Projekat 6

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Katarina Savičić
261/2016

avgust 2020.

1 Podaci

Podaci se nalaze u 7 datoteka: GSM3892570 – GSM3892576.

Svaku datoteku čine geni i ćelije, odnosno vrste i kolone (slika 1).

Vrednosti atributa su numeričke i trebalo bi da su nenegativne.

Index	AAACCTGAGCAGACTG-1	AAACCTGAGGTCGGAT-1	AAACCTGAGTGTACCT-1	AAACCTGAGTGTACTC-1	AAACCTGAGTTAAGTG-1
ENSG00000243485	0	0	0	0	0
ENSG00000237613	0	0	0	0	0
ENSG00000186092	0	0	0	0	0
ENSG00000238009	0	0	0	0	0
ENSG00000239945	0	0	0	0	0
ENSG00000239906	0	0	0	0	0
ENSG00000241599	0	0	0	0	0
ENSG00000279457	1	1	0	0	1
ENSG00000228463	0	0	0	0	0
.....	.....	.....	.....	.....	.....

Slika a

Cilj je pronaći klastere različitih vrsta mononuklearnih ćelija periferne krvi,

kojih ima ukupno 5.

2 Čišćenje podataka

Pre klasterovanja potrebno je očistiti datoteke od suvušnih podataka.

Potrebno je izbaciti sve gene koji nisu validni. To su geni koji se ne nalaze u datoteci common_human_list.csv

U .ipnb fajlu Zajednicki skup gena se nalazi postupak za pravljenje liste validnih gena i pronalaženje zajedničkog skupa gena za sve datoteke, što je potrebno naći jer će dolaziti do spajanja datoteka.

Takav skup čuva se u datoteci Valid_genes.csv.

Nakon toga potrebno je iz datoteka izbaciti sve gene i ćelije koje ne zadovoljavaju određene uslove.

U ovom slučaju bilo je potrebno izbaciti sve ćelije koje ukupno imaju manje od 500 pozitivnih vrednostipo fajlu i čiji zbir pozitivnih vrednosti nije veći ili jednak 1000 i sve gene koji u svim datotekama ukupno imaju manje od 1% pozitivnih vrednosti u odnosu na ukupan broj vrednosti za ćelije u svim datotekama.

Potrebno je i promeniti imena gena (u skladu sa datotekom common_human_list.csv) i ćelija u skladu sa nazivom datoteke.

Ovaj proces se nalazi u .ipnb fajlu Priprema podataka (izuzev promene imena gena) i takvi podaci se čuvaju u datotekama GSM3982570_.csv - GSM3982570_.csv.

Na kraju potrebno je transponovati datoteke, jer se spajaju po genima koji treba da budu kolone. Postupak za ovo, kao promenu imena gena se nalazi u fajlu Transponovanje.ipnb.

Ovako gotove datoteke (očišćene, preimenovane ćelije i geni, transponovane) su datoteke GSM3982570_r_t.csv - GSM3982576_r_t.csv

Svi .ipnb fajlovi potrebni za prečišćavanje podataka se nalaze u folderu Ciscenje_podataka, u kome se nalaze i posebni folderi sa odgovarajućim verzijama svake datoteke. Folderi su nazvani imenom datoteke.

Primer za folder GSM3982570:

GSM3892570_PBMC_DRESS1_filtered_gene_bc_matrices_h5.csv – originalan skup podataka

GSM3892570.csv – prečišćena datoteka (nema nevalidnih gena i nepotrebnih ćelija)

GSM3892570_.csv – ponovo prečišćena datoteka (menjan je skup validnih zajedničkih gena koji ne zadovoljavaju zadati uslov, pa se samim tim opet menjao i skup ćelija) i promenjena imena ćelija

GSM3892570_r.csv – preimenovani su i geni

GSM3892570_r_t.csv – transponovani su podaci i ovo je finalna datoteka spremna za klasterovanje

Sve datoteke spremne za klasterovanje se nalaze u folderu Ocisceni_podaci.

