 ANOVA tabela za linearnu regresiju

4.3.1.2 Višestruka linearna regresija

Ako se ispituje zavisnost jedne pojave od dve ili više nezavisnih pojava, onda se govori o višestrukoj regresiji. Zadatak regresije je da otkrije što više faktora (nezavisnih promenljivih) koji utiču na zavisnu promenljivu. Polazi se od pretpostavke da što je više nezavisnih varijabli u modelu, sve je manji uticaj latentne promenljive (standardne greške) ε_i , $i = 1, 2, \dots, n$. Veoma je bitno pažljivo birati promenljive koje će biti uključene u model.

Regresiona jednačina ima sledeći oblik:

$$b_1x_1 + b_2x_2 + \dots + b_nx_n + b_0 = \tilde{y},$$

gde je \hat{y} procenjena vrednost zavisne promenljive, b_0 presek regresione prave sa y -osom a b_1, b_2, \dots, b_n koeficijenti uz nezavisne promenljive.

4.3.2 Jednostruka regresija

4.3.2.1 Primena jednostrukih regresija

Kako je zadatak da se jednostrukim regresijom modeluje zavisnost *iznos kredita* atributa od *trajanje računa* atributa, nakon učitavanja početnog skupa podataka treba da se izdvoje ova dva atributa.

Nakon toga kao opcionalni korak inicijalni skup podataka se deli na trenirajući i na testni podskup. Ovo se vrši pomoću funkcije `train_test_split` iz biblioteke `sklearn.model_selection`. Nakon što su testni i obučavajući skupovi podataka izdvojeni pomoću funkcije `LinearRegression` iz `sklearn.linear_model` biblioteke. Izlaz ove funkcije je izgenerisani regresioni model. Dakle, on na osnovu podataka nalazi regresionu jednačinu koja služi za predikciju budućih slučajeva putem funkcije `model.fit` kojoj se prosleđuje obučavajući skupovi.

Sledeći operator koji se koristi je `model.predict` kome se prosleđuje testni skup podatka.

Zatim se vrše operacije nad dobijenim modelom da bi se došlo do parametara koji služe za merenje performansi regresionog modela. Izabrani su sledeći kriterijumi za evaluaciju: tipična greška procene (*root mean squared error*), koeficijent korelacije (*correlation*) i koeficijent determinacije (*squared correlation*).

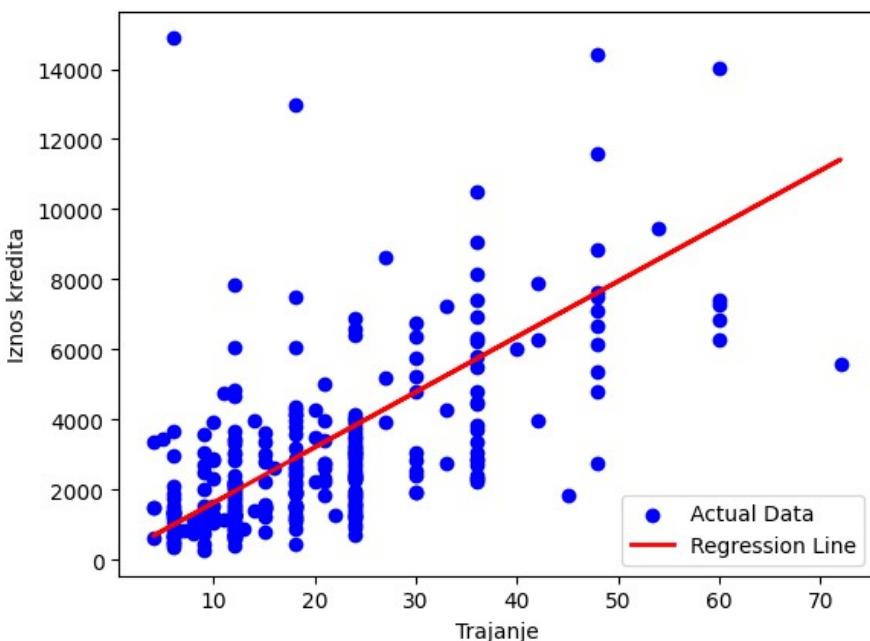
4.3.2.2 Proučavanje rezultata jednostrukog regresije

Na slici 4.15 može se videti grafik dobijenog modela, a sa slike 4.16, a na osnovu kolone *estimate* da se zaključiti da jednačina regresije glasi:

$$\hat{y} = 157.87 + 51.629x.$$

Dakle, koeficijenti regresije su:

- Nagib, $b_1 = 51.629$
- Presek sa y -osom $b_0 = 157.87$



Slika 4.15 Model dobijen primenom jednostrukog linearnog regresije

Na osnovu slike 4.16 se takođe može zaključiti da *p*-vrednost za promenljivu *trajanje* iznosi 0. Ova vrednost (manja od 0.05) ukazuje da postoji linearna veza između prediktor varijable *trajanje* i varijable *iznos kredita*.

Coefficients:				
	Estimate	Std. Error	t value	p value
_intercept	51.629469	139.714462	0.3695	0.711807
x1	157.869891	3.068807	51.4434	0.000000

Slika 4.16 Model dobijen primenom jednostrukne linearne regresije

Tipična greška procene, prikazana na slici 4.17, govori da se prilikom gore izvedene regresione jednačine (*iznos kredita* = 157.87 + 51.63**trajanje*) tipično greši 1974.7998 u proceni vrednosti atributa *iznos kredita*.

Root mean squared error 1974.799837766968

Slika 4.17 Root mean squared error

Na slici 4.18 prikazana je dobijena vrednost koeficijenta korelacije i ona iznosi 0.622, što predstavlja pozitivnu korelaciju. Ova vrednost je približna 1 što znači da posmatrani atributi jesu u delimičnoj korelaciji.

Correlation Coefficient: 0.6223747111078135

Slika 4.18 Correlation

Sa slike 4.19, može se zaključiti da je vrednost koeficijenta determinacije 0.329 što znači da je približno 33% varijabilnosti u vrednostima atributa *iznos kredita* objašnjeno pomoću vrednosti atributa *trajanje*.

Squared Correlation: 0.3288909360584207

Slika 4.19 Performance Vector – kriterijum squared correlation

U tabeli 4.9 su prikazani rezultati regresije dobijeni u Excel-u na osnovu podataka za atributе *iznos kredita* i *trajanje*.

Regression Statistics	
Multiple R	0,624713646
R Square	0,390267139
Adjusted R Square	0,389655572
Standard Error	9,41839716
Observations	999

Tabela 4.9 Rezultati regresije dobijeni u Excel-u – primer: *trajanje, iznos kredita*

U tabeli 4.10 su prikazani rezultati regresije ANOVA tabele dobijeni u Excel-u na osnovu podataka za atribute *trajanje* i *iznos kredita*.

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	56607,18283	56607,18283	638,1423125	3,1493E-109
Residual	997	88440,08644	88,70620506		
Total	998	145047,2693			

Tabela 4.10 ANOVA tabela za jednostruku linearnu regresiju u Excel-u – primer: *trajanje, iznos kredita*

Na osnovu gore izvedenih podataka može se zaključiti da atribut *trajanje* je delimično koristan za određivanje vrednosti *iznosa kredita* atributa ako se koristimo pretpostavkom da što je vremenski period duži, iznos tražene sume je veći.

Već je ranije rečeno da u regresionoj jednačini b_0 i b_1 predstavljaju statistike čije vrednosti zavise od odabranog uzorka. b_0 i b_1 su procene parametara β_0 i β_1 populacionog modela:

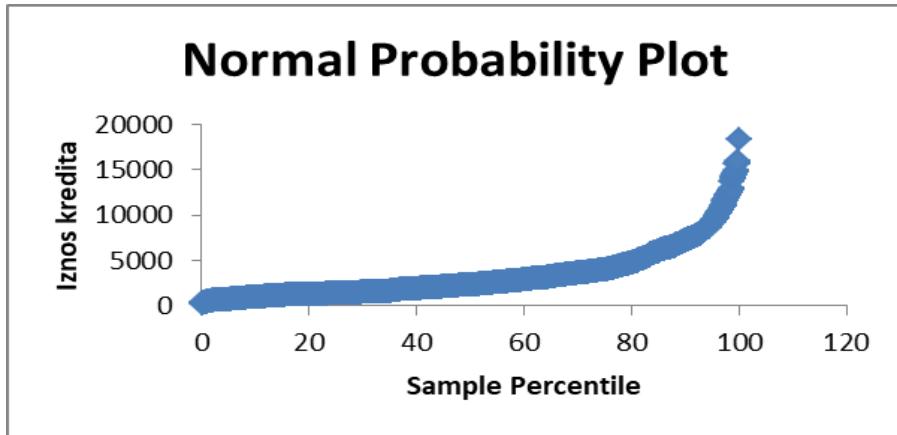
$$y = \beta_0 + \beta_1 x + \varepsilon$$

koji predstavlja stvaran odnos nezavisne i zavisne promenljive. Greška ε uvedena je zbog neodređenosti modela. Za grešku ε prepostavlja se da važe sledeće pretpostavke:

1. Normalnost greške. Greška je slučajna promenljiva sa normalnom distribucijom.
2. Verovatnoća raspodele greške ima konstantu varijansu
3. Vrednosti grešaka su statistički nezavisne
4. Vrednosti grešaka imaju srednju vrednost 0

Sve do sada navedene metode za rasuđivanje o parametrima regresije zavise od toga da li se podaci (uzorak) uklapaju u pretpostavke regresije. Postoje dve grafičke metode za proveru uklapanja podataka u pretpostavke:

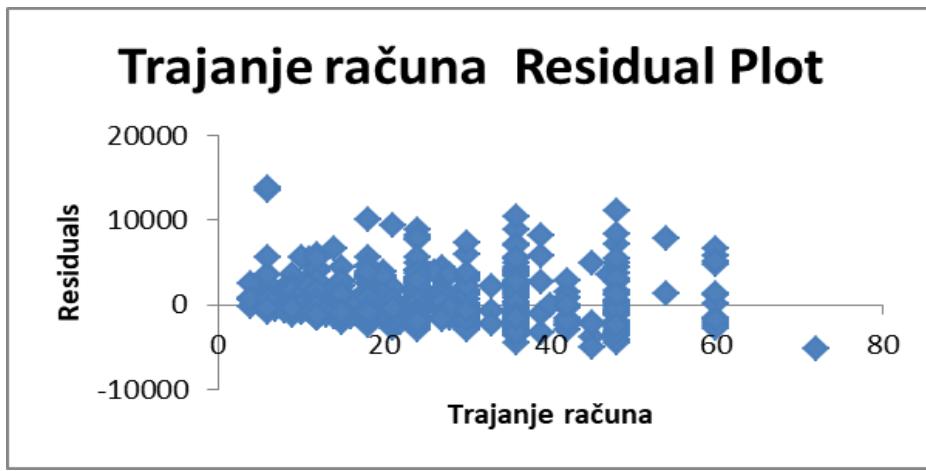
1. Grafik normalne distribucije reziduala (ostataka) $\hat{y}_i - y_i$
2. Grafik reziduala u odnosu na \hat{y} (vrednosti y dobijene pomoću regresione jednačine)



Slika 4.20 *Normal probability plot*

Na slici 4.20 prikazan je grafik normalne distribucije reziduala na osnovu koga možemo da zaključimo da se podaci delimično uklapaju u prepostavke regresije. Da bi prepostavka ostala neoborenja potrebno je da reziduali što više prate pravu liniju što vidimo da je ovde slučaj.

Sa grafika, prikazanog na slici 4.21, može da se zaključi da reziduali imaju konstantu varijansu kao i da imaju srednju vrednost 0. Dakle, možemo da zaključimo da se podaci uklapaju u prepostavku regresije.



Slika 4.21 *Residual plot*

4.3.3 Višestruka regresija

4.3.3.1 Primena višestruke regresije

Nakon učitavanja početnog seta podataka potrebno je obraditi podatke istim postupcima kao i kod jednostrukih regresija. Ovde će biti prikazan postupak za rešenje zadatka da se višestrukom regresijom modeluje zavisnost *iznos kredita* od ostalih atributa, u nastavku će biti opisan proces.

