

Zadaća 3

Amina Alagić, Nejra Rovčanin, Ema Rudalija

*Mašinsko učenje
Univerzitet u Sarajevu
Elektrotehnički fakultet Sarajevo, Odsjek računarstvo i informatika
Zmaja od Bosne bb, Kampus Univerziteta, 71000 Sarajevo*

Sažetak— Fokus ovog rada je primjena klasterizacije koristeći algoritme k-means i PAM k-medoids, te procjeni validacije klasteringa.

Ključne riječi— klasterizacija, tendencija, validacija, k-means, k-medoids

I. UVOD

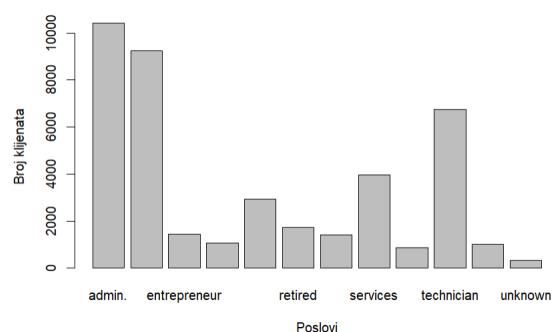
Dataset kojeg smo odabrale za analizu je dataset ‘bankmarketing’. Podaci se odnose na direktne marketinške kampanje jedne portugalske bankarske institucije. Sadrži 41 118 observacija i 21 varijablu od kojih je 10 kategoričkih atributa, 10 numeričkih i 1 atribut klase, te ovom strukturom naš dataset zadovoljava tražene specifikacije za daljni rad zadaće 3.

II. ANALIZA SETA PODATAKA

A. Odrediti tipove i distribuciju atributa;

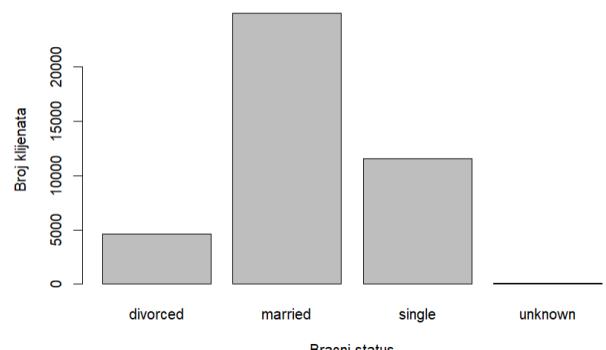
Distribuciju kategoričkih atributa smo odredili pomoću barplot-a. Deset kategoričkih atributa dataseta prikazano je u nastavku.

job - predstavlja tipove poslova i ima sljedeće vrijednosti:
(‘admin.’,’blue-collar’,’entrepreneur’,’housemaid’,’management’,’retired’,’self-employed’,’services’,’student’,’technician’,’unemployed’,’unknown’)



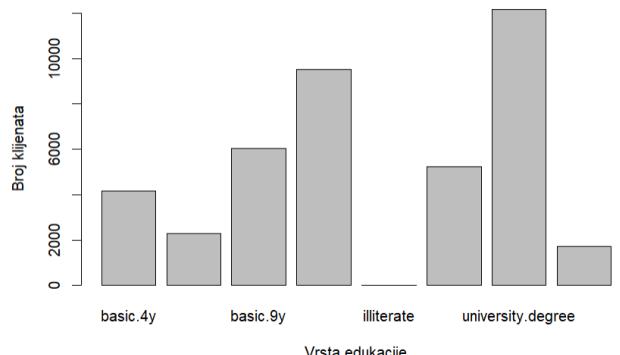
Sl. 1 Prikz barplota atributa job

marital - predstavlja bračni status i ima sljedeće vrijednosti: ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)



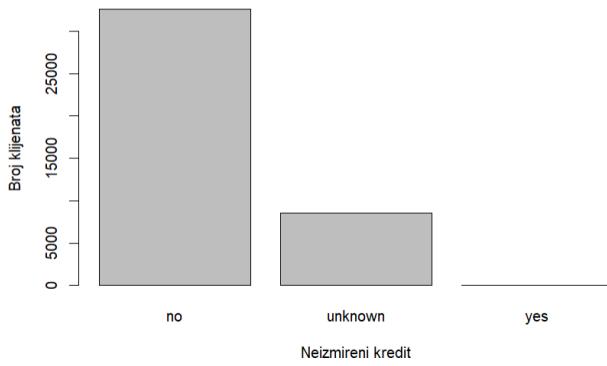
Sl. 2 Prikz barplota atributa marital

education - predstavlja tip edukacije koju osoba poseduje i ima sljedeće vrijednosti: ('basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')



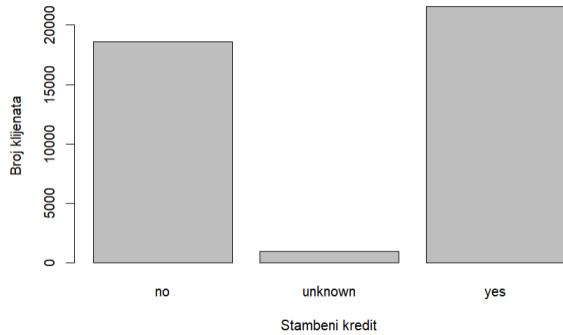
Sl. 3 Prikz barplota atributa education

default - predstavlja informaciju da li osoba ima kredit u kašnjenju/da li ima neizmireni kredit?, i ima sljedeće vrijednosti: ('no','yes','unknown')



