



# Klasterizacija

Bojan Tomić  
Jelena Jovanović

# Klasterizacija kao zadatak VI



- Osnove Python-a
  - Numpy, Pandas, Scikit-learn...
- Utvrđivanje zavisnosti i predviđanje
- Klasifikacija
- Klasterizacija
- Pretraživanje

# Klasterizacija kao zadatak VI



- Utvrđivanje zavisnosti i predviđanje
  - Korelacije, (linearna) regresija
- Klasifikacija
  - Stabla odlučivanja, KNN, neuronske mreže
- Klasterizacija
  - K-means, hijerarhijska klasterizacija
- Pretraživanje (search)
  - Breadth-first, depth-first, best-first

A yellow pencil and a pink eraser are positioned in the top right corner of the white paper, suggesting a classroom or study environment.

Šta je klasterizacija?

# Šta je klasterizacija?

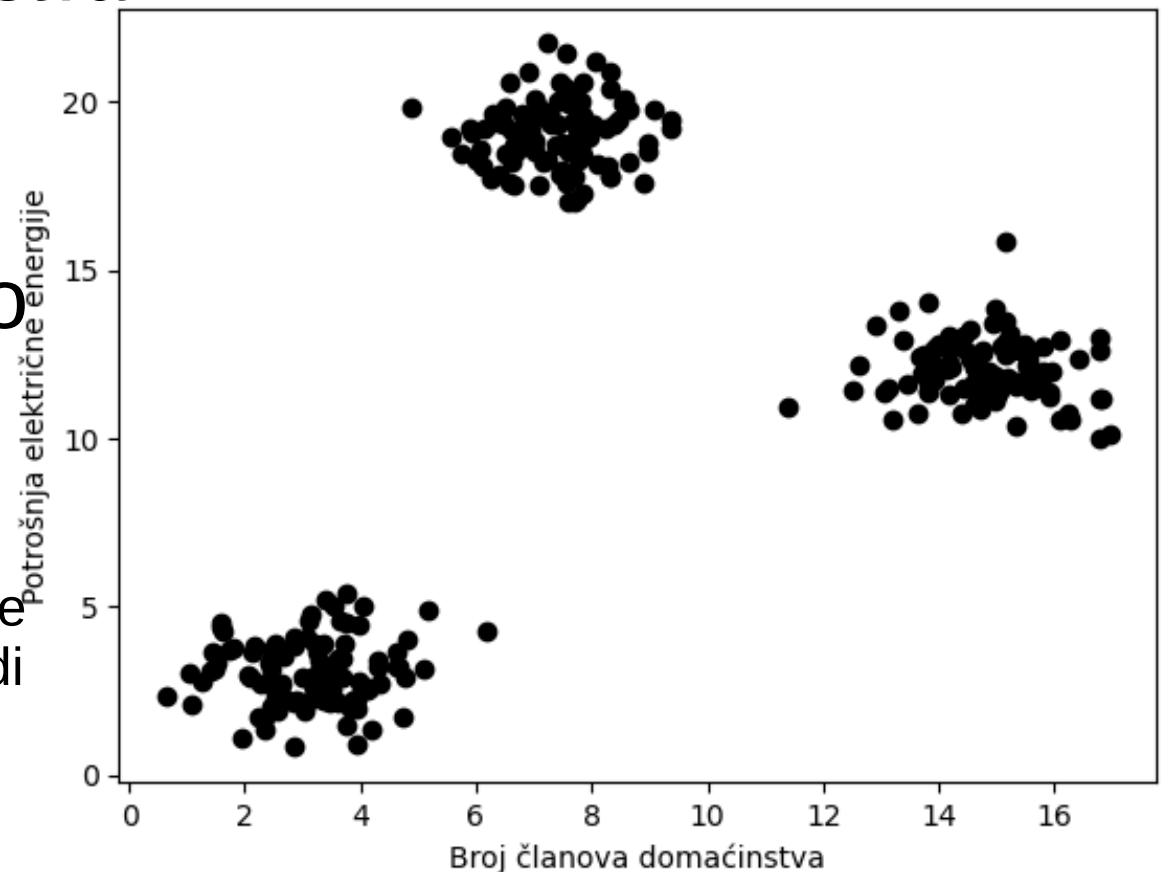
- Klasterizacija je zadatak grupisanja instanci, tako da za svaku instancu važi da je sličnija (bliža) instancama iz svoje grupe (klastera), nego instancama iz drugih grupa (klastera)
- Ciljevi
  - Identifikacija potencijalnih klastera sličnih instanci
  - Svrstavanje novih instanci u identifikovane klastere
- Nije klasifikacija – jer nisu poznate klase
- Nije regresija – jer se ne traži zavisnost, samo sličnost.



# Primer 1

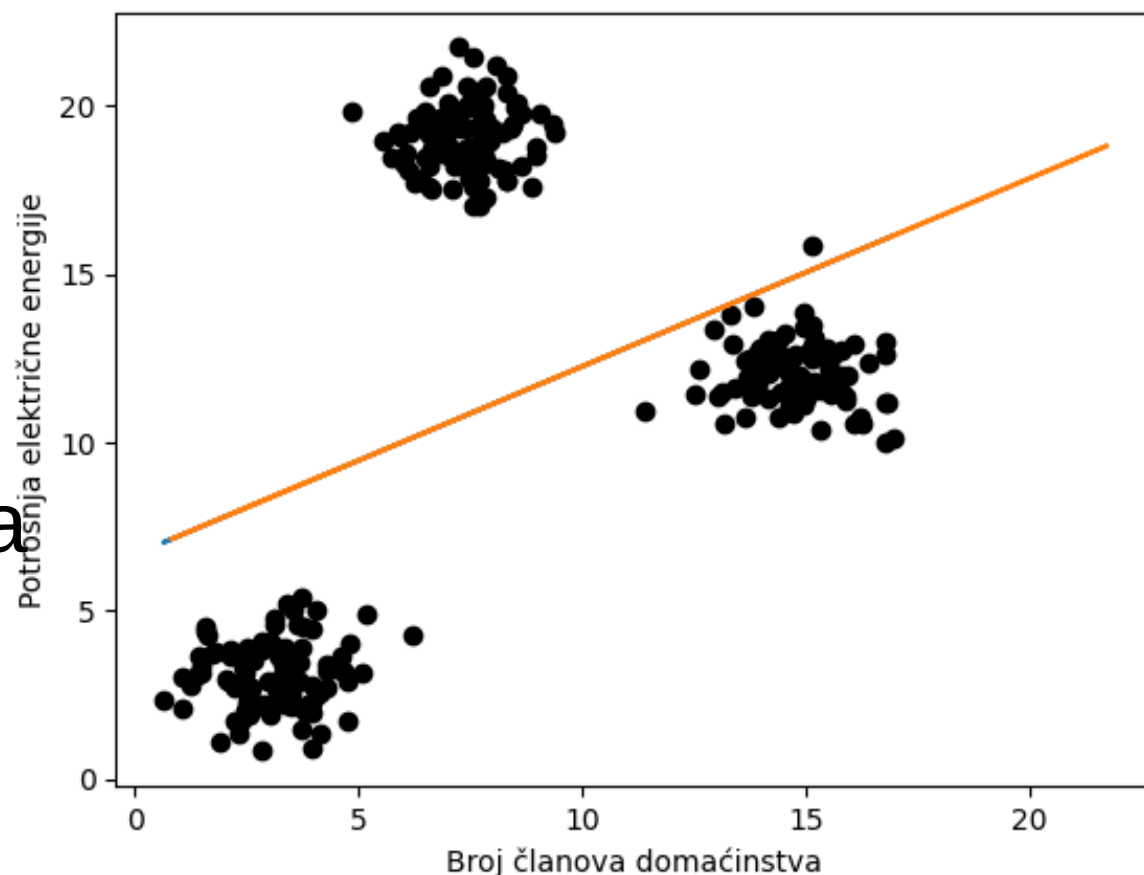
- Podaci:
  - Potrošnja električne energije domaćinstva
  - Broj članova domaćinstva\*
- Jedna tačka-jedno domaćinstvo

\* broj članova domaćinstva je uvek ceo broj, ali je ovde radi stvaranja „idealnog“ primera stavljeno drugačije



# Primer 1

- Ovo nije problem koji se rešava linearnom regresijom
- Nema (linearne) veze među promenljivima
- U tom smislu, „problem“ je grupa domaćinstava sa 5-10 članova