3 Klasterovanje

Klasterovanje je rađeno nad datotekama GSM3892570_r_t.csv - GSM3892575_r_t.csv i grupi datoteka GSM3892572_r_t.csv - GSM3892575_r_t.csv koji čine grupu 2.

Korišćeni su algoritmi: DBScan, Kmeans, Agglomerative clustering, Spectral clustering i Birch, sa izuzetkom klasterovanja grupe 2 gde nisu korišćeni Spectral i Birch zbog velike količine podataka.

Rezultati klasterovanja se nalaze u folderima GSM3892570 - GSM3892575 i folderu grupa_2, i predtsvaljaju slike dobijenih klastera. Pored rezultata nalaze se i slike dobijenog dijagrama za izbor vrednosti parametra epsilon kod algoritma DBScan, gde je koriščena knee metoda, i raspodele podataka korišćenjem koordinatama koje su dobijene TSNE algoritmom. Takođe nalaze se i datoteke sa tsne koordinatama i labelama klastera za svaki gen.

U folderu Grupa_2 nalazi se i dodatno datoteka sa spojenim podacima kad kojima se vršilo klasterovanje.

Pre korišćenja TSNE algoritma, korišćen je PCA algoritam kako bi se veliki broj atributa zemanio skupom od 300 reprezentativnih komponenti i zatim pomoću TSNE algoritma našle 2D koordinate, pri čemu parametar perplexity imao vrednost 30.

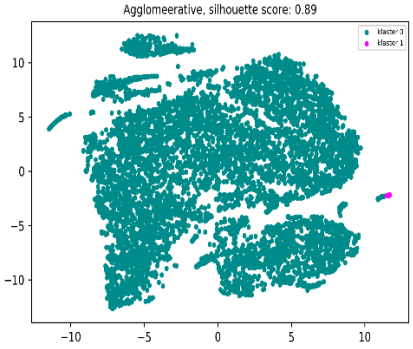
Za svaki algoritam na onovu vrednosti senka koeficijenta biran najbolji broj klastera i najbolje vrednosti dodatnih atributa ako su korišćeni.

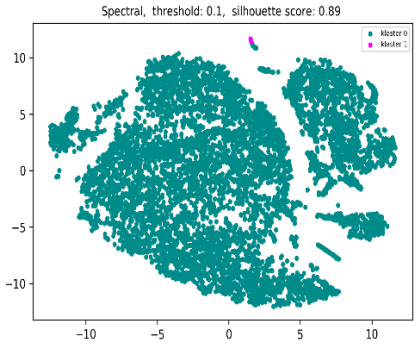
U nastavku će biti prikazani rezultati za svaku datoteku/grupu.

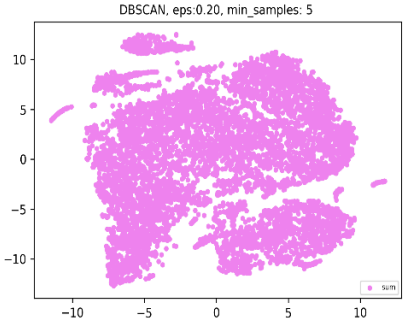
Napomena: greškom i slika za Birch i za Spectral algoritam imaju naziv Spectral, tako da će ispod oznake datoteke biti dat redosled imena algoritama na slikama.

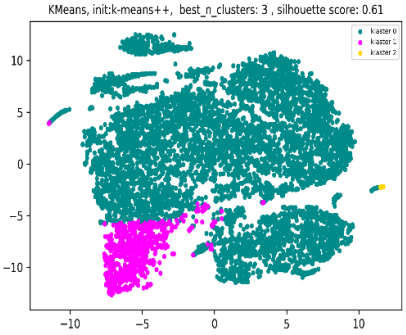
GSM3892570

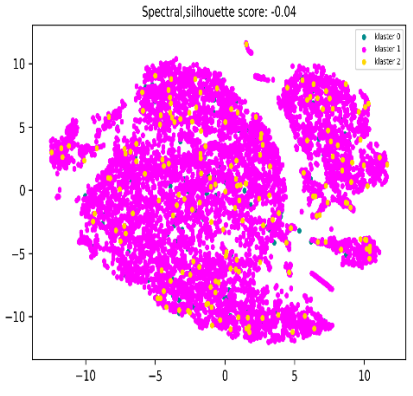
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