Nakon toga kao opcionalni korak inicijalni skup podataka se deli na trenirajući i na testni podskup. Ovo se vrši pomoću funkcije *train_test_split* iz biblioteke *sklearn.model_selection*. Nakon što su testni i obučavajući skupovi podataka izdvojeni pomoću funkcije *LinearRegression* iz *sklearn.linear_model* biblioteke ili *OLS* iz *statsmodels.api* biblioteke. Izlaz ove funkcije je izgenerisani regresioni model. Dakle, on na osnovu podataka nalazi regresionu jednačinu koja služi za predikciju budućih slučajeva putem funkcije *model.fit* kojoj se prosleđuje obučavajući skupovi. Sledeci operator koji se koristi je *model.predict* kome se prosleđuje testni skup podatka.

Zatim se vrše operacije nad dobijenim modelom da bi se došlo do parametara koji služe za merenje performansi regresionog modela. Izabrani su sledeći kriterijumi za evaluaciju: tipična greška procene (*root mean squared error*), koeficijent korelacije (*correlation*) i koeficijent determinacije (*squared correlation*).

4.3.3.2 Proučavanje rezultata višestruke regresije

Sa obzirom na to da skupovi podataka obično sadrže veliki broj atributa javlja se potreba za određivanjem koji su od njih korisni, a koji ne za linearu regresiju. Alat *RapidMiner* nudi takvu opciju u operatoru *LinearRegression*, kroz opcije za automatsku selekciju atributa (*feature_selection*) i eliminaciju koreliranih atributa (*eliminate_colinear_features*).

Sa slike 4.22, na kojoj je prikazan model dobijen primenom višestruke regresije, možemo da zaključimo da u njemu nisu prisutni svi atributi koji su imali vrednost težine nula. Težine atributa su prikazane na slici 4.40. Dakle, svi atributi koji imaju vrednost težine 0 eliminisu se iz modela automatskom selekcijom atributa.

OLS Regression Results							
<hr/>							
Dep. Variable:	Iznos kredita	R-squared:	0.220				
Model:	OLS	Adj. R-squared:	0.214				
Method:	Least Squares	F-statistic:	35.00				
Date:	Sat, 30 Sep 2023	Prob (F-statistic):	6.23e-49				
Time:	20:35:07	Log-Likelihood:	-9239.5				
No. Observations:	1000	AIC:	1.850e+04				
Df Residuals:	991	BIC:	1.854e+04				
Df Model:	8						
Covariance Type:	nonrobust						
<hr/>							
		coef	std err	t	P> t	[0.025	0.975]
const		2943.6663	251.251	11.716	0.000	2450.621	3436.711
Kamata - u procentima od tekućih prihoda		-740.9401	71.501	-10.363	0.000	-881.250	-600.630
Prebivalište (u godinama)		63.0221	72.554	0.869	0.385	-79.356	205.400
Postojeći kredit u ovoj banci		85.1967	139.109	0.612	0.540	-187.785	358.178
Izdržavani: broj lica koji zavise od prihoda aplikanta		96.0052	223.095	0.430	0.667	-341.787	533.797
Vrsta_posla_nema_zvanično_obrazovanje: ima stalan posao		-58.9854	207.794	-0.284	0.777	-466.752	348.781
Vrsta_posla_nezposlen/nema_zvanično_obrazovanje		-160.0479	429.224	-0.373	0.709	-1002.341	682.245
Vrsta_posla_radi_u_menadžmentu/samo-zaposlen/visoko_obrazovanje		2577.3740	225.847	11.412	0.000	2134.181	3020.567
Vrsta_posla_zaposlen_i_ima_zvanično_obrazovanje		585.3257	162.961	3.592	0.000	265.538	905.114
Telefon_ima		1921.6788	161.859	11.873	0.000	1604.054	2239.304
Telefon_nema		1021.9875	146.645	6.969	0.000	734.218	1309.757
<hr/>							
Omnibus:	309.624	Durbin-Watson:	1.975				
Prob(Omnibus):	0.000	Jarque-Bera (JB):	903.999				
Skew:	1.557	Prob(JB):	5.00e-197				
Kurtosis:	6.464	Cond. No.	1.58e+16				

Slika 4.22 Model dođen primenom višestruke linearne regresije

Većina statistika za višestruku regresiju tumači se na isti način kao i kod jednostrukih regresija. Na slici 4.23 su prikazani rezultati višestruke regresije dođeni u Excel-u.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4549.595862	662.178382	6.870649941	1.12844E-11	3250.161439	5849.030285	3250.161439	5849.030285
Kamata – u procentima od tekucih prihoda	-772.3552049	70.35346059	-10.97821199	1.53589E-26	-910.41424	-634.2961699	-910.41424	-634.2961699
Prebivalište (u godinama)	58.84012057	71.2099943	0.826290202	0.408838578	-80.89574462	198.5799858	-80.89574462	198.5799858
Postojeći kredit u ovoj banci	123.7064036	136.6648931	0.905180554	0.365589971	-144.4797397	391.892547	-144.4797397	391.892547
Izdržavani: broj lica koji zavise od prihoda aplikanta	82.48541864	218.961851	0.376711369	0.706468838	-347.1972381	512.1680754	-347.1972381	512.1680754
Odobren kredit	-1063.210761	170.5502043	-6.234004614	6.72021E-10	-1397.892188	-728.529333	-1397.892188	-728.529333
Vrsta_posla_nema_zvanično_obrazovanje; ima stalan posao	177.630633	555.7515565	0.319622376	0.749322045	-912.9557146	1268.216981	-912.9557146	1268.216981
Vrsta_posla_nezposlen/nema_zvanično_obrazovanje	0	0	65535	#NUM!	0	0	0	0
Vrsta_posla_radi_u_menadžmentu/samo-zaposlen/visoko_obrazovanje	2707.776064	574.8532438	4.710377984	#NUM!	1579.70527	3835.846858	1579.70527	3835.846858
Vrsta_posla_zaposlen_i_ima_zvanično_obrazovanje	794.5357469	536.6371639	1.480582785	0.13903595	-258.5412219	1847.612716	-258.5412219	1847.612716
Telefon_ima	961.7858393	176.6067137	5.445918896	6.50186E-08	615.219341	1308.352338	615.219341	1308.352338
Telefon_nema	0	0	65535	#NUM!	0	0	0	0

Slika 4.23 Rezultati višestruke regresije dođeni u Excel-u

Na osnovu p-vrednosti t-testa za promenljive *prebivalište u godinama*, *postojeći kredit u ovoj banci*, *izdržavani broj lica koji zavise od prihoda aplikanta*, *vrsta_posla = nema_zvanično_obrazovanje-ima_stalan_posao*, *vrsta_posla = ima_zvanično_obrazovanje/stalan_posao* (značajno veće od 0.05) vidimo da su ove promenljive beskorisne. Pored toga i vrednost 0 nalazi se u intervalu poverenja za koeficijent uz ove promenljive što takođe potvrđuje da nisu korisne.

Za razliku od jednostrukih kod višestruke regresije koeficijent determinacije nije dobra mera kvaliteta. Može se pokazati da se vrednost ovog koeficijenta uvek povećava sa dodavanjem novih nezavisnih promenljivih bez obzira da li su te promenljive korisne za regresiju. Iz tog razloga uvedena je nova statistika koja se naziva popravljeni koeficijent determinacije - R^2 *popravljen* (*Adjusted R-squared*). Kada se posmatra dobijena vrednost *Adjusted R-squared* nakon primene regresije, sa i bez posmatranog atributa, može se zaključiti da ukoliko je *Adjusted R-squared* manji sa posmatranim atributom tada taj atribut nije koristan za regresiju.

Regression Statistics	
Multiple R	0,499773
R Square	0,249773
Adjusted R Square	0,240933
Standard Error	2456,019297
Observations	1000

Tabela 4.11 Rezultati višestruke regresije u *Excel-u*

ANOVA					
	df	SS	MS	F	Significance F
Regression	11	1988165148	180742286,16	36,6224	4,28497E-66
Residual	990	5971710480	6032030,79		
Total	1001	7959875627			

Tabela 4.12 ANOVA tabela za višestruku linearnu regresiju u *Excel-u*

Tipična greška procene, prikazana na slici 4.43, nam govori da nakon primenjene višestruke regresije tipično grešimo 2440.981 u proceni vrednosti atributa *iznos kredita*.

Root mean squared error 2440.9806840694173

Slika 4.24 Root mean squared error

Na slici 4.23 prikazana je dobijena vrednost koeficijenta korelacije i ona iznosi 0.466. Ova vrednost nije blizu 1 što znači da posmatrani atributi nisu u korelaciji.

Correlation Coefficient: 0.4659701532250734

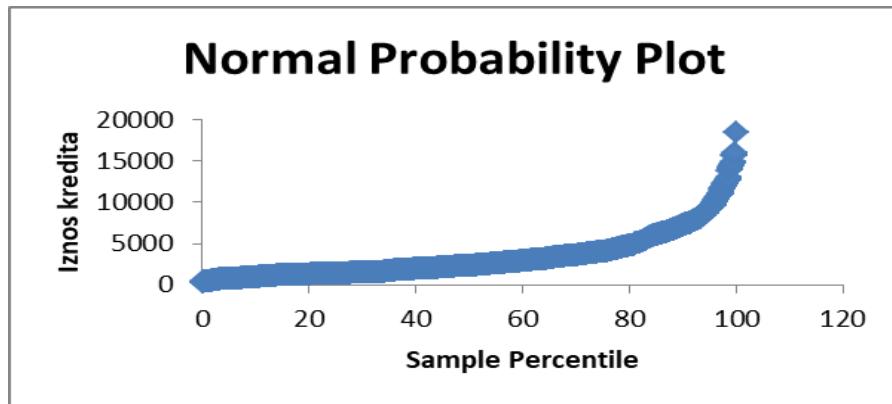
Slika 4.23 Correlation

Sa slike 4.24, možemo da zaključimo da je vrednost koeficijenta determinacije 0.145 što znači da je približno 14.5% varijabilnosti u vrednostima atributa *iznos kredita* objašnjeno pomoću posmatranih vrednosti atributa.

Squared Correlation: 0.14511338348339087

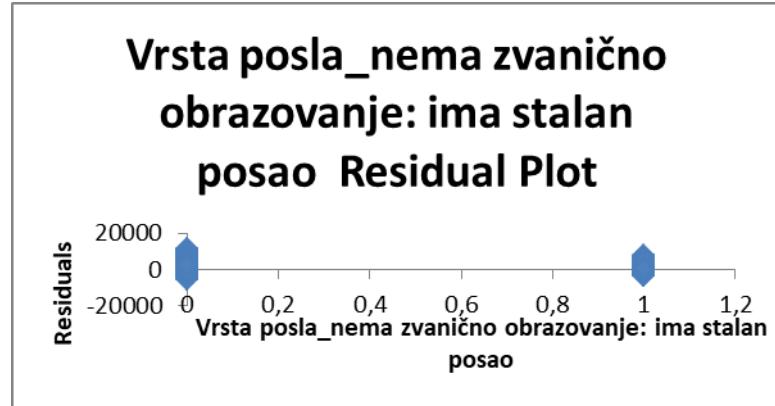
Slika 4.24 Squared correlation

Na slici 4.46 prikazan je grafik normalne distribucije reziduala na osnovu koga možemo da zaključimo da se podaci ne uklapaju u prepostavke regresije. Da bi prepostavka ostala neoborenata potrebno je da reziduali što više prate pravu liniju što vidimo da to ovde nije slučaj.



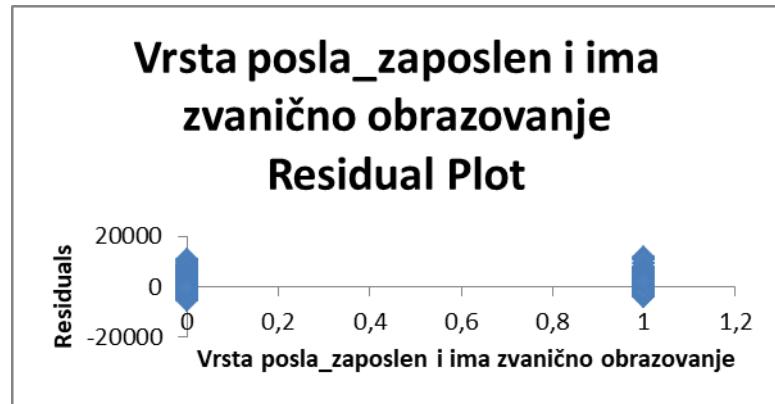
Slika 4.25 Normal Probability Plot

Sa grafika, prikazanog na slici 4.26, može da se zaključi da reziduali nemaju konstantu varijansu kao i da nemaju srednju vrednost 0. Dakle, možemo da zaključimo da se podaci ne uklapaju u prepostavku regresije.



