

Analiza skupa podataka za rak dojke

Branko Grbić - Matematički fakultet, Univerzitet u Beogradu

13. septembar 2023.

Sadržaj

1 Uvod	2
2 Opis skupa podataka	2
3 Eksplorativna analiza podataka	2
4 Klasifikacija	5
4.1 XGBoost	6
4.2 SVM	8
4.3 KNN	9
4.4 Poređenje modela	11
4.5 Ansambl	12
5 Klasterovanje	12
5.1 KMeans	13
5.2 Gaussian mešavina	15
5.2.1 PCA dataset - poređenje tipova kovarijanse	15
5.2.2 PCA dataset sa punim tipom kovarijanse	17
5.2.3 PCA dataset sa sfernim tipom kovarijanse	18
5.2.4 PCA dataset sa povezanim tipom kovarijanse	19
5.2.5 t-SNE dataset - poređenje tipova kovarijanse	20
5.2.6 t-SNE dataset sa punim tipom kovarijanse	21
5.2.7 t-SNE dataset sa sfernim tipom kovarijanse	22
6 Pravila pridruživanja	23
7 Zaključak i diskusija	23
8 Reference	23

1 Uvod

U ovom seminarskom radu, na skupu podataka za rak dojke, demonstrirana je: - eksplorativna analiza podataka, klasifikacija, klasterovanje i pravila pridruživanja ispisana u programu SPSS. Za klasifikaciju je korišćen:

- XGBoost
 - SVM
 - KNN

Za klasterovanje je korišćen:

- KMeans
 - Gaussian Mixture

Za pravila pridruđivanja je korišćen:

- Apriori

2 Opis skupa podataka

Skup podataka "Breast cancer gene expression - CuMiDa"[1], sadrži 151 uzorak, od kojih svaki ima 54676 atributa (gena) koji se klasifikuju u 6 klasa:

- Basal
 - HER
 - Cell line
 - Luminal A
 - Luminal B
 - Normal

S obzirom na ogroman broj atributa, nemoguće je prikazati šta koj atribut služi, ali ćemo u nastavku objasniti dublje model kroz EDA.

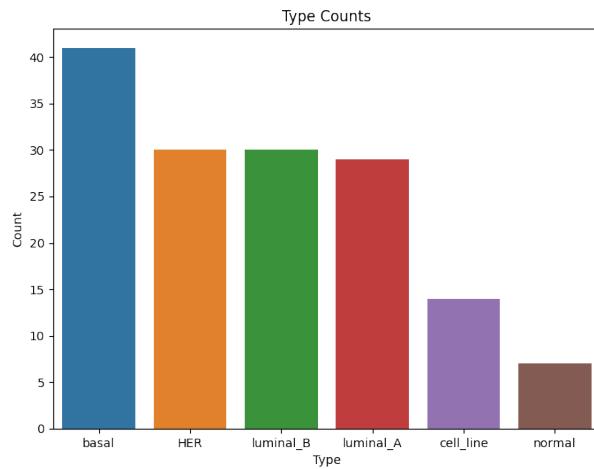
3 Eksplorativna analiza podataka

Prvo je neophodno proveriti karakteristike skupe podataka, odnosno broj atributa, klasa i nedostajuće vrednosti.

	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	...	AFFX-r2-Ec-bioD-3_at	AFFX-r2-Ec-bioD-5_at
samples													
84	basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.408328	8.870780	...	12.229711	11.852955
85	basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.584418	7.767646	...	12.178531	11.809408
87	basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.425195	9.417956	...	12.125108	11.725766
90	basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.567956	9.022345	...	12.111235	11.719215
91	basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.424426	9.400056	...	12.173642	11.861296

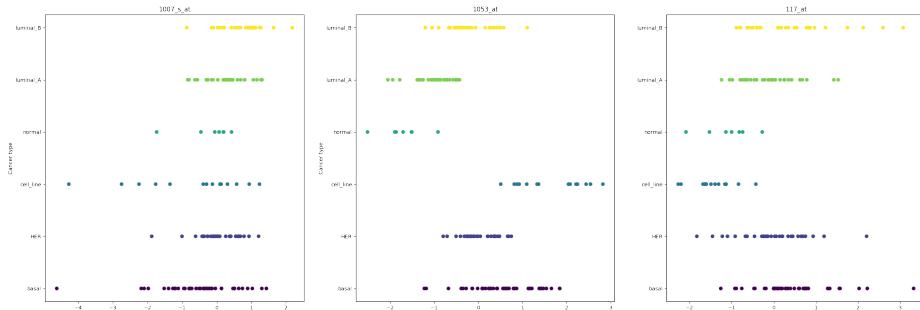
Slika 1. Prikaz skupa podataka.

Nakon kratkog utvrđivanja, može se pokazati i da su svi atributi postojeći, sa decimalnim vrednostima. Možemo takođe videti i koliko su klase pojedinačno reprezentovane. Iz naredne slike, možemo videti da su klase nebalansirane.



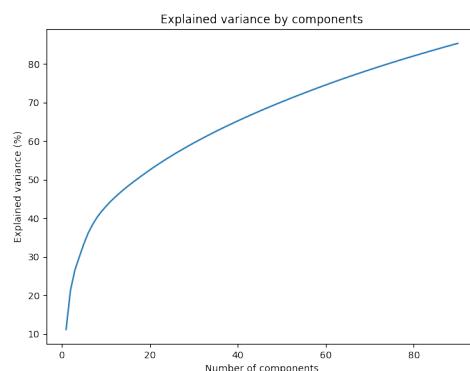
Slika 2. Balansiranost klase

Nakon normalizacije podataka, prikazana su prva 3 atributa u odnosu na klase. Može se primetiti da se ne vidi jasna korelacija između atributa i klasa, te se moraju još procesirati podaci.



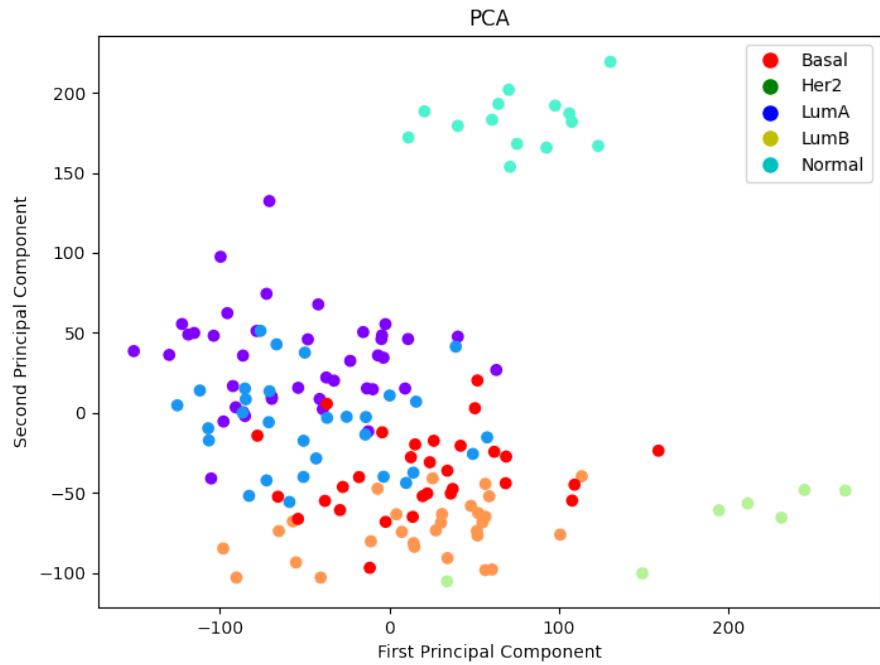
Slika 3. Prva 3 atributa

To navodi na pokušaj redukcije dimenzionalnosti, pa PCA (Principal Component Analysis) tehnika može doći jako korisna. Naime, za ukupnu objašnjenu varijansu od 85% koristi se 90 atributa, što nam može doći kasnije kao odlična metrika za redukovanje skupa podataka.

