



Kvalitet podataka

Prikupljanje i predobrada podataka za
Mašinsko učenje

Mentor: doc. dr Aleksandar Stanimirović

Student: Nastasija Stanković 1622



Sadržaj

01

Uvod

02

**Mere kvaliteta
podataka**

03

**Raspodela
podataka**

04

**Mere centralne
tendencije**

05

Korelacija

06

Varijansa

07

**Praktični deo
rada**

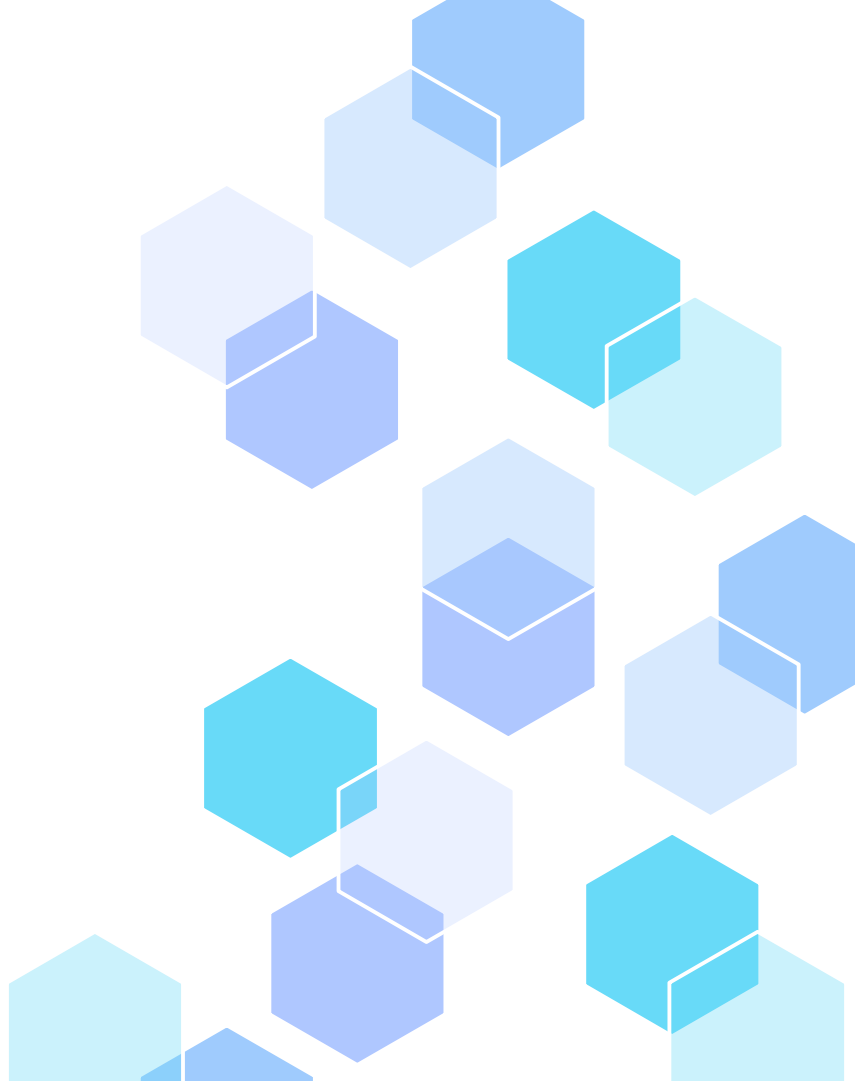
08

Zaključak



01

Uvod





Kvalitet podataka je ključni faktor u savremenom poslovanju i istraživanjima, ne samo tehnički, već i strateški. On utiče na svaki aspekt procesa, od optimizacije poslovnih procedura do donošenja odluka i zadovoljstva korisnika.



Visokokvalitetni podaci su oni koji ispunjavaju svoju svrhu i omogućavaju analizu performansi algoritama. Ocenjivanje kvaliteta svakog uzorka je suštinski korak u donošenju zaključaka o kvalitetu podataka.



Tokom preprocesiranja podataka za mašinsko učenje, ključno je provesti niz koraka kako bi se osiguralo da podaci budu pripremljeni za obučavanje modela.



02

Mere kvaliteta podataka

Ključne mere kvaliteta podataka uključuju:

○ **Tačnost**

○ **Kompletnost**

○ **Konzistentnost**

○ **Koherentnost**

○ **Aktuelnost**

○ **Relevantnost**

○ **Jasnoća**

Tačnost

Predstavlja meru kvaliteta podataka koja definiše vrednost odstupanja podataka od stvarne ili ispravne vrednosti originalnog podatka.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

A=odnos broja tačno predviđenih instanci u odnosu na ukupan broj instanci u testnom skupu podataka.



Kompletnost

Kompletnost podataka se odnosi na broj popunjenih vrednosti unutar skupa podataka što doprinosi celovitosti ili sveobuhvatnosti skupa podataka. Kada podaci nisu potpuni, to otežava analitičke procese i može dovesti do zaključaka koji nisu zasnovani na svim relevantnim informacijama.

Da bi se nepotpuni podaci popunili mogu se primeniti različite tehnike obrade podataka. Ove tehnike uključuju *interpolaciju* ili *imputaciju*, gde se nedostajući podaci popunjavaju procenjenim vrednostima na osnovu dostupnih podataka, ili eliminaciju, gde se redovi ili kolone sa nedostajućim vrednostima uklanjaju iz analize.

```
data.isna().sum()
```

Area	0
Perimeter	0
MajorAxisLength	0
MinorAxisLength	0
AspectRatio	0
Eccentricity	0

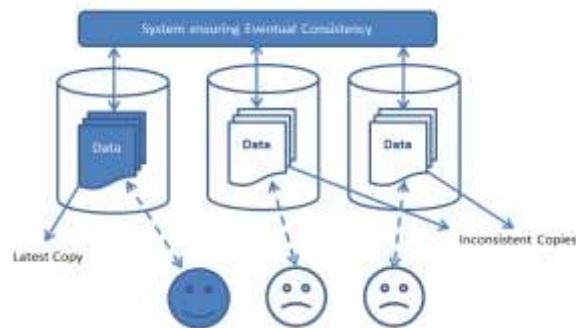
Primer provere nedostajućih podataka

```
imputer = KNNImputer(n_neighbors=1, missing_values=np.nan)  
data2 = pd.DataFrame(imputer.fit_transform(data2), columns=data2.columns)
```

Primer popunjavanja nedostajućih podataka

Konzistentnost

Ovaj pojam opisuje stepen u kojem podaci ostaju uniformni, dosledni i bez konflikata širom različitih sistema, aplikacija i baza podataka u kojima se koriste.



Koherentnost

Koherentnost podataka odnosi se na stepen u kojem su podaci logički usklađeni, dosledni i precizni kroz različite setove podataka unutar organizacije. Koherentni podaci treba da održavaju jedinstvenu strukturu, format i definiciju, omogućavajući da se podaci iz različitih izvora mogu lako kombinovati, uporediti i analizirati bez konflikta ili nejasnoća.

Aktuelnost

Aktuelnost podataka predstavlja meru kvaliteta podataka koja se odnosi na dostupnost i ažuriranost podataka u određenom vremenskom trenutku.