Slika 4.26 Residual Plot za atribut $vrsta\ posla = nema\ zvanično\ obrazovanje: ima\ stalan\ posao$

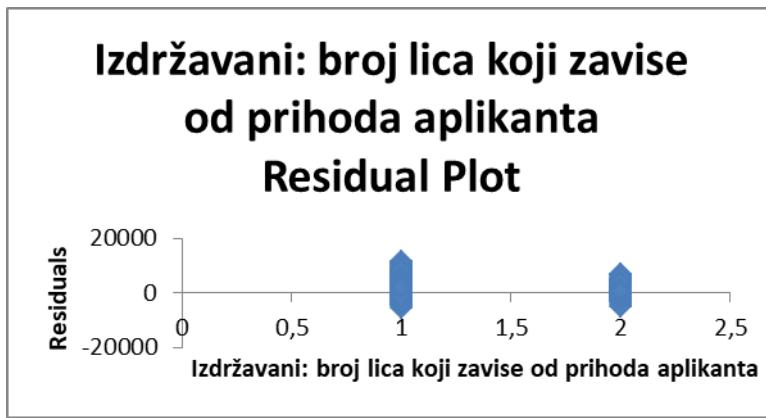
Na slikama 4.27, 4.28, 4.29, 4.30 i 4.31, su prikazani grafici reziduala za još par atributa, respektivno. Za svaki od njih važe gore navedene tvrdnje za atribut $vrsta\ posla = nema\ zvanično\ obrazovanje: ima\ stalan\ posao$.



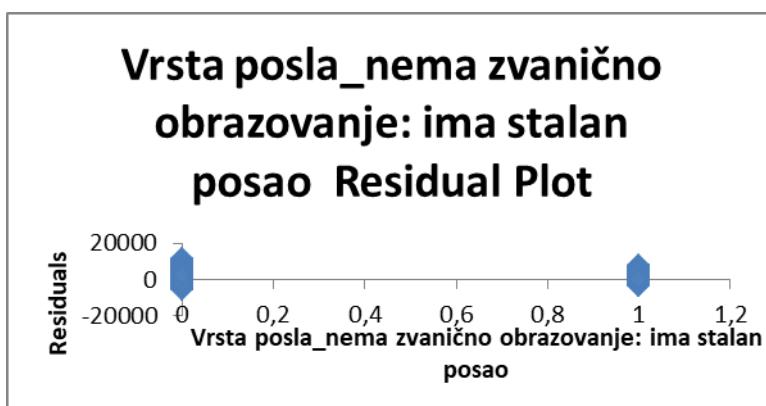
Slika 4.27 Residual Plot za atribut $vrsta\ posla = zaposlen\ i\ ima\ zvanično\ obrazovanje$



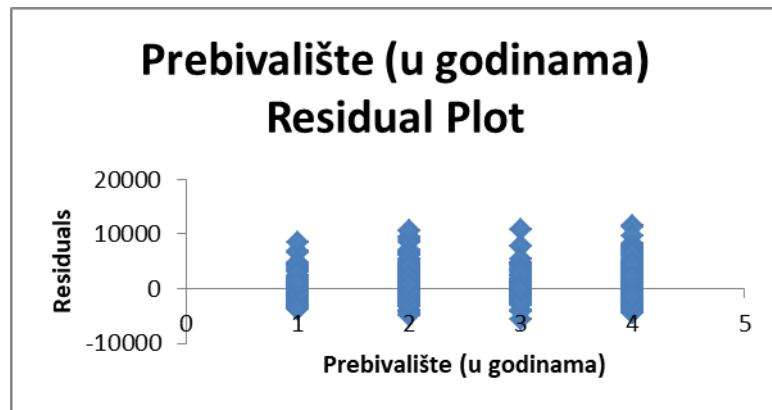
Slika 4.28 Residual Plot za atribut *postojeći kredit u ovoj banci*



Slika 4.29 Residual Plot za atribut *Izdržavani: broj lica koji zavise od prihoda aplikanta*



Slika 4.30 *Residual Plot* za atribut *Vrsta posla_nema zvanično obrazovanje: ima stalani posao*



Slika 4.31 *Residual Plot* za atribut *Prebivalište (u godinama)*

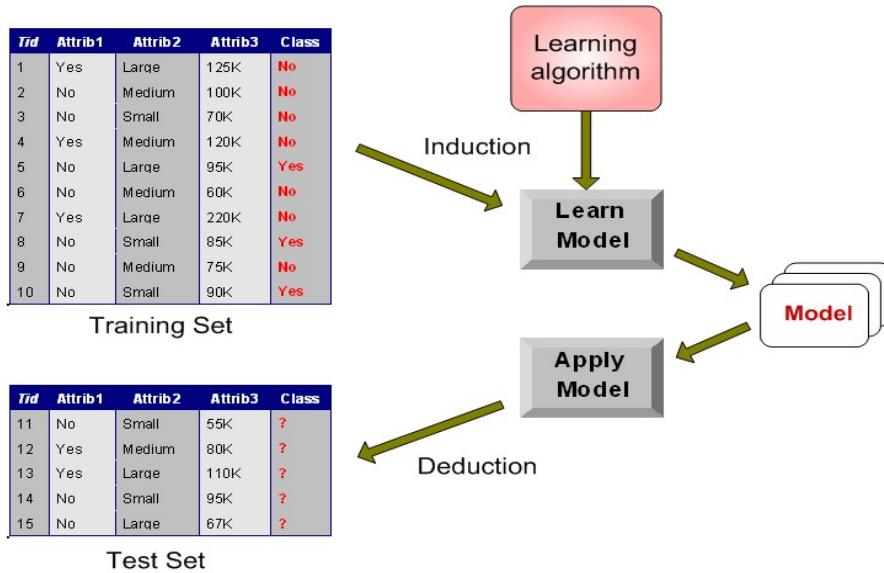
4.4 Klasifikacija

U ovom delu faze modelovanja biće formirani modeli za svaki od sledećih klasifikatora:

- Stabla odlučivanja (DTREE)
- k -najbližih suseda (KNN)
- Naivni Bayes (NB)
- Maštine potpornih vektora (SVM)

Takođe će za svaki od njih biti izvršena optimizacija parametara kao i selekcija osobina. Performanse klasifikatora biće evaluirane na test skupu dok će se optimizacija parametara i selekcija osobina vršiti na *train* skupu pri čemu će se performanse modela izračunavati pomoću 3-tostrukne unakrsne validacije.

Klasifikacija, prikazana na slici 4.32, predstavlja određivanje klase novim sloganima (onim koji nemaju vrednost atributa klase) što pouzdanije. Dakle, zadatak je da se pomoću datog skupa sloganova formira model kod kojeg se vrednost klasnog atributa izražava kao funkcija vrednosti ostalih atributa. Test skup je skup neklasifikovanih sloganova koji služi da se odredi pouzdanost modela.



Slika 4.32 Klasifikacija

Pre samog formiranja modela i evaluacije potrebno je odrediti način testiranja modela. To podrazumeva kreiranje posebnog test skupa iz celog skupa podataka koji će služiti za evaluaciju dok će ostatak biti obučavajući skup za formiranje modela. Odnos koji je u ovom radu korišćen je 75% za obučavajući i 25% za test skup.

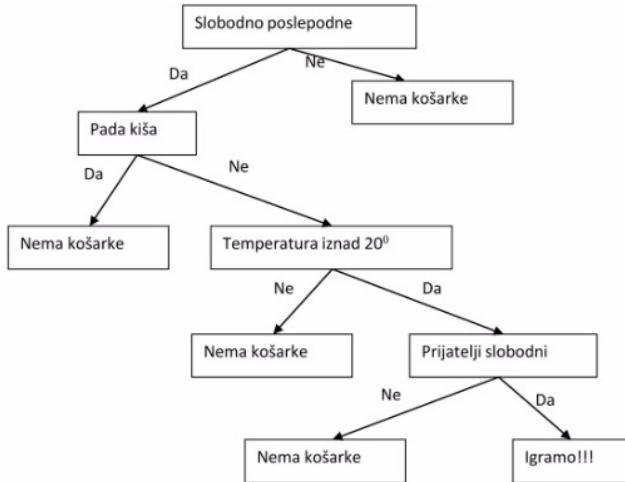
4.4.1 Stabla odlučivanja

4.4.1.1 Teorijske osnove

Ilustracija 4.33 predstavlja stabla odlučivanja koja omogućavaju strukturiran prikaz znanja na hijerarhijski način. Stabla odlučivanja se sastoje od čvorova i grana koje povezuju nadređene čvorove sa njihovim podređenim čvorovima. Početni čvor stabla, koji nema nadređeni čvor, naziva se "korenski čvor". Nasuprot tome, čvorovi koji nemaju podređene čvorove nazivaju se "listovi". U okviru stabla odlučivanja, listovi prikazuju sve moguće rešenja za dati problem i mogu se nazivati i "čvorovima odgovora". Svi preostali čvorovi unutar stabla se nazivaju "čvorovi odlučivanja" i služe za postavljanje pitanja ili donošenje odluka.

Tokom procesa odlučivanja, kada se odgovara na postavljena pitanja ili se donose odluke, određuje se konkretna grana stabla odlučivanja koju treba

dalje pratiti. Odgovori na postavljena pitanja mogu biti "da" ili "ne", ili čak izbori između različitih vrednosti ili raspona vrednosti



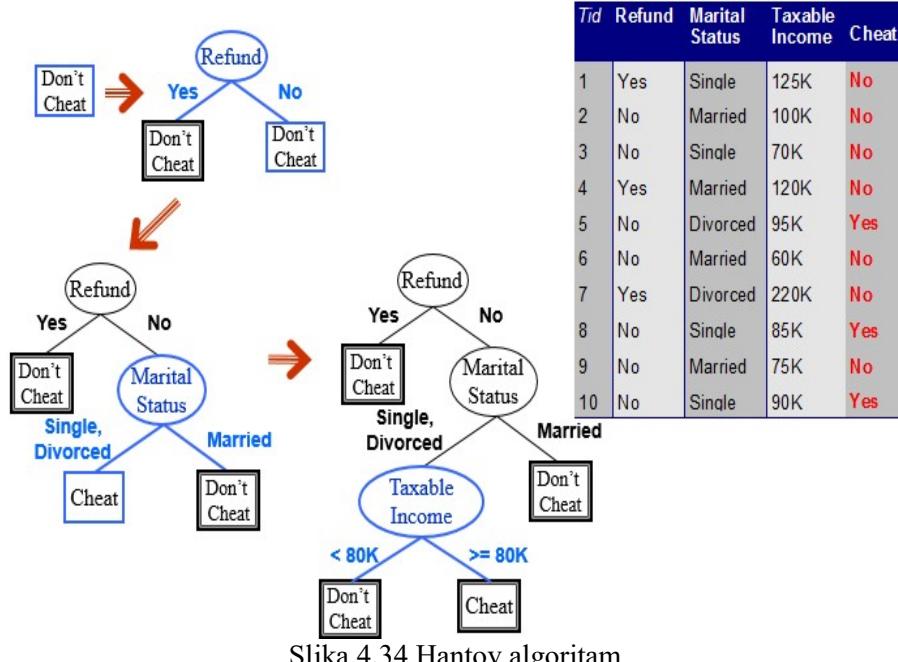
Slika 4.33 Primer stabla odlučivanja

Hantov algoritam

Algoritam, prikazan na slici 4.34, koji je osnova većini algoritama za formiranje stabla odlučivanja kao što su *CART*, *C4.5* i *ID3* je Hantov (*Hunt*) algoritam. Neka je D_t skup slogova koji su unapred klasifikovani i vezani su za čvor t , a $y = \{y_1, y_2, \dots, y_c\}$ skup klasa. Hantov algoritam funkcioniše na sledeći način:

Korak 1. Ako svi slogovi u D_t pripadaju istoj klasi y_t , onda t postaje list koji obeležen klasom y_t .