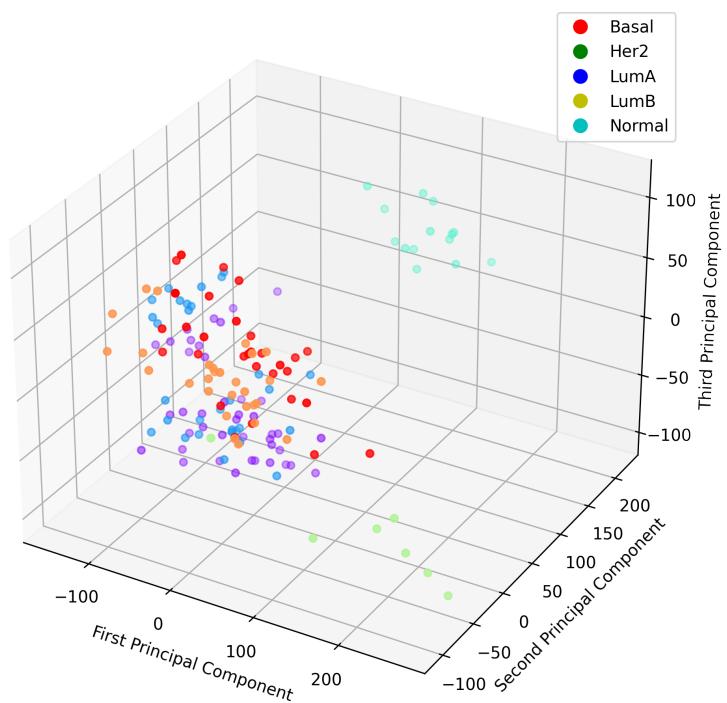


Slika 4. Suma varijanse kroz komponente

Za prve 3 izvučene komponente, ne može se uočiti jasna razlika između klasa.

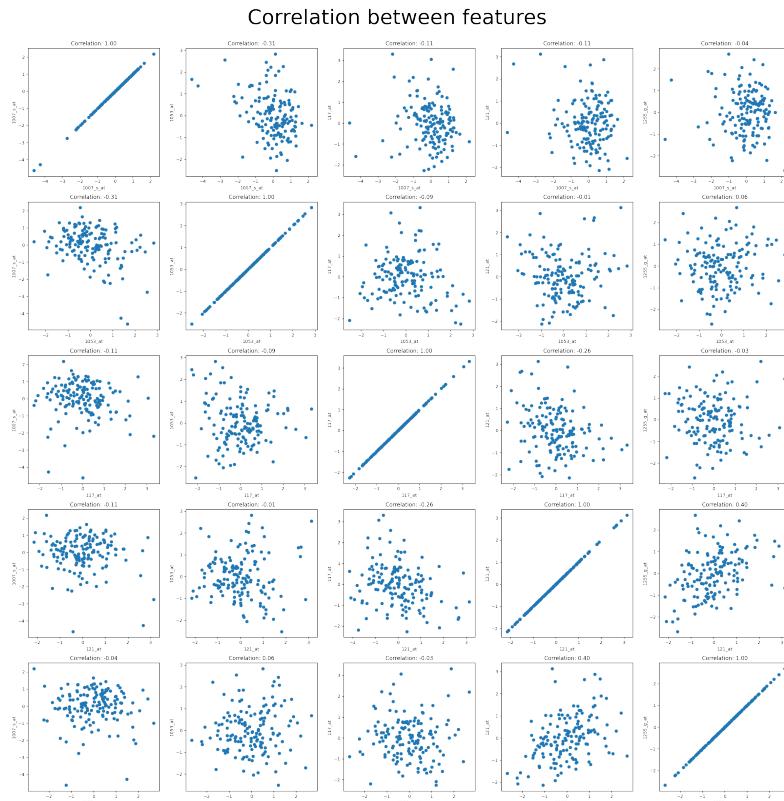


Slika 5. PCA prve 2 komponente



Slika 6. PCA prve 3 komponente

To nas navodi da treba proveriti korelaciju atributa. S obzirom na ogroman broj atributa, ne može se otkriti korelacija između svakog para jednostavno, te je ovo iskorišćeno samo u demonstracione svrhe. Naime, jako redak broj atributa kroz ceo skup podataka ima korelaciju $|corr| > 0.9$

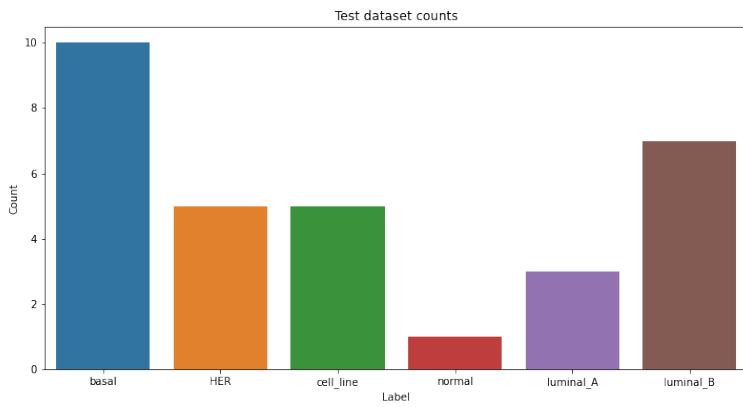


Slika 7. Matirce korelacije prvih 5 atributa

4 Klasifikacija

Kao što je i ranije napomenuto, gledaće se 3 modela, XGBoost, SVM i KNN, ali takođe će se modeli uporediti. Na kraju, napraviće se ansambl sa 3 najbolja modela od svaka od ova 3 tipa i proveriti uspešnost kombinovanja modela.

Klasifikaciju smo za sva 3 modela vršili sa i bez redukcije dimenzionalnosti (PCA) i sa i bez nadsemplovanja (SMOTE). Svi podaci su u startu normalizovani. Kao metrika tačnosti uzet je F1 rezultat, ali takođe je urađena i tačnost modela. Vršena je podela skupa na trening (80%) i test (20%) koja je ista kroz sve modele, da bi se mogla uporediti tačnost.



Slika 8. Broj klasa u test skupu

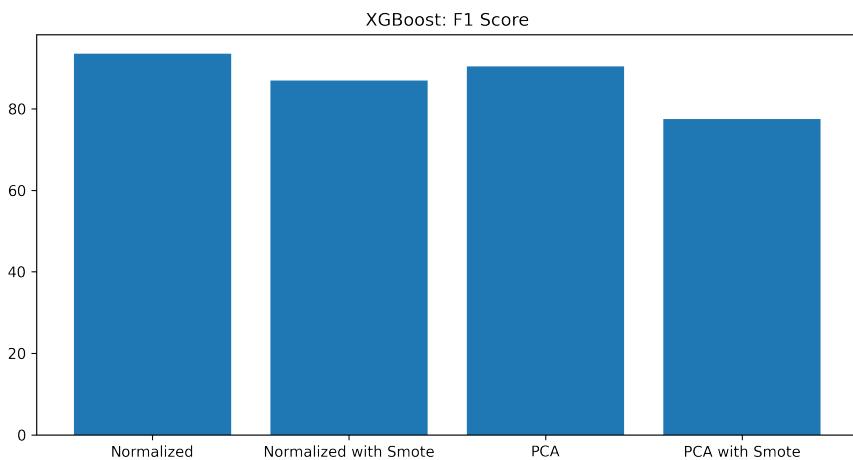
4.1 XGBoost

XGBoost pripada grupi stabla odlučivanja koja koristi "gradijentno pojačanjeđa nađe najbolji rezultat. Za razliku od modela koji radi na sličan način, modela nasumičnih šuma (eng. Random Forest), XGBoost se fokusira na minimizaciju pristrasnosti i uklanjanje neprilagođavanja. On radi tako da stabla koja daju loš rezultat, kombinuje u nadi da će iz njih izrasti kombinovano bolji model nego ranije.

