Data mining - Breast Cancer

Branko Grbić - Matematički fakultet, Univerzitet u Beogradu
mi20002@alas.matf.bg.ac.rs

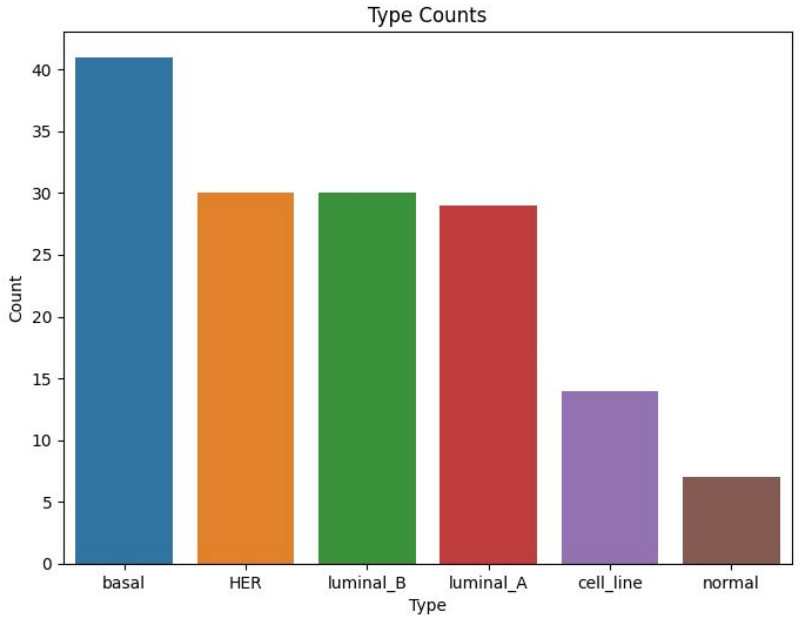
O projektu

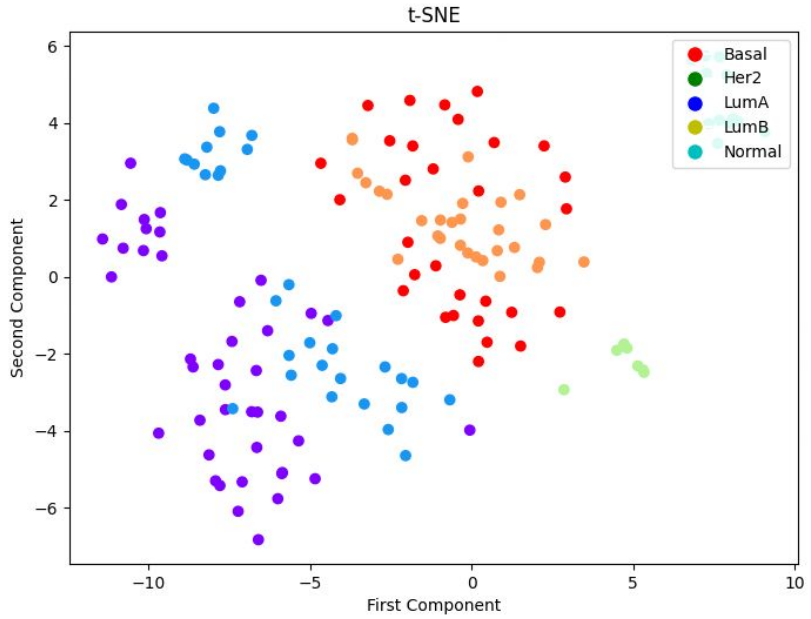
- CuMiDa dataset - 151 entiteta, 54676 atributa
- EDA
  - Informacije o podacima
  - Prvi uvid u zavisnosti i značaj atributa
- Klasifikacija
  - XGBoost
  - SVM
  - KNN
  - Ansambl
- Klasterovanje
  - KMeans
  - Gausova mešavina
- Pravila pridruživanja
  - Apriori

