Izveštaj
Domaći 3-klasterizacija

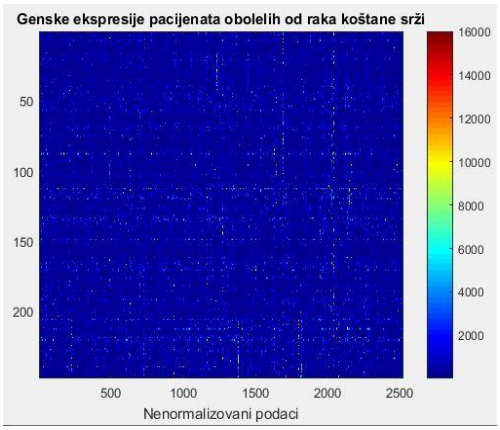
Ljiljana Popović, EE72/2014, biskvitcokolada@gmail.com

A. Baza podataka

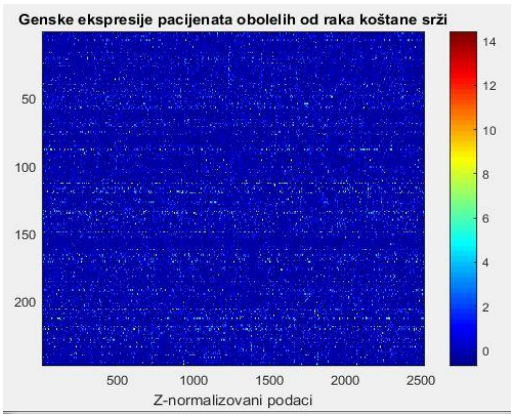
Baza podataka pod nazivom „yeoh_data“ predstavlja podatake genskih ekspresija pacijenata obolelih od raka koštane srži. Date su takođe i labele koje ukazuju na 6 podtipova raka. Baza se sastoji od 248 uzoraka i 2526 atributa. Atributi predstavljaju gene i svi su numerički. Baza podataka je potpuna, nema nedostajućih podataka. Ovakva baza može poslužiti u medicinske svrhe, za lakše tumačenje, vizuelizaciju ili predviđanje bolesti raka.

B. Zadatak 1-Vizuelizacija podataka

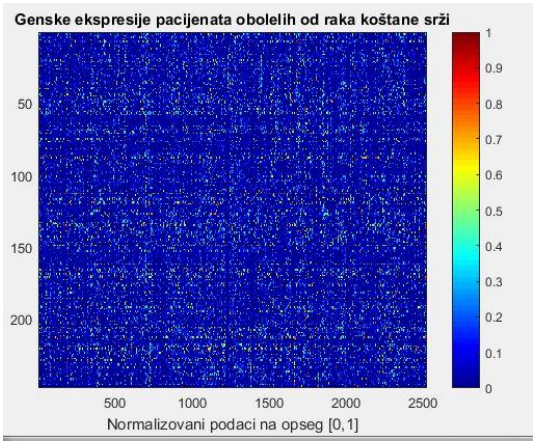
Potrebno je nad datim setom podataka izvršiti normalizaciju i time napraviti tri zasebna skupa podataka. Zatim prikazati te podatke kao slike radi lakše vizuelizacije, i to, prikazati originalan skup nenormalizovanih podataka, zatim, skup koji sadrži z-normalizovane podatke i skup koji sadrži podatke normalizovane na opseg [0,1]. To je ilustrovano sledećim slikama.



Sl. 1. Nenormalizovani(originalni) podaci.



Sl. 2. Z-normalizovani podaci.



Sl. 3. Podaci normalizovani na opseg [0,1].

Podaci su dobijeni čipom *Affymetrix* i imaju mnogo neujednačene i male opsege vrednosti ekspresija. Stoga, dobijene slike nisu preterano ilustrativne pa se ne može jasno odrediti neka pravilnost među podacima, te ova vizuelizacija nije značajna.