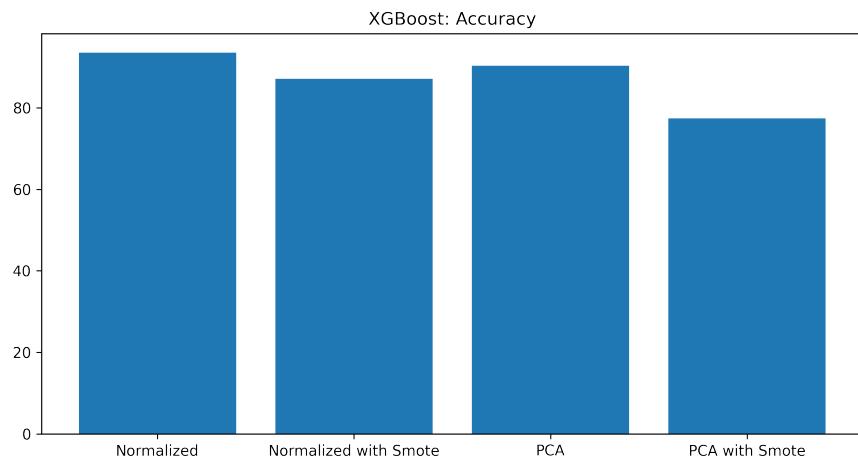
Korišćenjem Grid Search-a, nađeni su sledeći hiperparametri

- Normal dataset - learning_rate: 0.1, max_depth=6, n_estimators=100
- PCA dataset - learning_rate: 0.1, max_depth=5, n_estimators=500
- Normal dataset with SMOTE - learning_rate: 0.1, max_depth=6, n_estimators=100
- PCA dataset with SMOTE - learning_rate: 0.1, max_depth=5, n_estimators=500

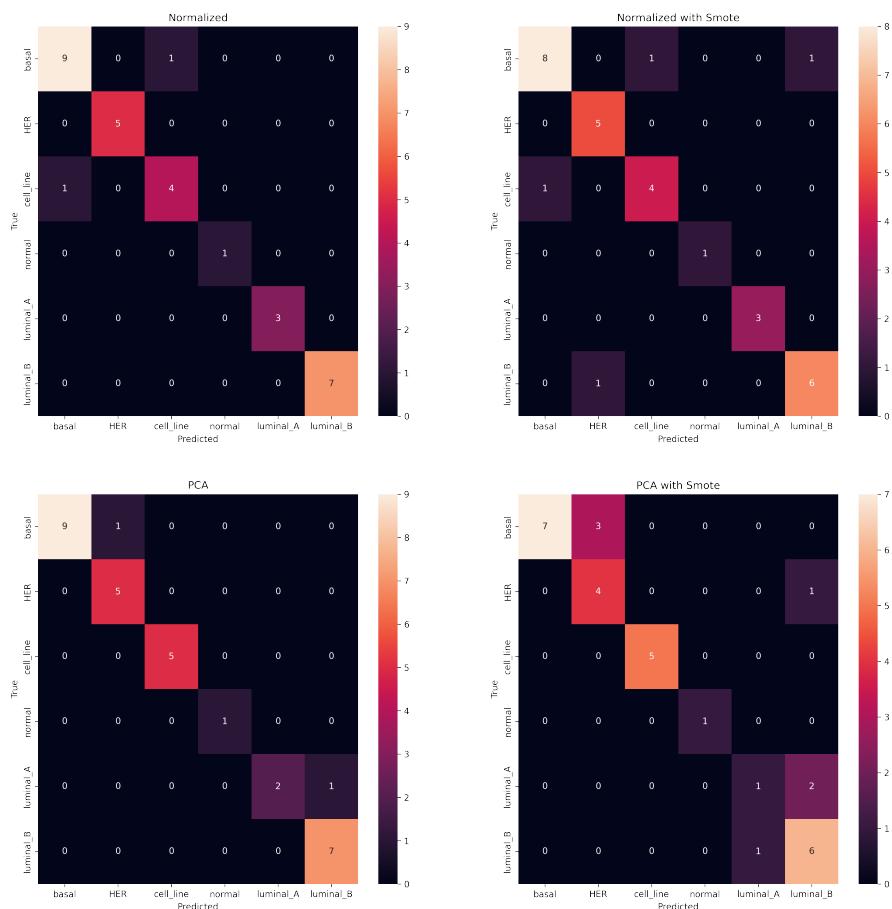
U narednim slikama možemo videti rezultate:



Slika 9. XGBoost grafik F1 rezultata



Slika 10. XGBoost grafik tačnosti



Slika 11. XGBoost Matrica kofnuzije

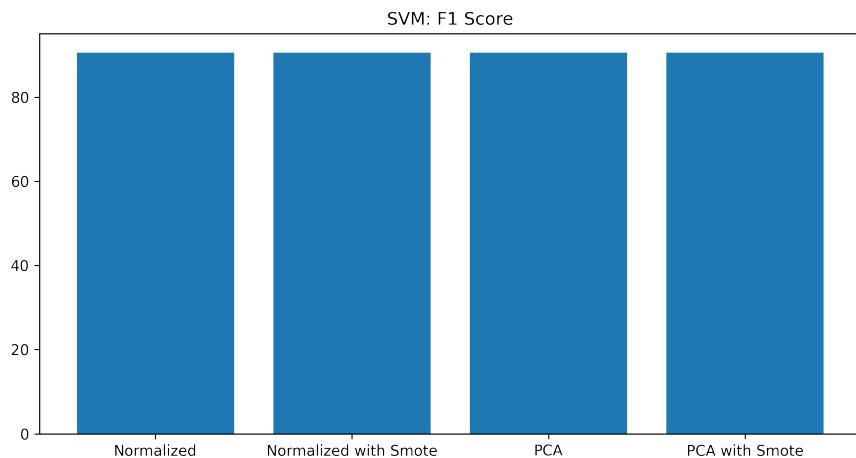
4.2 SVM

Metoda potpornih vektora je model nadgledanog učenja, korišćen kako i u klasifikacionim, tako i u regresionim problemima. SVM karakteriše tip kernela s kojim radi. Linearni kernel je jako pogodan za uzorak s visokom dimenzionalnošću (zbog male šanse preprilagođavanja), što se može pokazati i traženjem najboljih hiperparametara.

Korišćenjem Grid Search-a, nađeni su sledeći hiperparametri

- Normal dataset - C: 0.1, gamma: 1, kernel: linear
- PCA dataset - C: 1, gamma: scale, kernel: sigmoid
- Normal dataset with SMOTE - C: 0.1, gamma: 1, kernel: linear
- PCA dataset with SMOTE - C: 1, gamma: scale, kernel: sigmoid

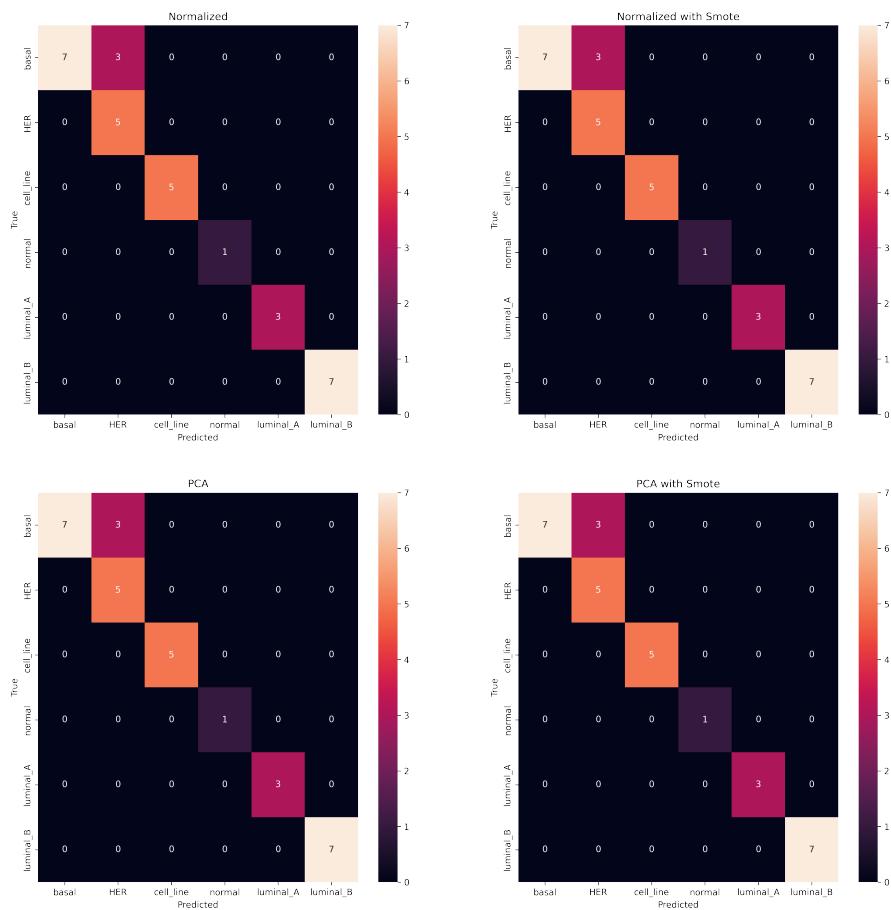
U narednim slikama možemo videti rezultate:



Slika 12. SVM grafik F1 rezultata



Slika 13. SVM grafik tačnosti



Slika 14. Matrica konfuzije SVM-a

Može se primetiti da ne samo što imaju istu tačnost i F1 rezultat, već su iste i promašene klase, te možemo zaključiti da modeli, iako imaju malo drugačije parametre u zavisnosti od ulaza, na kraju identično uče.