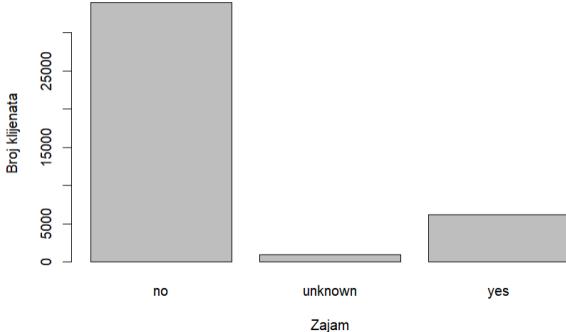
Sl. 4 Prikaz barplota atributa default

housing - predstavlja podatak da li klijent ima stambeni kredit?, i ima sljedeće vrijednosti: ('no', 'yes','unknown')



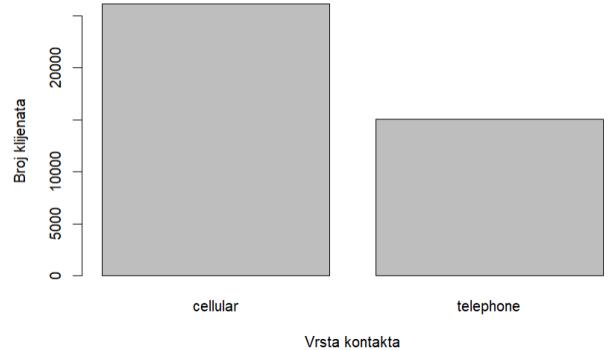
Sl. 5 Prikaz barplota atributa housing

loan - predstavlja podatak da li klijent ima lični kredit?, i ima sljedeće vrijednosti: ('no', 'yes', 'unknown')



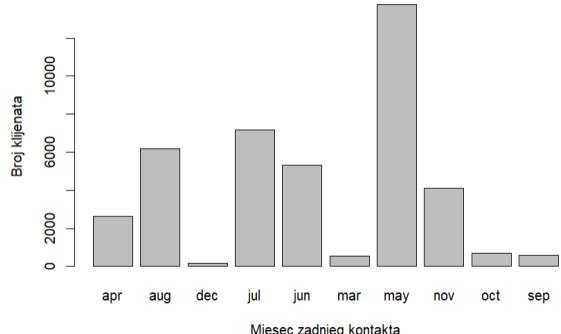
Sl. 6 Prikaz barplota atributa loan

contact - predstavlja tip komunikacije banke i klijenta i ima sljedeće vrijednosti: ('cellular', 'telephone')



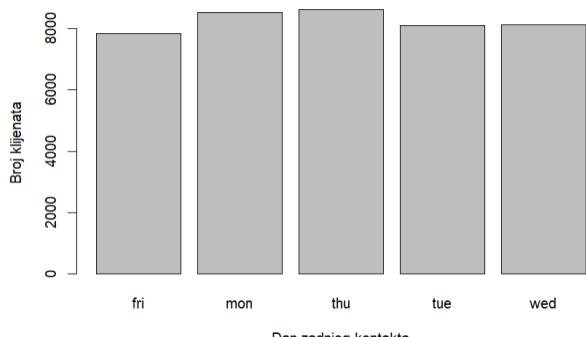
Sl. 7 Prikaz barplota atributa contact

month - predstavlja posljednji kontakt mjesec u godini i ima sljedeće vrijednosti: ('jan', 'feb', 'mar', ..., 'nov', 'dec')



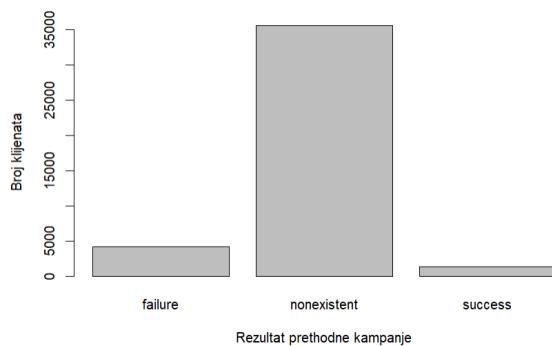
Sl. 8 Prikaz barplota atributa month

day_of_week - predstavlja posljednji dan kontakta sa klijentom i ima sljedeće vrijednosti: ('mon','tue','wed','thu','fri')



Sl. 9 Prikaz barplota atributa day_of:week

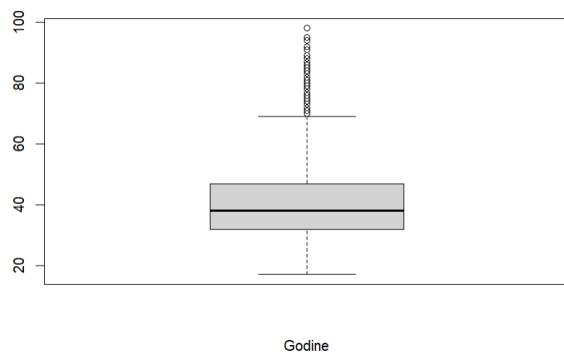
poutcome - predstavlja rezultat prethodne marketinške kampanje i ima sljedeće vrijednosti ('failure','nonexistent','success')



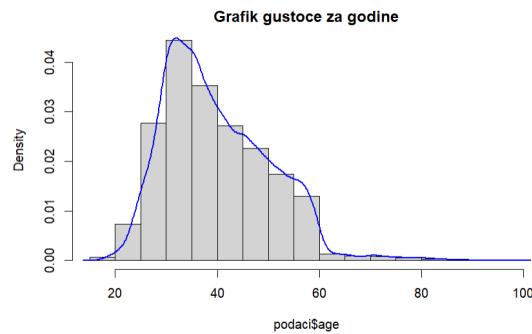
Sl. 10 Prikaz barplota atributa poutcome

Distribuciju numeričnih atributa smo odredili pomoću boxplot-a i histograma. Desert numeričkih atributa dataseta prikazano je u nastavku.