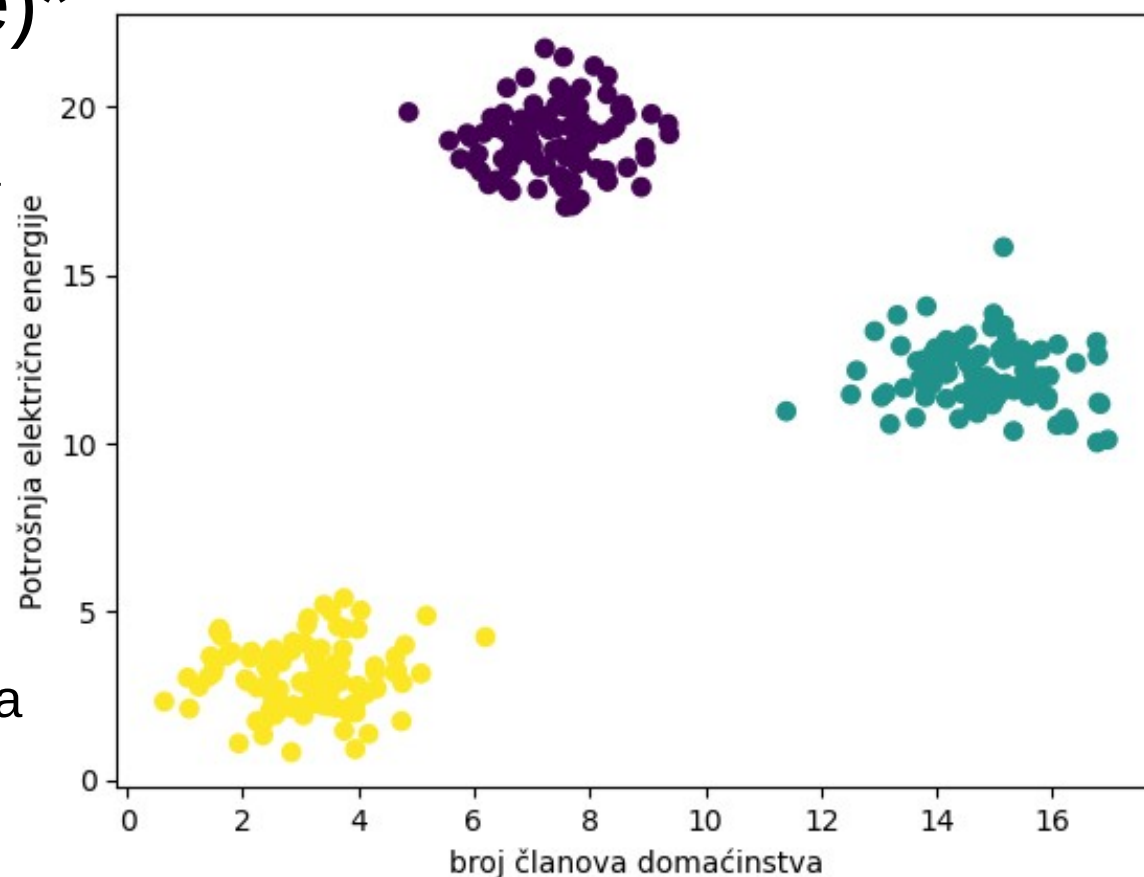


# Primer 1

- Uradimo klasterizaciju
- Tri klastera (grupe)\*

- 1) Mala domaćinstva sa malom potrošnjom - žuta (stanovi?)
- 2) Srednja domaćinstva sa visokom potrošnjom - ljubičasta (gradske porodične kuće?)
- 3) Velika domaćinstva sa prosečnom potrošnjom - zelena (seoska porodična imanja?)

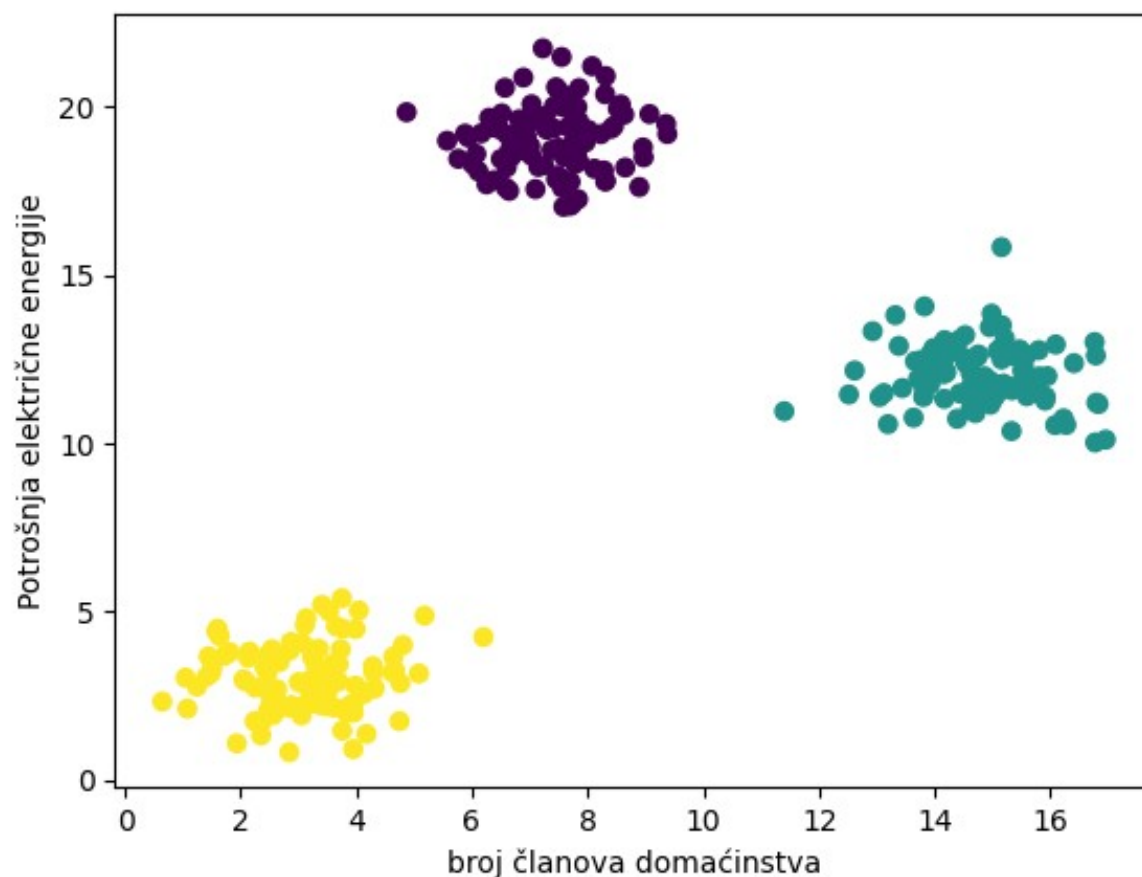
\* *Subjektivno tumačenje*





# Primer 1

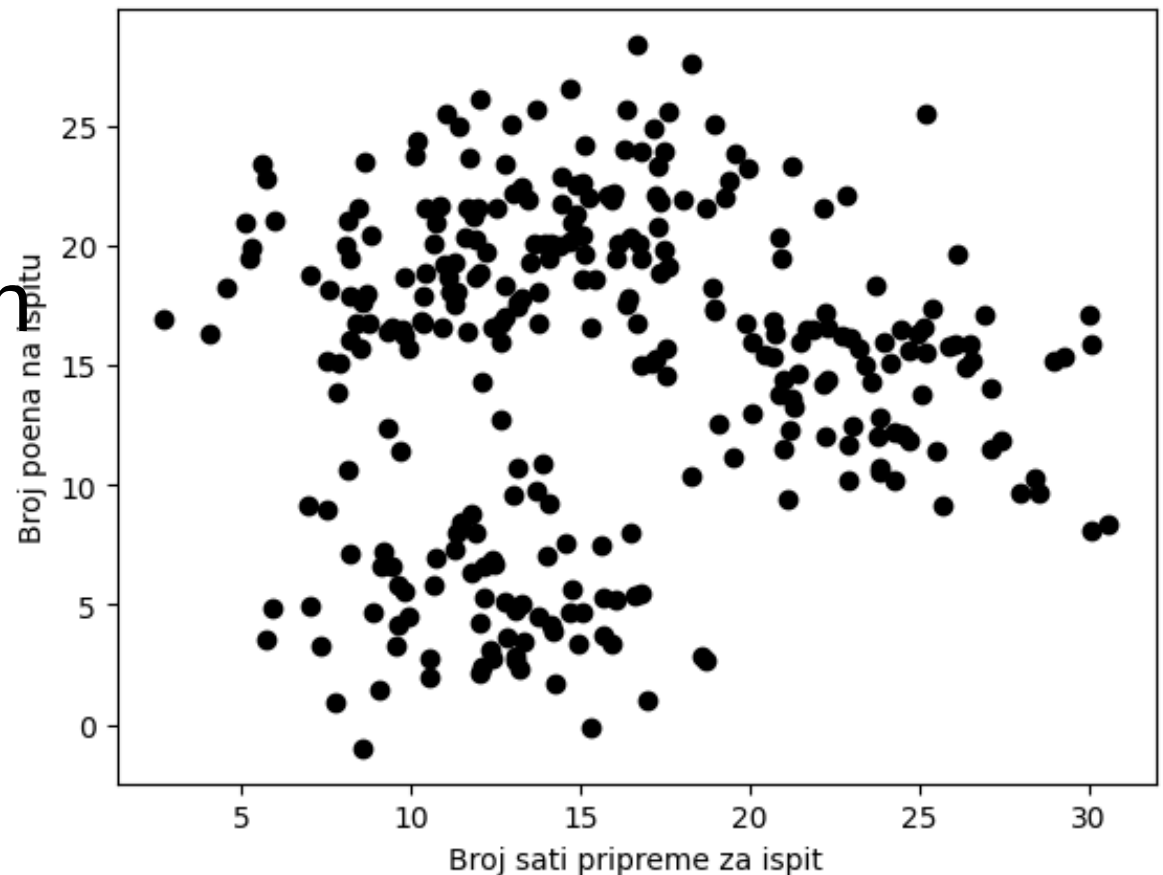
- Samo jedno rešenje (3 klastera)
- Dve promenljive
- Idealni klasteri
  - Gusto zbijeni
  - Međusobno razdvojeni
  - Iste veličine
- Idealan, nerealan primer (jednorog)