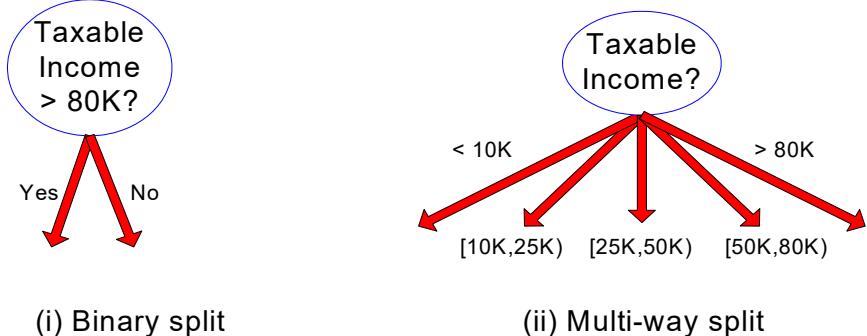
Korak 2. Ako svi slogovi u D_t ne pripadaju istoj klasi, formira se test uslov za neki od atributa kako bi se D_t podelio na manje podskupove. U zavisnosti od rezultata testa slogovi se distribuiraju u podskupove. Po jedan podskup za svaki mogući ishod testa. Svaki od podskupova postaje novi čvor koji je potomak čvora t . Algoritam se nakon toga primenjuje na svaki od novonastalih podskupova.



Slika 4.34 Hantov algoritam

Jedno od bitnih pitanja prilikom određivanja načina razdvajanja elemenata obučavajućeg skupa jeste „Kako specificirati uslov za atributski test?“. Test uslovi zavise od dva kriterijuma na osnovu kojih se bira test uslov. To su tip atributa (nominalni, ordinalni i kontinualni) i način razdvajanja (binarni i višestrani). **Nominalni** atributi podrazumevaju da se prilikom višestrukog razdvajanja formira onoliko particija koliko ima vrednosti atributa, dok se kod binarnog razdvajanja formiraju dve particije. **Ordinalni** atributi imaju ista svojstva kao i nominalni sa napomenom da je prilikom binarnog razdvajanja potrebno održati redosled. **Kontinualni** atributi imaju dva pristupa, prikazana na slici 4.35, za razdvajanje i to su:

- diskretizacija - interval u kome atribut uzima vrednosti deli se na podintervale i
- binarno odlučivanje - bira se vrednost v i formiraju se dve particije u odnosu na tu vrednost. Prilikom odabira v u obzir se uzimaju sve vrednosti koje kontinualni atribut može da uzme. Zato je ovaj pristup računski zahtevan.



Slika 4.35 Razdvajanje po kontinualnim atributima

Druge bitne pitanje koje se postavlja pri određivanju načina razdvajanja elemenata obučavajućeg skupa jeste „Kako odrediti najbolje razdvajanje?“. U tu svrhu uvodi se potrebna mera „nečistoće“ čvora (*node impurity*). Kao mere nečistoće čvora koriste se: *GINI* indeks, Entropija i Greška klasifikacije.

GINI indeks

Gini indeks za čvor t:

$$GINI(t) = 1 - \sum_j [p(j | t)]^2$$

gde je $p(j | t)$ relativna frekvencija klase j u čvoru t . **Maksimum** ($1-1/n_c$), kada su elementi uniformno raspodeljeni po svim klasama, ukazuje na najmanje interesantne informacije, dok **minimum** (0.0), kada svi elementi pripadaju istoj klasi, ukazuje na najinteresantnije informacije.

Da bi utvrdili kvalitet razdvajanja potrebno je da uporedimo nivo nečistoće čvora koji se deli sa nivoom nečistoće svih particija na koje će se on podeliti. Ako se čvor p deli na k particija onda se kvalitet razdvajanja naziva dobitak (*gain*) i definije sa:

$$GAIN = I(p) - \sum_{i=1}^k \frac{n_j}{n} I(j)$$

gde je I mera nečistoće čvora, n broj slogova u p , a n_i broj slogova iz p koji će se posle razdvajanja naći u particiji i . Što veća vrednost $GAIN$ to je

razdvajanje bolje. U slučaju da se kao mera nečistoće koristi *entropija* mera *GAIN* naziva se dobitak na informaciji (*information gain*).

Mere nečistoće, kao što su *Gini* indeks, prilikom odabira atributa za razdvajanje favorizuje attribute koji imaju veliki broj vrednosti. Međutim ako bi svaka vrednost bila list, stablo odlučivanja bilo bi beskorisno tj. ne bi bilo nikakve generalizacije nad podacima. Da bi se ovaj nedostatak prevazišao uvedena je nova mera koja se naziva koeficijent dobitka (*gain ratio*). Ako se čvor p deli na k particija koeficijent dobitka definiše se sa:

$$\text{GainRatio} = \frac{\text{GAIN}}{\text{SplitInfo}} \quad \text{SplitInfo} = -\sum_{i=1}^k \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

gde je n broj slogova u p , a n_i broj slogova iz p koji će se posle razdvajanja naći u particiji i . *SplitInfo* je mera entropije razdvajanja. Razdvajanja sa većim brojem particija imaju veću entropiju i samim tim manju vrednost za *GainRatio*.

Entropija

Entropija čvora t :

$$\boxed{\text{Entropy}(t) = -\sum_j p(j | t) \log p(j | t)},$$

gde je $p(j | t)$ relativna frekvencija klase j u čvoru t . **Maksimum** ($\log n_c$), kada su elementi uniformno raspodeljeni po svim klasama, ukazuje na najmanje interesantne informacije, dok **minimum** (0.0), kada svi elementi pripadaju istoj klasi, ukazuje na najinteresantnije informacije. Sračunavanja na bazi entropije su slična sračunanjima na bazi *GINI* indeksa.

Greška klasifikacije

Greška klasifikacije u čvoru t :

$$\boxed{\text{Error}(t) = 1 - \max_i P(i | t)}$$

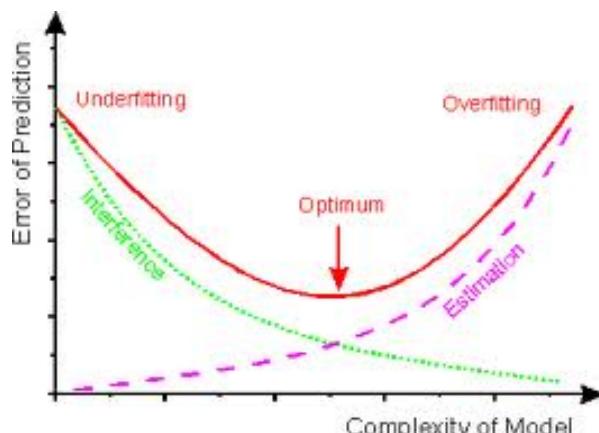
Maksimum ($1 - 1/n_c$), kada su elementi uniformno raspodeljeni po svim klasama, ukazuje na najmanje interesantne informacije, dok **minimum**

(0.0), kada svi elementi pripadaju istoj klasi, ukazuje na najinteresantnije informacije.

Podbacivanje (*underfitting*) i prebacivanje (*overfitting*)

Podbacivanje se odnosi na fenomen kod koga klasifikator na obučavajućem skupu ne daje zadate rezultate klasifikacije (slično kao najbolja aproksimacija). Sa druge strane, prebacivanje se odnosi na fenomen kada klasifikator „slepo“ sledi obučavajući skup pri klasifikaciji (slično interpolaciji). U osnovi, radi se o sposobnosti generalizacije na bazi obučavajućeg skupa:

- da se dobro klasificuje obučavajući skup
- da se dobro klasifikuju i novi primeri koji ne moraju da budu isti kao oni iz obučavajućeg skupa



Slika 4.36 Podbacivanje i prebacivanje

Sa slike 4.36 se može zaključiti da podbacivanje nastaje kada je model previše jednostavan a prebacivanje kada je model kompleksan dok su greška obučavanja i greška evaluacije prevelike u oba slučaja. Postavlja se pitanje kako tretirati ove pojave, pa u tom smislu se javljaju dva nova termina *pre-pruning* i *post-pruning*, kao načini rešavanja *overfitting* problema.

Pre-pruning

Cilj je smanjiti veličinu stabla dok se formira. To se postiže tako što se algoritam zaustavlja ranije. Neki od uslova zaustavljanja za čvor su

zaustaviti algoritam ako čvor ima manje od predefinisanog broj slogova ili ako dalje razdvajanje čvora ne poboljšava čistoću.

Post-pruning

Post-pruning funkcioniše na sledeći način. Skup podataka podelimo na obučavajući skup i test skup. Formiramo stablo bez orezivanja pomoću obučavajućeg skupa. Od listova prema korenu redom razmatramo granu po granu da li da je orežemo. Grana se orezuje ako smanjuje grešku klasifikacije. Odredimo grešku pomoću test skupa za stablo sa tom granom i za stablo bez te grane. Ako stablo bez grane daje manju, grešku grana se orezuje. Postoje dva načina za određivanje greške:

- Optimistična - klasična greška klasifikacije ($\text{broj_pogrešno_klasifikovanih_slogova} / \text{ukupan_broj_slogova}$)
- Pesimistična - na klasičnu grešku klasifikacije za svaki čvor dodaje se vrednost α . Pošto se vrednost α dodaje na grešku za svaki čvor stabla sa većim brojem čvorova su kažnjena tj. daju veću grešku.

4.4.1.2 Primena i proučavanje rezultata metode stabla odlučivanja

Da bi se primenila metoda stabla odlučivanja i da bi se dobio što precizniji rezultat pre svega je potrebno obraditi set podataka, koristeći *Chi-Square test* dolazimo do informacija koje od kategoričkih atributa su u korelaciji sa ciljanim atributom *odobren kredit*. Iz ranije analize je ustaljeno da numerički atributi *Iznos kredita*, *Godine starosti* i *Trajanje računa* jesu u korelaciji sa ciljanim atributom tako da se pridružuju skupu atributa dobijenih iz *Chi-Square test-a*. Sledeći korak za pripremu podataka jeste mapiranje vrednosti atributa u numeričke attribute, ovo se postiže pomoću funkcije *replace* u kombinaciji sa *get_dummies* iz *Pandas* biblioteke. Zatim se odvaja ciljani atribut *Odobren kredit* od prediktorskih varijabla. Sledeći korak je deljenje skupa podataka u set za treniranje modela i set za testiranje tačnosti modela. Izabrano je da set pude podeljen na sledeći način 70% podatak je trenirajući a 30% je testni deo. Ova podela se vrši pomoću *train_test_split* funkcije iz *sklearn.model_selection* biblioteke.

Jednom kada je set podataka podeljen na obučavajući i testni može da se krene u postupak kreiranja modela pomoću funkcije *tree.DecisionTreeClassifier* iz *sklearn* biblioteke. Modelu se prosleđuju podaci za obučavanje. Da bi dobili rezultat obučenog modela, prosleđuje

mu se testni skup podataka, zatim pomoću funkcije *metrics.f1_score* možemo da proverimo tačnost dobijenog modela.

Na slikama 4.37, 4.38, 4.39 i 4.40 prikazani su dobijeni rezltati za svaki od kriterijuma: *gini* i *entropy* respektivno.

```
Decision Tree with gini criterion
precision    recall   f1-score   support
          0       0.53      0.57      0.54      92
          1       0.80      0.77      0.79     208
accuracy                           0.71      300
macro avg       0.66      0.67      0.67      300
weighted avg    0.72      0.71      0.71      300

[[ 52  40]
 [ 47 161]]
Accuracy of the model on Testing Sample Data: 0.71

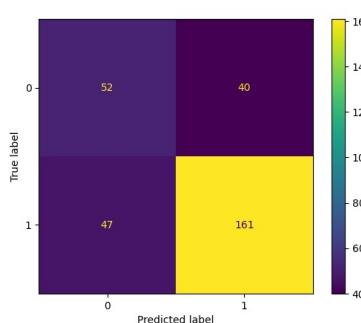
Accuracy values for 10-fold Cross Validation:
[0.7026648  0.68529412  0.72463235  0.72798574  0.65160986  0.68503119
 0.71751767  0.73734823  0.73089802  0.6565323 ]]

Final Average Accuracy of the model: 0.7
```

Slika 4.37 Rezultat metode stable odlučivanja – *gini criterion*

Sa slike 4.38 može da se vidi da je za *gini criterion* profit sledeći: $161*1962+40*(-5232)+47*(-1962)+52*0=28122$ DM

Što je profit od 93.74DM po klijentu.