Relevantnost

Relevantnost podataka je dimenzija koja određuje koliko su informacije sadržane u skupu podataka značajne za specifične ciljeve analize ili odlučivanja. Stepenu u kojem podaci odgovaraju i pomažu u ispunjavanju konkretnih informacionih potreba direktno utiče na njihovu korisnost i vrednost.



Jasnoća

Jasnoća podataka omogućava korisnicima da razumeju podatke bez zabune ili pogrešnog tumačenja. Ona se odnosi na lakoću sa kojom se podaci mogu interpretirati, i to ne samo od strane analitičara, već i od strane svih koji se oslanjaju na te podatke za donošenje odluka.



Jedinstvenost

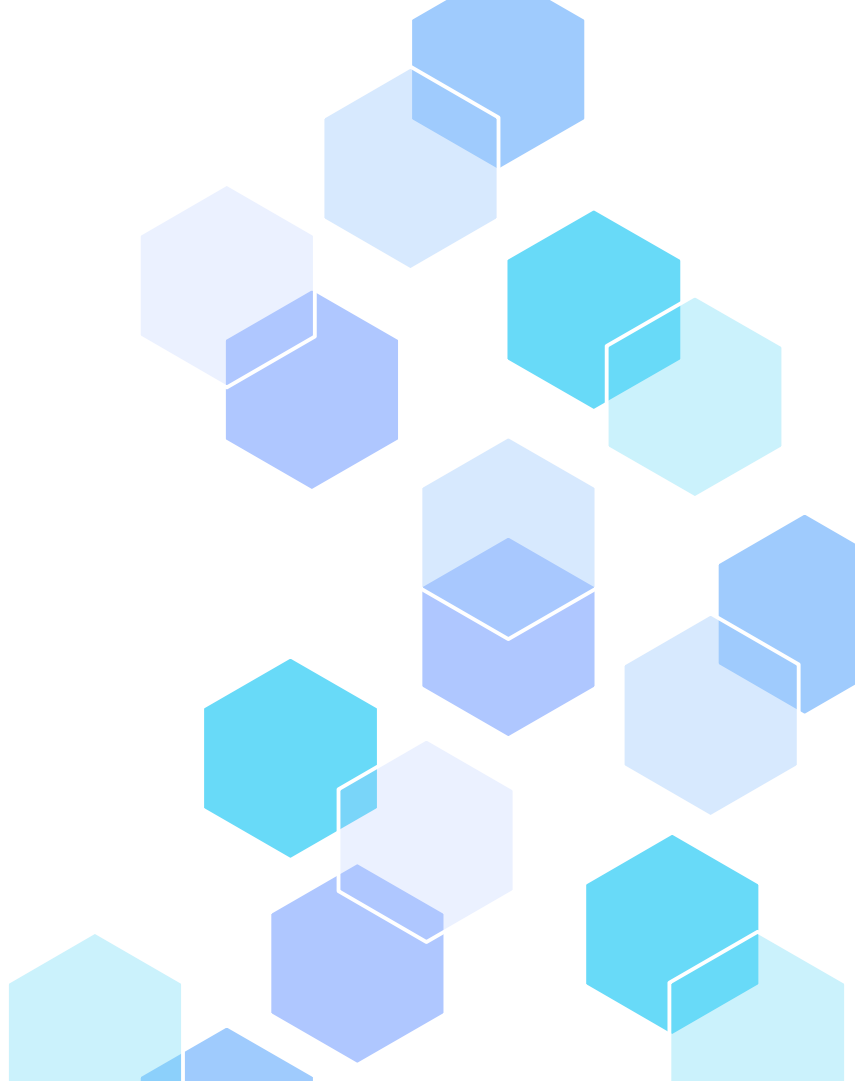
Jedinstvenost predstavlja osobinu podataka koja se odnosi na svaku pojedinačnu stavku u podacima gde se većim kvalitetom podrazumeva i veća količina jedinstvenih podataka.

Jedinstvenost jeste suprotnost multiplikativnosti podataka u tabeli podataka. Multiplikativnost dovodi do povećanja obima skupa podataka bez unošenja varijabilnosti.



03

Raspodela podataka



Raspodela podataka

Raspodela podataka je ključni statistički koncept koji ilustruje kako su vrednosti u skupu podataka raspoređene i učestale od najnižih do najviših vrednosti



Vrste raspodela podataka

U zavisnosti od tipova podataka koji se obrađuju, raspodelu podataka je moguće podeliti u dve grupe:



Diskretne raspodele

- Bernulijeva raspodela
- Binomialna raspodela
- Poasonova raspodela

Kontinualne raspodele

- Normalna raspodela
- Eksponencijalna raspodela

Diskretne raspodele

Bernulijeva raspodela

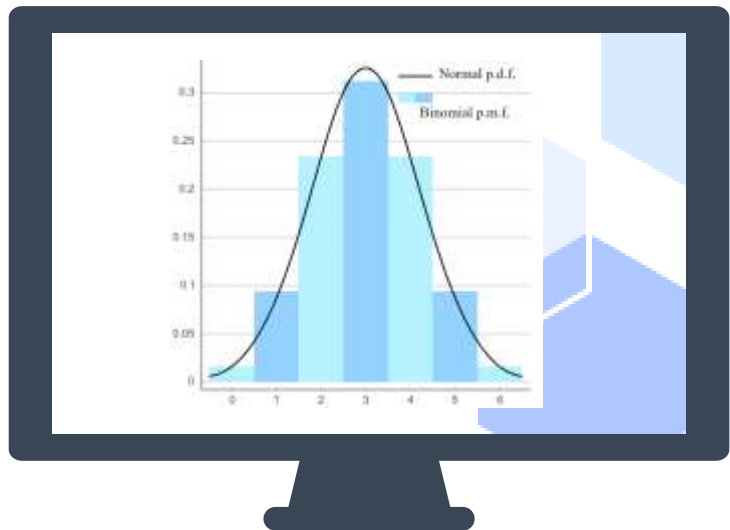
Bernulijeva raspodela je najjednostavnija diskretna raspodela i modelira slučajeve u kojima postoji samo dva moguća ishoda nekog eksperimenta ili procesa, obično označeni kao "uspeh" i "neuspeh"



Diskretne raspodele

Binomialna raspodela

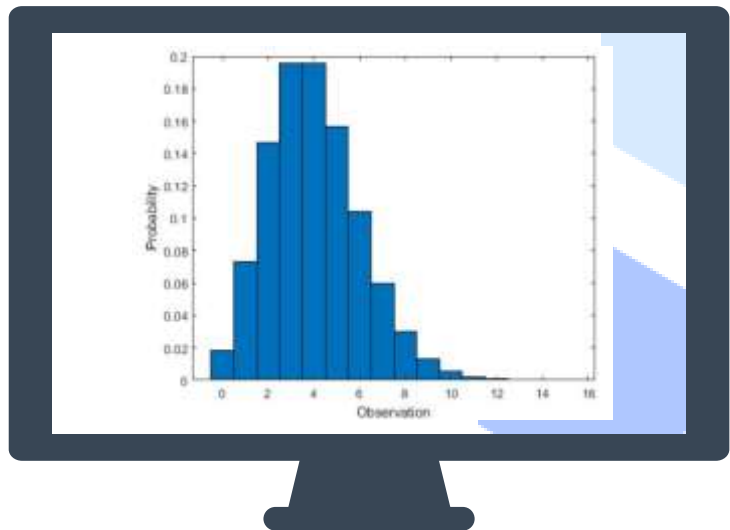
Binomialna raspodela je diskretna raspodela koja generalizuje Bernulijevu raspodelu za niz nezavisnih i identičkih ispitivanja. Koristi se kada je interesovanje usmereno na brojanje uspeha u fiksnom broju ponavljanja nekog slučajnog eksperimenta.