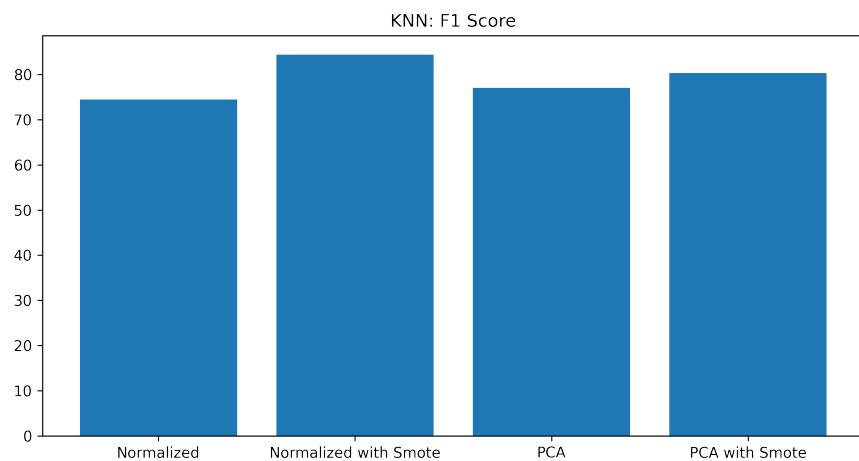
4.3 KNN

KNN je model koji klasificiše tako što za jedan uzorak nalazi k najблиžih suseda, pa metodom glasanja odluči kojoj klasi taj uzorak pripada.

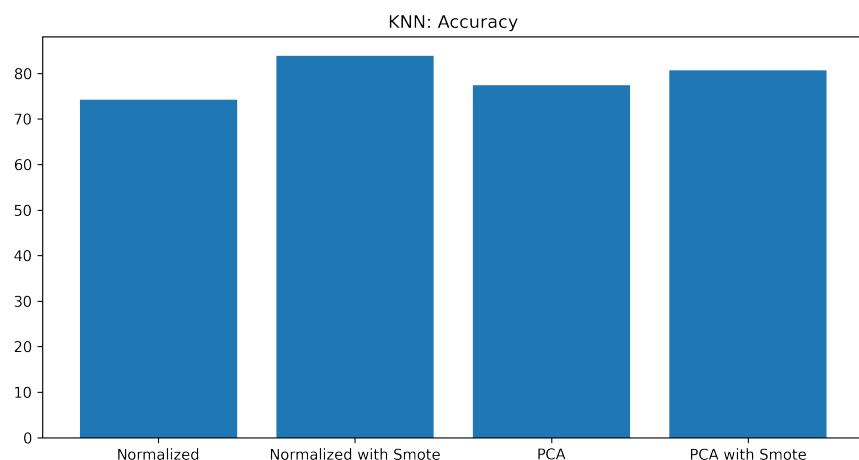
Korišćenjem Grid Search-a, nađeni su sledeći hiperparametri

- Normal dataset - n_neighbors: 3, p: 1, weights: distance
- PCA dataset - n_neighbors: 1, p: 8, weights: uniform
- Normal dataset with SMOTE - n_neighbors: 3, p: 1, weights: uniform
- PCA dataset with SMOTE - n_neighbors: 2, p: 4, weights: uniform

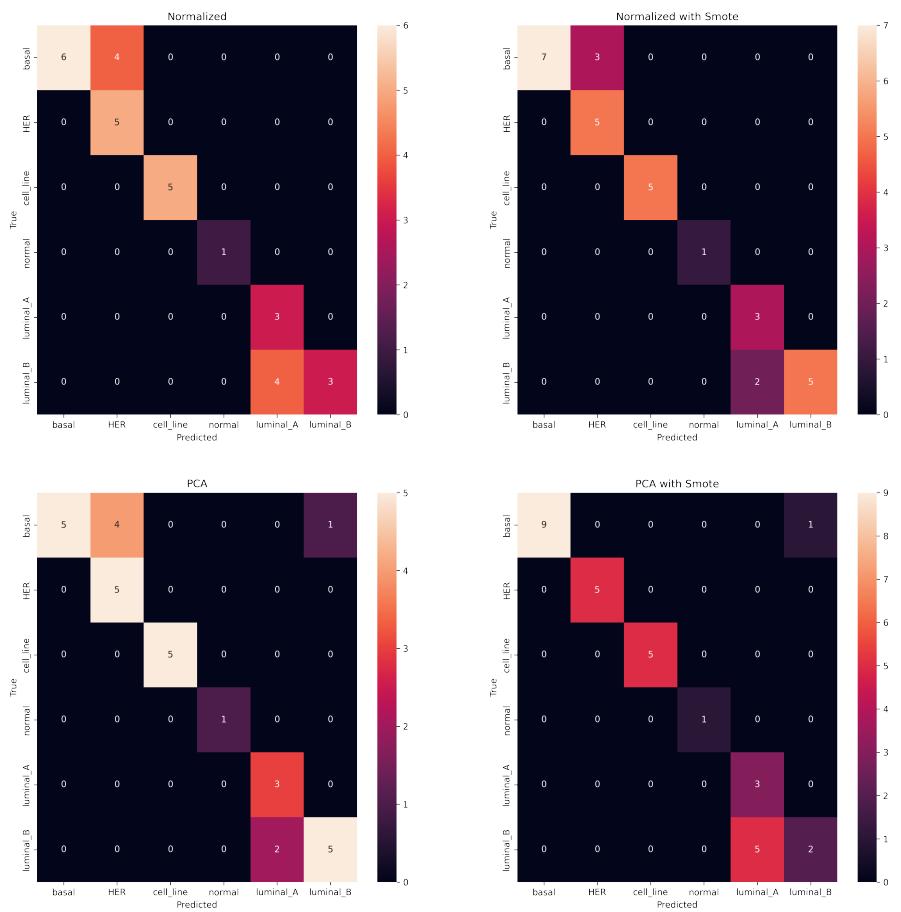
U narednim slikama možemo videti rezultate:



Slika 15. KNN grafik F1 rezultata



Slika 16. KNN grafik tačnosti

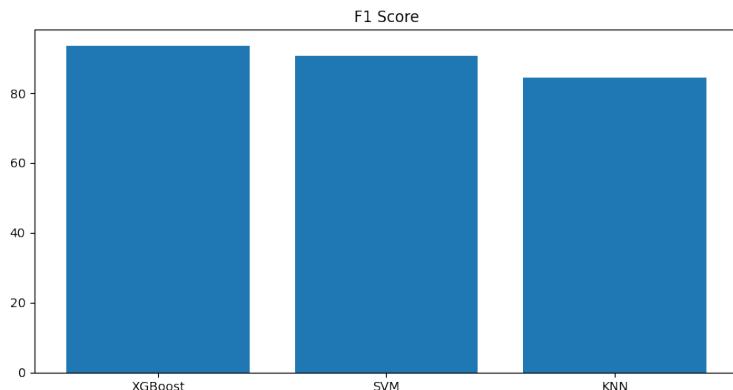


Slika 17. KNN Matrica konfuzije

4.4 Poređenje modela

Koristićemo model koji se najbolje pokazao u F1 metrikama od svih 3 modela.

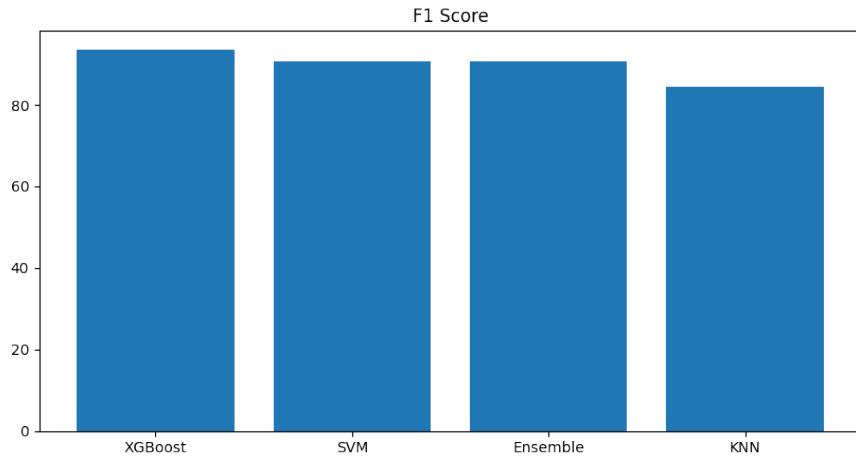
Najjači model je XGBoost sa 96%, zatim SVM sa 90.59%, dok KNN ima 84.40%.