age - godine klijenta

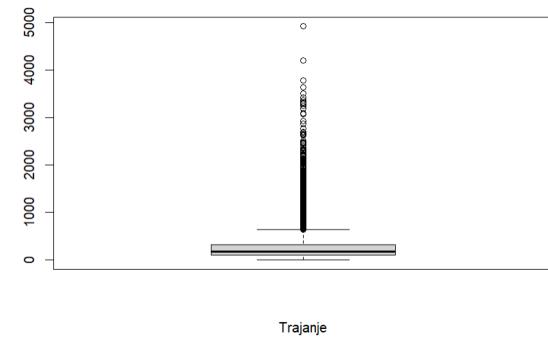


Sl. 11 Prikaz boxplota atributa age

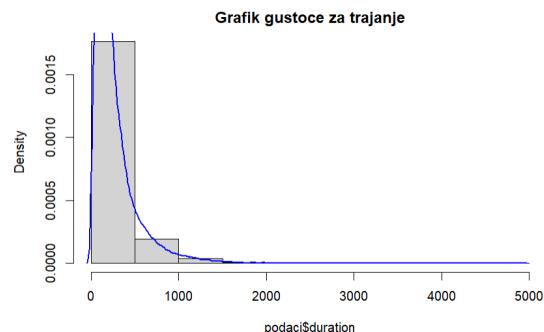


Sl. 12 Prikaz histograma atributa age

duration - trajanje posljednjeg kontakta, u sekundama.

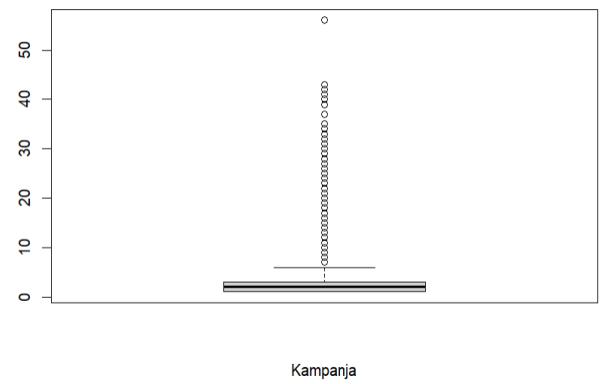


Sl. 13 Prikaz boxplota atributa duration

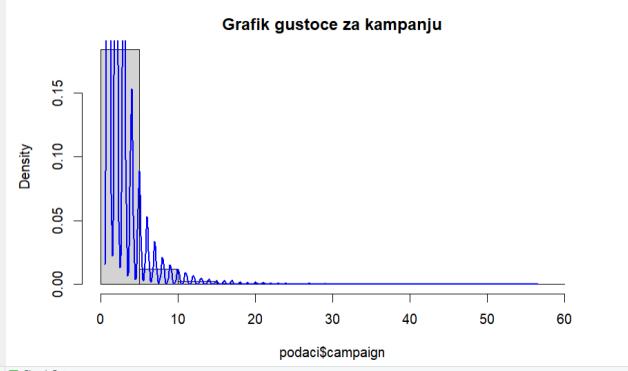


Sl. 14 Prikaz histograma atributa duration

campaign - predstavlja broj ostvarenih kontakata tokom ove kampanje za ovog klijenta, pri čemu sadrži i uključuje zadnji kontakt

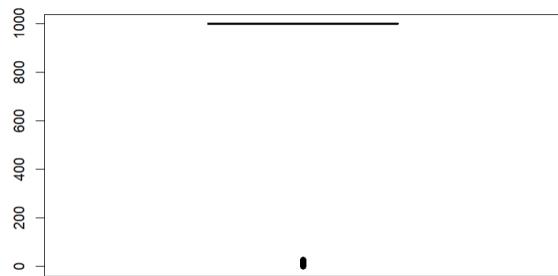


Sl. 15 Prikaz boxplota atributa campaign



Sl. 16 Prikaz histograma atributa campaign

pdays - predstavlja broj dana koji je prošao nakon što je klijent posljednji put kontaktiran iz prethodne kampanje (vrijednost 999 znači da klijent nije prethodno kontaktiran)

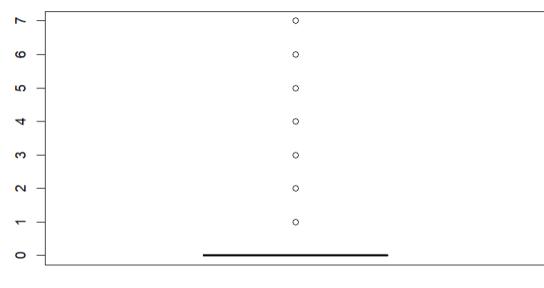


Sl. 17 Prikaz boxplota atributa pdays



Sl. 18 Prikaz histograma atributa pdays

previous - predstavlja broj ostvarenih kontakata prije ove kampanje za ovog klijenta