Diskretne raspodele

Poasonova raspodela

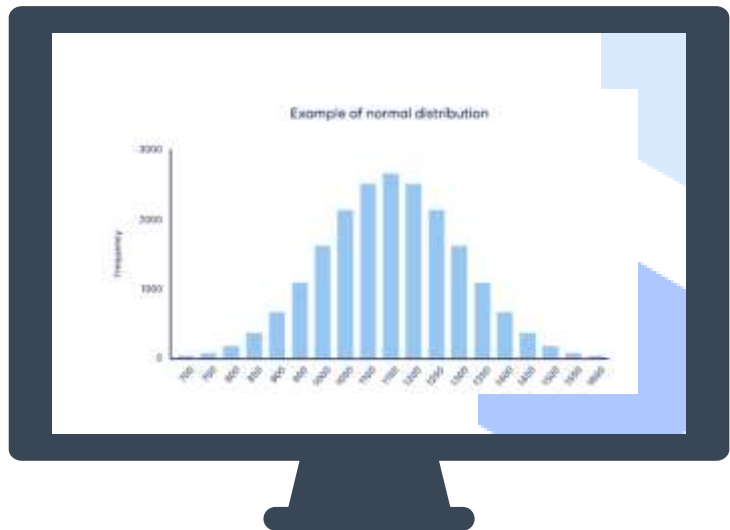
Poasonova raspodela se koristi za modeliranje broja puta koji se neki događaj dešava u fiksanom vremenskom intervalu, prostoru ili skupu.



Kontinualne raspodele

Normalna raspodela

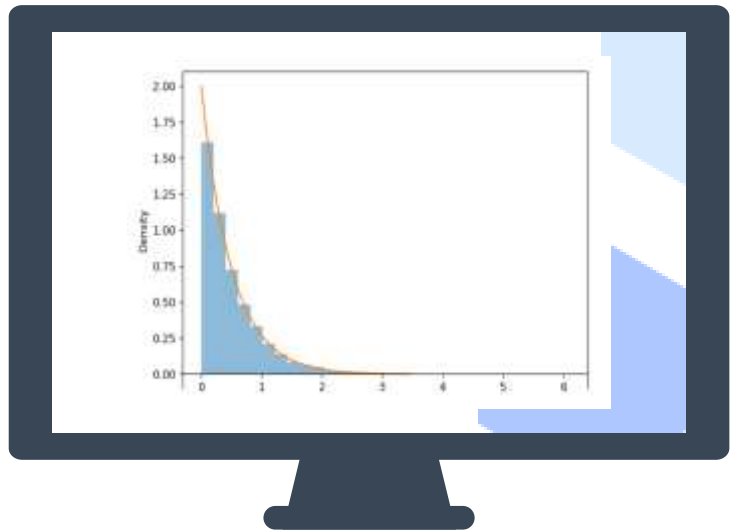
Glavna karakteristika podataka koji su predstavljeni ovom raspodelom jeste da su mere centralne tendencije srednja vrednost, medijana i modus jednake.



Kontinualne raspodele

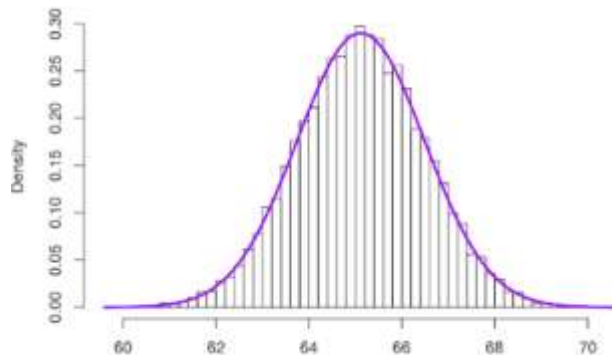
Ekspionencijalna raspodela

Ekspionencijalna raspodela se koristi za modeliranje vremena između nezavisnih događaja koji se dešavaju sa konstantnom stopom.