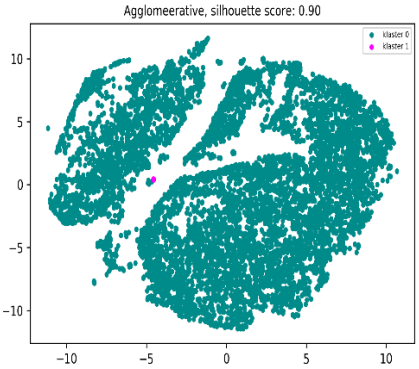


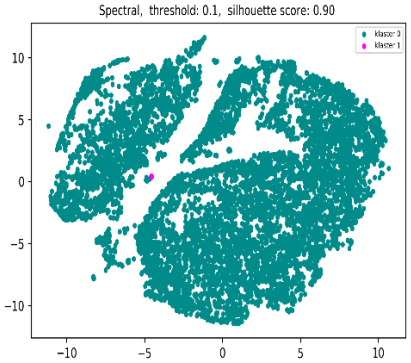


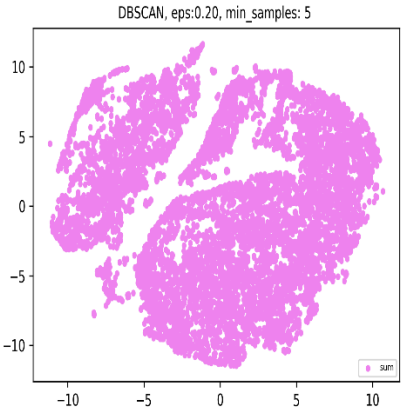


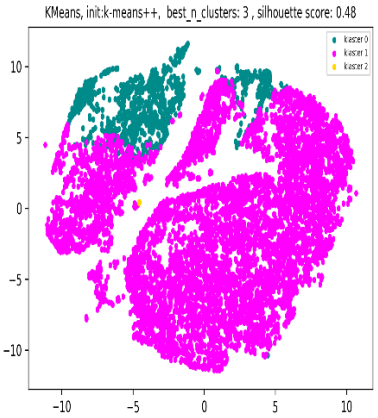
GSM3892571

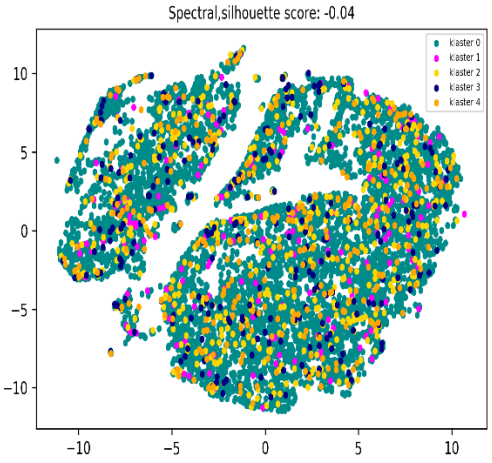
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