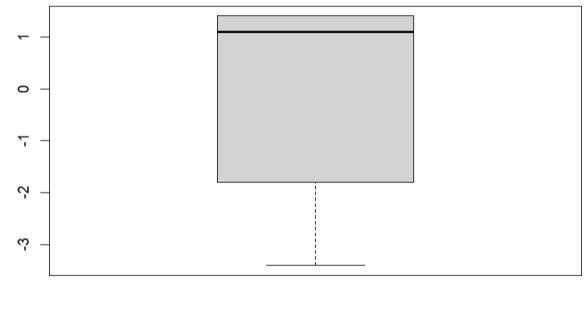


Sl. 19 Prikaz boxplota atributa previous

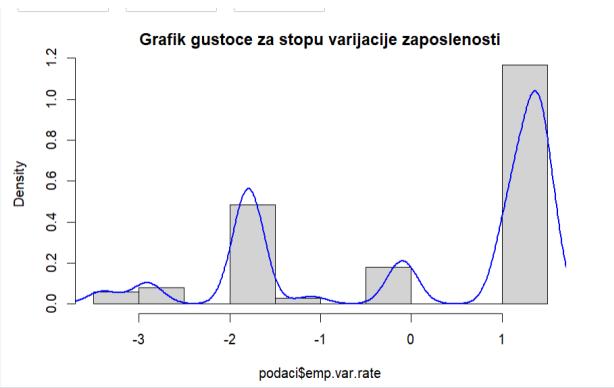


Sl. 20 Prikaz histograma atributa previous

emp.var.rate - stopa varijacije zaposlenosti, kvartalni indikator

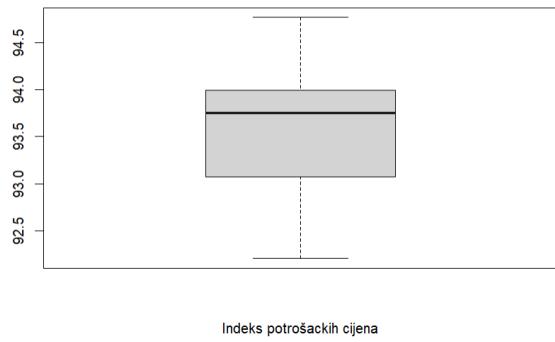


Sl. 21 Prikaz boxplota atributa emp.var.rate

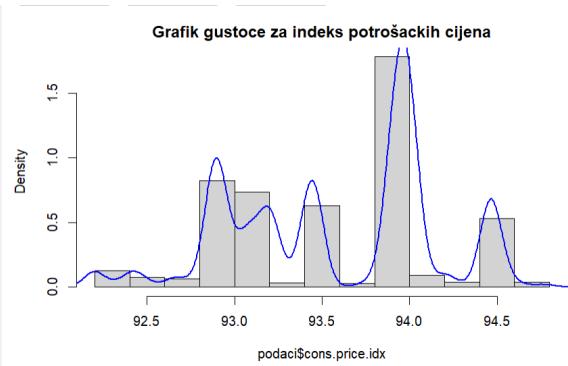


Sl. 22 Prikaz histograma atributa emp.var.rate

cons.price.idx - indeks potrošačkih cijena, mjesечni indikator

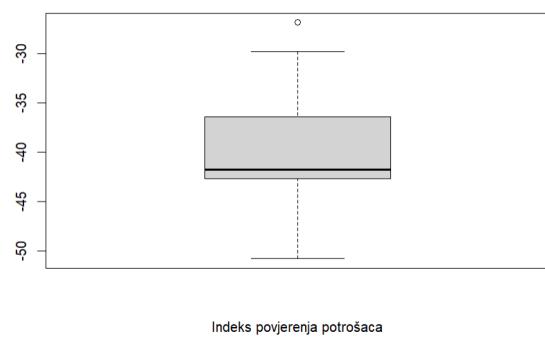


Sl. 23 Prikaz boxplota atributa cons.price.idx

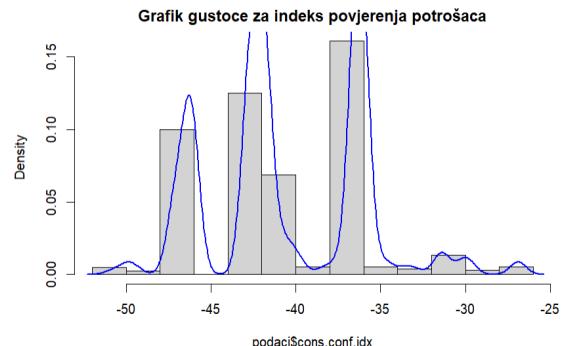


Sl. 24 Prikaz histograma atributa cons.price.idx

cons.conf.idx - indeks povjerenja potrošača, mjesечni indikator

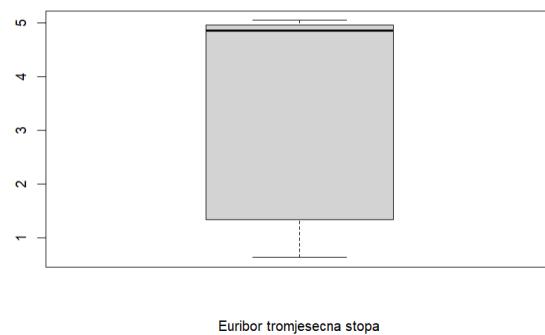


Sl. 25 Prikaz boxplota atributa cons.conf.idx

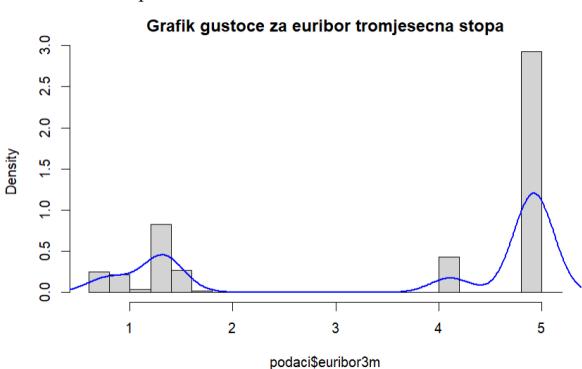


Sl. 26 Prikaz historama atributa cons.conf.idx

euribor3m - euribor tromjesečna stopa, dnevni indikator

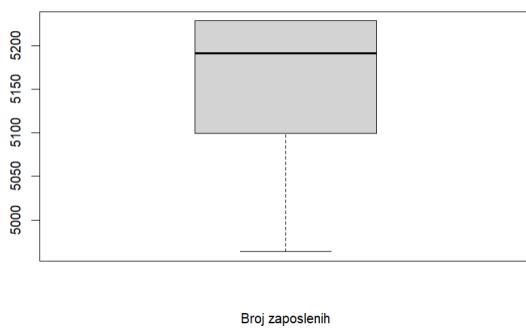


Sl. 27 Prikaz boxplota atributa euribor3m

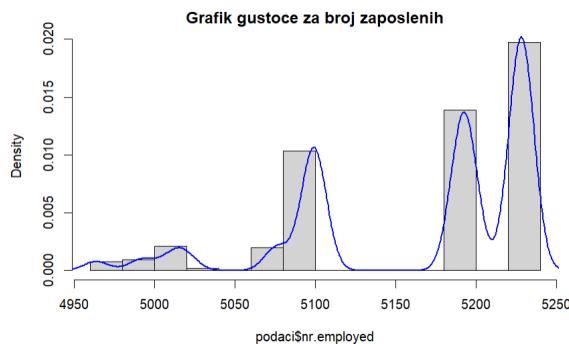


Sl. 28 Prikaz histograma atributa euribor3m

nr.employed - broj zaposlenih, kvartalni pokazatelj



Sl. 29 Prikaz boxplot atributa nr.employed



Sl. 30 Prikaz histogram atributa nr.employed

B. Ispuniti nedostajuće i ukloniti nepodobne vrijednosti ukoliko je potrebno;