Slika 4.38 confusion matrix *gini criterion*

Na slici 4.39 može da se vide parametri i rezultati metode stable odlučivanja za *entropy criterion*.

```
Decision Tree with entropy criterion
      precision    recall   f1-score   support
          0       0.46     0.42     0.44      92
          1       0.75     0.78     0.77     208

      accuracy                           0.67      300
      macro avg       0.61     0.60     0.60      300
  weighted avg       0.66     0.67     0.67      300

[[ 39  53]
 [ 46 162]]
Accuracy of the model on Testing Sample Data: 0.67

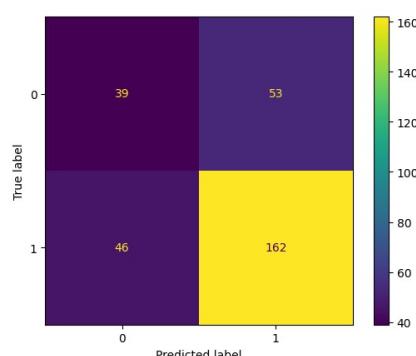
Accuracy values for 10-fold Cross Validation:
[0.68912656 0.674244  0.70857074 0.68       0.63632831 0.70541038
 0.6785546  0.74230949 0.67301587 0.6444608 ]]

Final Average Accuracy of the model: 0.68
```

Slika 4.39 Rezultat metode stable odlučivanja – *entropy criterion*

Zatim do računice za profit dolazimo pomoću podatka sa slike 4.40:
 $162*1962+53*(-5232)+46*(-1962)+39*0=-49704$ DM

Što je gubitak od 165.68 po klijentu.



Slika 4.40 confusion matrix *entropy criterion*

Iz dobijenih rezultata vidi se da je gubitak veći za *entropy criterion*. Da bi poboljšali tačnost rezultata vrši se optimizacija parametara modela. Neki od parametara koji utiču na preciznost su *max_depth*, *min_samples_split* i *min_samples_leaf*. Da bi došli do njihovih optimalnih vrednosti koristimo *GridsearchCV* objekat koji za dato stablo odlučivanja pronađe najbolje vrednosti za navedene parametre. Dobijeni rezultati za parametre su: '*min_samples_leaf*: 5, '*min_samples_split*: 2 i '*max_depth*': 5. Kreiranjem novog modela koristeći gore dobijene parameter dobijaju se sledeći rezultati koji su vidljivi na slikama 4.41 i 4.42.

```
Decision Tree with gini criterion
      precision    recall   f1-score   support
          0         0.59     0.48     0.53      92
          1         0.79     0.85     0.82     208

      accuracy                           0.74      300
      macro avg       0.69     0.66     0.67      300
  weighted avg       0.73     0.74     0.73      300

[[ 44  48]
 [ 31 177]]
Accuracy of the model on Testing Sample Data: 0.73

Accuracy values for 10-fold Cross Validation:
[0.75755221 0.60267559 0.71776316 0.71859838 0.68344988 0.70952831
 0.71019787 0.72866931 0.64375   0.69317015]

Final Average Accuracy of the model: 0.7
```

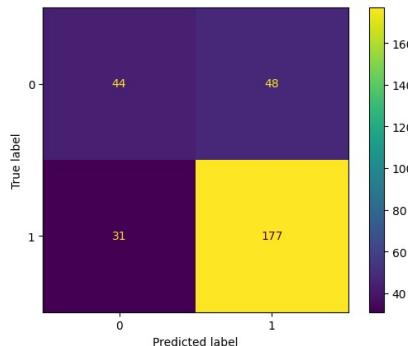
4.41 Rezultat modela stabla odlučivanja sa optimizovanim parametrima *gini criterion*

Rezultat modela stabla odlučivanja sa parametrima se minimalno popravio kao što se vidi na slici 4.41. Računica za profit se dobija pomoću slike 4.42:

$$177*1962+48*(-5232)+31*(-1962)+44*0=35316\text{DM}$$

Što je gubitak od 117.72 po klijentu.

Ovakav model je za 44 klijenata tačno, a za 31 pogrešno predviđao da im neće biti odobren zahtev za kredit banke. Takođe je tačno predviđao da će 177 dobiti pozitivan odgovor od banke dok je za 48 klijenta za isti uslov pogrešno predviđao.



Slika 4.42 confusion matrix *gini criterion*

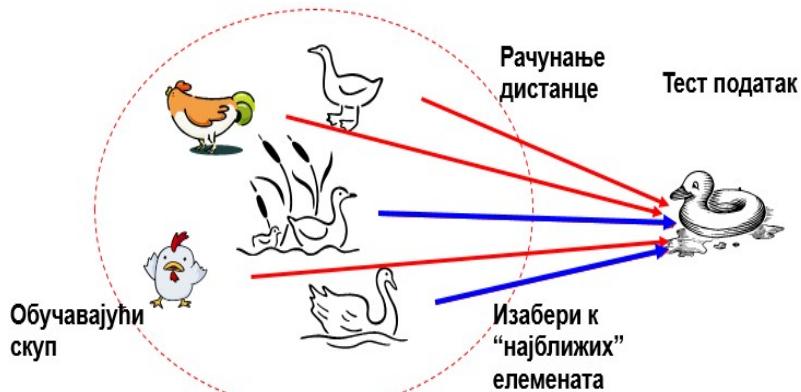
Na osnovu ovih podataka možemo izračunati i tačnost, preciznost i odziv.

- tačnost = $(44+177)/300 = 0,7366 = 73,66\%$
- preciznost (*false*) = $44/(44+31) = 0,5867 = 58,66\%$
- preciznost (*true*) = $177/(177+48) = 0,7867 = 78,67\%$
- odziv (*false*) = $44/(44+48) = 0,4782 = 47,82\%$
- odziv (*true*) = $177/(177+31) = 0,8509 = 85,09\%$

4.4.2 K-najbližih suseda

4.4.2.1 Teorijske osnove

Osnovna ideja ove vrste klasifikacije jeste da se na osnovu prepoznatih klasa, prema određenim specifičnostima definisanih u okviru obučavajućeg skupa, novi testni podatak može svrstati u neku od postojećih klasa na osnovu svojih karakteristika. Prosto rečeno, na osnovu ilustrativnog primera prikazanog na slici 4.43, ako nešto hoda kao patka, gače kao patka onda je to verovatno patka.



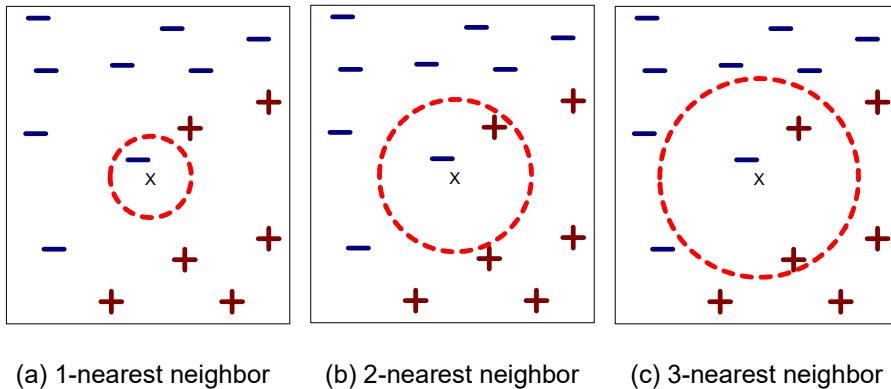
Slika 4.43 Ilustrativan primer klasifikacije k -najbližih suseda

Model za klasifikaciju se formira kada je potrebno klasifikovati nepoznati slog. Za formiranje modela potreban je pre svega obučavajući skup, skup skladištenih objekata, zatim metrika kojom se meri rastojanje među objektima – mera za izračunavanje „sličnosti“ između slogova, kao i vrednost za k koji predstavlja broj najbližih suseda pomoću kojih se klasificuje nepoznati slog. Za klasifikovanje novog objekta potrebno je ispratiti sledeće korake:

1. Sračunati rastojanje od objekata u obučavajućem skupu
2. Identifikovati k najbližih suseda
3. Koristiti oznake klasa najbližih suseda za određivanje klase nepoznatog objekta (npr. klasa kojoj pripada većina suseda)

Definicija: k -najbližih susedi objekta x su k tačaka koje su od tačke x udaljene manje od ostalih tačaka iz skupa.

Slika 4.44 prikazuje primer klasifikacije objekata putem K -najbližih suseda za različite odabране vrednosti k . Ako je k suviše malo, dolazi do osetljivosti na šum (*noise*) u podacima. Šum predstavljaju podaci u obučavajućem skupu kojima je iz različitih razloga pogrešno dodeljena klasa. Ako je k preveliko, skup suseda može da sadrži previše tačaka iz drugih klasa što dovodi do pogrešne klasifikacije.



Slika 4.44 K -najbližih suseda za $k=1$ (a), $k=2$ (b), $k=3$ (c)

Sračunavanje rastojanja između dve tačke vrši se putem Euklidskog rastojanja.

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

Određivanje klase nepoznatog objekta se vrši tako što se uzima klasa koju ima većina k najbližih suseda. S druge strane, može da se desi da sve klase imaju jednak broj suseda i tada se na slučajan način bira klasa za nepoznati objekat. Međutim, prilikom određivanja klase za novi slog treba voditi računa i o tome da li svi k susedi treba da imaju isti uticaj na proces, pa se zbog toga uvodi mera težinski faktor:

$$w = 1/d^2$$

Na osnovu ove mере omogućeno je da bliži susedi imaju veći uticaj prilikom odabira klase za nepoznati objekat od daljih suseda.

Takođe, u velikom broj slučajeva potrebno je skalirati (normalizovati) vrednosti atributa da bi sprečili da jedan atribut utiče više od drugih na meru sličnosti samo zbog jedinica u kojima se meri tj. intervala u kome se njegove vrednosti nalaze.

Prednosti metode k -najbližih suseda:

- Jednostavna implementacija
- Ne zahteva učenje
- Može da modeluje kompleksne odnose između podataka i klase
- Ne dolazi do gubitaka informacija prilikom formiranja modela
- Ima veoma dobre performanse ako postoji velika količina podataka u obučavajućem skupu

Mane metode *k*-najbližih suseda:

- Brzina – prilikom određivanja kategorije za instancu potrebno je izračunati udaljenost od nje od svih ostalih instanci
- Osetljivost na atribute koji nisu od značaja
- Nije uvek lako odrediti *k* i meru koja će se koristiti

4.4.2.2 Primena i proučavanje rezultata metode *k*-najbližih suseda

Proces klasifikacije primenom metode *k*-najbližih suseda se razlikuje od procesa klasifikacije primenom metoda stabla odlučivanja u tome što se umesto operatora *DecisionTreeClassifier* koristi operator *neighbors.KNeighborsClassifier* iz biblioteke *sklearn*. Ovaj operator generiše *k*-najbližih suseda model na osnovu ulaznog seta podataka.

Proces primene metode *k*-najbližih suseda se sastoji od obrade podataka pre samog obučavanja modela. Ovaj korak je isti kao i za metodu stabla odlučivanja tako da neće biti opisan ovde.

Jednom kada su skupovi podataka pripremljeni poziva se funkcija *neighbors.KNeighborsClassifier* kojoj se prosleđuje parametar *n_neighbors* da bi se izradio model

Na osnovu slike 4.79 može se zaključiti da je mera profita:
 $142*1962 + 35*(-5232) + 66*(-1962) + 57*0 = -34008$.

Zatim se mogu izračunati sledeće vrednosti:

- tačnost = $(57+142)/300 = 0.6633 = 66,33\%$
- preciznost (*false*) = $57/(57+66) = 0,4634 = 46,34\%$
- preciznost (*true*) = $142/(142+35) = 0,8023 = 80,23\%$
- odziv (*false*) = $57/(57+35) = 0,6196 = 61,96\%$
- odziv (*true*) = $142/(142+66) = 0,6827 = 68,27\%$

```
KNeighborsClassifier(n_neighbors=2)
      precision    recall   f1-score   support
          0       0.46     0.62     0.53      92
          1       0.80     0.68     0.74     208

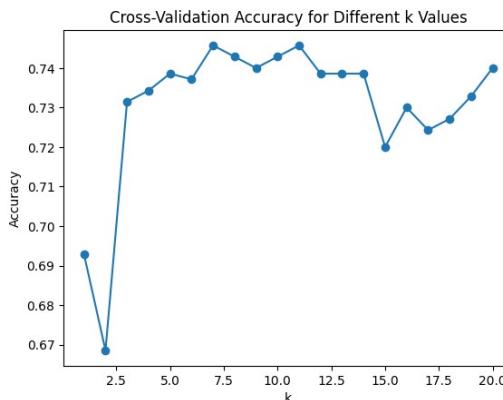
      accuracy                           0.66      300
     macro avg       0.63     0.65     0.63      300
weighted avg       0.70     0.66     0.67      300

[[ 57  35]
 [ 66 142]]
Accuracy of the model on Testing Sample Data: 0.67

Accuracy values for 10-fold Cross Validation:
[0.69303905 0.69249249 0.65186813 0.64527505 0.67385399 0.75059423
 0.57833333 0.65186813 0.66969697 0.655      ]
Final Average Accuracy of the model: 0.67
```

Slika 4.45 Performanse modela k -najbližih suseda za $k=2$

Kako bi se pronašao optimalni k , nakon učitanih podataka, koristi se operator *GridSearchCV* iz biblioteke *sklearn.model_selection*. Ovaj operator pronalazi optimalne vrednosti za selektovane parametre operatora koji se nalaze u okviru njegovih potprocesa. Na slici 4.46 može da se vidi rezultat potrage za optimalnim k .