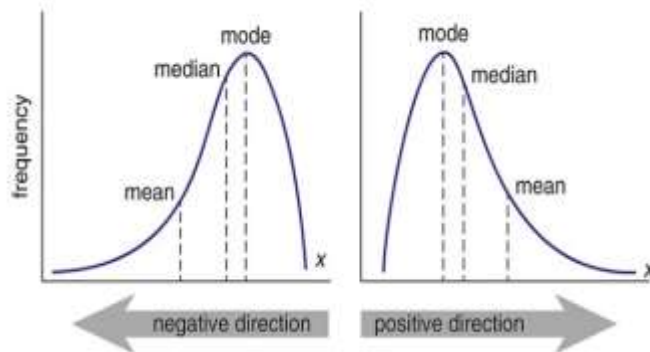


Tipovi raspodele I njihova vizualizacija

Simetrična raspodela

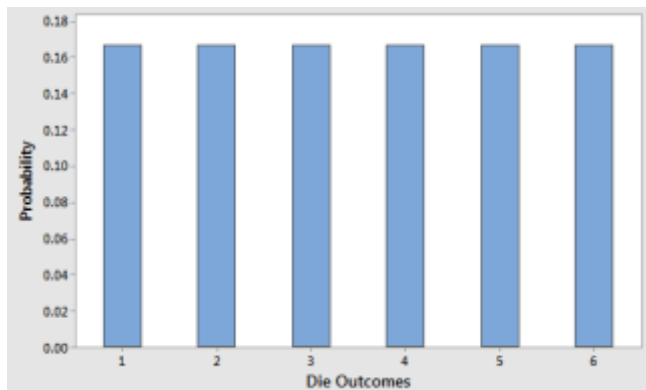


Asimetrična raspodela

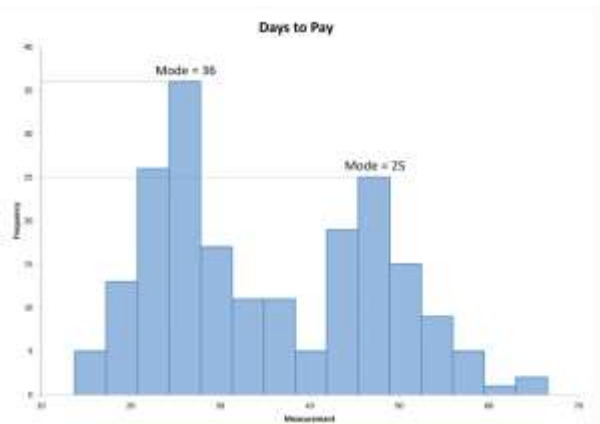


Tipovi raspodele I njihova vizualizacija

Uniformna raspodela



Bimodalna raspodela





04

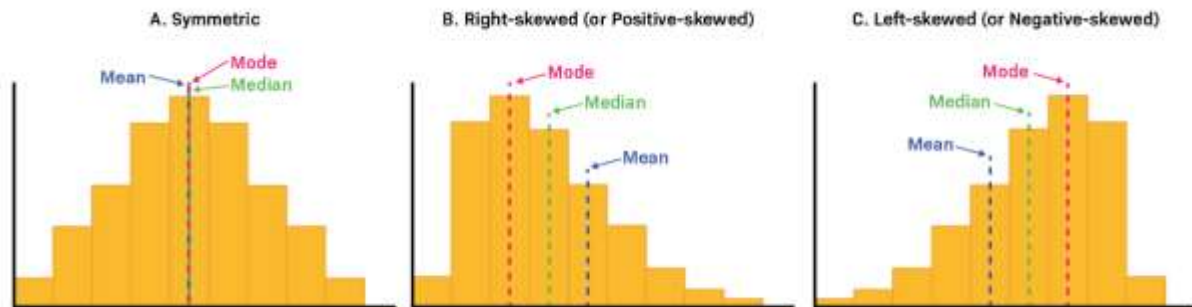
Mere centralne tendencije

Mere centralne tendencije

Mere centralne tendencije su statistički indikatori koji pružaju sažet pregled seta podataka označavajući jednu vrednost koja je reprezentativna za celokupan skup. Ove mere su ključne za sumiranje velikih količina podataka, olakšavajući razumevanje i interpretaciju podataka u jednostavnijem obliku.

Osnovne mere centralne tendencije su:

- **Srednja vrednost**
- **Medijana**
- **Moduo**



Srednja vrednost

Aritmetička srednja vrednost predstavlja najzastupljeniju i najviše korišćenu meru centralne tendencije. Aritmetička srednja vrednost uzima u obzir sve vrednosti iz skupa podataka prilikom procesa računanja konačne vrednosti. Upravo iz razloga što uzima sve vrednosti iz skupa, može dovesti do nepravilnih zaključivanja o skupu podataka zato što se u procesu izračunavanja koriste i granične „outlier“ vrednosti.

Formula za izračunavanje:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \text{ gde je } x_i \text{ vrednost svakog pojedinačnog podatka u skupu, a } n \text{ je ukupan broj podataka.}$$

Medijana

Medijana određuje vrednost koja deli skup podataka tako da ima jednak broj vrednosti ispod i iznad sebe kada su podaci sortiran. Medijana je važna zato što pruža jasan uvid u "sredinu" skupa podataka i efikasna je u situacijama kada skup podataka sadrži ekstremne vrednosti ili "outlier"-e, tako što ih eliminiše u toku procesa izračunavanja.

Moduo

Odlikuje se kao vrednost ili vrednosti koje se pojavljuju najčešće u datom skupu podataka. Ova mera centralne tendencije se najčešće koristi kod fičera koji mogu imati manji broj mogućih vrednosti kao što su kategorički podaci.

```
data2.mean(numeric_only=True)
```

age	40.023800
duration	258.315815
campaign	2.567879
pdays	962.464810
previous	0.173013
emp.var.rate	0.081922

```
data2.median(numeric_only=True)
```

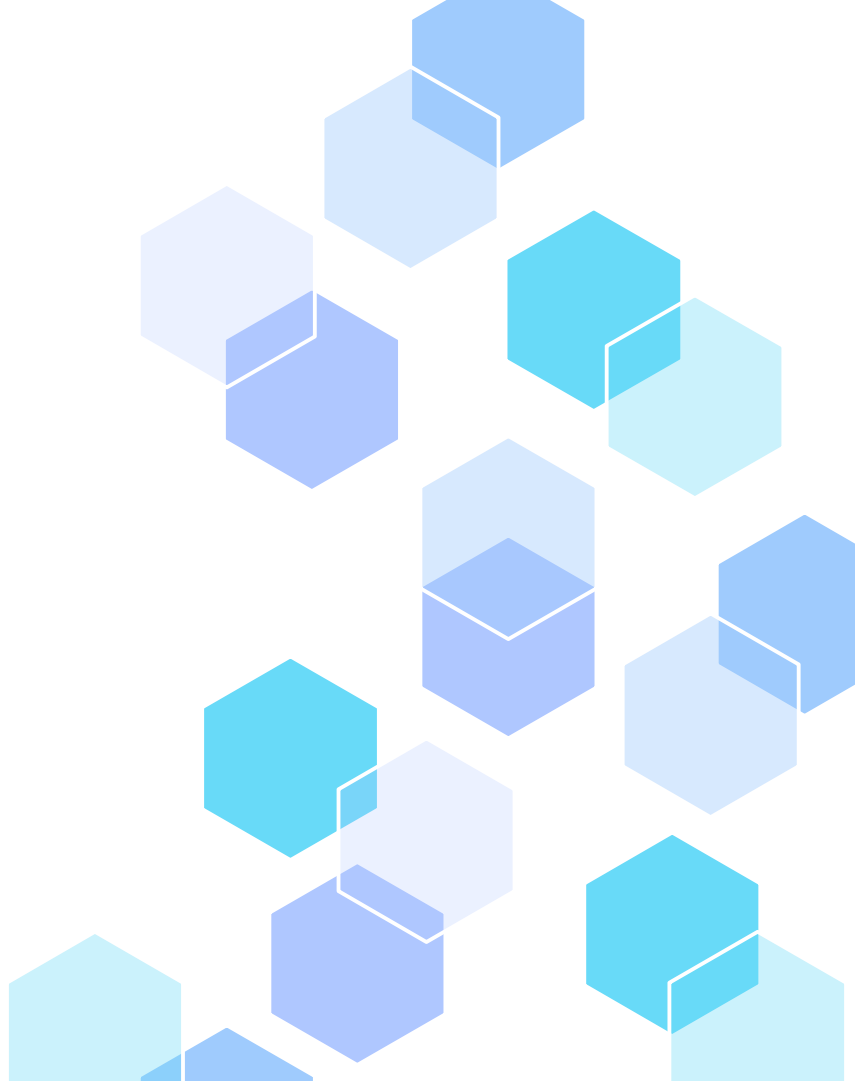
age	38.000
duration	180.000
campaign	2.000
pdays	999.000
previous	0.000
emp.var.rate	1.100

```
data2.mode(axis=0).head(1)
```

	age	job	marital	education	default	housing	loan
0	31.0	admin.	married	university.degree	no	yes	no