# Primer 2

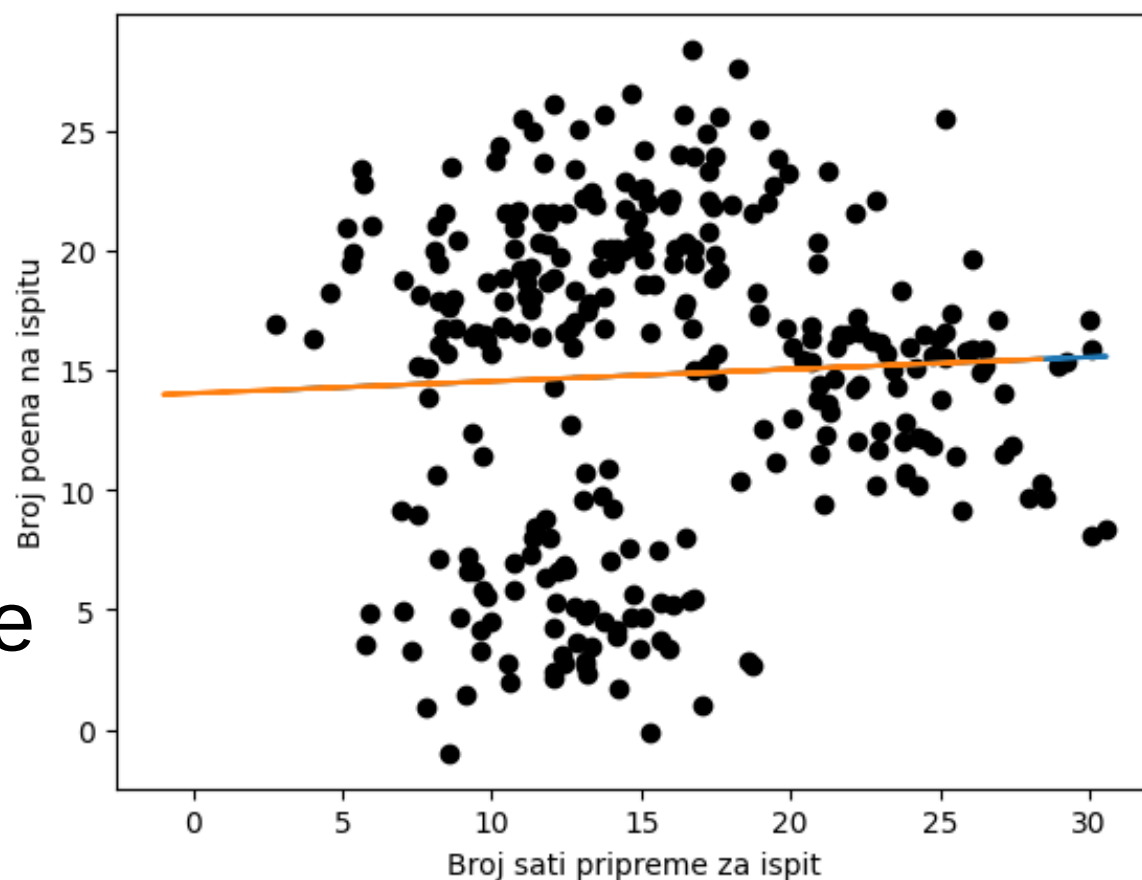


- Podaci:
  - Broj sati pripreme za ispit
  - Broj ostvarenih poena na ispitu
- Jedna tačka-jedan student
- Malo realniji primer



# Primer 2

- Ni ovo nije problem koji se rešava regresijom
- Nema (linearne) veze među promenljivima
- Prevelika je „raštrkanost“ podataka oko linije regresije



# Primer 2

- Uradimo klasterizaciju

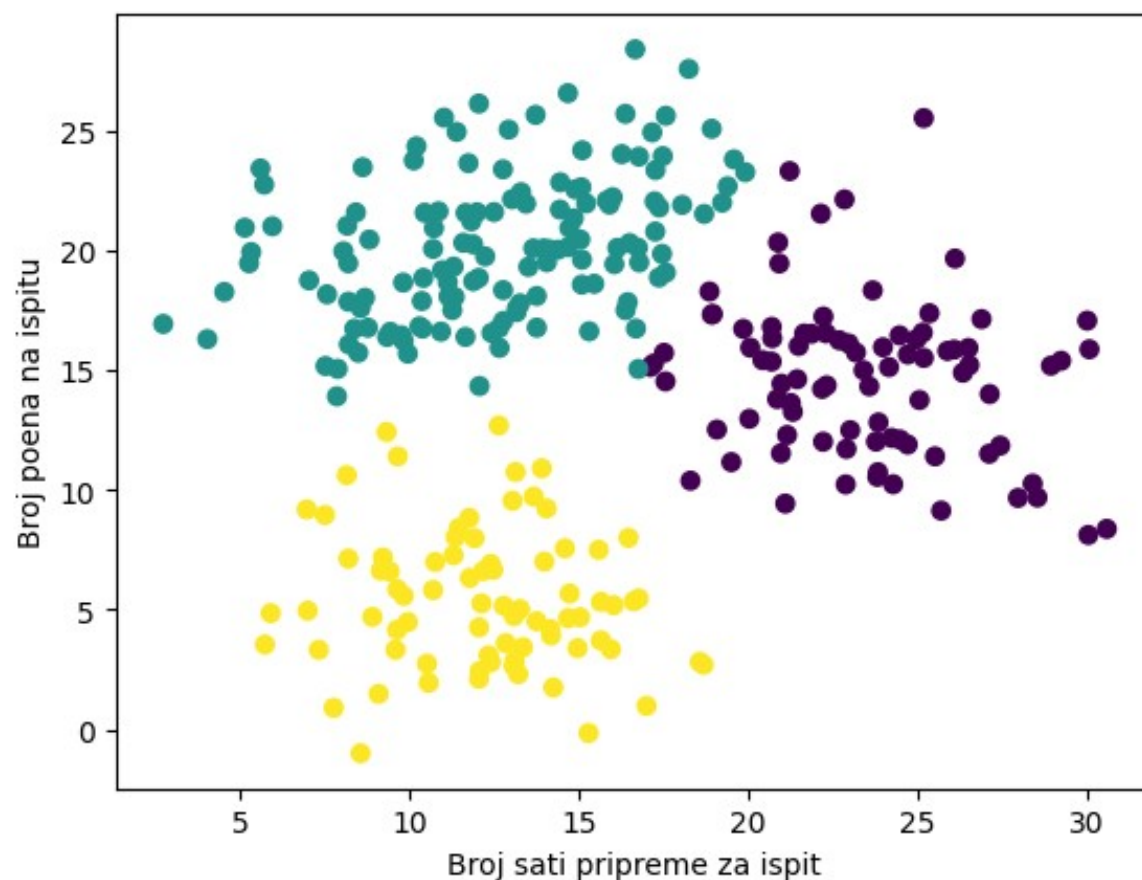
- Tri klastera (možda)?\*

1) Malo učili, slabo uradili (žuta)

2) Malo ili srednje učili, dobro uradili (zelena)

3) Puno učili, osrednje uradili (ljubičasta)