C. Zadatak 2- K-means

K-means algoritam je jednostavan algoritam klasterizacije, koji na iterativan način minimizuje varijacije unutar klastera. Ovaj algoritam zavisi od početne inicijalizacije i neophodno je definisati broj klastera na početku, na koje će se uzorci deliti. Shodno tome, za broj klastera se postavlja unapred poznat broj klasa-6. Za ovaj algoritam se koristi ugradjena MATLAB-ova funkcija *kmeans*. Algoritam primenjujemo nad svim napravljenim skupovima podataka (3), kroz 15 iteracija, za dve vrste rastojanja: *sqeuclidean* i *cosine* i dobijamo ukupno 90 particija. Neophodno je čuvati sve dobijene particije jer svaka od njih, za svaki uzorak, čuva oznaku labele klastera. Rezultati dobijeni ovim algoritmom biće ilustrovani box plot slikama i diskutovani u delu E.

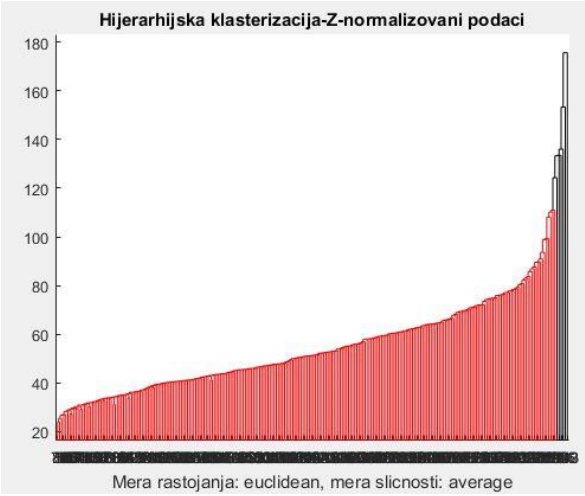
D. Zadatak 3- Hijerarshijska klasterizacija

Metoda hijerarhijske klasterizacije koja se u ovom zadatku primenjuje je poznata kao *Aglomerativna* klasterizacija ili metoda *odozdo na gore* i zasnovana je na spajanju klastera. Algoritam počinje od toga da svaki uzorak predstavlja zaseban klaster, a onda grupisanjem najbližih, dolazi do jedinstvenog klastera u kom su svi uzorci. Potom se bira pogodno mesto za „presecanje veza“ i dobija se određen broj klastera.

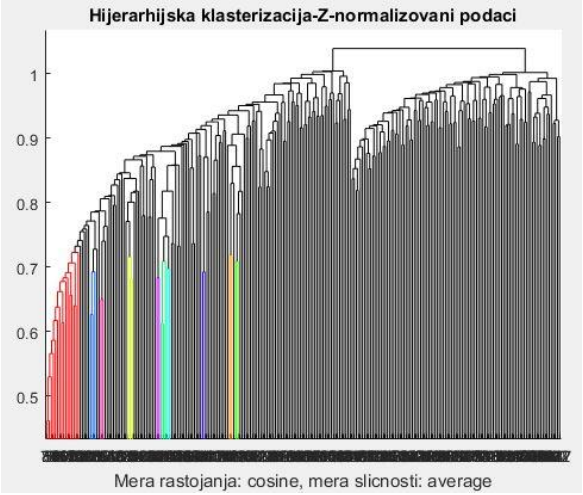
Po zahtevu zadatka, algoritam je takođe primenjen na sva tri skupa podataka. Kao bliskost između dva klastera

koriste se mere sličnosti- *linkage* i zahteva se klasterovanje za *average*(prosečnu) i *single*(minimalnu) meru sličnosti, kao i za dve mere rastojanja: *euclidean* i *cosine.* Kao rezultat, dobijeno je 12 različitih particija, koje se takođe čuvaju. Rezultati će biti ilustrovani i diskutovani u delu E.

Hijerarhijska klasterizacija može biti prikazana pomoću dendrograma. To je binarno stablo koje ukazuje na strukturu klastera i pored toga obezbeđuje i meru sličnosti među klasterima. Na slici 4 i 5 su ilustrovani neki od zanimljivijih dendrograma koji su dobijeni ovim algoritmom.