Slika 4.46 Potraga za optimalnim k

Nakon sliči 4.46 može da se vidi da je za dati model optimalan broj za k jest 7. Kada se prosledi optimalni broj modelu dobija se rezultat prikazan na sliči 4.47.

```
Optimal k: 7
KNeighborsClassifier(n_neighbors=7)
      precision    recall   f1-score   support
          0       0.55     0.29     0.38      92
          1       0.74     0.89     0.81     208

      accuracy                           0.71      300
   macro avg       0.65     0.59     0.60      300
weighted avg       0.68     0.71     0.68      300

[[ 27  65]
 [ 22 186]]
Accuracy of the model on Testing Sample Data: 0.68

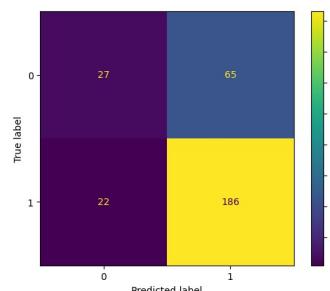
Accuracy values for 10-fold Cross Validation:
[0.69806452 0.65333333 0.71225296 0.72266667 0.71859838 0.72266667
 0.67434211 0.70602911 0.72867133 0.73089802]

Final Average Accuracy of the model: 0.71
```

Slika 4.47 Rezultat modela za optimalnu vrednost $k=7$

Sa Slike 4.47 možemo da vidimo da se model popravio sa novom vrednošću za k. Dakle sa podacima iz matrice sa slike 4.48, za najbolji kriterijum $k=7$ mera profit-a je:

$$186*1962 + 65*(-5232) + 22*(-1962) + 27*0 = -18312.$$



Slika 4.48 confusion matrix

Na osnovu rezultata, prikazanih na slici 4.79, možemo izračunati i tačnost, preciznost i odziv.

- tačnost = $(27+186)/300 = 0.71 = 71\%$
- preciznost (*false*) = $27/(27+22) = 0,5510 = 55,10\%$
- preciznost (*true*) = $186/(186+65) = 0,7410 = 74,10\%$
- odziv (*false*) = $27/(27+65) = 0,2935 = 29,35\%$
- odziv (*true*) = $186/(186+22) = 0,8942 = 89,42\%$

Izvršenom optimizacijom povećale su se vrednosti za tačnost kao i za preciznost za *false* i odziv za *true*. Prezinzost za *true* i odziv za *false* su se smanjile. Profit iako je ostao negativan, manji je nego za k=2.

4.4.3 Naivna Bayes-ova metoda

4.4.3.1 Teorijske osnove

Naivna *Bayes*-ova metoda predstavlja metodu za klasifikaciju koja se oslanja na teoriju verovatnoće tj. na *Bayes*-ovu teoremu. Veoma je korisna u slučajevima u kojima ne postoji deterministička veza između atributa sloga i klase. Koristi se u situacijama kada obučavajući skup sadrži greške ili kada postoje još neki dodatni, nama nepoznati faktori, koji utiču na klasifikaciju. Na primer iako znamo da zdrava ishrana i vežbanje smanjuju rizik od srčanog udara ne možemo samo na osnovu podataka o ishrani i vežbanju da tačno klasifikujemo ljude. Postoji mnogo drugih faktora koji utiču na rizik kao što su genetika, uzimanje alkohola itd.

Bayes-ova teorema glasi:

Neka su X i Y slučajne promenljive. $P(X=x, Y=y)$ je verovatnoća da će u isto vreme X uzeti vrednost x , a Y uzeti vrednost y . $P(X=x|Y=y)$ je verovatnoća da će X uzeti vrednost x ako već imamo informaciju da je Y uzeo vrednost y . To je uslovna verovatnoća. Veza između zajedničke i uslovne verovatnoće je sledeća:

$$P(X, Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$$

iz prethodne jednakosti izvodi se *Bayes*-ova teorema:

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

Dalje je potrebno objasniti kako funkcioniše naivni Bayes-ov klasifikator. Ako se prepostavi da ne postoji deterministička veza između atributa sloga i klase, onda se klasa i vrednosti atributa sloga mogu posmatrati kao slučajne promenljive. Neka X predstavlja attribute sloga, a Y klasu sloga. Verovatnoća da će slog sa vrednostima atributa X imati klasu Y je $P(Y|X)$. Tokom faze obučavanja klasifikatora potrebno je iz obučavajućeg skupa izračunati $P(Y|X)$ za sve moguće kombinacije atributa X i sve moguće klase Y . Kada se klasificuje novi slog koji ima kombinaciju atributa X' klasificuje se u klasu Y' koja maksimizuje verovatnoću $P(Y'|X')$.

Znači, kada je dat slog X i potrebno ga je klasifikovati, računamo:

$$P(Y | X) = \frac{P(Y) \prod_{i=1}^d P(X_i | Y)}{P(X)} \quad (1)$$

Za svaku klasu menja se samo Y dok je $P(X)$ konstanto tako da ne utiče na klasifikaciju. Znači slog X klasifikujemo u klasu Y koja maksimizuje

$$P(Y) \prod_{i=1}^d P(X_i | Y).$$

Za kategorički atribut X_i verovatnoća $P(X_i=x_i|Y=y)$ računa se kao deo (razlomak) slogova koji imaju klasu y i x_i vrednost atributa X_i .

Za kontinualne attribute postoje dva pristupa:

- Diskretizacija – deli se kontinualni interval na podintervale. Neka je data neka vrednost x_i kontinualnog atributa X_i i neka klasa y . Da bi se izračunala verovatnoća $P(X_i=x_i|Y=y)$ odredi se u kom se podintervalu nalazi x_i i onda se izračuna koliki je deo slogova klase y u tom podintervalu.
- Pretpostavimo da vrednosti tog kontinualnog atributa prate neku statističku distribuciju. Najčešće se uzima Gausova (normalna) distribucija. Parametre distribucije izračunavamo iz obučavajućeg skupa. Gausova distribucija zahteva dva parametra, srednju vrednost μ i standardnu devijaciju σ^2 . Za klasu y i vrednost xatributa X_i verovatnoća $P(X_i=x_i|Y=y)$ računa se po formuli:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Karakteristike naivnog Bayes-ovog klasifikatora:

- Robusni su na greške (dobijene prikupljanjem podataka) ili nedostajuće vrednosti atributa u obučavajućem skupu. Greške ne utiču mnogo na verovatnoće pošto su verovatnoće prosečne vrednosti, dok se nedostajući podaci jednostavno ignoriraju prilikom izračunavanja verovatnoća.
- Robusni su na nevažne attribute. Ako je atribut nebitan onda su njegove vrednosti skoro uniformno distribuirane po klasama. Dakle verovatnoće tog atributa jednakо utiču na uslovne verovatnoće za sve klase tj. atribut nema uticaja na klasifikaciju.
- Atributi koji su u jakoj korelaciji mogu da degradiraju performanse klasifikatora pošto tada prepostavka o nezavisnosti atributa ne važi.

4.4.3.2 Primena i proučavanje rezultata naivne Bayes-ove metode

Proces klasifikacije primenom Bayes-ove metode suseda se razlikuje od procesa klasifikacije primenom metoda stabla odlučivanja u tome što se umesto operatora *DecisionTreeClassifier* koristi operator *neighbors.GaussianNB* iz biblioteke *sklearn*. Ovaj operator generiše k -najbližih suseda model na osnovu ulaznog seta podataka.

Proces primene metode k -najbližih suseda se sastoji od obrade podataka pre samog obučavanja modela. Ovaj korak je isti kao i za metodu stabla odlučivanja tako da neće biti opisan ovde.

Rezultati Bayes-ove metode su prikazani na slici 4.49. A na osnovu podataka sa slike 4.50 može se doći do sledećih zaključaka:

Da je mera profita:

$$149*1962 + 33*(-5232) + 59*(-1962) + 59*0 = 3924.$$

Zatim se mogu izračunati sledeće vrednosti:

- tačnost = $(59+149)/300 = 0.69333 = 70\%$
- preciznost (*false*) = $59/(59+59) = 0,5 = 50\%$

- preciznost (*true*) = $149/(149+33) = 0,8187 = 81,87\%$
- odziv (*false*) = $59/(59+33) = 0,6413 = 64,13\%$
- odziv (*true*) = $149/(149+59) = 0,7163 = 71,63\%$

```
GaussianNB()
    precision    recall   f1-score   support
          0         0.50      0.64      0.56       92
          1         0.82      0.72      0.76      208

    accuracy                           0.69      300
   macro avg       0.66      0.68      0.66      300
weighted avg       0.72      0.69      0.70      300

[[ 59  33]
 [ 59 149]]
Accuracy of the model on Testing Sample Data: 0.7

Accuracy values for 10-fold Cross Validation:
[0.72923077 0.71751767 0.72        0.71578348 0.74230949 0.77295794
 0.65405405 0.73347237 0.75321515 0.76553596]

Final Average Accuracy of the model: 0.73
```

Slika 4.49 Performanse modela naivne Bayes-ove metode

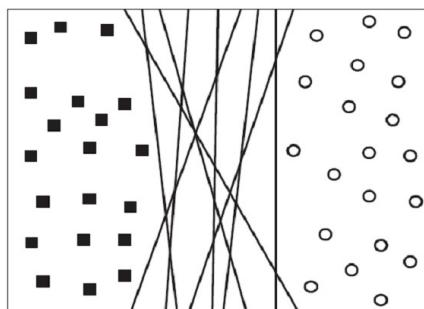
Dobijena je zarada a ne trošak, pa će se model dalje koristi.

4.4.4 Mašine potpornog vektora

4.4.4.1 Teorijske osnove

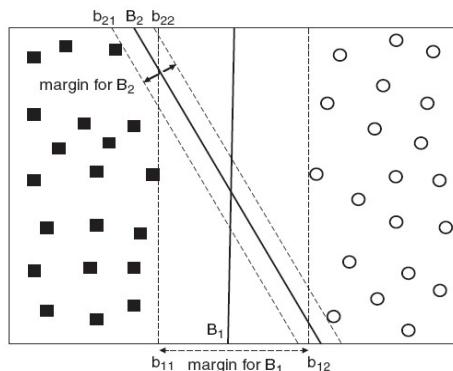
Mašine potpornog vektora (*Support vector machines*, SVM) predstavljaju skup povezanih nadgledanih metoda koje analiziraju podatke i uočavaju obrazce. Ova metoda klasifikacije ima veliku primenu jer pruža jako dobre rezultate naročito na podacima koji imaju veliki broj atributa. SVM je linearni klasifikator koji pronalazi hiperravan koja razdvaja dve klase. Hiperravan koju pronalazi SVM je hiperravan sa najvećom marginom separacije (*maximum margin hyperplane*). Pojam hiperravan sa najvećom marginom separacije ilustrovaće se na sledećim primerima. Predpostavlja se

da postoje dve klase kao na slici 4.50. Prva klasa je označena kvadratima dok je druga klasa označena krugovima. Klase su linearно separabilne.



Slika 4.50 Hiperravnji koje razdvajaju dve klase

Na slici 4.50 se vidi da postoji beskonačno mnogo hiperravnih koje razdvajaju dve klase. Vidi se da svaka od hiperravnih savršeno klasificiše obuhavajući skup (date tačke). Pitanje je koja od njih će najbolje klasifikovati nove tačke tj. test primere. Najbolja je ona koja ima najveću marginu separacije kao na slici 4.51.