\* *Subjektivno tumačenje*



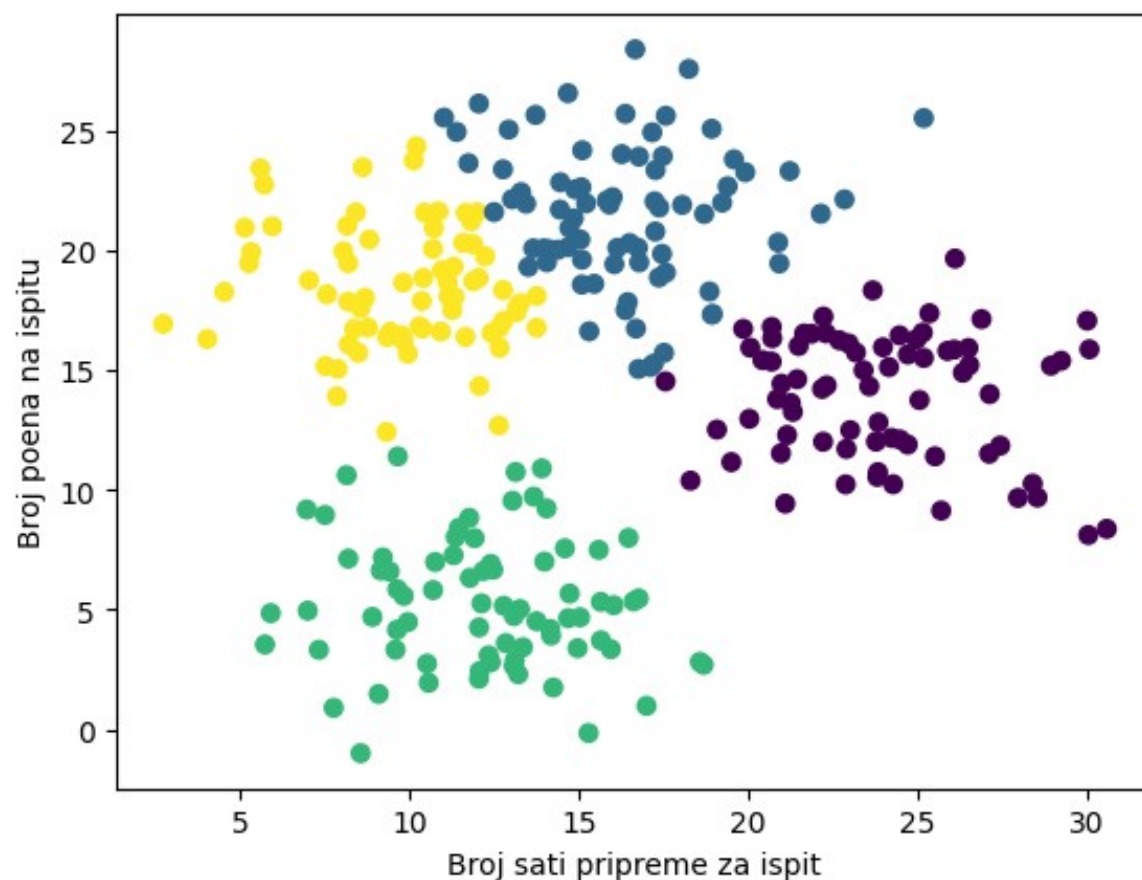


# Primer 2

- Četiri klastera (možda)?\*

- 1) Malo učili, slabo uradili (zelena)
- 2) Malo učili, dobro uradili (žuta)
- 3) Srednje učili, dobro uradili (plava)
- 4) Puno učili, osrednje uradili (ljubičasta)

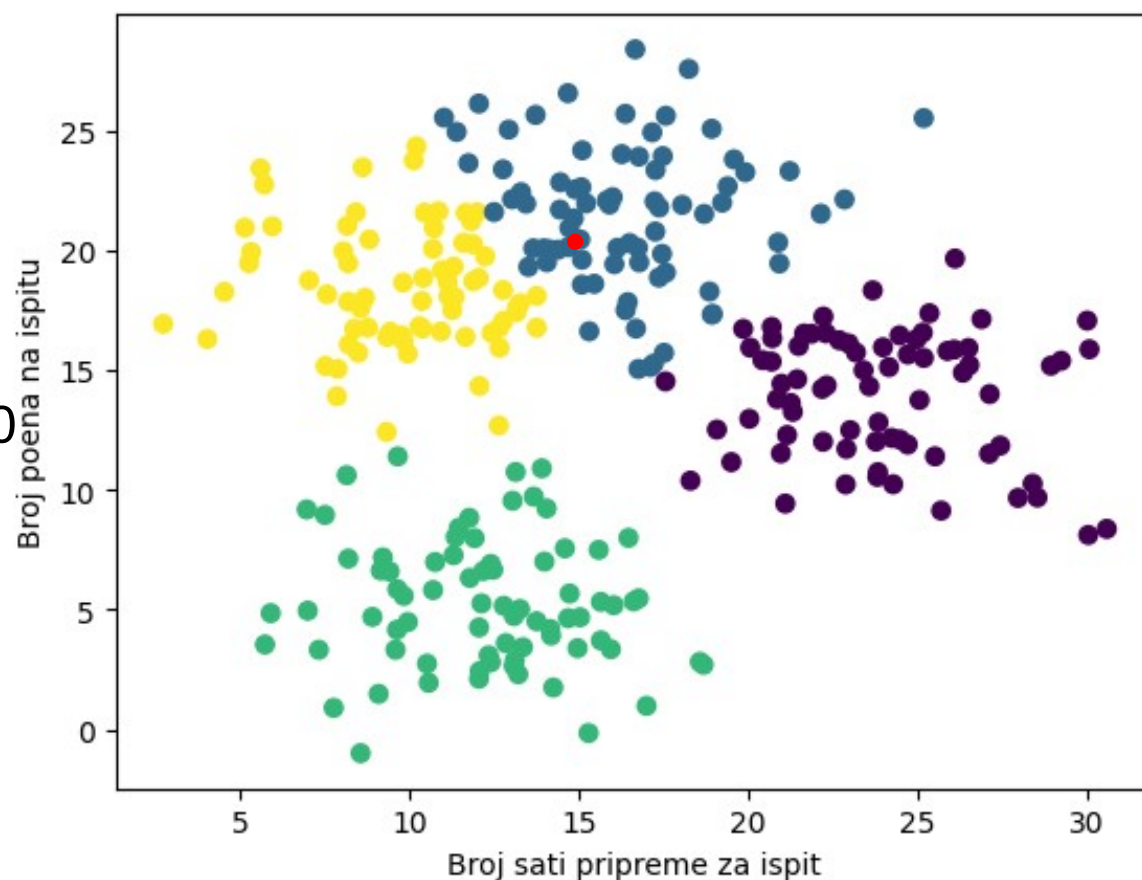
\* *Subjektivno tumačenje*





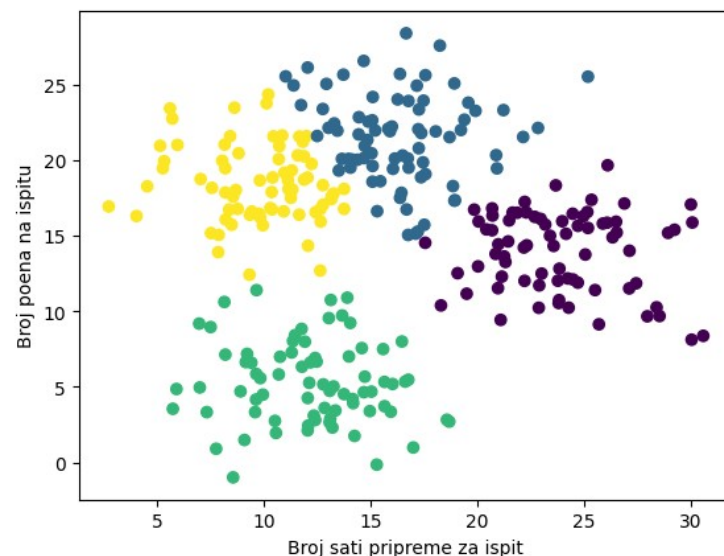
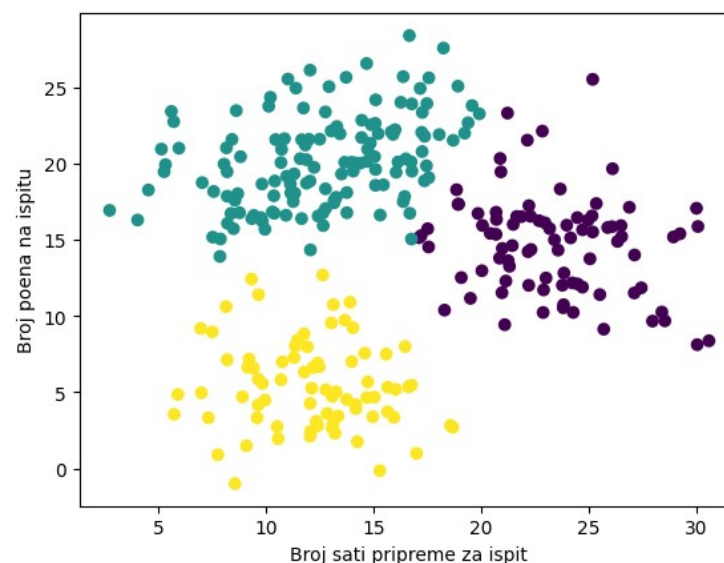
# Primer 2

- Dodatni zadatak klasterizacije
- Svrstavanje novih instanci u nađene klustere:
  - Kom klasteru pripada student koji je 15 sati pripremao ispit i dobio 20 poena (crvena tačka)?
  - Prema modelu sa 4 klastera, pripada plavom klasteru „Srednje učili, dobro uradili“.