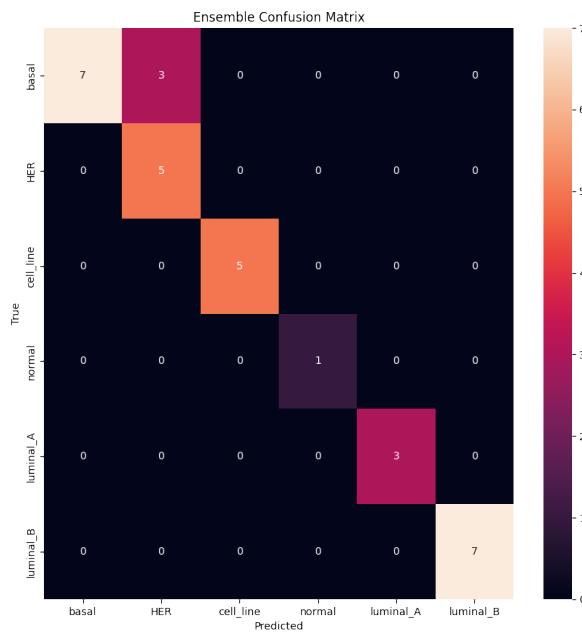
Slika 18. Poređenje najboljih modela

4.5 Ansambl

Ansambl se lošije pokazao nego najbolji model, što opravdavamo sličnosti u razlikama modela, pa greške u KNN-u i SVM-u utiču gore na rezultat ansambl modela, davajući tačnost od 90.58% što je lošije od najboljeg modela, XGBoost-a.



Slika 19. Poređenje modela sa ansamblom



Slika 20. Matrica konfuzije ansambl modela

5 Klasterovanje

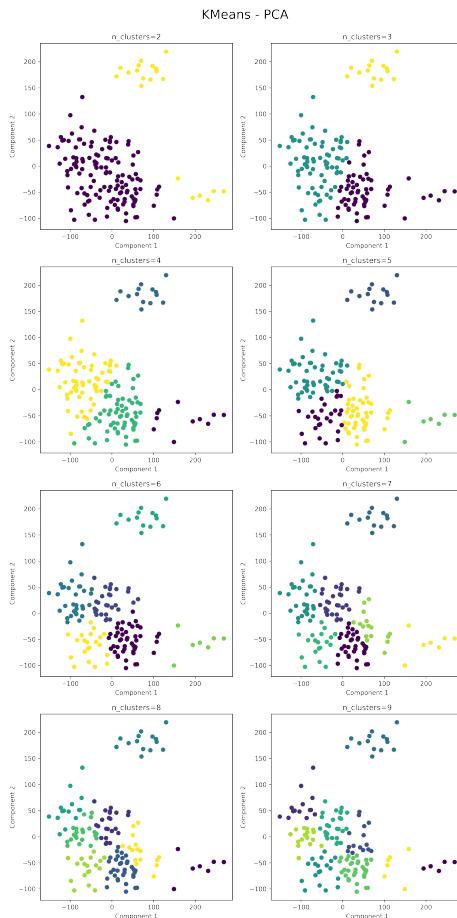
Klasterovanje je rađeno korišćenjem 2 algoritma:

- KMeans
- Gaussian Mixture

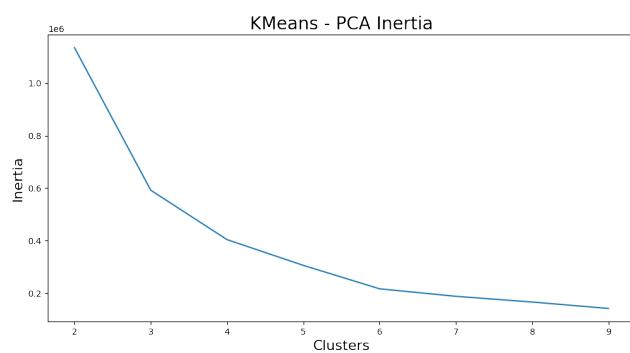
Oba algoritma zbog vizualizacije koriste 2 komponente ekstraktovane pomoću 2 algoritma za redukciju dimenzionalnosti: PCA i t-SNE.

5.1 KMeans

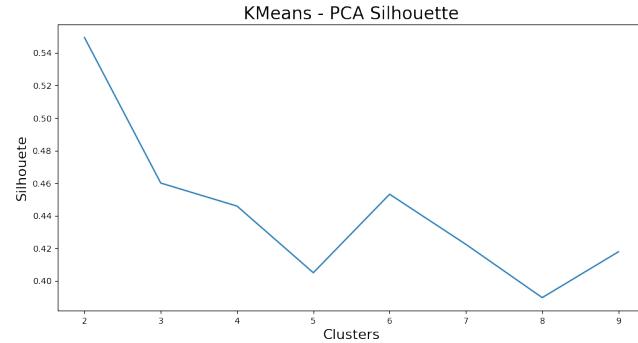
KMeans algoritam je jedan od najstarijih i najpopularnijih algoritama za klasterovanje podataka, gde svaki uzorak se iterativno spaja nekom klasteru u zavisnosti od K najbližih klastera.



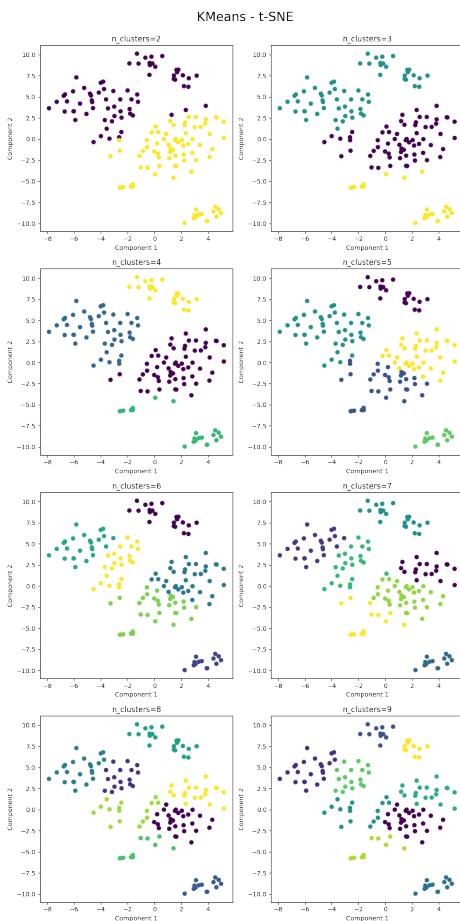
Slika 21. KMeans algoritam na PCA datasetu



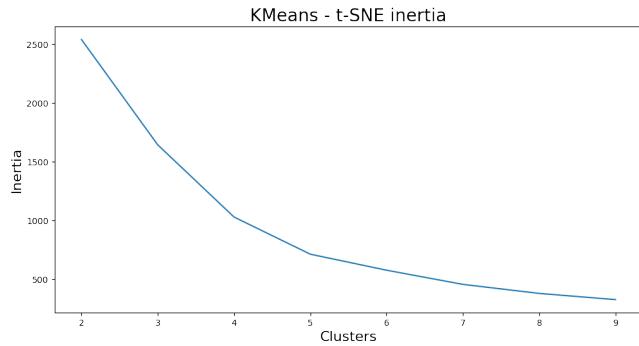
Slika 22. KMeans inercija na PCA datasetu



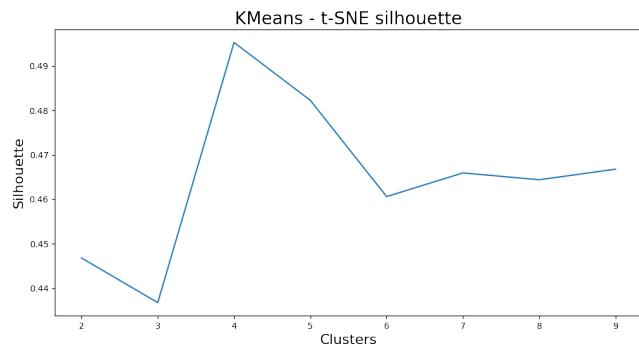
Slika 23. KMeans skor siluete na PCA datasetu



Slika 24. KMeans algoritam na t-SNE datasetu



Slika 25. KMeans inercija na t-SNE datasetu



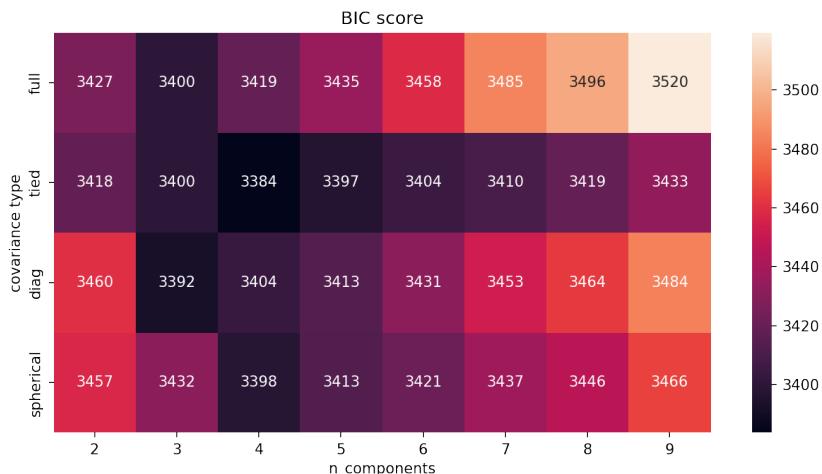
Slika 26. KMeans skor siluete na t-SNE datasetu

Može se primetiti da nijedan ni drugi oblik redukcije dimenzionalnosti ne pomaže u dobrom klasiranju. Podaci su takođe pokušani za klaster sa 90 PCA komponenti (85% suma varijanse) ali bez uspeha.