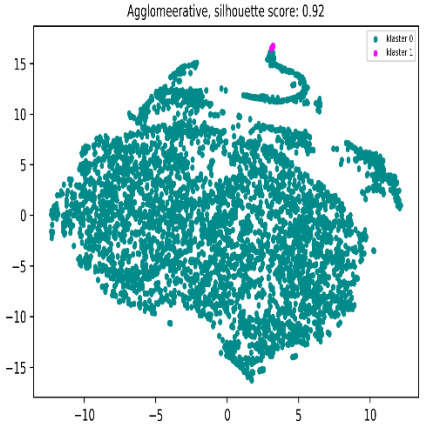


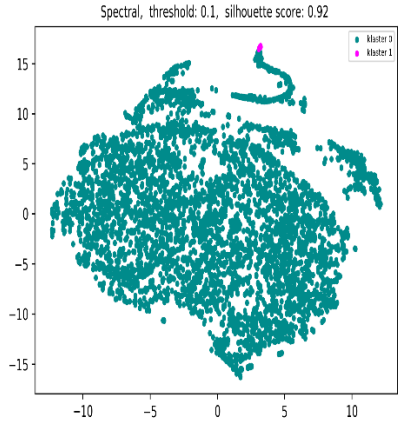


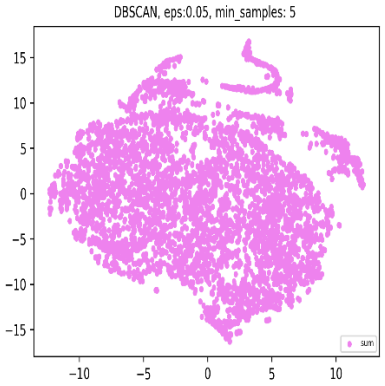


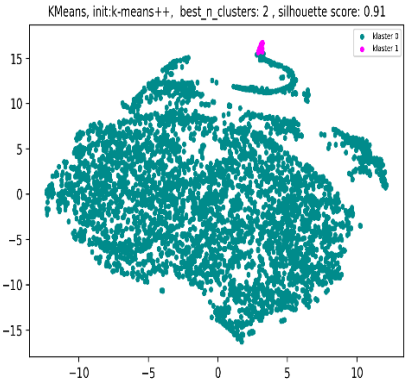
GSM3892572

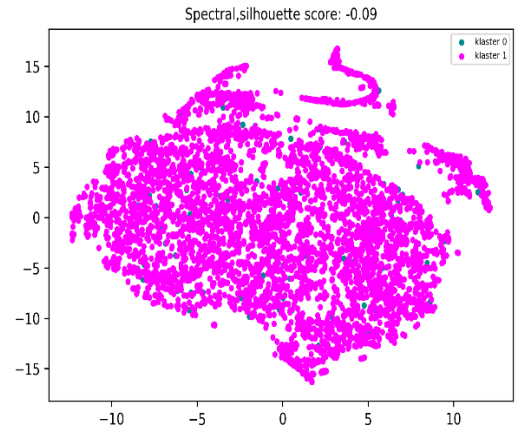
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