# Primer 2

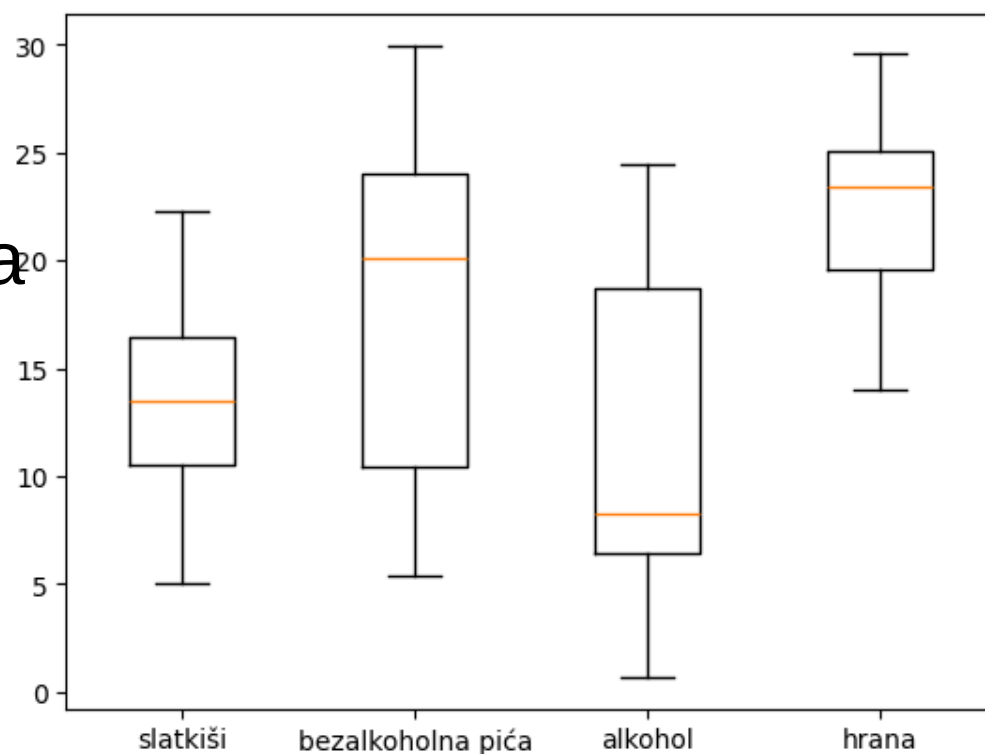
- Dva rešenja koja mogu imati smisla (3 ili 4 klastera)
- Dve promenljive
- Klasteri nisu potpuno odvojeni
- Malo realniji primer



# Primer 3

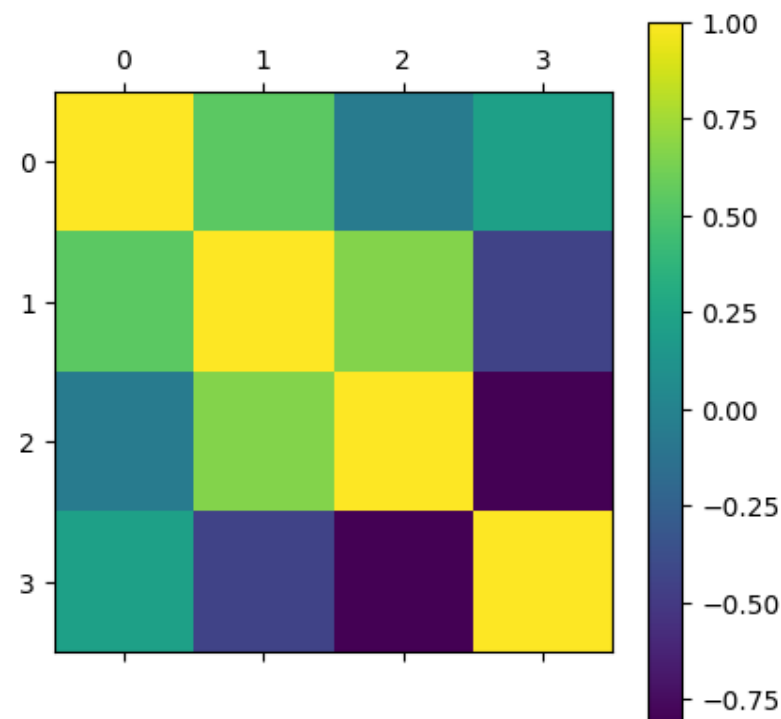


- Podaci:
  - Potrošnja novca na namirnice za više porodica
  - Slatkiši, bezalkoholna pića, alkohol, hrana
- Višedimenzijski podaci
- Još realniji primer



# Primer 3

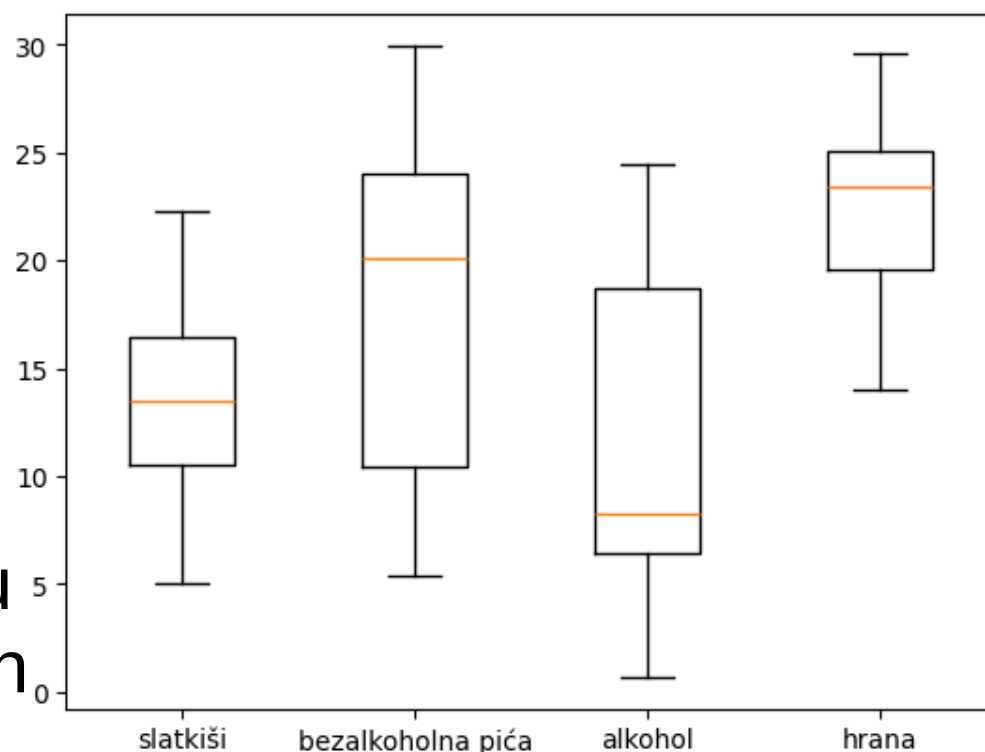
- Probamo korelaciju da vidimo da li su neke promenljive u vezi
- Nažalost, ne daje dobre rezultate
- Velika (negativna) korelacija je samo između hrane (2) i alkohola (3)



# Primer 3



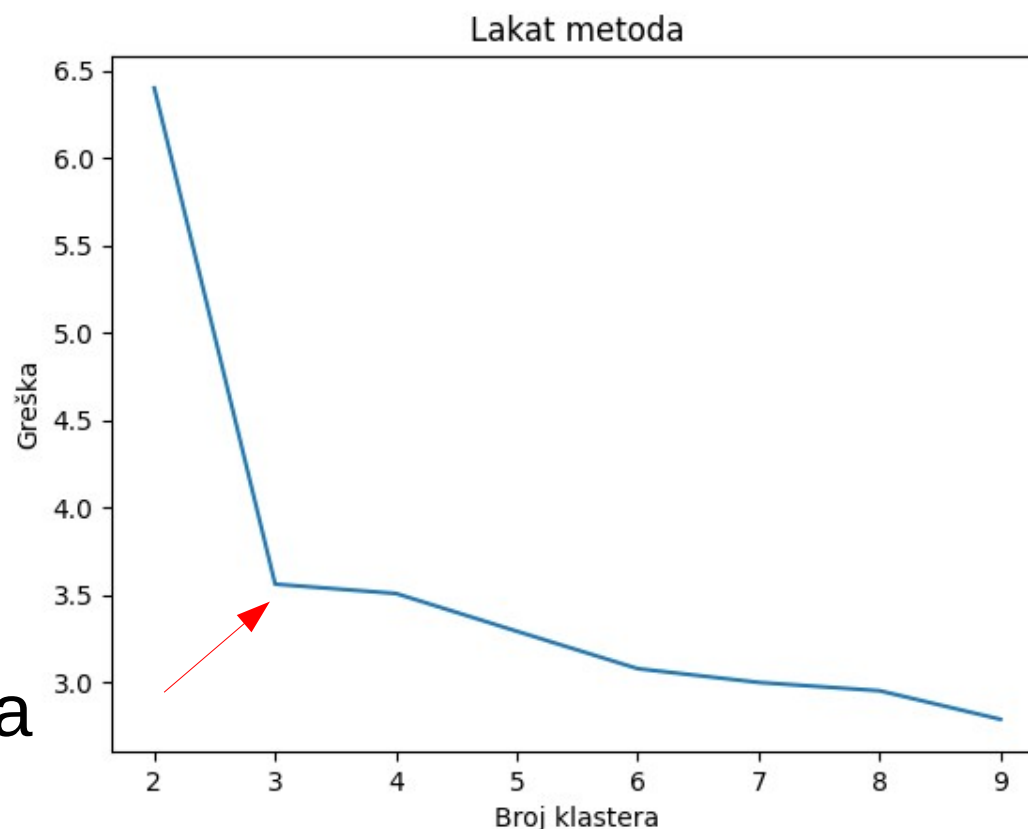
- Uradimo klasterizaciju
- Problem: klasteri se ne mogu uočiti okom
  - Nije moguće nacrtati jedan scatterplot
  - Koristi se više boxplot-ova za uvid u vrednosti promenljivih
- Koliko klastera je optimalno?