Sl. 4. Dendrogram z-normalizovanih podataka za *average linkage* i meru rastojanja *euclidean*.



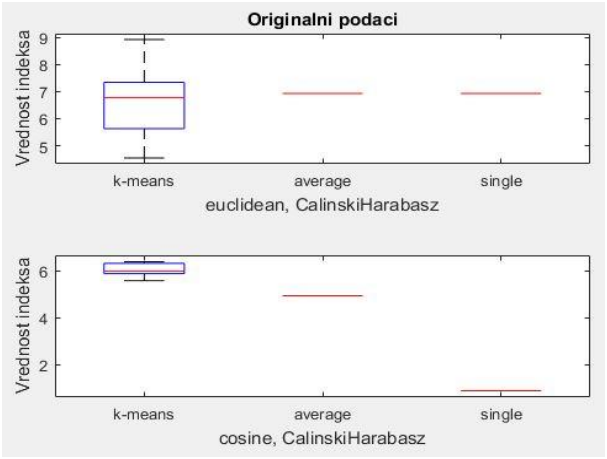
Sl. 5. Dendrogram z-normalizovanih podataka za *average linkage* i meru rastojanja *cosine*.

Može se zaključiti da je *cosine* rastojanje „bolje“, u smislu da za njega dobija realniji dendogram. Za *euclidean* rastojanje je dobijeno da ima 5 klastera sa po jednim uzorkom i da svi ostali uzorci pripadaju 6. klasteru, što najverovatnije i nije dobar rezultat. Dok se za *cosine* rastojanje dobija naizgled bolje rešenje. Klasteri su malo bolje raspoređeni.

Prednost algoritma hijerarhijske klasterizacije u odnosu na k-means se ogleda u ponovljivosti i činjenici da ne zahteva unapred broj klastera, ali je kompleksniji.

E. Zadatak 4-Interne mere validacije

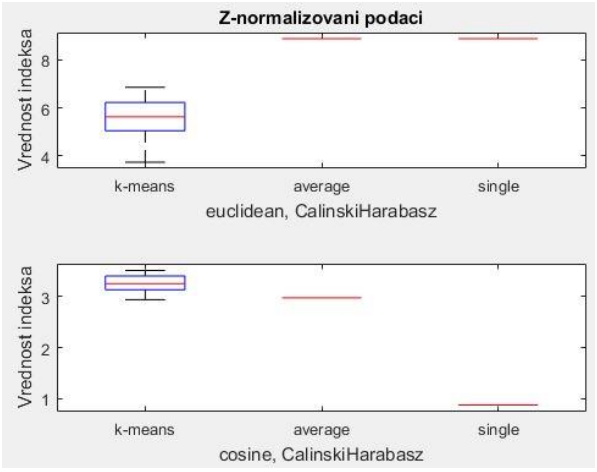
Potrebno je za svaku dobijenu particiju iz prethodna dva zadatka (ukupno 90+12), izračunati interne mere validacije, i to: *CalinskiHarabasz*, *DaviesBouldin* i *Silhouette*. Ovo se postiže ugrađenom funkcijom *evalclusters.* Upoređeni su rezultati dobijeni pomoću tri različita algoritma za sva tri skupa podataka, različite interne indekse i različite mere rastojanja. Poređenja su ilustrovana sledećim box plot graficima.



Sl. 6. Originalni podaci - *CalinskiHarabasz* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

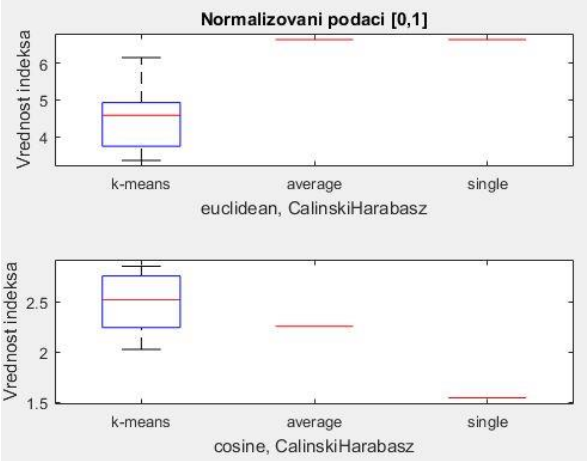
CalinskiHarabasz indeks koristi pretpostavku da dobra klasterizacija ima karakteristiku male sume kvadrata rastojanja u okviru klastera, odnosno velike između različitih klastera tj. veće vrednosti su bolje.