05

Korelacija







Korelacija



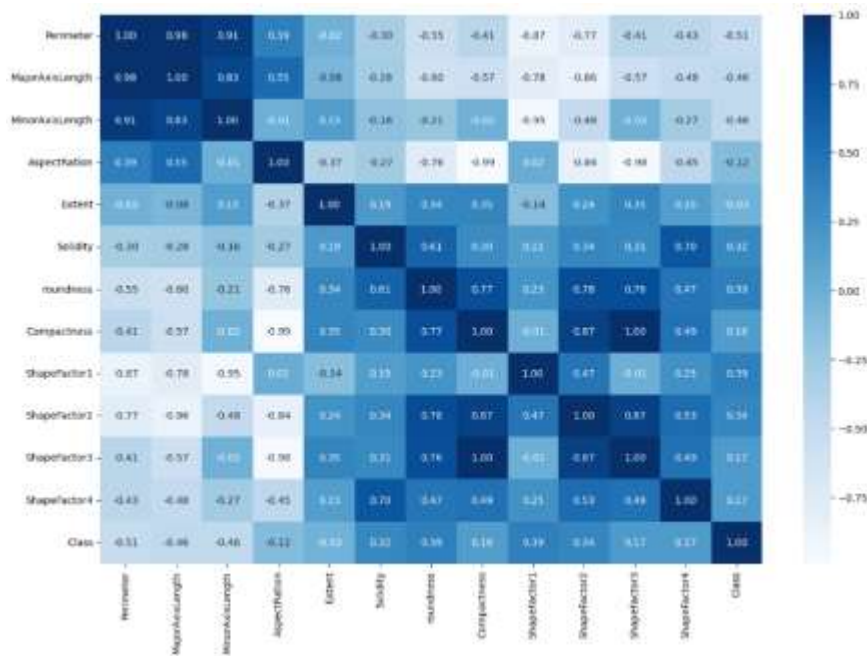
Korelacija predstavlja meru koja opisuje stepen međusobne veze između dve ili više promenljivih. Korelacija može ukazivati na to kako promena vrednosti jedne promenljive utiče na vrednost druge promenljive.

Vrste korelacija:

- **Pozitivna korelacija:** Kada vrednost jedne promenljive raste, vrednost druge promenljive takođe raste.
 - **Negativna korelacija:** Kada vrednost jedne promenljive raste, vrednost druge promenljive opada
 - **Nulta korelacija:** Ne postoji uočljiva veza između promenljivih
- 
- 

Korelacija

Prilikom izračunavanja vrednosti korelacije na nivou celokupnog skupa podataka, primenom ugrađenih funkcija, dobija se matrica korelacije – simetrična matrica koja za vrste i kolone ima ulazne fičere posmatranog skupa podataka.



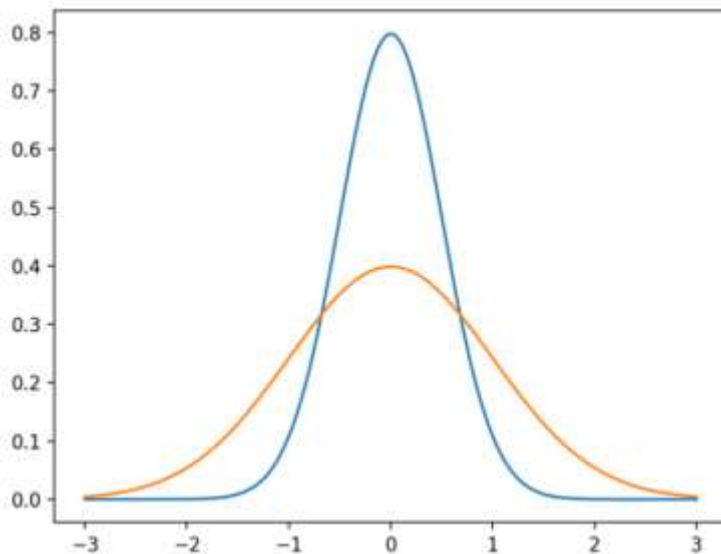


06

Varijansa

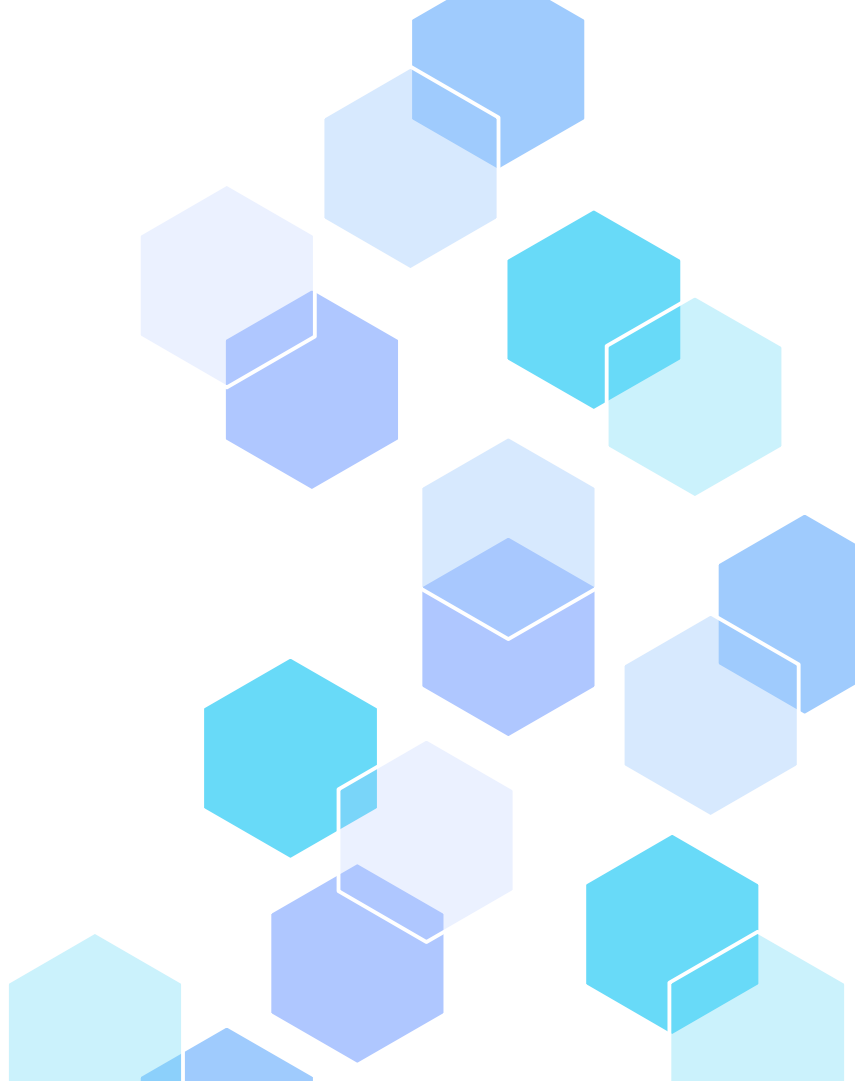
Varijansa

Varijansa podataka je mera koja opisuje rasprostranjenost ili disperziju vrednosti u skupu podataka u odnosu na njihovu srednju vrednost. Drugim rečima, varijansa pokazuje koliko se vrednosti u datasetu razlikuju jedna od druge i od srednje vrednosti. **Veća varijansa** ukazuje na to da su vrednosti više rasprostranjene oko srednje vrednosti, dok **manja varijansa** ukazuje na to da su vrednosti bliže srednjoj vrednosti.