Nedostajućih i NA vrijednosti nema u našem datasetu. Što se tiče nepodobnih vrijednosti, uklonili smo outlier iz sljedećih numeričkih kolona: duration (uklonili sve vrijednosti veće od 500), zatim campaign (uklonili sve vrijednosti veće od 2) i pdays (uklonili sve vrijednosti manje od 999). Također uklonili smo i kolonu vrijednost klase.

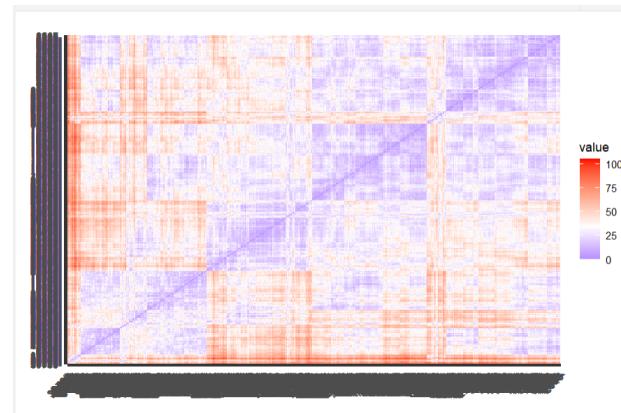
C. Uraditi transformaciju i skaliranje atributa u skladu sa njihovom distribucijom i algoritmima koji će se koristiti u narednim koracima;

Od transformacija podataka, najprije smo sve kategoričke kolone faktorizirali, a zatim pretvorili u numeričke, koristeći funkcije factor i as.numeric respektivno. Zatim smo prateći boxplotove odlučili

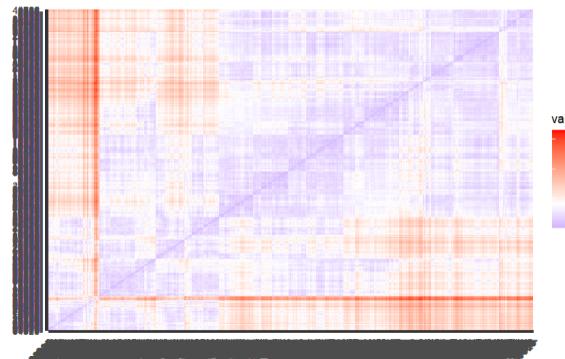
na sljedeće transformacije: decimalna skaliranja kolona: duration, pdays, cons.price.idx, i nr.employed, te min-max normalizaciju kolona: emp.var.rate i cons.conf.idx. Na kraju smo naš dataset odlučili smanjiti, radi lakšeg izvršavanja klasterizacije. Kako bismo dobili randomizirano selektiranje koji redovi ostaju a koji ne iz dataseta, korištena je houldout metoda, s odnosom 90:10, pri čemu je manji data set, naš novi dataset. Nakon cjelokupnog preprocesiranja podataka, sada umjesto 41188 observacija i 21 kolonu, imamo 1496 observacija i 20 atributa.

D. Odrediti da li set podataka ima klastering tendenciju;

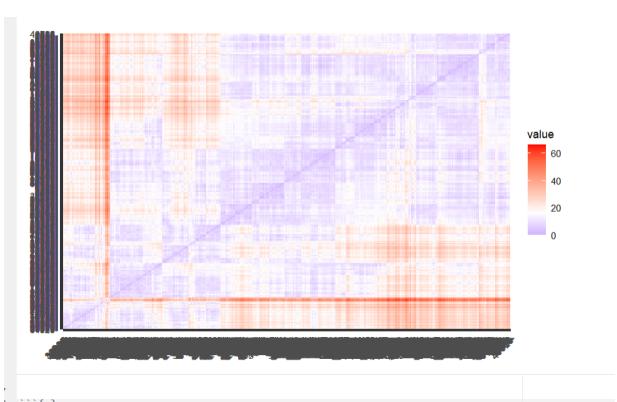
Klastering tendenciju smo odredili koriteći VAT algoritam (Visual Assessment of Tendency). Ispod je izvršen pirkaz nad originalnim podacima, četiri najčešće korištene metrike (Manhattan, Euklidska, Minkowski i Pearson distanca).