Slika 4.51 Margina separacije

Hiperravnji b_{11} i b_{12} dobijaju se tako što se pomeraju ravni B_1 paralelno dok ne dodirnu jednu od tačaka. Margina separacije je rastojanje između hiperravnih b_{11} i b_{12} . Vidimo da je margina veća za hiperravan B_1 . Klasifikatori koji imaju veliku marginu separacije daju bolje rezultate na novim primerima tj. imaju bolju generalizaciju. Vidi se da mala pomeranja hiperravnih B_2 mogu mnogo da utiču na klasifikaciju, što ne važi za B_1 .

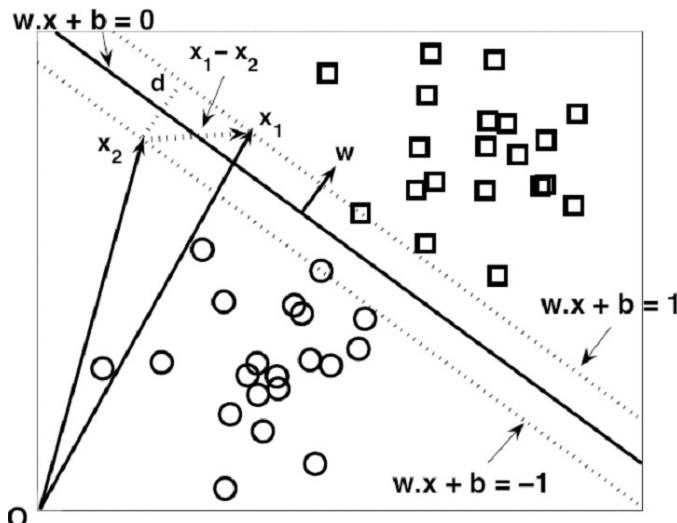
Linearni SVM

Linearni SVM klasifikator funkcioniše tako što za zadati obučavajući skup (dve klase) pronalazi hiperravan separacije sa maksimalnom marginom separacije. Pretpostavlja se da postoji obučavajući skup sa N uzoraka koji su klasifikovani u dve klase. Svaki uzorak reprezentovan je kao (x_i, y_i) gde je $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ skup atributa za i -ti uzorak (slog), a y_i klasa koja mu odgovara. Klase će se označiti sa $y_i \in \{-1, +1\}$. Hiperravan koja razdvaja klase (pomoću koje se vrši klasifikacija) zove se granica odluke (*decision boundary*) i označava se jednačinom:

$$w^*x + b = 0$$

gde su w i b parametri modela.

Na slici 4.52 date su dve klase tačaka označene kvadratima i krugovima. Takođe se vidi i hiperravan koja razdvaja te dve klase.



Slika 4.52 Linearni SVM

Na primer ako su tačke x_a i x_b na granici odluke onda za njih važi:

$$w^*x_a + b = 0$$

$$w^*x_b + b = 0$$

Ako se oduzme druga jednačina od prve dobija se:

$$w^*(x_a - x_b) + b = 0$$

gde je $x_a - x_b$ vektor paralelan granici odluke i usmeren od x_a ka x_b . Pošto je skalarni proizvod ovog i vektora w jednak 0, vektor w mora biti normalan na granicu odluke kao na slici 4.52.

Za svaki kvadrat x_s koji je lociran iznad granice odluke može se pokazati da važi:

$$w^*x_s + b = k$$

gde je $k > 0$. Isto tako za svaki krug ispod granice odluke važi:

$$w^*x_s + b = k'$$

gde je $k' < 0$. Ako se označe kvadri sa 1 klasom a krugove sa -1 klasom onda se svaka nova tačka z može klasifikovati na sledeći način:

$$y = \begin{cases} +1, & \text{ako je } w \cdot z + b > 0 \\ -1, & \text{ako je } w \cdot z + b < 0 \end{cases}$$

Granica odluke može se skaliranjem parametara w i b dovesti u takav oblik da za hiperravnji b_{i1} i b_{i2} (hiperravnji koje su paralelne granici odluke i kojima pripadaju tačke najbliže granici odluke) važi:

$$b_{i1} : w^*x + b = 1$$

$$b_{i2} : w^*x + b = -1$$

Margina granice odluke je rastojanje između ovih hiperravnji. Može se pokazati da je to rastojanje:

$$d = \frac{2}{\|w\|}$$

Obučavanje linearog SVM sastoji se od izračunavanja vektora w i b tako da važi:

$$w \cdot x_i + b \geq 1 \quad \text{ako je } y_i = 1$$

$$w \cdot x_i + b \leq -1 \quad \text{ako je } y_i = -1$$

gde je $i=1,\dots,N$, a N je broj uzoraka u obučavajućem skupu.

Ovi uslovi znače da svi uzorci koji imaju klasu 1 moraju da se nalaze na ili iznad hiperravnji $w^*x + b = 1$, a svi uzorci koji imaju klasu -1 moraju da se nalaze na ili ispod hiperravnji $w^*x + b = -1$.

Pored ovih uslova potrebno je odrediti w i b tako da margina bude maksimalna. Ovaj problem se rešava metodom *Lagranžovih množilaca*. Upotreboom ove metode dobija se nova funkcija koja se minimizuje:

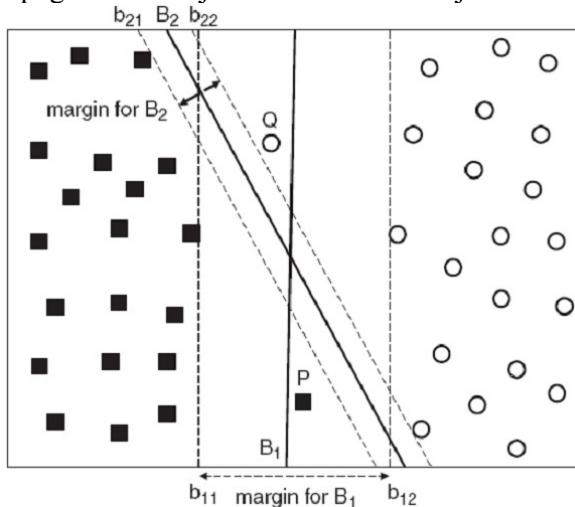
$$Lp = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \lambda_i (y_i(w \cdot x_i + b) - 1)$$

gde se parametri λ_i zovu *Lagranžovi množioci*.

Vektori x_i za koje važi $y_i^*(w^*x_i + b) = 1$ nazivaju se **vektori potpore** (*support vectors*). To su vektori koji se nalaze na hiperravnima b_{i1} i b_{i2} (hiperravni koje su paralelne granici odluke i kojima pripadaju tačke najbliže granici odluke).

Linearni SVM – neseparabilan slučaj

Linearno separabilni podaci su idealan slučaj. U realnosti postoje greške u podacima npr. pogrešno dodeljena klasa itd. Primer je slike 4.53.

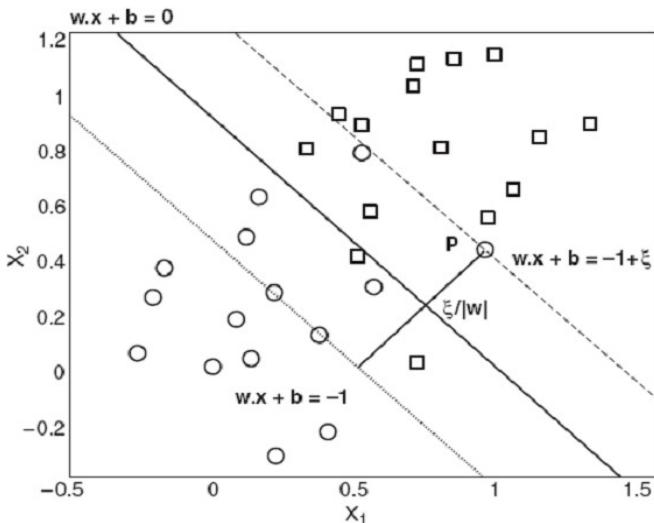


Slika 4.53 Neseparabilan slučaj za linearni SVM

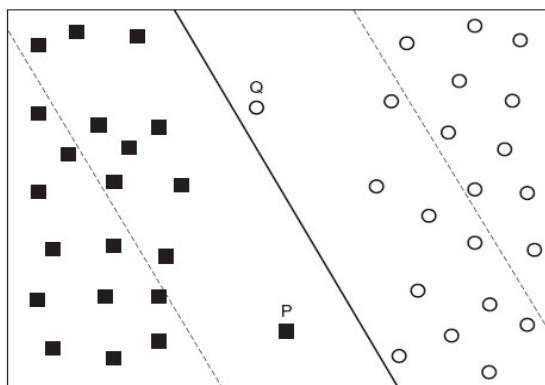
U ovom slučaju postoje dva nova primera P i Q . Vidi se da hiperplan B_1 tj. SVM klasifikator ne bi dobro klasifikovao nove primere. Iako bi ravan B_2 tačno klasifikovala P i Q , to ne znači da je treba uzeti za klasifikator. Ne treba odustati od hiperplanova za klasifikator jer B_1 ima veću marginu, a P i Q mogu samo biti šum (greške u podacima). Problem je u tome što uz prisustvo P i Q u obučavajućem skupu ne bi se mogao izračunati B_1 jer P i Q ne bi zadovoljili uslove iz prethodnog odeljka. Zato je potrebno relaksirati te uslove. Tako se dobija granica odluke koja ima relaksiranu marginu (*soft margin*). Novi uslovi imaju sledeći oblik:

$$\begin{aligned} w \cdot x_i + b &\geq 1 - \varepsilon_i && \text{ako je } y_i = 1 \\ w \cdot x_i + b &\leq -1 + \varepsilon_i && \text{ako je } y_i = -1 \end{aligned}$$

gde je $\varepsilon_i > 0$ $i=1,\dots,N$ (broj uzoraka). Vrednosti ε_i zovu se promenljive relaksacije (*slack variables*) ili fiktivne promenljive – slika 4.54. Sada smo promenili uslove za obučavanje linearног SVM. Ako funkcija ostane ista kao u prethodnom odeljku moguće je da će metod pronaći SVM sa jako velikim vrednostima ε_i tj. sa širokom marginom ali koji pogrešno klasificiše veliki broj tačaka – slika 4.55. Da bi se taj problem rešio uvodi se parametar C koji će kažnjavati granice odluke sa velikim vrednostima ε_i .



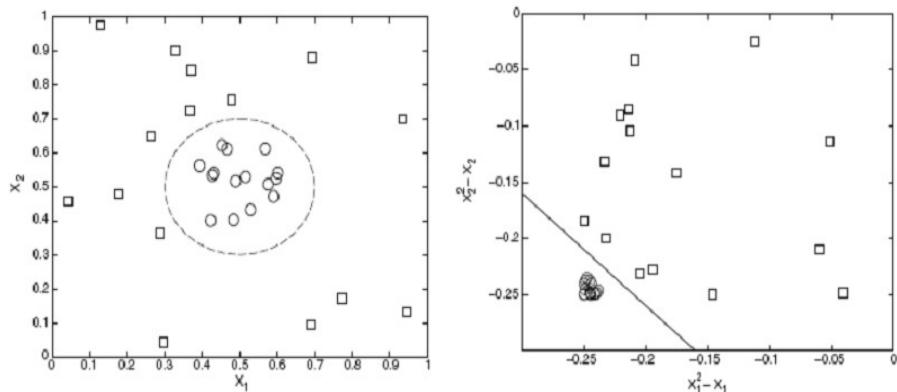
Slika 4.54 SVM sa relaksiranim marginom



Slika 4.55 SVM sa velikim vrednostima ε_i

Nelinearni SVM

Prethodno opisani SVM klasifikatori formiraju linearu granicu odluke između dve klase. Postoje i SVM klasifikatori sa nelinearnom granicom odluke. Kao primer, predpostavlja se da je dat obučavajući skup u kome su dve klase razdvojene nelinearnom granicom. SVM klasifikator za ovaj skup formira se tako što se transformišu tačke iz obučavajućeg skupa u prostor u kome će te tačke biti linearno separabilne. U tom novom prostoru pronalazimo linearni SVM pomoću postupka koji je već objašnjen. Slika 4.56 prikazuje dat prostor i prostor dobijen trasformacijom.



Slika 4.56 Granice odluke u datom prostoru i u prostoru dobijenom trasformacijom

Funkcija koja služi za transformaciju označava se sa $\Phi(x)$. Prilikom implementacije nelinearnih SVM dolazi do više problema. Kako formirati $\Phi(x)$ se sigurnošću da će tačke u novom prostoru biti linearno separabilne? Takođe ako je novi prostor jako velike dimenzionalnosti izračunavanje skalarnog proizvoda $\Phi(x_i) * \Phi(x_j)$ je vremenski i računski veoma zahtevno.