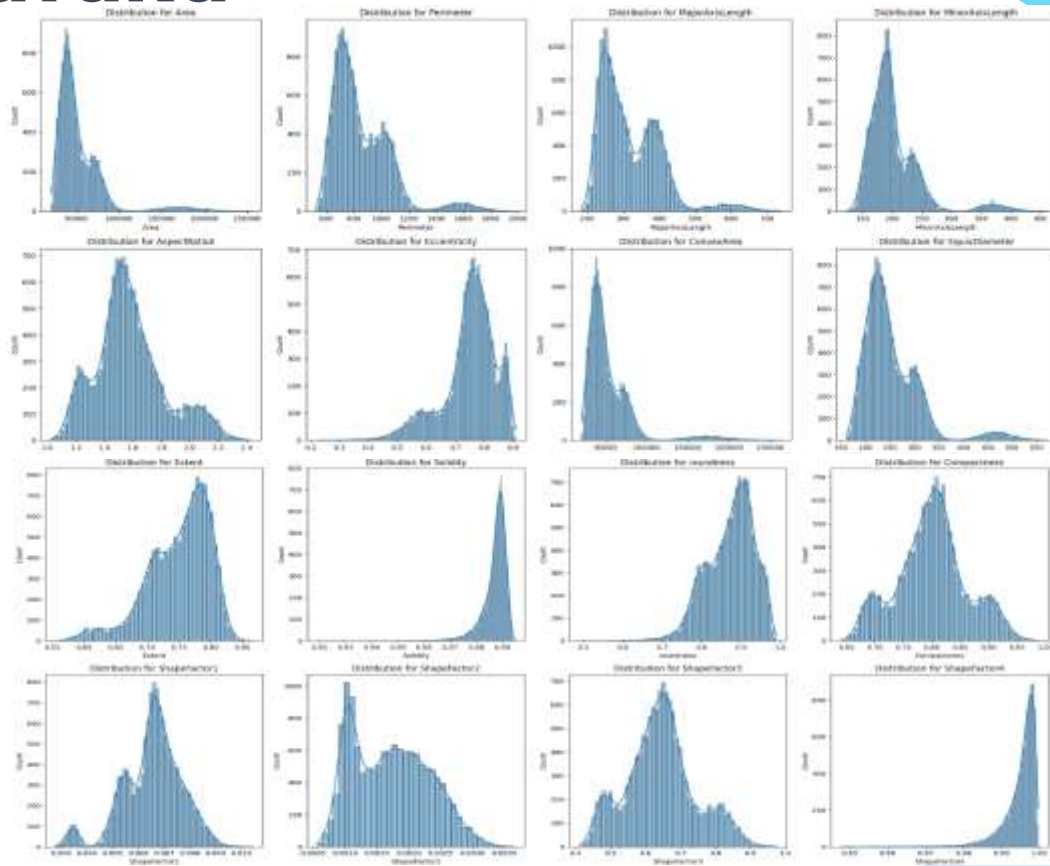
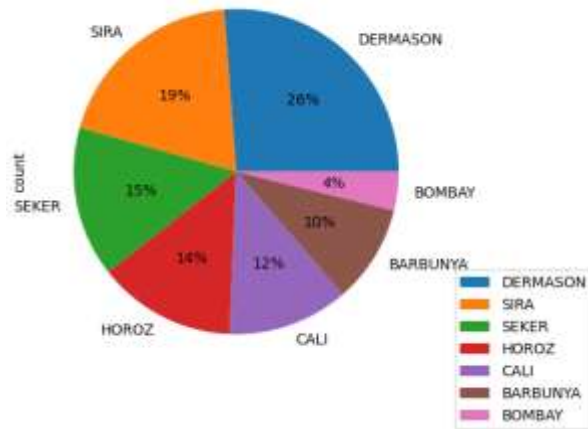


07

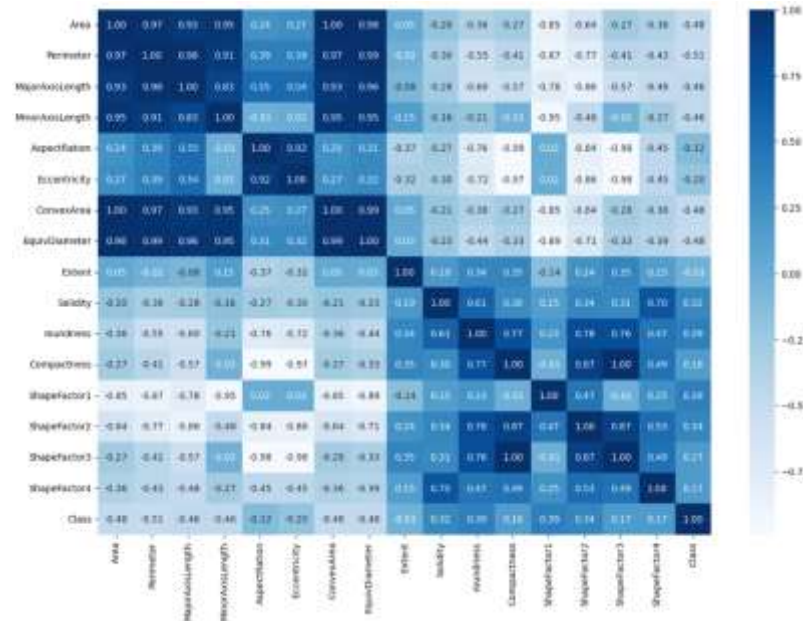
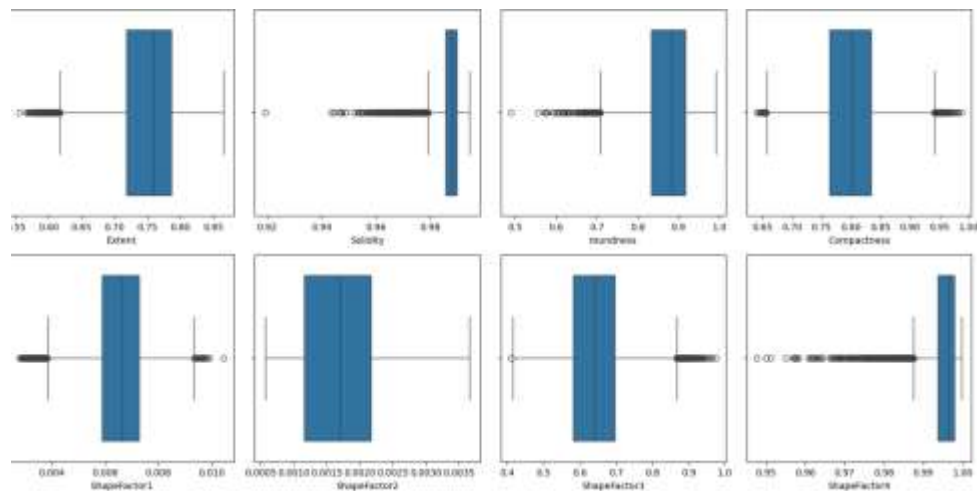
Praktični deo rada