Shodno tome, vidi se da je za skup originalnih podataka *euclidean* rastojanje dalo bolji rezultat klasterovanja jer su vrednosti indeksa veće. Za sva tri algoritma su preko 6, dok su za *cosine* rastojanje 6 ili ispod. Takođe se primećuje da za *euclidean* rastojanje hijerarhijska klasterizacija *average* i *single* daje približne vrednosti medijani k-means-a, ali da je opseg vrednosti koji uzima k-means značajano veliki i da podaci dosta osciluju, dok kod *cosine* rastojanja to nije slučaj. Kod ovog rastojanja je k-means algoritam dao bolji rezultat klasterizacije.



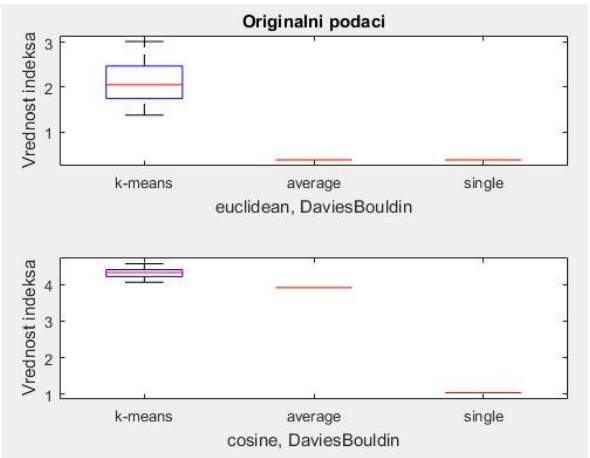
Sl. 7. Z-normalizovani podaci - *CalinskiHarabasz* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

Za skup z-normalizovanih podataka takođe je *euclidean* rastojanje generalno dalo bolji rezultat. Particije dobijene hijerarhijskom klasterizacijom *average* i *single* daju istu vrednost internog indeksa, koja je veća od čitavog opsega vrednosti koje daje k-means algoritam, te se zaključuje da je hijerarhijska klasterizacija ovde rezultovala boljom klasterizacijom. Za razliku od toga, kod kosinusnog rastojanja k-means daje bolji rezultat.



Sl. 8. Normalizovani [0,1] podaci - *CalinskiHarabasz* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

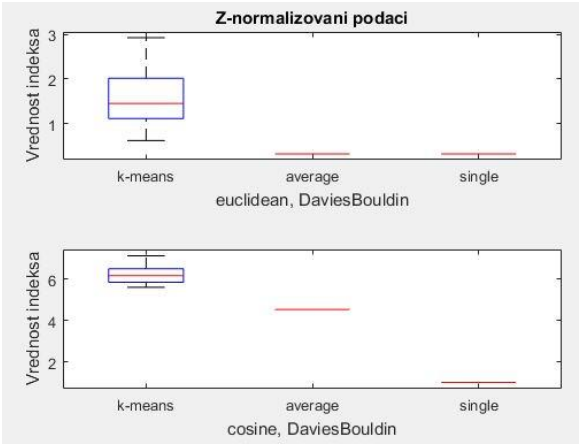
Što se tiče podataka koji su normalizovani na opseg [0,1], hijerarhijska klasterizacija *average* i *single* kod *euclidean* rastojanja, rezultuje istom vrednošću indeksa, većom od celog opsega vrednosti za k-means, i generalno većom od ostalih indeksa, kod oba rastojanja, te je zaključak da je bolja takva vrsta klasterizacije. Ukoliko posmatramo samo kosinusno rastojanje, k-means daje bolji rezultat, iako vrednosti opsega variraju.