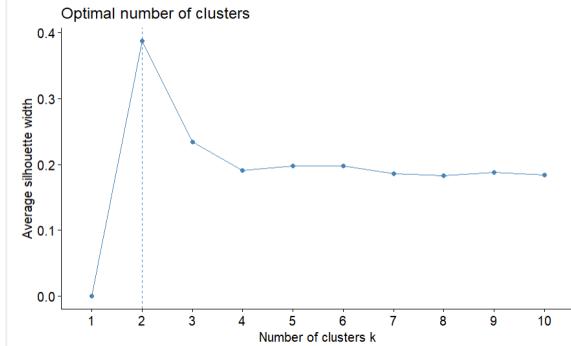
Sl. 31 Prikaz manhattan korištene metrike



Sl. 32 Prikaz euklidske korištene metrike

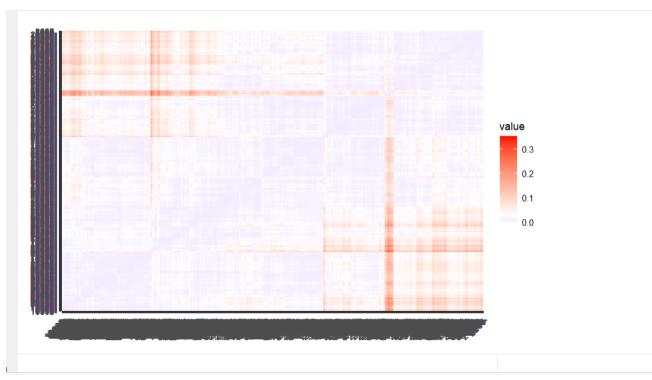


Sl. 32 Prikaz minkowski korištene metrike



Sl. 33 Prikaz silhouette metode

Vidimo da je optimalan broj klastera $k = 2$. Zatim smo primjenili PAM k-medoids algoritam i izvršili vizualizaciju klastera.



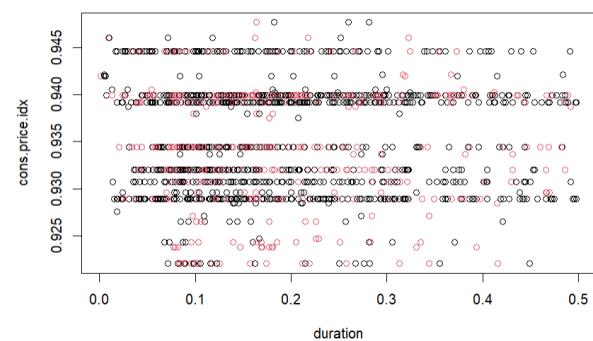
Sl. 33 Prikaz pearson korištene metrike

E. Utvrditi da li je potrebno izvršiti dodatne analize za podatke, kao i dodatne transformacije, skaliranja i sl.

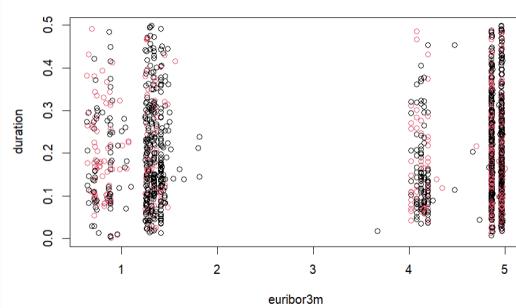
Primijetimo da metrika manhattan daje najbolju tendenciju za klasterizaciju. Nije izvršeno dalje analiziranje podataka.

III. PRIMIJENITI PAM K-MEDOIDS ALGORITAM

Koristeći VAT algoritam, primjećujemo da nam je najbolje koristiti Manhattan distancu. Optimalan broj klastera smo dobili koristeći Silhouette metoda.



Sl. 34 Prikaz plota za PAM k-medoids

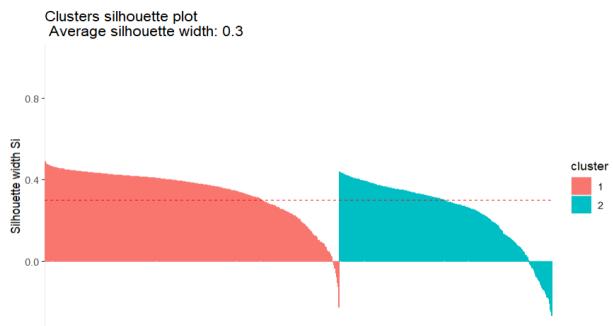


Sl. 35 Prikaz plota za PAM k-medoids

Validaciju PAM k-medoidsa smo vršili pomoću Silhouette koeficijenta i čistoće klastera.

Srednja vrijednost silhouette metrike za PAM (manhattan metrika): 0.06914166

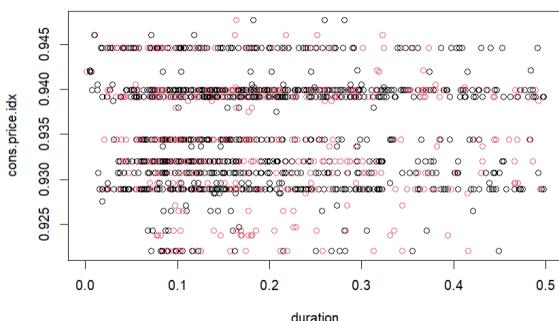
Sl. 36 Prikaz silhouette koeficijenta za PAM k-medoids



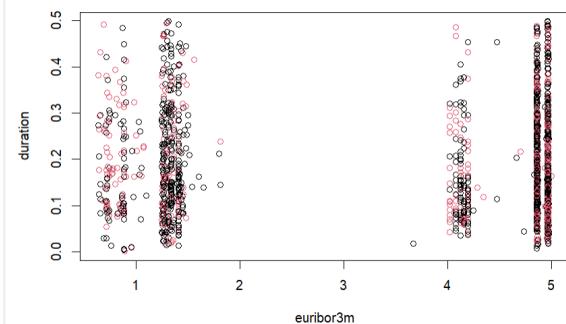
Sl. 37 Prikaz čistoće klastera za PAM k-medoids

IV. PRIMIJENITI K-MEANS ALGORITAM

Za ovaj algoritam smo također iskoristili manhattan distancu i broj klastera 2. Rezultat klasterizacije pokazali smo na sljedeća dva plota.