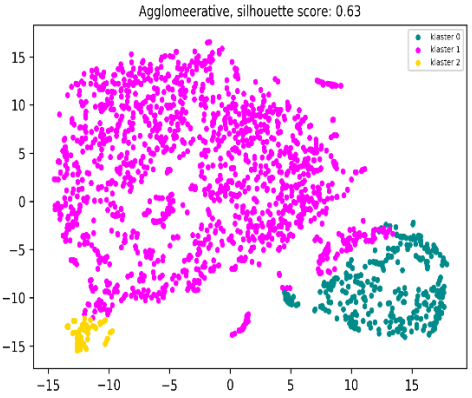


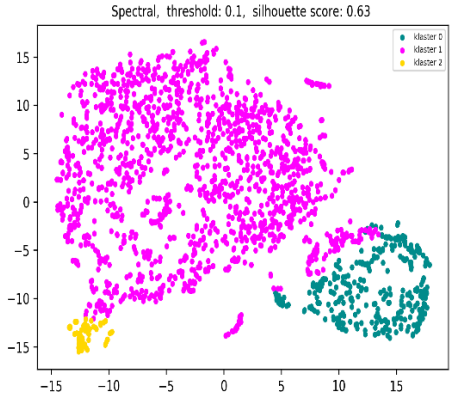


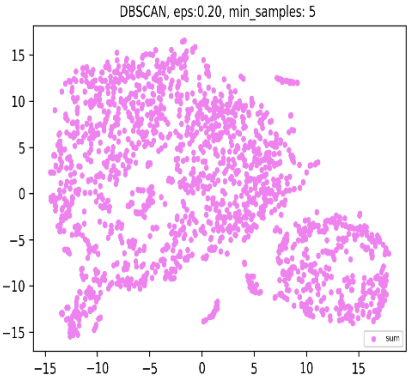


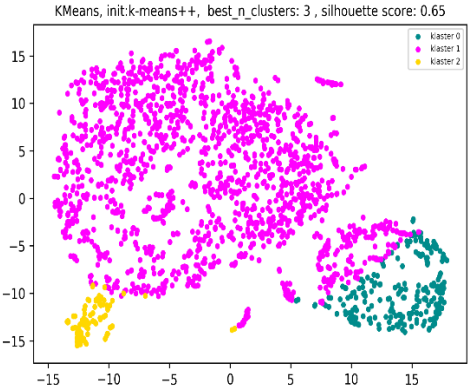
GSM3892573

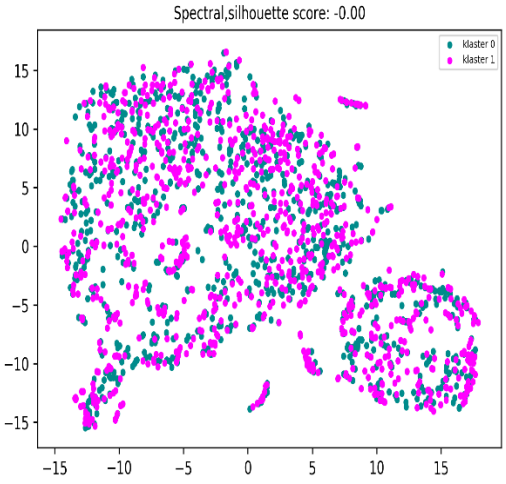
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