Navedeni problemi rešavaju se upotrebom *funkcija jezgara K (kernel function)*. Nije potrebno eksplicitno poznavanje funkcije $\Phi(x)$ već samo znanje koliko je skalarni proizvod $\Phi(x_i) * \Phi(x_j)$ za sve tačke iz originalnog prostora (obučavajućeg skupa). Znači ako se može nekako izraziti skalarni proizvod kao funkcija skalarnog proizvoda $x_i * x_j$ ne mora se znati funkcija $\Phi(x)$. Funkcije koje daju vezu između skalarnog proizvoda $x_i * x_j$ i $\Phi(x_i) * \Phi(x_j)$ su *funkcija jezgara*. Za funkciju jezgara K važi:

$$K(x_i, x_j) = \Phi(x_i) * \Phi(x_j)$$

Neke od često korišćenih *funkcija jezgra* su polinomijalne $K(x,y)=(x^*y)^d$ i $K(x,y)=(x^*y + 1)^d$.

4.4.4.2 Primena i proučavanje rezultata metode mašine potpornog učenja

Za proces klasifikacije primenom metode mašine potpornog učenja se koristi operator *SVC* iz biblioteke *sklearn.svm*. Ovaj operator implementira metodu mašine potpornog učenja i kao rezultat daje izgenerisan model.

Proces primene metode mašine potpornog učenja se sastoji od obrade podataka pre samog obučavanja modela. Ovaj korak je isti kao i za metodu stabla odlučivanja tako da neće biti opisan ovde. Dodatno se korsiti *StandardScaler* iz biblioteke *sklearn.preprocessing* za dodatnu obradu skupa podataka pre samog kreiranja modela.

Na osnovu slike 4.57 može se zaključiti da je mera profita:
 $196*1962 + 77*(-5232) + 12*(-1962) + 15*0 = -41856$.

Zatim se mogu izračunati sledeće vrednosti:

- tačnost = $(196+15)/300 = 0.7033 = 70,33\%$
- preciznost (*false*) = $15/(15+77) = 0,1630 = 16,30\%$
- preciznost (*true*) = $196/(196+77) = 0,7179 = 71,79\%$
- odziv (*false*) = $15/(15+77) = 0,1630 = 16,30\%$
- odziv (*true*) = $196/(196+12) = 0,9423 = 94,23\%$

SVC(C=5, gamma=0.1)				
	precision	recall	f1-score	support
0	0.56	0.16	0.25	92
1	0.72	0.94	0.81	208
accuracy			0.70	300
macro avg	0.64	0.55	0.53	300
weighted avg	0.67	0.70	0.64	300
 [[15 77] [12 196]]				
Accuracy of the model on Testing Sample Data: 0.64				

Slika 4.57 Performanse modela metode mašine potpornog učenja

Kao mogući način poboljšanja performansi modela izvršava se operator *GridSearchCV* kome se prosleđuje model, i niz mogućih parametara. Ovaj operator se koristi za pronalaženje optimalnih vrednosti parametara koji daju najbolje rezultate.

Performanse modela klasifikacije primenom metode mašine potpornog učenja sa najboljom parametrima prikazane su na slici 4.58.

<code>SVC(C=10, gamma=0.01)</code>				
	<code>precision</code>	<code>recall</code>	<code>f1-score</code>	<code>support</code>
0	0.67	0.46	0.54	92
1	0.79	0.90	0.84	208
<code>accuracy</code>			0.76	300
<code>macro avg</code>	0.73	0.68	0.69	300
<code>weighted avg</code>	0.75	0.76	0.75	300
 [[42 50] [21 187]]				
<code>Accuracy of the model on Testing Sample Data: 0.75</code>				

Slika 4.58 Performanse modela metode mašine potpornog učenja

Zaključuje se da nakon primene selekcije osobina mera profita iznosi:
 $187*1962 + 50*(-5232) + 21*(-1962) + 42*0 = 54282$

Na osnovu rezultata prikazanih na slici 4.88 mogu se izračunati vrednosti za:

- tačnost = $(187+42)/300 = 0.7466 = 75\%$
- preciznost (*false*) = $42/(42+21) = 0,6667 = 66,67\%$
- preciznost (*true*) = $187/(187+50) = 0,7890 = 78,9\%$
- odziv (*false*) = $42/(42+50) = 0.4565 = 45,65\%$
- odziv (*true*) = $187/(187+21) = 0.8990 = 89,9\%$

Dakle, nakon primene optimalnih parametara tačnost je porasla za 5%, odziv za *false* kao i preciznost i za *true* i za *false* su povećani a odziv za *true* se smanjio.

4.5 Evaluacija

Zadatak ovog seminarског rada jeste da reši problem dodele rejtinga klijentima banke. U prvoj fazi definisana je matrica troškova za ovaj problem - tabela 2.1. Na osnovu prirode problema izvodi se zaključak da postoje dva osnovna modela kao moguća rešenja:

1. Odobriti zahtev svim klijentima
2. Ne odobriti zahtev nijednom klijentu.

Potrebno je postaviti ova dva modela u kontekst troškova i profita i uzeti onaj koji je bolji kao model sa polaznim performansama.

Da bi se izračunale polazne performanse koristi se test skup na osnovu koga se računaju ukupne cene za prethodna dva modela. U ovom slučaju skup sadrži:

- 180 klijenta kojima je rejting dobar i
- 70 klijenata kojima je rejting loš.

Na osnovu prethodnog i matrice troška formiramo tabelu matrice troška - tabela 4.13.

Model	TN	TP	FN	FP	Ukupna cena
Ne odobriti nikome	70		180		-358560 dM (-1434,24 DM po klijentu)
Odobriti svima		180		70	-65400 DM (-261,6 DM po klijentu)

Tabela 4.13 Tabela troškova

Na osnovu tabele 4.13 vidi se da je bolji model "odobriti kredit svima" jer takav model obezbeđuje manji gubitak od 261,6DM po klijentu. Znači pri formiranju modela cilj je da se premaže performanse od 261,6DM gubitka po klijentu.

Pri evaluaciji modela stabla odlučivanja, uz pomoć funkcije `accuracy_score` iz biblioteke `sklearn.metrics`, dobijaju se performanse prikazane na slici 4.59:

Decision Tree with gini criterion				
	precision	recall	f1-score	support
0	0.58	0.63	0.60	70
1	0.85	0.82	0.84	180
accuracy			0.77	250
macro avg	0.71	0.73	0.72	250
weighted avg	0.77	0.77	0.77	250
 [[44 26] [32 148]]				
Accuracy of the model on Testing Sample Data: 0.77				

Slika 4.59 Rezultati evaluacije modela stabla odlučivanja

Sa slike 4.59 dovodi se do zaključka da je profit:

$148*1962+26*(-5232)+32*(1962)+44*0=91560\text{DM}$ što dovodi do cifre od 366,24DM zarade po klijentu. Kako je ova vrednost veća od zarade definisane osnovnim modelom, može se reći da je model stabla odlučivanja bolji od početnog modela.

Za *true* kategoriju preciznost iznosi 85% a odziv 82%. Dakle performanse modela se mogu smatrati za dobre jer su vrednosti preciznosti i odziva za kategoriju koja je od interesa visoke.

U osnovu na performanse onsovnog modela, model k-najbližih suseda daje bolje performanse odnosno veći profit i to u iznosu od 3,99DM po klijentu –slika 4.60. Međutim, iako je bolji od početnog modela ovaj model daje lošije performanse u odnosu na model stabla odlučivanja.

KNeighborsClassifier(n_neighbors=7)				
	precision	recall	f1-score	support
0	0.76	0.46	0.57	70
1	0.82	0.94	0.88	180
accuracy			0.81	250
macro avg	0.79	0.70	0.72	250
weighted avg	0.80	0.81	0.79	250
[[32 38] [10 170]]				
Accuracy of the model on Testing Sample Data: 0.79				

Slika 4.60 Performanse modela k-najbližih suseda

Sa slike 4.60 vidimo da je profit:

$$170*1962+38*(-5232)+10*(-1962)=115104 \text{ DM} \text{ što iznosi } 460,42 \text{ DM po klijentu.}$$

Za *true* kategoriju preciznost iznosi 82% a odziv 94%. Dakle, iako je tačnost celog klasifikatora velika 79% performanse modela se mogu smatrati za dobre jer su vrednosti preciznosti i odziva za kategoriju koja je od interesa visoke.

Za Bayes-ov model rezultati su prikazani na slici 4.61 dala je sledeći rezultat za profit:

$$136*1926+24*(-5232)+44*(-1962)+46*0=50040 \text{ DM} \text{ što tačno iznosi profit od } 200,16 \text{ DM.}$$

GaussianNB()				
	precision	recall	f1-score	support
0	0.51	0.66	0.57	70
1	0.85	0.76	0.80	180
accuracy			0.73	250
macro avg	0.68	0.71	0.69	250
weighted avg	0.76	0.73	0.74	250
[[46 24] [44 136]]				
Accuracy of the model on Testing Sample Data: 0.74				

Slika 4.61 Performanse Bayes-ovog modela

Za *true* kategoriju preciznost iznosi 85% a odziv 76%. Dakle, iako je tačnost celog klasifikatora velika 74% performanse modela se mogu smatrati za dobre jer su vrednosti preciznosti i odziva za kategoriju koja je od interesa visoke. U odnosu na performanse osnovnog modela, model naivnog *Bayes-a* daje bolje performanse odnosno veći profit po klijentu.

Konačno pocesom evaluacije modela formiranog pomoću SVM metode dolazi se do sledećih rezultata prikazanih na slici 4.62:

SVC(C=10, gamma=0.01)					
	precision	recall	f1-score	support	
0	0.83	0.76	0.79	70	
1	0.91	0.94	0.92	180	
accuracy			0.89	250	
macro avg	0.87	0.85	0.86	250	
weighted avg	0.89	0.89	0.89	250	
[[53 17] [11 169]]					
Accuracy of the model on Testing Sample Data: 0.89					

Slika 4.62 Performanse SVM modela

Kako se može videti sa slike 4.62 profit koji se može izračunati naspram podataka je sledeći:

$$169*1962+17*(-5232)+11*(-1962)+53*0=221052 \text{ DM} \text{ što iznosi } 884.21 \text{ DM po klijentu.}$$

Za *true* kategoriju preciznost iznosi 91% a odziv 94%. Dakle, performanse modela se mogu smatrati za dobre jer su vrednosti preciznosti i odziva za kategoriju koja je od interesa visoke. U odnosu na performanse osnovnog modela, SVM model daje bolje performanse odnosno veći profit po klijentu.

5. ZAKLJUČAK

U ovom radu dat je prikaz rešenja problema odobravanja zahteva za kredit banke u Nemačkoj. Cilj je odobriti zahtev onim klijentima koji će najverovatnije uspešno vratiti novac koji su zahtevali a to dalje implicira povećanje profita banke. Najpre je izvršeno istraživanje dostupnih podataka na osnovu kojih je kasnije izvršeno formiranje modela a čiji zadatak je da sa što većim procentom uspeha određuju pogodne klijente kojima treba odobriti zahtev za kredit.

Za istraživanje podataka je iskorišćen postupak eksplorativne analize. Kao rezultat ove analize izvučeni su zaključci da klijentima koji su stariji i koji imaju račun u banci manje od 30 meseci, a koji potražuju kredit na manji iznos novca u većini slučaja bude odobren kredit. Što dovodi do logičnog zaključka da klijentima koji su stariji i potražuju kredit na manje iznose bude u više slučaja odobren nego klijentima koji pripadaju mlađoj starosnoj grupi i koji potražuju kredite na veće iznose od 7500DM.

Za formiranje modela korišćene su metode klasifikacije: stablo odlučivanja, *Bayes-ova*, *k*-najbližih suseda i maštine potpornog vektora. Iako su svi modeli dali pozitivne rezultate - postoji zarada, najbolje rezultate je dala Metoda maština potpornog vektora (884.12DM) po aplikantu. Metoda *k*-najbližih suseda takođe je dala sasvim zadovoljavajuće rezultate sa profitom od 460,42DM po aplikantu.

LITERATURA

- [1] "Metode istraživanja podataka u proceni rizika u bankarstvu",
<http://elibrary.matf.bg.ac.rs/bitstream/handle/123456789/1889/Metode%20Istr.pdf?sequence=1>
- [2] <http://www.informatika.ftn.ns.ac.yu/SIAP/>