# Primer 3

- Proba se na više načina, sa npr. 2 do 9 klastera
- Za svako od rešenja se izračuna greška pri klasterizaciji
- Grafik sa greškom klasterizacije
  - „Lakat“(elbow)metoda
  - Optimalno 3 klastera



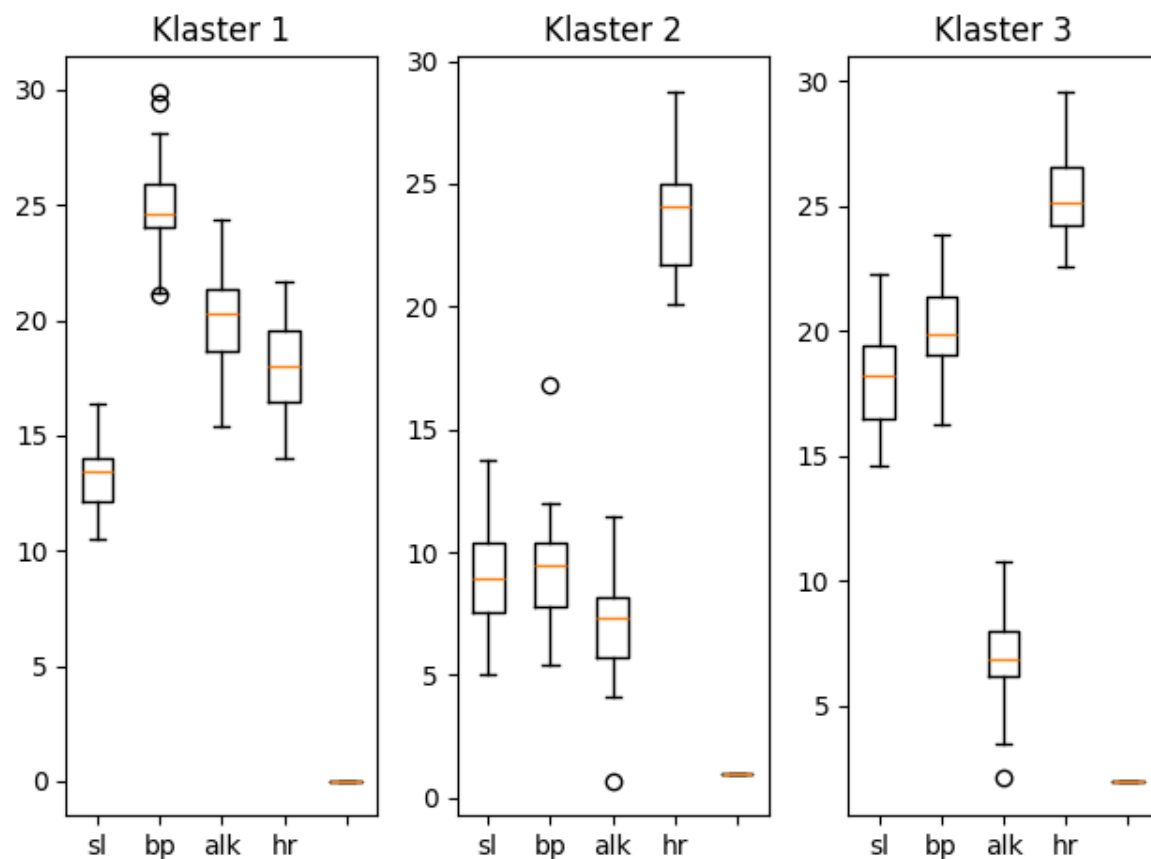
# Primer 3

- Tri klastera

1. Porodice koje puno piju (klaster 1)
2. Porodice koje se zdravo hrane (klaster 2)
3. Porodice koje vole puno da jedu i piju (klaster 3)

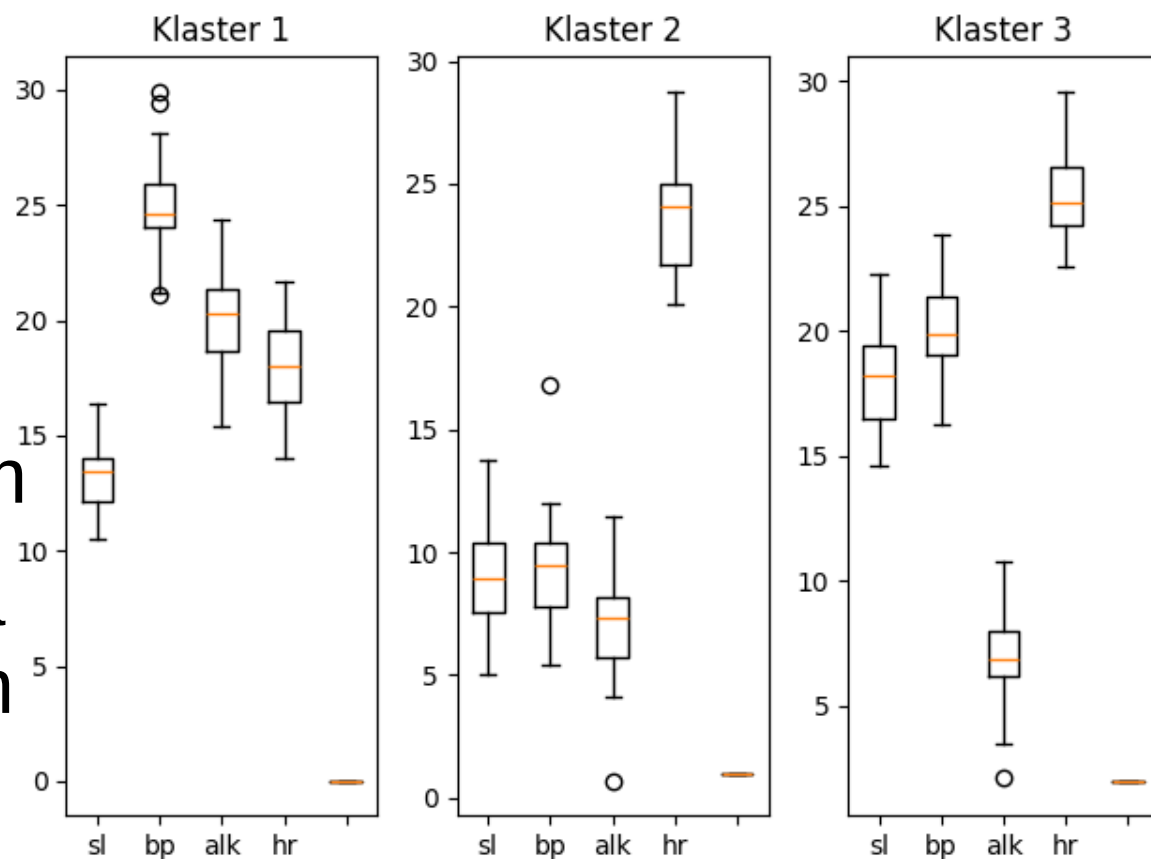
- Prikaz

- Nije moguć jedan scatterplot
- više uporednih boxplot-ova




# Primer 3

- Više promenljivih
- Više rešenja
- Klasteri nisu potpuno odvojeni
- Klasterne nije moguće uočiti okom
- Neophodna analiza klastera algoritmom
- Prilično realan primer




# Klasterizacija - karakteristike



- Eksploratorna analiza podataka
  - Tek se utvrđuje da li postoje neke grupe ili ne
  - Obično se koristi da se neko upozna sa podacima
- Nenadgledano mašinsko učenje (unsupervised)
  - Ne postoji „tačno“ ili „uzorno“ rešenje
  - Ne postoji trening set i set za validaciju

# Klasterizacija - ograničenja



- Klasterizaciju ima smisla primeniti kada su:
  - Numerički podaci u pitanju
  - Višedimenzijski podaci (bar dve promenljive)
  - Podaci neistraženi (nisu poznate klase, zakonitosti)
- Često postoje i neke pretpostavke ili iskustveni predosećaji koje želimo proveriti klasterizacijom
  - Npr. pretpostavljamo da postoje tri različite grupe gostiju u nekom restoranu



# Oblasti primene



- Segmentacija tržišta
- Uočavanje grupa u društvenim mrežama
- Identifikacija korisnika koje karakterišu slični oblici interakcije sa sadržajima nekog Web sajta/aplikacije
- Grupisanje objekata (npr., slika/dokumenata) radi lakše i efektivnije pretrage
- ...