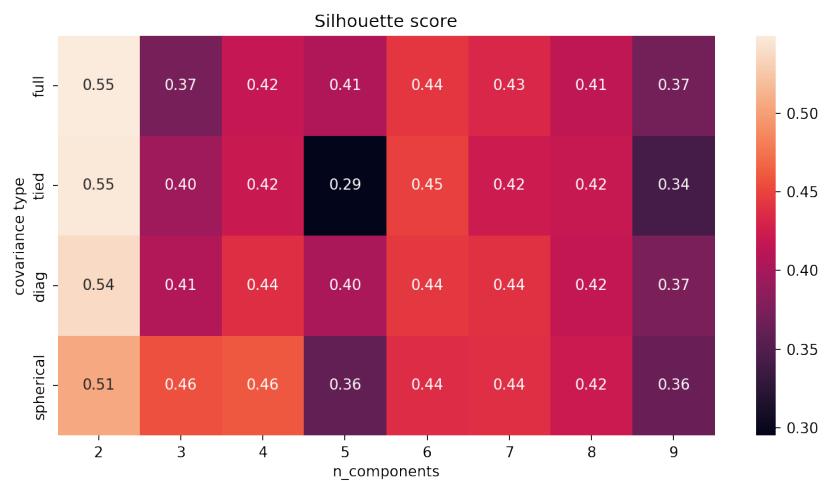
5.2 Gaussian mešavina

Gausova mešavina je model koji prepostavlja da su podaci izvedeni iz jedne ili više normalnih distribucija.