Sl. 9. Originalni podaci - *DaviesBouldin* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

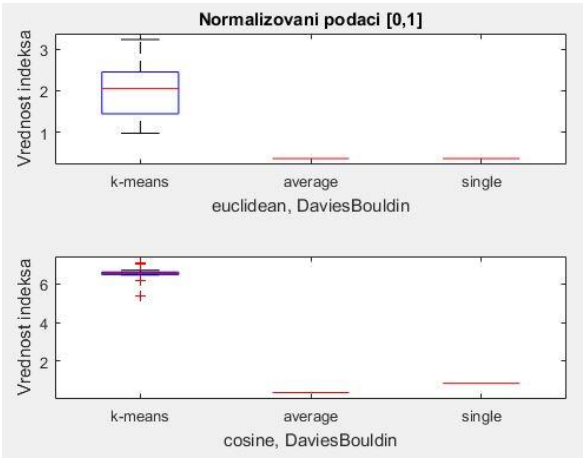
DaviesBouldin koristi pretpostavku da klasteri treba da budu kompaktni i da su međusobno što bolje razdvojeni tj. da su manje vrednosti bolje.

Shodno tome, ponovo zaključujemo da kod originalnog skupa podataka, *euclidean* rastojanje daje bolji rezltat, jer su sve dobijene vrednosti manje. Takođe se vrednosti za *average* i *single* poklapaju i manje su od čitavog opsega vrednosti za k-means, pa hijerarhijska klasterizacija rezultuje boljim klasterovanjem. Kod *cosine* rastojanja se pokazuje da je *single* algoritam rezultovao znatno boljom vrednošću indeksa od *average* metode i k-means-a, pa i boljom klasterizacijom. Takođe se ovde vidi da je opseg vrednost k-means-a jako uzan, kao i da je *average* vrednost približna medijani k-means opsega.



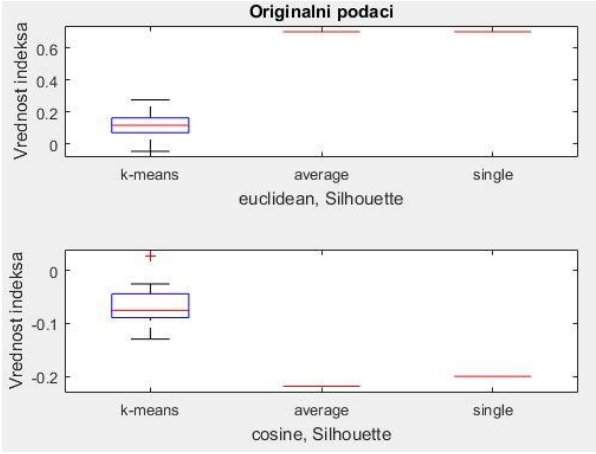
Sl. 10. Z-normalizovani podaci - *DaviesBouldin* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

Kod z-normalizovanih podataka sledi isti zaključak kao i kod *CalinskiHarabasz* indeksa za isti skup podataka, što se tiče euklidskog rastojanja. Međutim, za kosinusno rastojanje je *single* algoritam dao najbolji rezultat.



Sl. 11. Normalizovani [0,1] podaci - *DaviesBouldin* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