A yellow pencil and a pink eraser are positioned in the top right corner of the white paper, suggesting a drawing or writing activity.

Kako klasterizacija funkcioniše?

# Kako klasterizacija funkcioniše?



- „Slične“ instance – šta to znači?
- Pojam udaljenosti
  - Euklidska, Menhetn (city block), kosinusna...
- Metode klasterizacije
  - KMeans
  - Hijerarhijska klasterizacija
  - ...

# Šta to znači „slične“ instance?



- Sličnost se izračunava korišćenjem neke mere udaljenosti ili sličnosti:
  - Udaljenost dve instance (Euklidska ili Manhattan)
  - Sličnost dve instance (kosinusna sličnost ili koeficijent korelacije)

# Euklidska i Manhattan udaljenost

- Euklidska udaljenost (crvena)

$$d = \sqrt{\sum_{j=1}^n (x_{sj} - x_{tj})^2}$$

- Manhattan udaljenost (plava)

$$d = \sum_{j=1}^n |x_{sj} - x_{tj}|$$

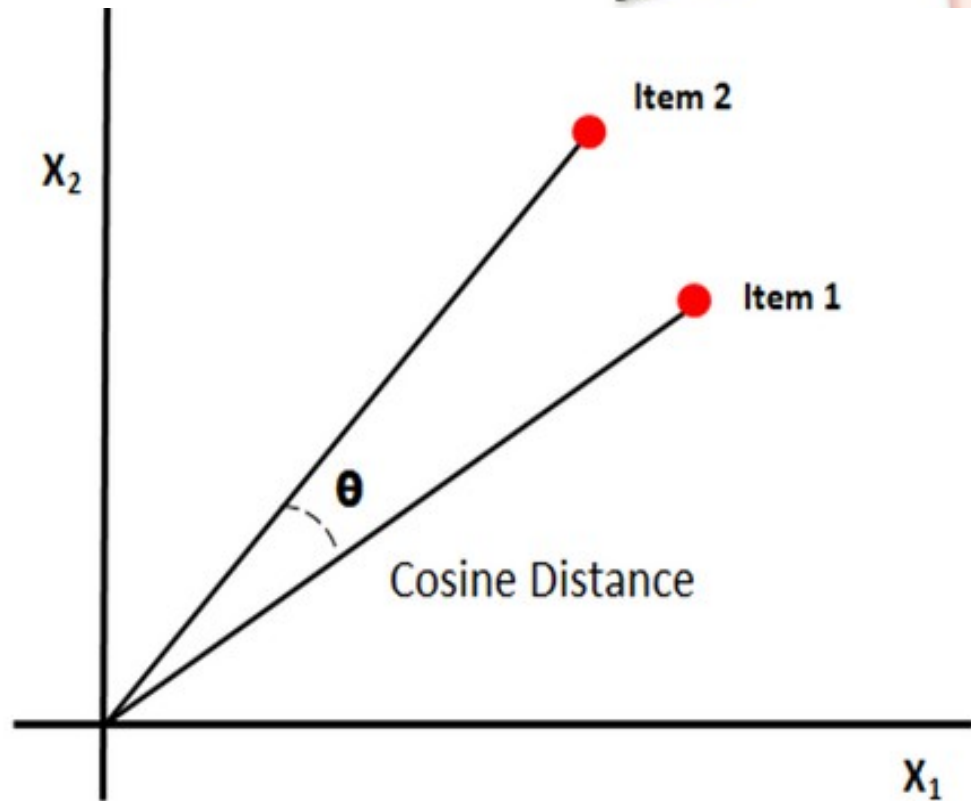
n – broj atributa kojima su instance opisane





# Kosinusna sličnost

- Kosinus ugla između vektora povučениh od koordinatnog početka do dve instance (tačke)
- Kosinusna sličnost
- Kosinusna udaljenost (1-kosinusna slič.)



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

# Procena rezultata klasterizacije



- Ocena uspešnosti modela je dosta teža nego kod nadgledanog mašinskog učenja
- Ovde nemamo precizne metrike koje nedvosmisleno ukazuju na to koliko je model “dobar”

# Procena rezultata klasterizacije



- Pod “dobrim” rešenjem se podrazumeva model koji:
  - Dobro deli instance u nepreklapajuće grupe (klasterne) (objektivna procena)
  - Koristan je za dati zadatak / problem zbog koga se klasterovanje i radi (subjektivna procena)

# Procena rezultata klasterizacije



- Neki od objektivnih kriterijuma za procenu kvaliteta klastera:
  - Međusobna udaljenost težišta
    - što su težišta dalje jedno od drugog, to je stepen preklapanja klastera manji, i njihov kvalitet viši
  - Max udaljenost instanci u okviru istog klastera
  - Min udaljenost instanci iz različitih klastera
  - Suma kvadrata unutar klastera
    - suma kvadrata odstupanja instanci u okviru klastera od težišta klastera
  - *Veličina svakog klastera (broj instanci)?*

# Procena rezultata klasterizacije



- Problem: ne postoje metrike koje ukazuju na to koliko je neko rešenje sveukupno dobro, odnosno korisno za dati zadatak
- Subjektivna procena korisnosti klastera za dati domen i zadatak je značajnija od opisanih objektivnih metrika
- Domensko znanje presudno za evaluaciju, tj. izbor optimalnog skupa klastera



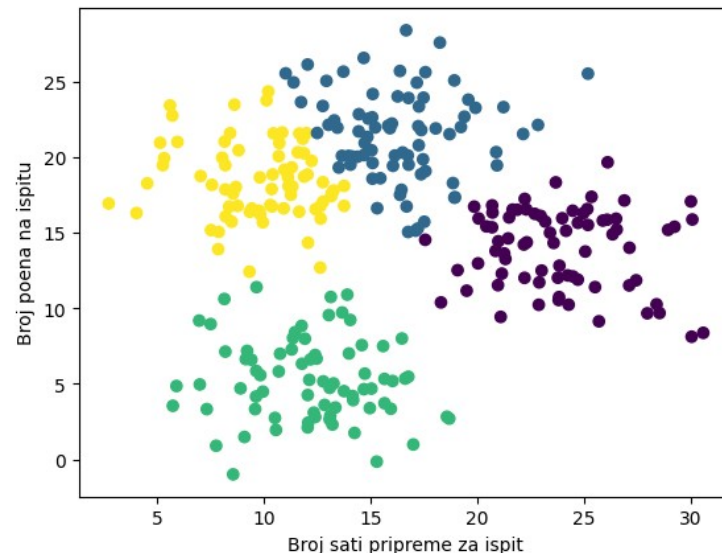
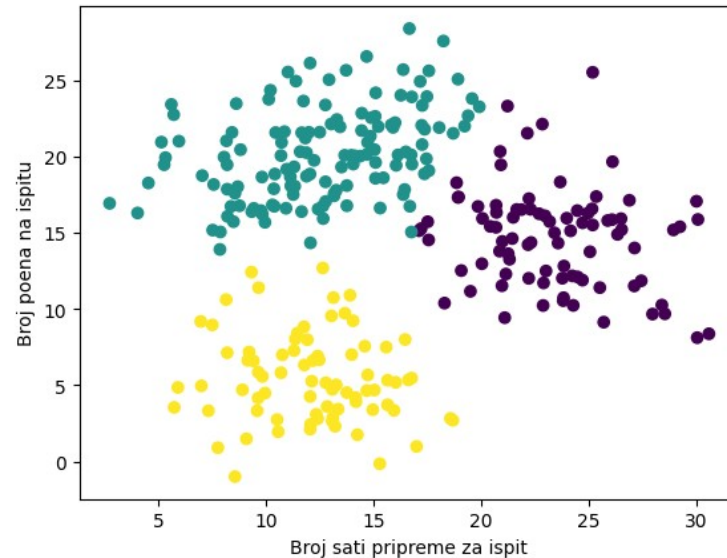
# Procena rezultata klasterizacije

- Rešenje sa tri klastera ili

- 1) Malo učili, slabo uradili (žuta)
- 2) Malo ili srednje učili, dobro uradili (zelena)
- 3) Puno učili, osrednje uradili (ljubičasta)

- Rešenje sa četiri klastera?