5.2.1 PCA dataset - poređenje tipova kovarijanse

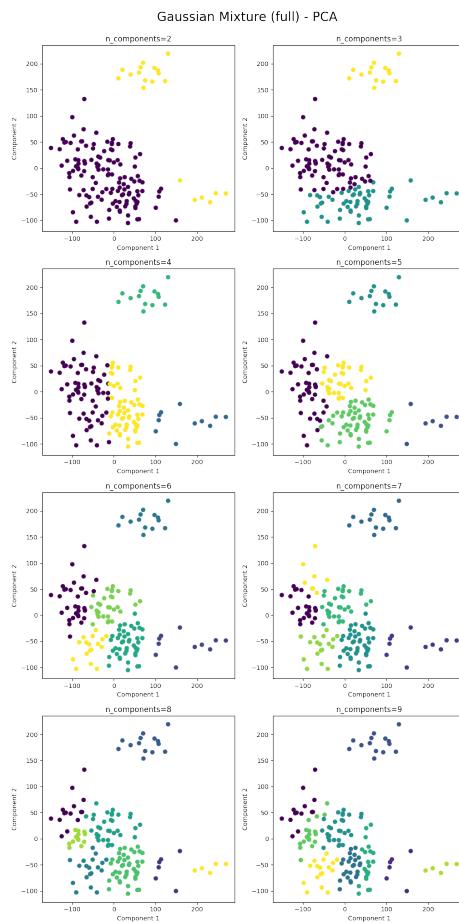


Slika 27. BIC skor u zavisnosti od tipa kovarijanse na PCA datasetu

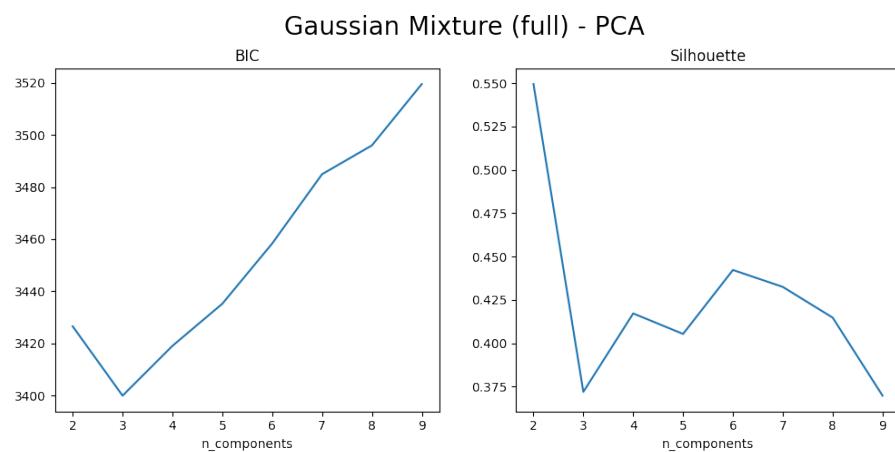


Slika 28. Skor siluete u zavisnosti od tipa kovarijanse na PCA datasetu

5.2.2 PCA dataset sa punim tipom kovarijanse

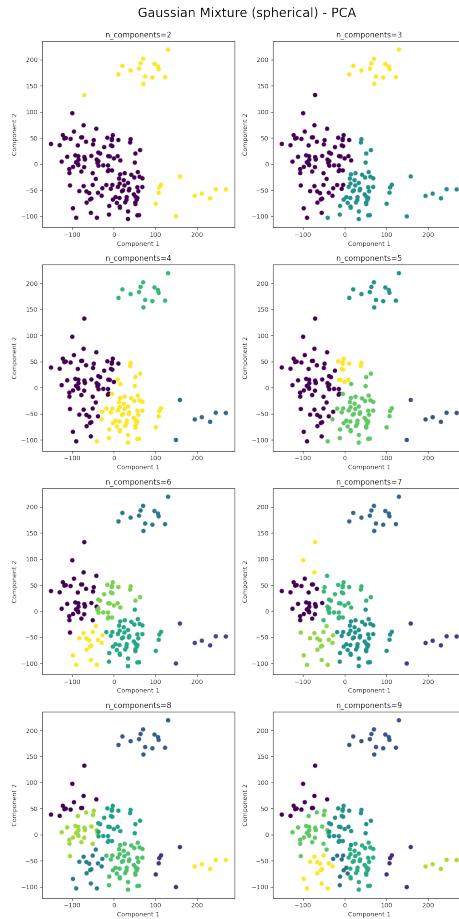


Slika 29. PCA full

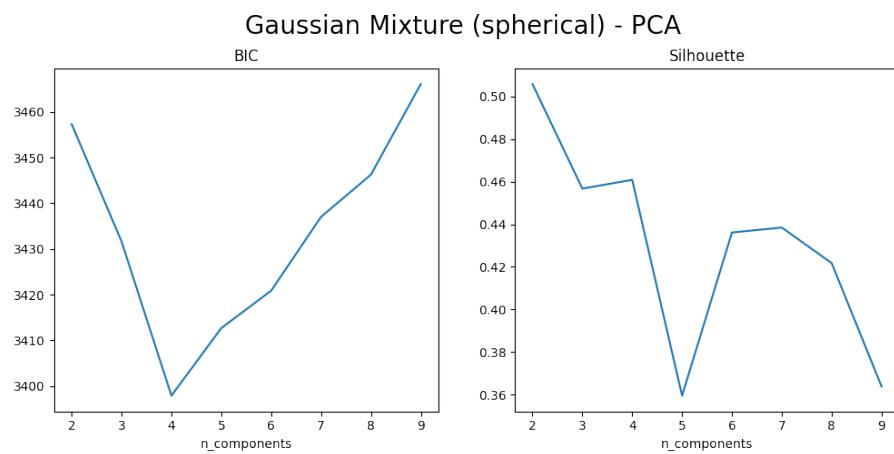


Slika 30. PCA full

5.2.3 PCA dataset sa sfernim tipom kovarijanse

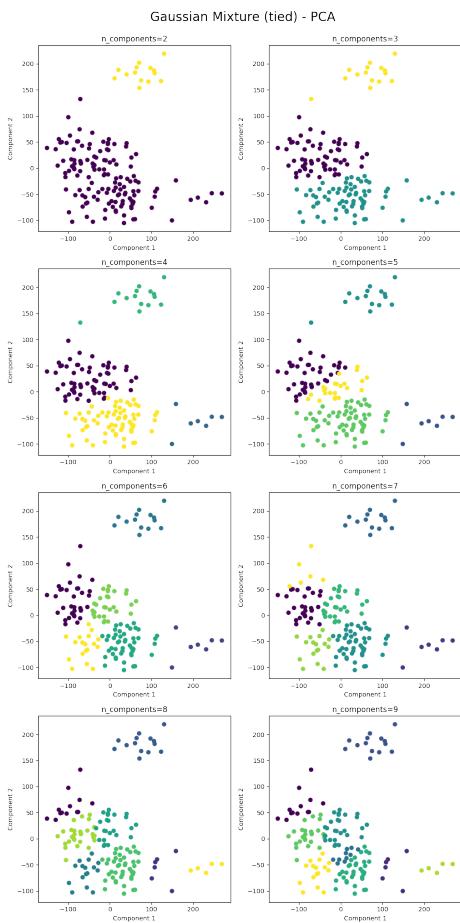


Slika 31. PCA spherical

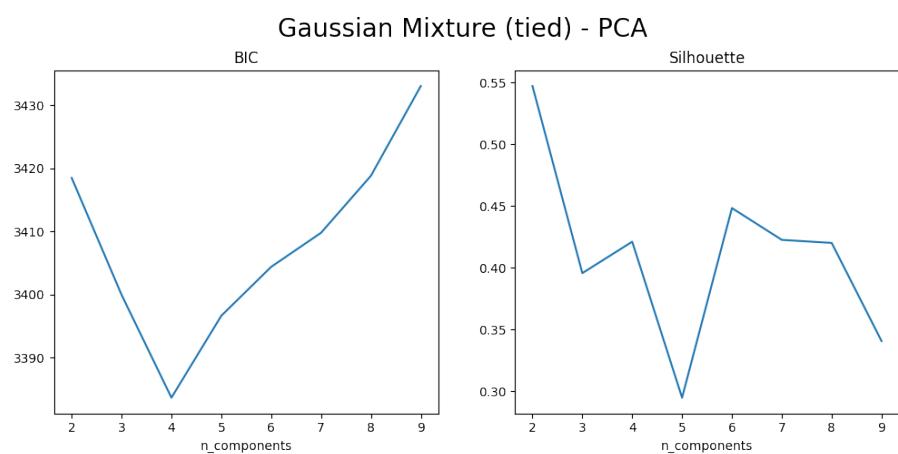


Slika 32. PCA spherical

5.2.4 PCA dataset sa povezanim tipom kovarijanse

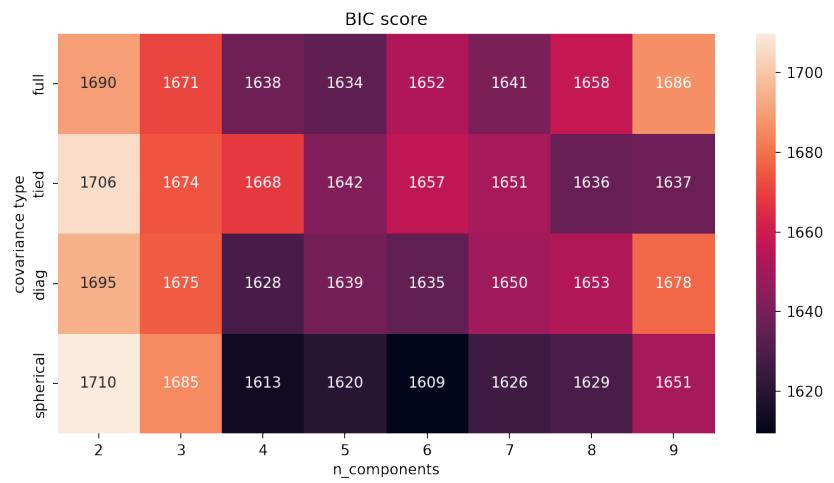


Slika 33. PCA tied

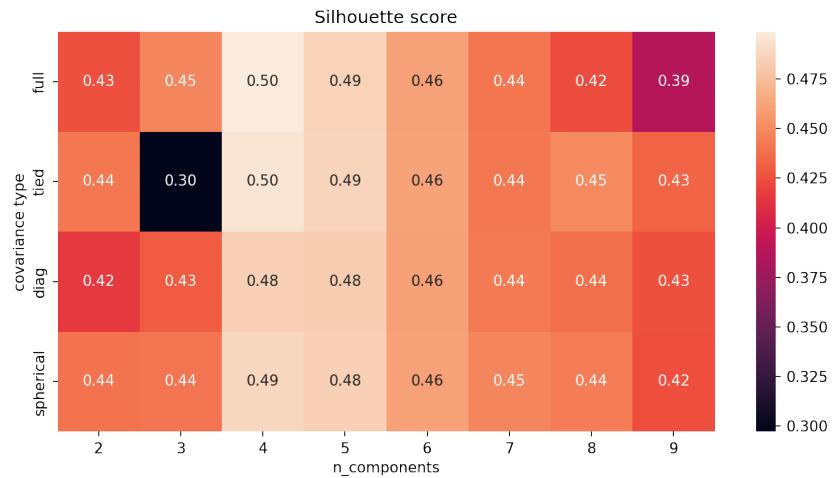


Slika 34. PCA tied

5.2.5 t-SNE dataset - poređenje tipova kovarijanse

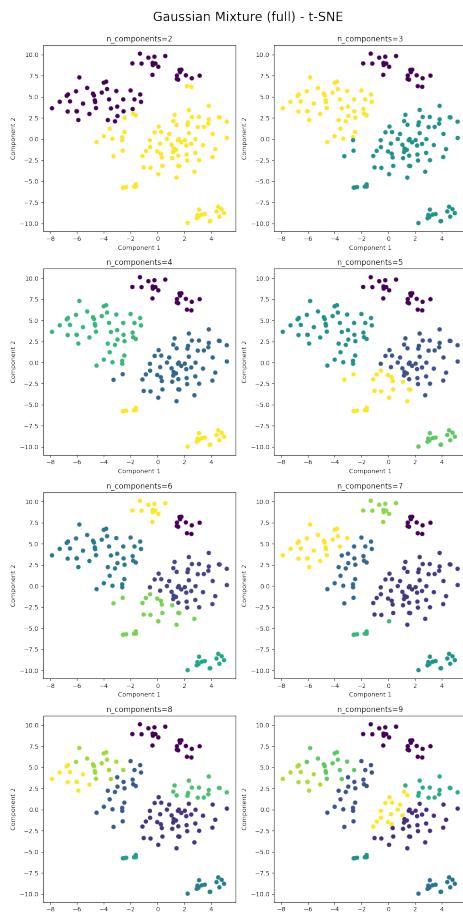


Slika 35. BIC skor u zavisnosti od tipa kovarijanse na t-SNE datasetu

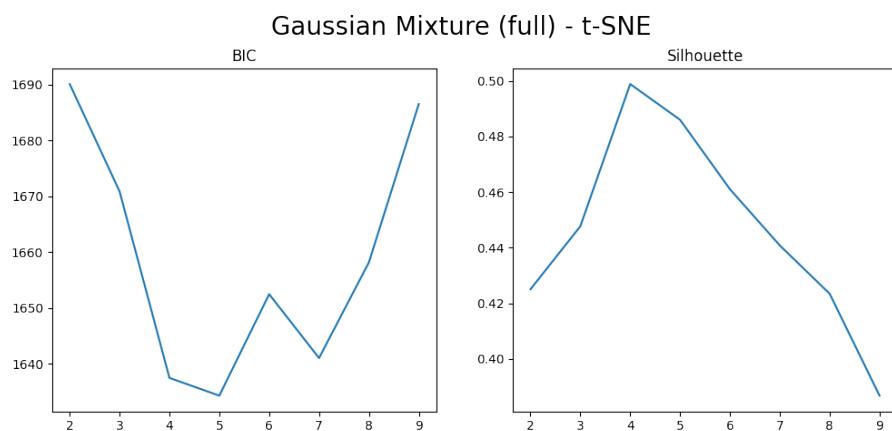


Slika 36. Skor siluete u zavisnosti od tipa kovarijanse na t-SNE datasetu

5.2.6 t-SNE dataset sa punim tipom kovarijanse

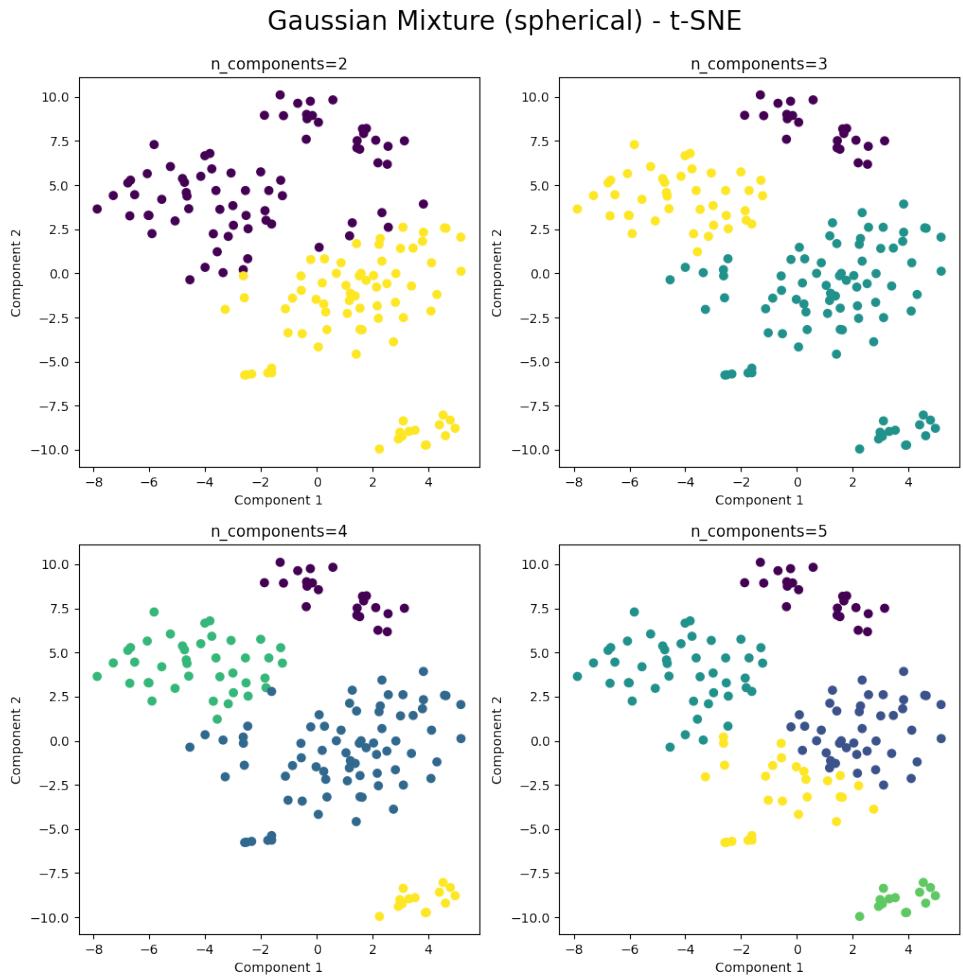


Slika 37. t-SNE full

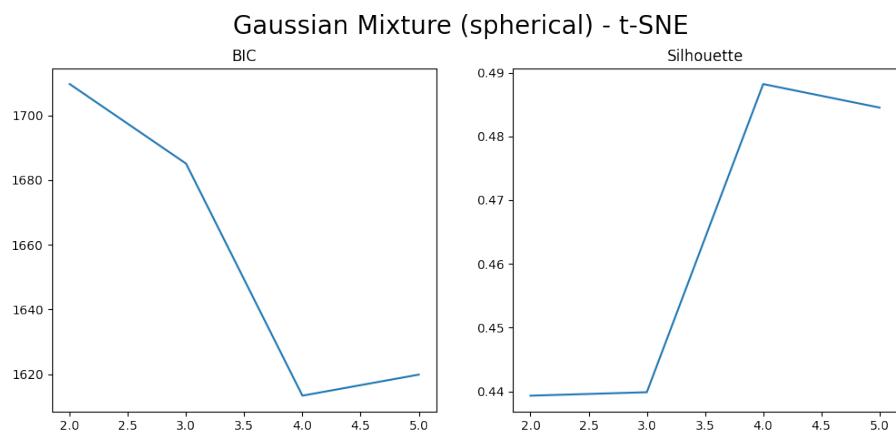


Slika 38. t-SNE full

5.2.7 t-SNE dataset sa sfernim tipom kovarijanse



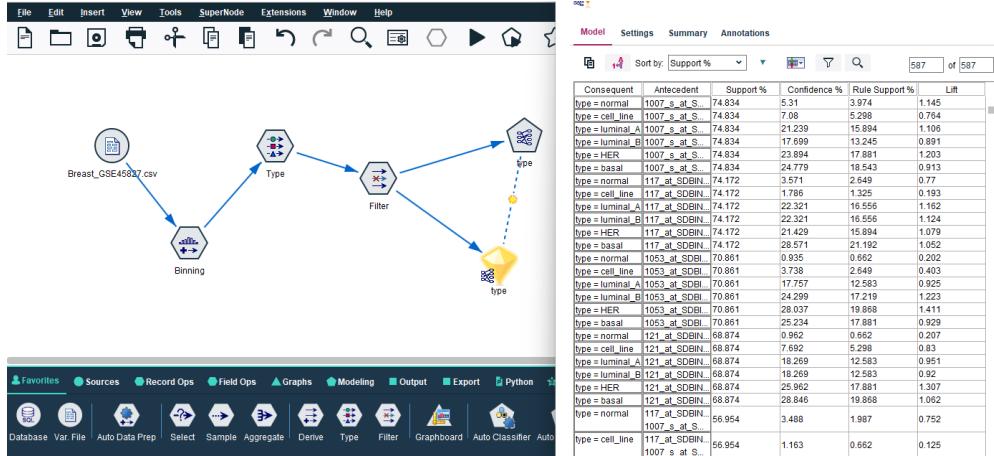
Slika 39. t-SNE spherical



Slika 40. t-SNE spherical

6 Pravila pridruživanja

Korišćen je algoritam Apriori na IBM-ovom programu SPSS. S obzirom da skup podataka sadrži samo kontinualne vrednosti, neophodno je pretvoriti te vrednosti u kategoričke atribute. To radimo koristeći binning koji podatke deli na rang sa ± 2 standardne devijacije. Radi jednostavnosti modela i prikaza izlaza, selektovaćemo samo 4 atributa.