Prvi skup podataka



Prvi skup podataka



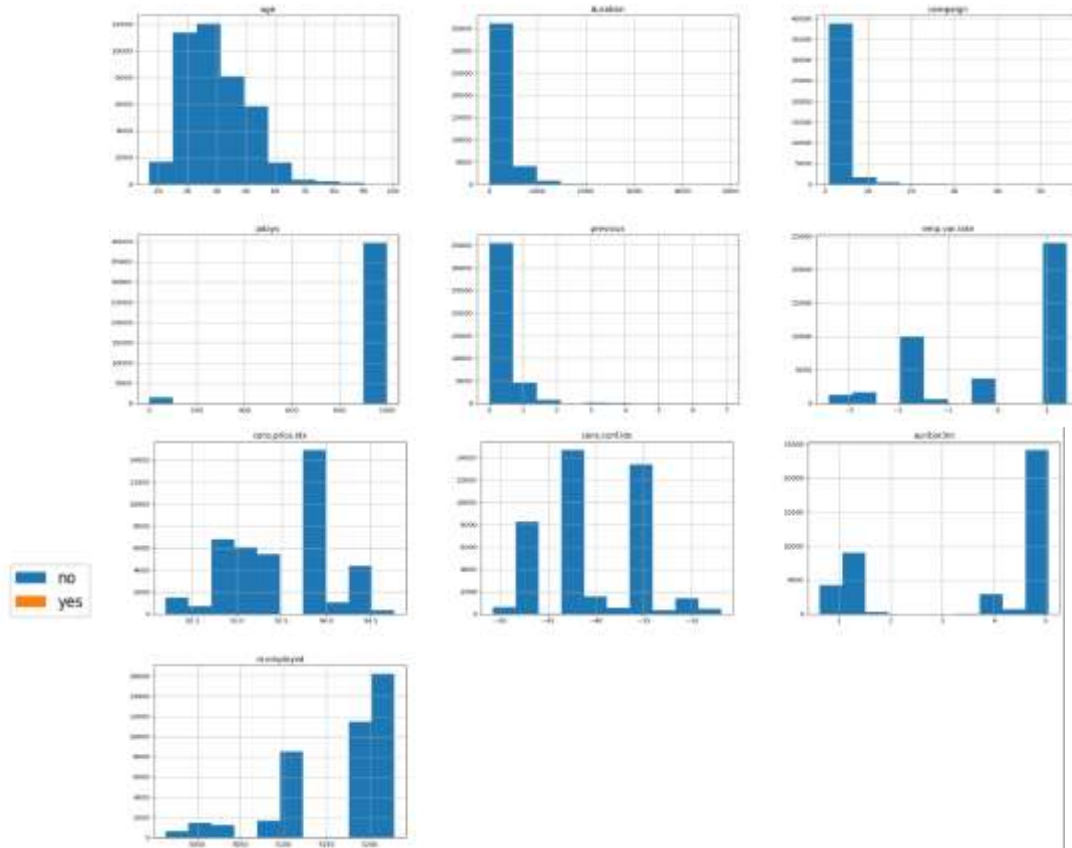
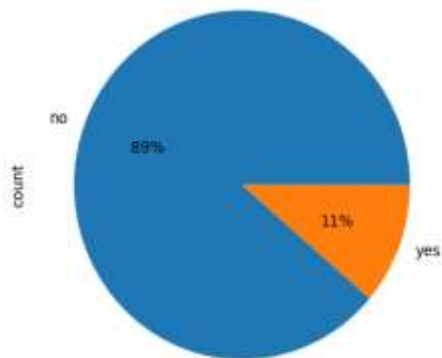
Prvi skup podataka

Da bismo videli kako sve ove transformacije utiču na rezultate algoritama mašinskog učenja, primenićemo dva algoritma za klasifikaciju Support Vector Classifier (SVC) i Random Forest algoritam.

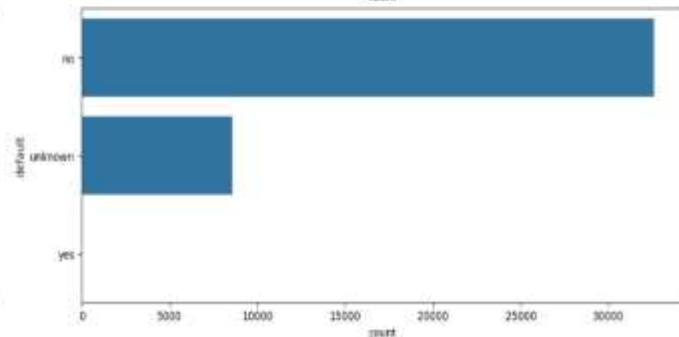
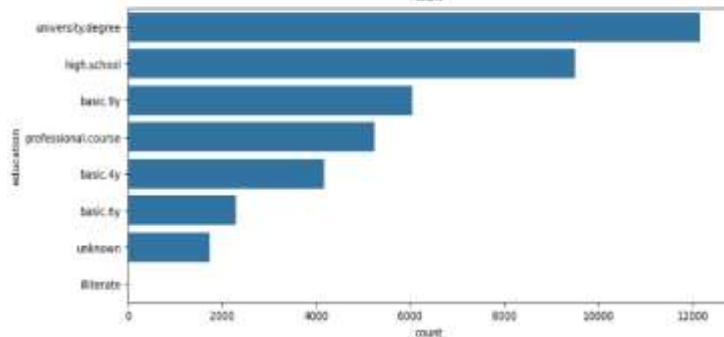
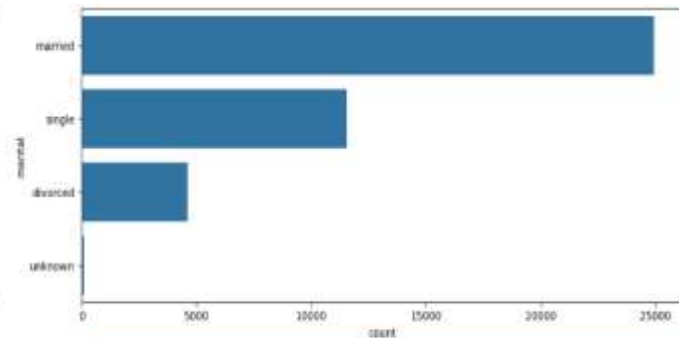
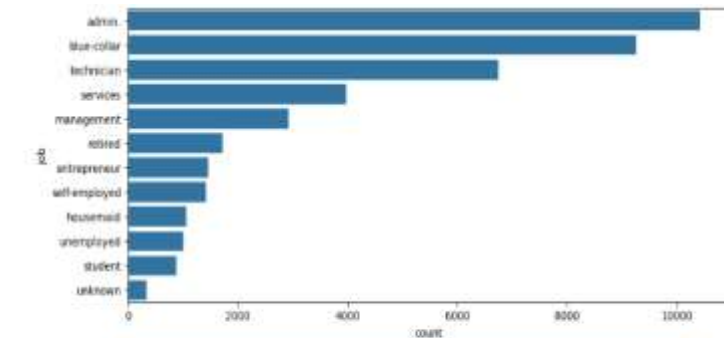
Tip podele	SVC accuracy	Random Forest accuracy
Osnovni skup	0.64	0.92
Balansirani	0.77	0.99
Raspodela podataka	0.76	0.92
Outlieri	0.64	0.92
Korelacija	0.88	0.93

Iz ovih rezultata možemo zaključiti da Random Forest algoritam generalno postiže bolje rezultate od SVC algoritma u svim scenarijima. Balansiranje skupa podataka, primena log transformacije i uklanjanje visoko korelisanih atributa imaju pozitivan uticaj na performanse klasifikacije, posebno za SVC algoritam. Međutim, uklanjanje outlier-a ne pokazuje značajne promene u tačnosti klasifikacije.