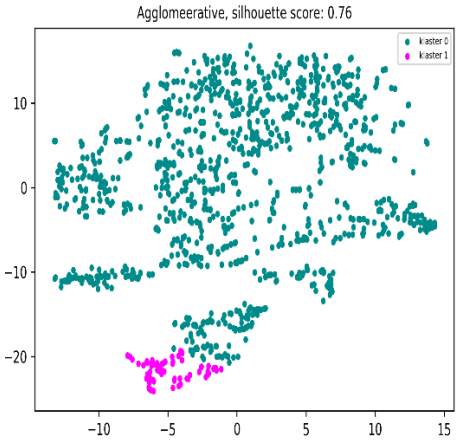


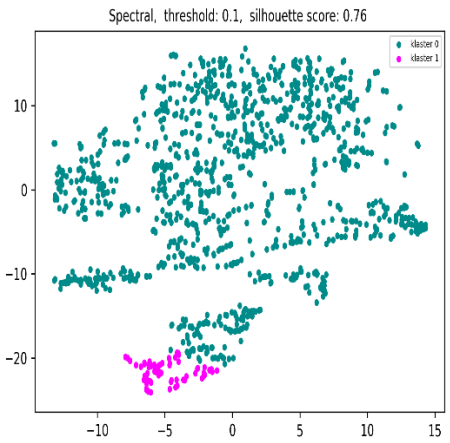


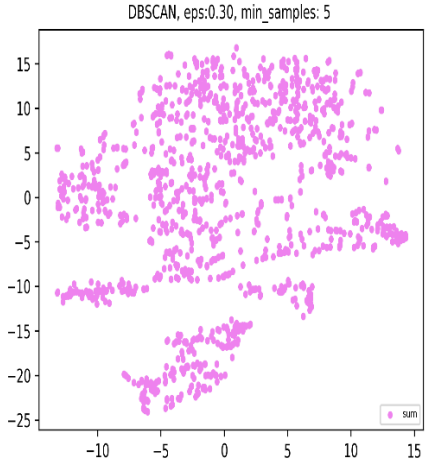


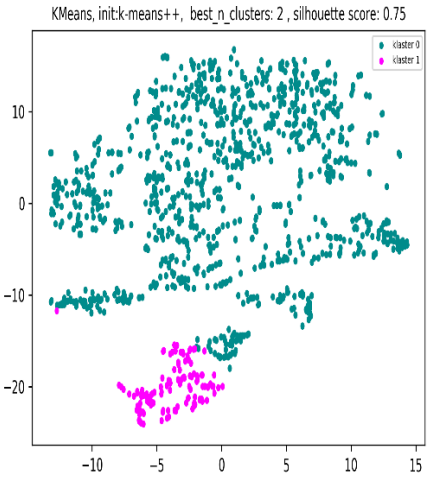
GSM3892574

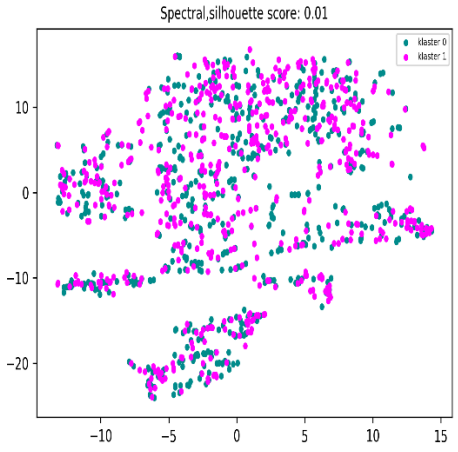
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





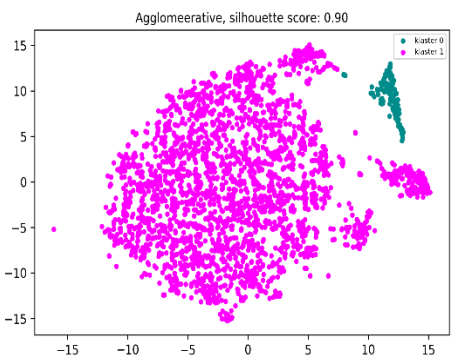


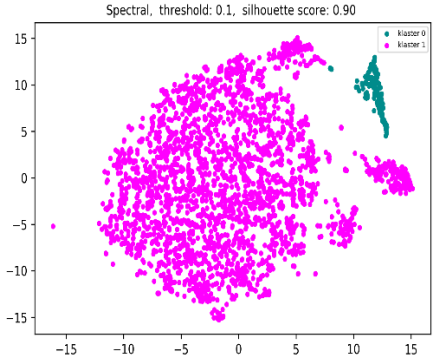


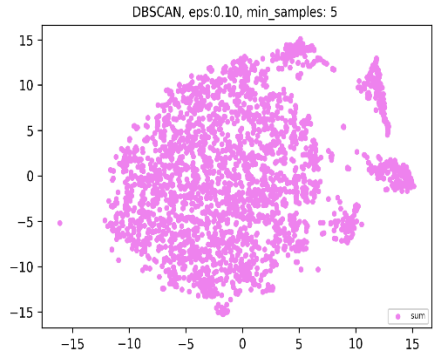


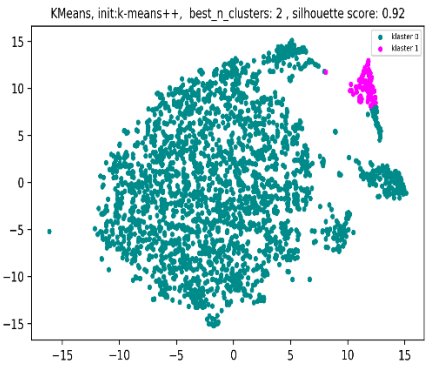
GSM3892575

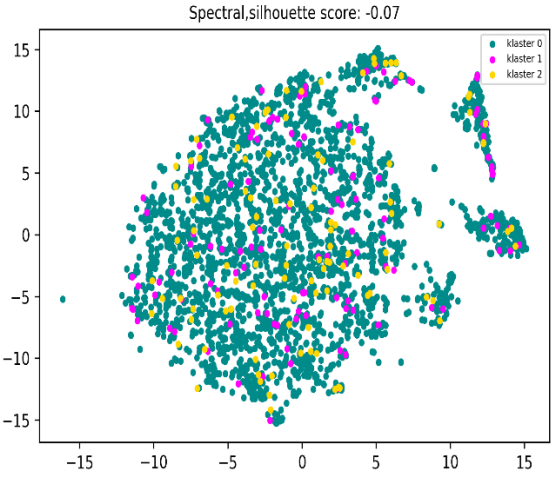
(Agglomerative, Birch, DBScan, Kmeans, Spectral)