Slika 41. Rezultat Apriori algoritma

Model nažalost ne pokazuje dobra pravila, davajući za neke grupe visok lift, a malu podršku, a na nekim visoku podršku ali mali lift. Može se zaključiti da model nije našao dobra pravila pridruživanja u ovom skupu podataka.

7 Zaključak i diskusija

Klasifikacija je urađena uspešno, očigledno je da sva 3 modela uče. Ni KNN, koji je po rezultatu najgori model, se nije pokazao bezuspešno, davajući rezultat od 84.40%. Modeli su većinski mešali 2 para gena:

- HER i Basal
- Luminal A i Luminal B

Ansambel metoda se pokazala u rangu sa SVM-om, što možemo opravdati gledanjem u matricu konfuzije, videći da su greške slične kroz modele te glasanjem, SVM i KNN izglasaju XGBoost.

Što se klasterovanja tiče, ni KMeans ni gausova mešavina se nisu pokazali dobro u klasterovanju, što nas navodi da nije moguće pronaći dobру opciju za klasterovanje na ovom skupu podataka. Gausova mešavina predstavlja komplikovaniju metodu za klasterovanje, s obzirom da ne prepostavlja oblik klastera, gde zapravo jedan uzorak sadrži verovatnoće da se nalazi u nekom klasteru. Metrike kao što su BIC i skor siluete takođe nisu pomogle u nalaženju dobrih klastera, te možemo prepostaviti da nema smisla raditi klasterovanje na ovom skupu podataka.

U ovom radu prikazane su razne tehnike i razni algoritmi za nadgledano i nenadgledano učenje. Rad je napisan u pratnji sa GitHub repozitorijumom koji sadrži implementaciju svega navedenog.

8 Reference

- [1] Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. (2019) CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology, 26 (4), 376-386.
- [2] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research