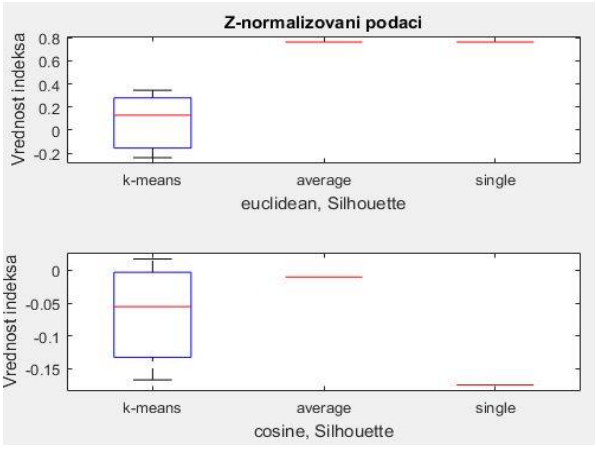
Za podtake normalizovane na opseg [0,1] se vidi da vrednosti za k-means sa euklidskim rastojanjem dosta variraju, dok za kosinusno pripadaju jako uskom opsegu. Osim toga, za njega postoje i outlier vrednosti. Ali generalno gledano se zaključuje da je hijerarhijsko klasterovanje za obe mere sličnosti rezultovalo boljom klasterizacijom, za oba rastojanja.



Sl. 12. Originalni podaci - *Silhouette* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

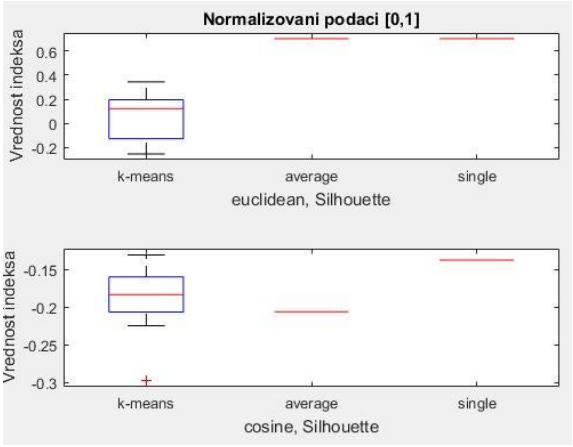
Silhouette indeks koristi pretpostavku da uzorci međusobno treba da budu bliskiji sa uzorcima iz svog kalstera u poređenju sa njihovom blizinom do uzoraka iz najbližeg susednog klastera. Računa se za svaki uzorak i kreće se u opsegu [-1,1]. Za celu particiju vrednost indeksa se dobija kao srednja vrednost indeksa za sve uzorke i veće vrednosti su bolje.

Na osnovu toga, za originalni skup podataka, euklidsko rastojanje daje bolji rezultat jer su sve vrednosti pozitivne, dok su za kosinusno rastojanje sve negativne, izuzev jednog outlier-a koji se javlja kod k-means algoritma. Generalno gledano, za euklidskko rastojanje su mere sličnosti hijerarhijskog klasterovanja dale veću vrednost indeksa i time i bolju klasterizaciju, dok za kosinusno, k-means daje veći indeks.



Sl. 13. Z-normalizovani podaci - *Silhouette* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

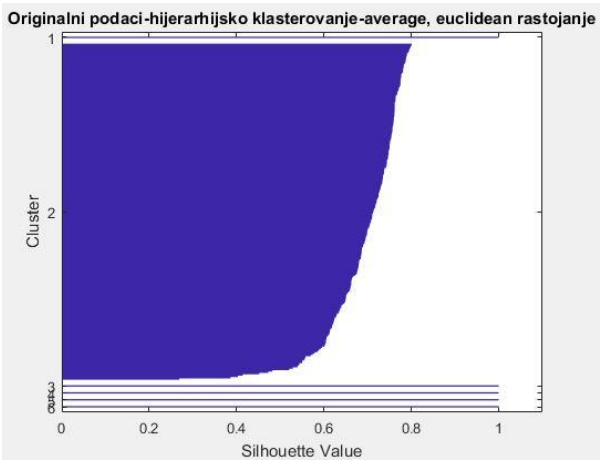
Kod z-normalizovanih podataka se primećuje da vrednosti za k-means algoritam variraju, ali da za euklidsko rastojanje *average* i *single* klasterizacija imaju iste vrednosti internog indeksa i veće od opsega vrednosti koje uzima k-means, pa shodno tome, daju i bolji rezultat. Za kosinusno rastojanje, vrednost indeksa koji daje *avergae* je približno jednak nuli i odstupa od vrednosti koje su dale ostala dva algoritma.