- 1) Malo učili, slabo uradili (zelena)
- 2) Malo učili, dobro uradili (žuta)
- 3) Srednje učili, dobro uradili (plava)
- 4) Puno učili, osrednje uradili (ljubičasta)



# Klasterizacija - problemi



- Kategorijski podaci (nisu numerički) i razdaljina
  - Ordinalni, uređeni (npr. zadovoljan, onako, nezadov.)
  - Nominalni, neuređeni (pol, boja očiju...)
- Moguća rešenja (prednosti i mane):
  - Ordinalni - pretvaranje u ordinalnu numeričku skalu
    - Zadovoljan – 2, onako – 1, nezadaovoljan – 0
  - Nominalni – one hot encoding (0 vektor sa jednom 1)
    - 2 vrednosti → binarna numerička: ženski – 1, muški – 0
    - Više (N) vrednosti → one hot encoding (vektor dimenzije N)
    - Npr. boja očiju: plave [0,0,1], zelene [0,1,0], braon [1,0,0]

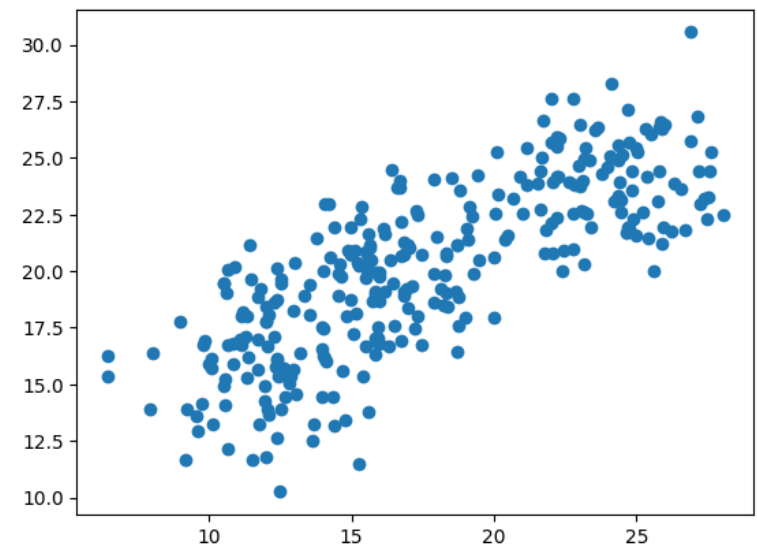
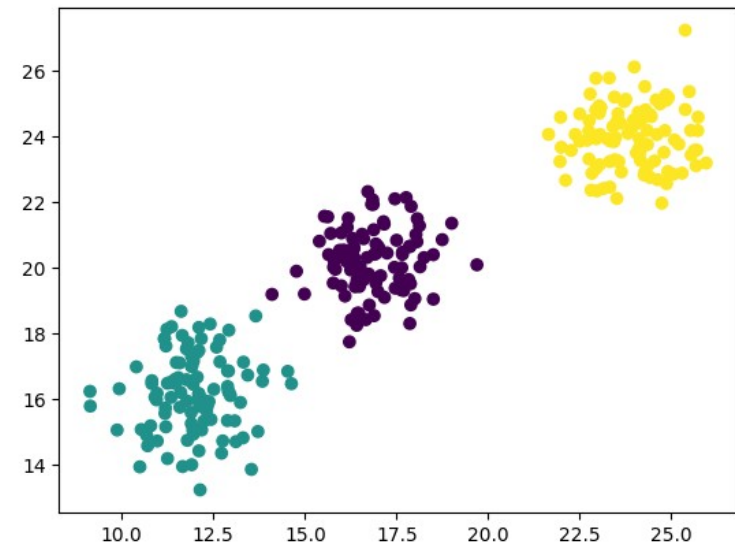
# Klasterizacija - problemi



- Nedostajući podaci (NaN)
  - Zbog lošeg merenja, dizajna istraživanja, više sile...
  - Nije moguće odrediti udaljenost instance od drugih
- Moguća rešenja (prednosti i mane):
  - Rad sa parcijalnim skupom podataka
    - Izbacivanje celih instanci (redova) sa NaN
    - Izbacivanje promenljivih koje imaju mnogo NaN
  - Ubacivanje vrednosti umesto NaN
    - Prosečna vrednost (mean) ili medijana umesto NaN
    - Imputacija vrednosti

# Klasterizacija - problemi

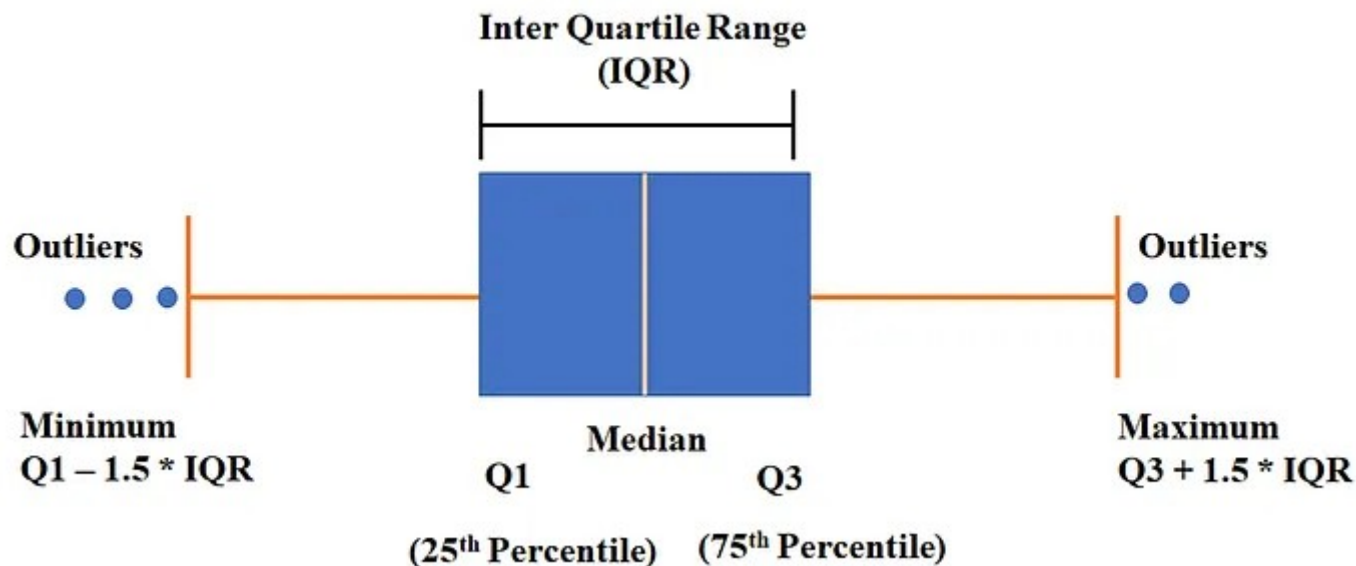
- Korelacija promenljivih
  - Ako je par promenljivih visoko korelisan, to utiče na udaljenost (može da bude problem)
  - U ekstremnom slučaju, dobija se samo jedan klaster
- Moguće rešenje:
  - Izbacivanje jedne visoko korelisane promenljive (iz svakog para) iz analize





# Klasterizacija - problemi

- Netipične/ekstremne vrednosti (eng. „outliers“)
  - Ekstremna vrednost neke promenljive u nekoj instanci može da „povuče“ ceo klaster na neku stranu





# Klasterizacija - problemi



- Moguće rešenje:
  - Izbacivanje promenljive (ako ima npr.  $> 10\%$  outlier-a)
  - Zamena ekstremnih vrednosti nekim drugim (winsorize/winsorization metoda)
    - Obično se biraju percentili kao donje i gornje granice (npr. 5% i 95% ako ima outliera na obe strane)
    - Pronađu se vrednosti iz skupa podataka koje odgovaraju tim percentilima.
    - Outlier-i koji su iznad se zamene vrednošću 95% percentila
    - Outlier-i koji su ispod se zamene vrednošću 5% percentila
    - Ponovo se proveriti da li ima outlier-a i, ako ih ima, ponovi se ceo proces sa drugim percentilima

# Klasterizacija - problemi



- Različite skale promenljivih
  - Ako je jedna promenljiva u rasponu od 1 do 100 a druga od 0 do 1, vrednost prve promenljive će dominantno uticati na ukupnu udaljenost.
- Moguće rešenje:
  - Normalizacija (svođenje na skalu 0 do 1).

# Klasterizacija - postupak



1) Učitavanje podataka

2) Inicijalni izbor promenljivih (objektivan i subjektivan)

3) Priprema podataka

1) Transformacija kategorijskih promenljivih u numeričke

2) Provera nedostajućih vrednosti (NaN)

- Ako ih ima, izbacivanje celih instanci ili zamena nedostajućih vrednosti (više načina)

3) Provera korelacije promenljivih

- Ako ima korelacije, izbacivanje po jedne promenljive iz svakog para

4) Provera ekstremnih vrednosti promenljivih (outliers)

- Ako ih ima, izbacivanje celih instanci ili zamena ekstremnih vrednosti (više načina)

5) Provera raspona vrednosti promenljivih

- Ako su rasponi različiti (ili u svakom slučaju) uraditi normalizaciju

# Klasterizacija - postupak



4) Izbor metode klasterizacije i parametara

5) Izvršavanje izabrane metode klasterizacije

6) Procena rezultata klasterizacije

1) Procena prema objektivnim kriterijumima

- Međusobna udaljenost težišta, obično metodom sume kvadrata unutar klastera

2) Procena prema subjektivnim kriterijumima

- Koliko dobijeni klasteri imaju smisla (na osnovu prethodnog znanja i iskustva)

3) Vraćanje na korake 4 i 5 ako rezultati procene nisu dobri (druga metoda i/ili parametri)





This work is licensed under a Creative Commons  
Attribution-ShareAlike 3.0 Unported License.  
It makes use of the works of Mateus Machado Luna.