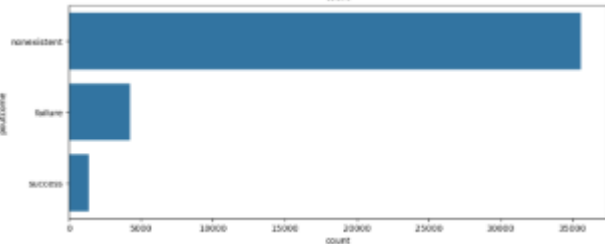
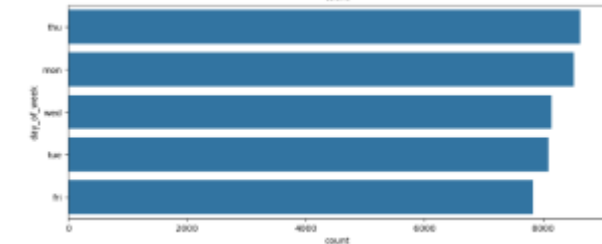
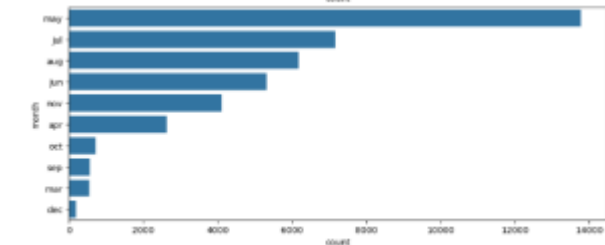
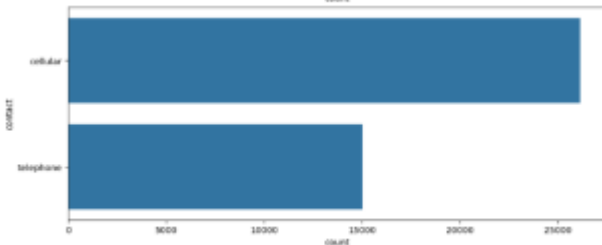
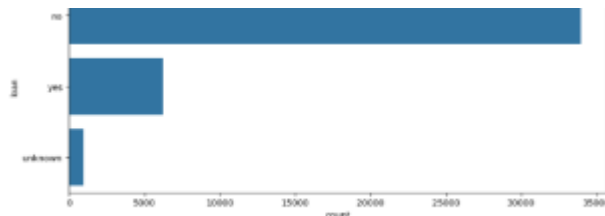
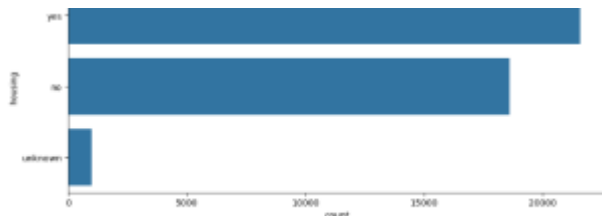
Drugi skup podataka



Drugi skup podataka



Drugi skup podataka



	age	job	marital	education	tenure	loan	contact	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	eurbor3m	reemployed	y	year_quarter	day_of_week_fri	day_of_week_mon	day_of_week_thu	day_of_week_tue	day_of_week_wed	outcome_failure	outcome_nonexistent	outcome_success
age	1.00	-0.01	0.19	-0.10	0.00	-0.01	-0.01	-0.00	0.00	-0.01	0.02	-0.00	0.00	-0.01	0.02	0.00	0.12	0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
job	-0.01	1.00	0.02	-0.13	0.21	-0.21	0.03	-0.07	0.01	-0.03	0.02	-0.21	-0.30	-0.05	-0.02	0.02	0.01	0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
marital	0.19	0.02	1.00	0.19	0.03	-0.02	-0.05	0.01	0.01	0.04	0.04	0.08	0.00	0.03	0.03	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
education	-0.10	-0.13	0.19	1.00	0.02	0.01	0.12	0.01	0.00	0.04	0.04	-0.05	-0.10	0.06	-0.04	-0.04	0.00	0.12	0.00	0.01	0.01	-0.01	-0.01	-0.01	-0.01	0.04
tenure	0.00	0.01	0.03	0.02	1.00	0.03	0.00	0.01	0.01	0.01	0.01	0.02	-0.06	-0.08	-0.03	-0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
loan	-0.01	-0.21	-0.02	0.01	0.03	1.00	0.01	-0.00	0.01	0.01	-0.00	0.00	-0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
contact	0.01	0.01	-0.05	-0.12	0.00	0.01	1.00	0.01	0.00	0.12	0.21	-0.39	-0.39	0.25	0.40	0.27	0.01	0.42	0.04	0.02	0.04	0.00	0.01	0.21	0.24	0.11
duration	0.00	-0.01	0.01	0.01	0.01	0.00	0.01	1.00	0.07	-0.05	0.02	-0.03	0.01	-0.01	-0.01	0.04	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.01	0.01	0.01
campaign	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.07	1.00	0.05	0.08	0.11	0.13	0.01	0.11	0.01	0.07	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
pdays	-0.04	-0.03	0.04	-0.04	0.01	0.00	-0.13	0.05	0.05	1.00	0.59	0.27	0.00	0.00	0.30	0.37	0.32	0.06	0.31	0.00	-0.01	0.01	0.00	0.01	0.00	0.00
previous	0.02	0.02	0.04	0.04	0.02	0.00	0.21	0.02	0.08	0.59	1.00	0.42	0.29	0.05	0.45	0.50	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
emp.var.rate	-0.00	-0.01	0.00	0.05	0.00	0.01	-0.39	0.05	0.15	0.27	0.42	1.00	0.70	0.30	0.97	0.91	0.30	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01
cons.price.idx	0.00	-0.02	0.00	-0.10	0.00	0.01	-0.50	0.01	0.13	0.00	0.20	0.70	1.00	0.06	0.89	0.52	0.14	-0.10	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00
cons.conf.idx	0.13	0.05	0.03	0.00	0.03	0.01	-0.25	0.01	0.12	0.00	0.05	0.30	0.06	1.00	0.20	0.10										

Drugi skup podataka

Tip podele	SVC accuracy	Random Forest accuracy
Osnovni skup	0.90	0.92
Balansirani	0.91	0.98
Raspodela podataka	0.89	0.91
Outlieri	0.90	0.91
Korelacija	0.89	0.92

Na osnovu manjih varijacija u rezultatima za oba algoritma, možemo zaključiti da je ovaj skup podataka kvalitetniji. Konzistentnost u rezultatima ukazuje na to da oba modela dobro rade na ovom datasetu, što može biti rezultat kvalitetnijih i ujednačenijih podataka.



08

Zaključak



Razmatrajući suštinske koncepte kvaliteta podataka i procesa pripreme podataka u kontekstu mašinskog učenja, ovo istraživanje ističe ključne elemente neophodne za uspješno razvijanje modela. Kvalitetni podaci su osnovni temelj koji omogućava izgradnju pouzdanih i preciznih modela mašinskog učenja, a njihova analiza pruža ključne uvide za postizanje efikasnih rešenja.



Hvala na pažnji