Sl. 14. Normalizovani [0,1] podaci - *Silhouette* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

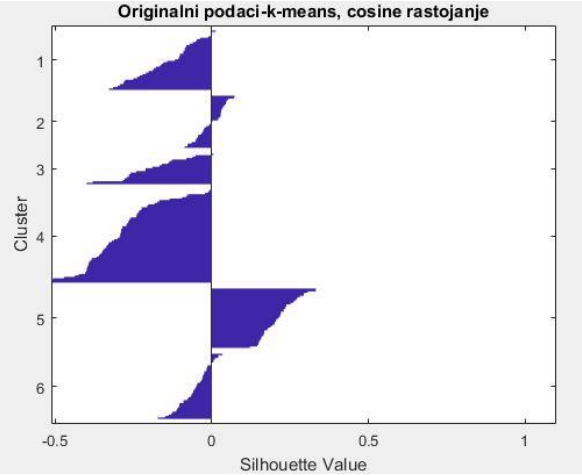
Kod podataka normalizovanih na opseg [0,1], za euklidsko rastojanje, vrednosti k-means-a dosta osciluju, ali hijerarhijsko klasterovanje daje veće vrednosti. Kod kosinusnog rastojanja, za k-means algoritam postoji jedan outlier koji daje najmanju vrednost i *single* mera sličnosti koja rezultuje najvećom vrednošću.

Potrebno je takođe dati slike *Silhouette* koeficijenata za sve dobijene particije. Na sledećim slikama su ilustrovani neki od dobijenih rezultata.



Sl. 15. Prikaz *Silhouette* koeficijenata za originalne podatke, algoritmom hijerarhijske klasterizacije-*average* i *euclidean* rastojanjem.

Primećuje se da su sve dobijene vrednosti pozitivne, što ukazuje da su ti uzorci dobro usklađeni sa svojim klasterima a slabo usklađeni sa susednim klasterima pa je i rešenje klasterovanja odgovarajuće, što se poklapa i sa rezultatima dobijenim za iste mere na slici 12. Međutim, hijerarhijsko klasterovanje daje neujednačene klastere, što smo videli i na dendrogramu prikazanom na slici 4. Dobija se da se u prvom, trećem, četvrtom i petom klaseru nalazi samo jedan uzorak, dok su svi ostali uzorci raspoređeni u 2. klaster što baš i nije najverovatniji slučaj.



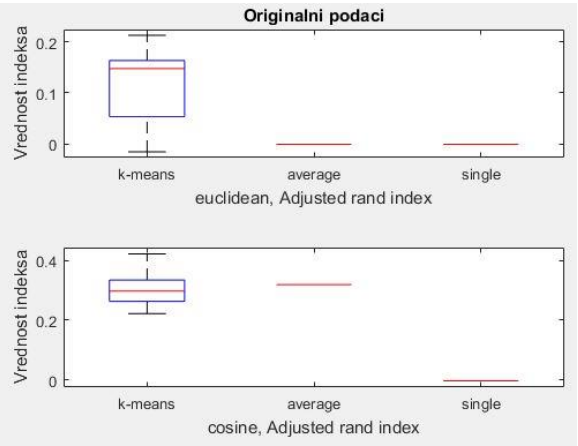
Sl. 16. Prikaz *Silhouette* koeficijenata za originalne podatke, k-means i *cosine* rastojanjem.

Sa slike vidimo da je za k-means algoritam većina vrednosti negativna, što odgovara uskom opsegu vrednosti dobijenom na slici 12 za isti slučaj. Stoga, ovakvo rešenje klasterovanja nije odgovarajuće. Međutim, k-means algoritam je dao bolje raspoređene klastere kao što je prikazano na dendrogramu na slici 5.