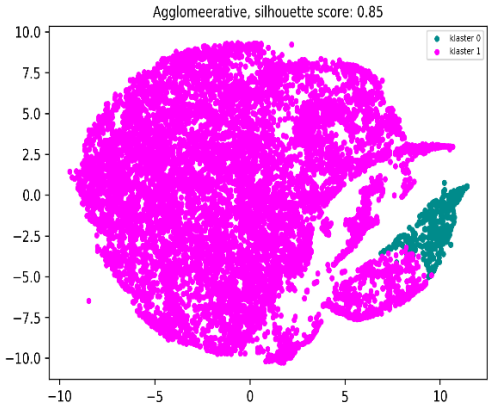


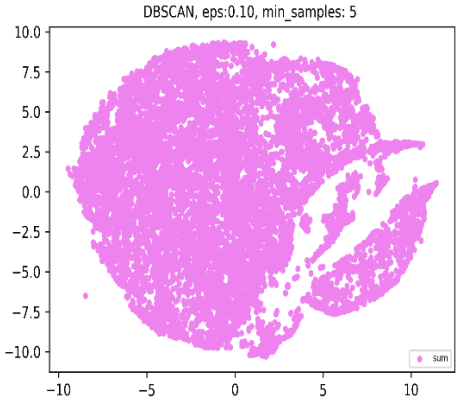


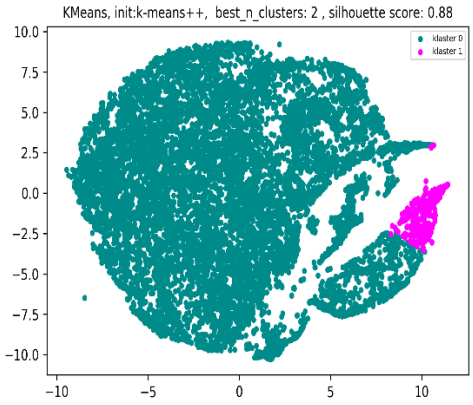


GRUPA_2

(Agglomerative, DBScan, Kmeans)







4 Zaključak

Od svih korišćenih algoritama DBScan je dao najgore rezultate, obzirom na to da ni jednom nije našao klastere, već je sve tačke obeležio kao šum.

Moglo bi se reći da je KMeans dao uopšteno najbolje rezultate, mada i to zavisi od samih datoteka, ali uglavnom je najjasnije razdvajao klastere, dok je Spectral algoritam svaki put dao nejasne klastere i nalazio je više klastera nego ostali algoritmi.

Birch i Agglomerative algoritmi su davali iste rezultate, koji su ili bili slični rezultatima Kmeans, ili su davali dva klastera takva da svega nekoliko tačaka pripada jednom klasteru.

Treba uzeti u obzir da ni jedan algoritam, sem DBScan, nije imao opciju pronalaženja jednog klastera, već je uvek najmanji razmatrani broj za parametar n_clusters bio 2, a najveći 5, tako da bi verovatno Birch i Agglomerative našli jedan klaster.

Ako izuzmemo algoritme DBScan, koji svaki put nadje samo šum, i Spectral, koji svaki put daje nejasne rezultate, može se reći da se u rezultatima za GSM3892572 - GSM3892575 naziru klasteri, posmatrajući slike koje su identične za preostala 3 algoritma.

Grupu 2 ne možemo skroz porediti sa prethodnim rezultatima, jer nisu korišćeni svi algoritmi, ali ako uzmemo u obzir da Birch i Agglomerative daju identične rezultate, i da bi Spectral verovatno i ovde dao nejasne klastere, kao što je DBScan našao i ovde samo šumove, može se isto reći da se naziru neki klasteri, što i ima smisla jer grupu 2 čine prethodno pomenute datoteke (GSM3892572 - GSM3892575).

Kako je za odabir najboljih parametara za model bio korišćen senka koeficijent, i tražena njegova najveća vrednost, uzimajući sve rezultate u obzir, možda negde to i nije bilo najbolje rešenje i možda bi se dobili bolji rezultati sa manjim koeficijentom. Opet, zavisi i od same hipoteze kako bi trebalo birati vrednost senka koeficijenta za najbolji model.