Sl. 38 Prikaz plota za k-means

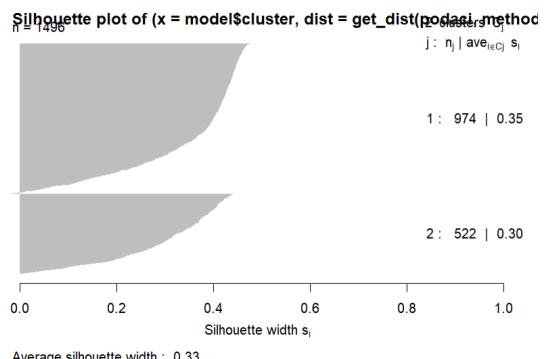


Sl. 39 Prikaz plota za k-means

Validaciju kmeans algoritma smo također vršili pomoću Silhouette koeficijenta i čistoće klastera.

Srednja vrijednost silhouette metrike za kmeans: 0.3316732

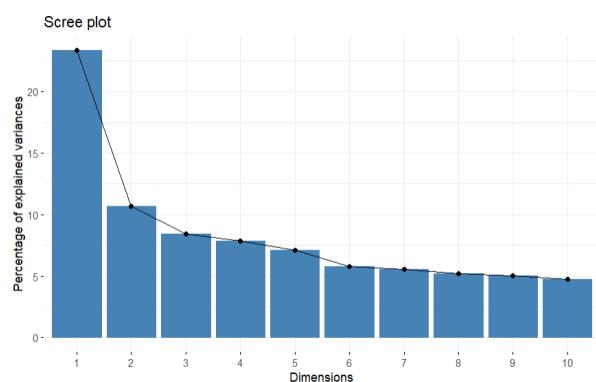
Sl. 40 Prikaz silhouette koeficijenta za k-means



Sl. 41 Prikaz čistoće klastera za k-means

V. SMANJITI DIMENZIONALNOST SETA PODATAKA KORISTEĆI PCA ANALIZU, A ZATIM PRIMIJENITI PAM KMEDOIDS I K-MEANS ALGORITME

Dimenzijsalnost algoritma smo postigli kao i na vježbama. Koristeći PCA metodu, dobivamo 19 principalnih komponenti. Na screeplotu ispod vidimo da nam prva komponenta nosi najveću količinu informacija i znanje.



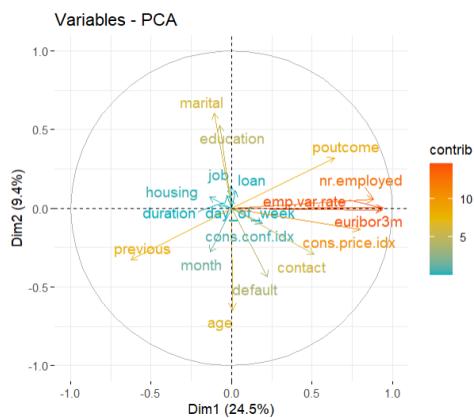
Sl. 41 Prikaz plota principijalnih komponenti

Koristeći funkciju summary, dobivamo analizu bitnosti svake od komponenti prikazanih na sljedećoj slici. Vidimo da kumulativnu proporciju od 83% dobivamo već nakon 10 komponenti. Također vidimo da komponentu 19 možemo i izbaciti iz datasets bez gubljenja i jedne informacije.

	PC9	PC10	PC11	PC12	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	0.9805	0.94756	0.93257	0.83548	2.1064	1.4262	1.2663	1.22261	1.16407	1.04536	1.02947	0.99352
Proportion of Variance	0.2335	0.1071	0.0844	0.07867	0.07132	0.05751	0.05578	0.05195	0.0506	0.04726	0.04577	0.03674
Cumulative Proportion	0.2335	0.3406	0.4250	0.50363	0.57495	0.63247	0.68825	0.74020	0.7908	0.83805	0.88382	0.92056
	PC13	PC14	PC15	PC16	PC17	PC18	PC19					
Standard deviation	0.78761	0.74897	0.48103	0.26211	0.1444	0.08425	4.358e-16	0.03265	0.02952	0.01218	0.00362	0.0011
Proportion of Variance	0.03265	0.02952	0.01218	0.00362	0.0011	0.00037	0.000e+00	0.95321	0.98274	0.99491	0.99853	0.9996
Cumulative Proportion	0.95321	0.98274	0.99491	0.99853	0.9996	1.00000	1.000e+00					

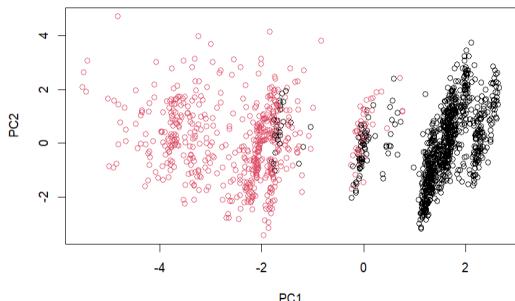
Sl. 42 Prikaz sažetka značenja principijalnih komponenti

Koristeći funkciju fviz_pca_var vidljivo je da najveći značaj pri formiranju principalnih komponenti, a samim tim i najmanju korelaciju s drugim atributima imaju atributi emp.var.rate i euribor3m.