F. Zadatak 5-Eksterne mere validacije

U ovom zadatku se zahteva da se izračunaju eksterne mere validacije, takođe nad svim dobijenim particijama. Za ovakav proračun je iskorišćen priloženi kod „rand_index“, koji poredi originalne labele uzoraka sa labelama dobijenim klasterizacijom. Korišćen je *Adjusted rand index*, koji je korigovan *rand index* kako bi se izbegla slučajna podudaranja.

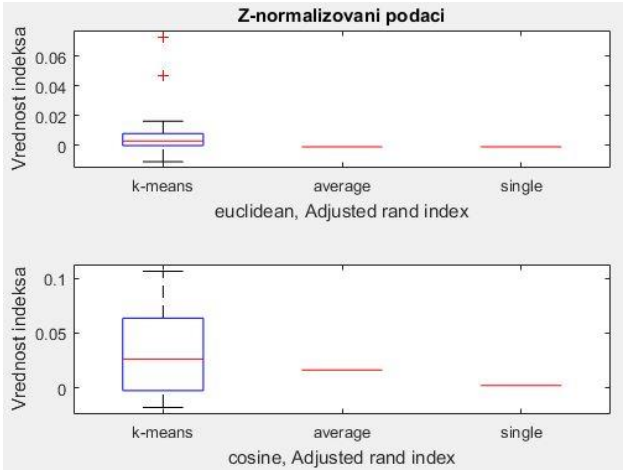
Ovaj indeks ukazuje na to koliko se dobro klasteri poklapaju sa stvarnim klasama. Rezultati su prikazani na box plot graficima ispod.



Sl. 17. Originalni podaci – *Adjusted rand index* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

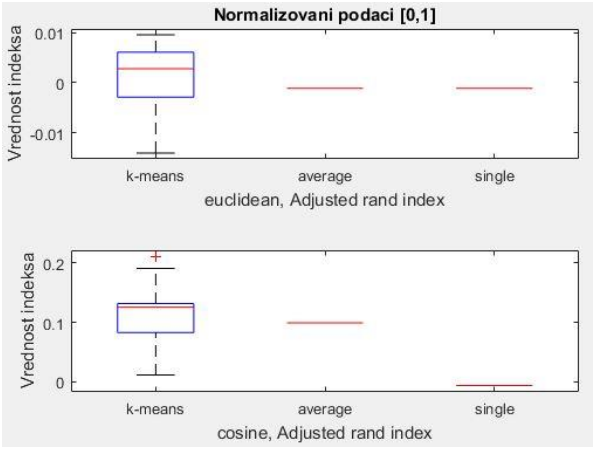
Za ovaj indeks, kod originalnih podataka se *single* mere poklapaju i imaju vrednost oko nule, dok *average* algoritam daje veći indeks za *cosine* rastojanje koji se poklapa sa medijanom k-means algoritma.

Opsezi vrednosti za k-means algoritam kod oba rastojanja se poklapaju ali su veće vrednosti kod *cosine* rastojanja. Sve vrednosti indeska su pozitivne.



Sl. 18. Z-normalizovani podaci – *Adjusted rand index* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

Za z-normalizovane podatke se dobijaju slični rezultati. Kod oba rastojanja, *average* i *single* vrednosti se poklapaju sa medianom k-means-a. Takođe k-means algoritam uzima malo veći opseg vrednosi sa kosinusnim rastojanjem u odnosu na onaj sa euklidskim ali euklidski sadrži i dve outlier vrednosti. Zaključuje se da su sve vrednosti slične i variraju oko nule.



Sl. 19. Normalizovani [0,1] podaci – *Adjusted rand index* a)*Euclidean* rastojanje b)*Cosine* rastojanje.

Što se tiče podataka normalizovanih na opseg [0,1], kod euklidskog rastojanja se pojavljuju negativne vrednosti, u jako malom opsegu, ali uglavnom variraju oko nule. Indeksi eksterne mere validacije za hijerarhijsko klasterovanje se poklapaju sa medianom k-means algoritma, što za kosinusno rastojanje nije slučaj. Kod njega su sve vrednosti pozitivne. Postoji pozitivan outlier, i *sinlge* algoritam daje najmanju vrednost, približno jednaku nuli.