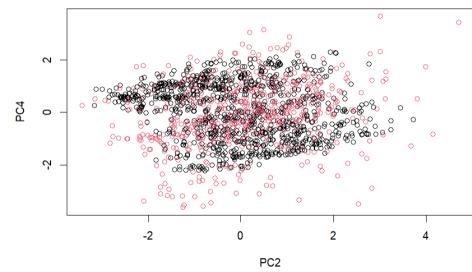


Sl. 43 Prikaz značaja atributa pri formiranju PCA komponenti

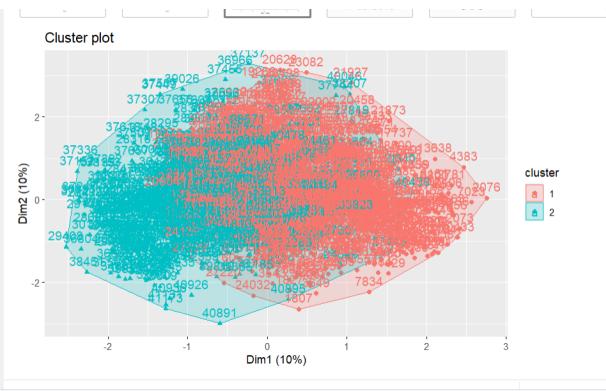
PAM k-medoids algoritam primjenjen nad principalnim komponentama



Sl. 44 Prikaz PAM k-medoids algoritma nad dvije komponente



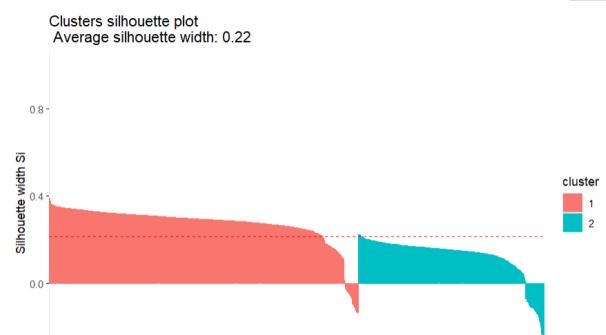
Sl. 44 Prikaz PAM k-medoids algoritma nad dvije komponente



Sl. 45 Prikaz kreiranih klastera

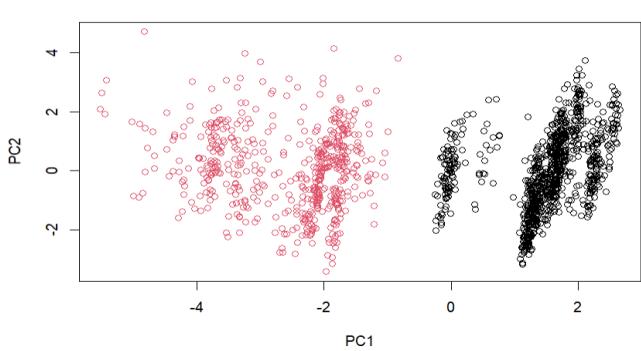
Srednja vrijednost silhouette metrike za PAM (euclidean metrika): 0.2161278

Sl. 46 Prikaz silhouette metrike za PAM k-medoids

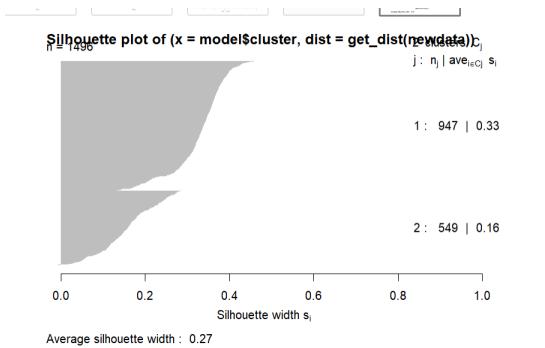


Sl. 47 Vizualni prikaz silhouette metrike za PAM k-medoids

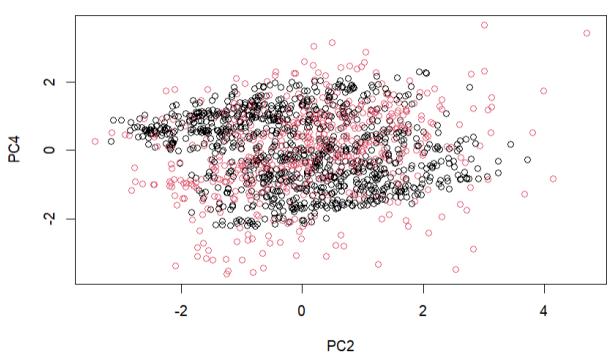
K-means algoritam primjenjen nad principalnim komponentama



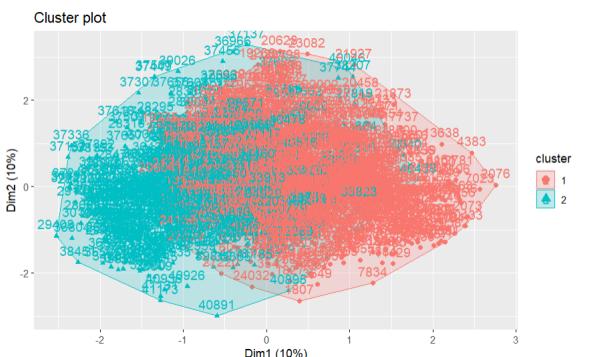
Sl. 48 Prikaz k-means algoritma nad dvije komponente



Sl. 52 Vizualni prikaz silhouette metrike za k-means



Sl. 49 Prikaz k-means algoritma nad dvije komponente



Sl.50 Prikaz kreiranih klastera

Srednja vrijednost silhouette metrike za kmeans: 0.2660183

Sl. 51 Prikaz silhouette metrike za k-means

Nakon što smo prikazali izgled klasterizacije nad principalnim komponentama, sa slika se može vidjeti da su se kreirala dva klastera, te da postoji preklapanje među njima.

Također, vrijednosti silhouette metrike su nam veće od 0, te ukazuju na pozitivnu tendenciju clusteringa i veliku udaljenost od centroida drugih clustera. Ovim zaključujemo da su klasteri uspješno formirani.