Eksplorativna analiza podataka

- Šest nebalansiranih klasa
- Uvid u atribute - nema naznaka zavisnosti pojedinačnih atributa od klasa
- PCA analiza
  - Objašnjiva varijansa - 90 komponenti = 85% varijanse
- t-SNE analiza
  - Demonstrira veću razliku atributa sa 2 komponente - model ima nelinearne zavisnosti
- Korelacija atributa
  - Vrlo mali broj atributa visoko korelisan
  - Ogroman broj atributa pa se ne može pokrenuti O(n^2) algoritam za uklanjanje atributa

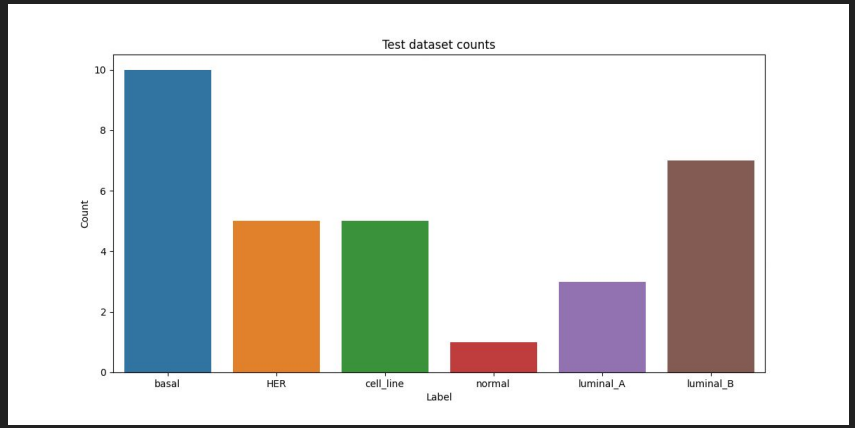
type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	...	AFFX-r2-Ec-bioD-3_at	AFFX-r2-Ec-bioD-5_at	AFFX-r2-P1-cre-3_at	AFFX-r2-P1-cre-5_at	AFFX-ThrX-3_at
basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079	4.963740	4.408328	8.870780	...	12.229711	11.852955	13.658701	13.477698	6.265781
basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283	5.129607	4.584418	7.767646	...	12.178531	11.809408	13.750086	13.470146	6.771853
basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369	4.813449	4.425195	9.417956	...	12.125108	11.725766	13.621732	13.295080	6.346952
basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344	5.490982	4.567956	9.022345	...	12.111235	11.719215	13.743108	13.508861	6.610284
basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457	5.372469	4.424426	9.400056	...	12.173642	11.861296	13.797774	13.542206	6.414354


Type Counts
Count
0
5
10
15
20
25
30
35
40
basal
HER
luminal_B
luminal_A
cell_line
normal
Type


t-SNE
Basal
Her2
LumA
LumB
Normal
Second Component
6
4
2
0
−2
−4
−6
−10
−5
0
5
10
First Component

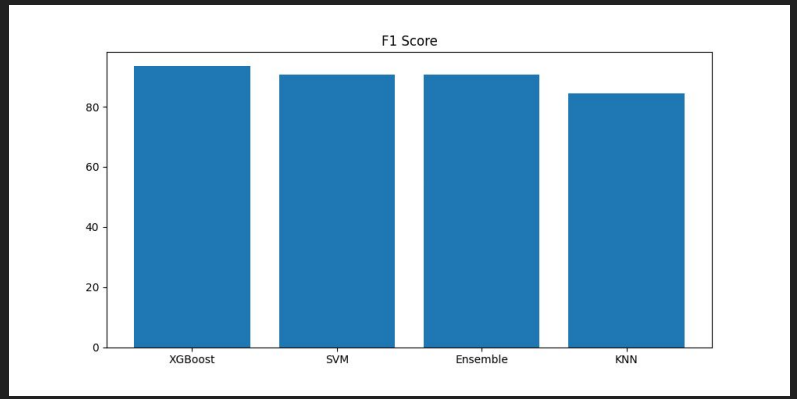
Klasifikacija

- Podela skupa - 80% / 20%
- Nadsemplovanje?
- Više varijanti skupa podataka
  - Običan dataset (Normalized dataset)
  - Dataset sa prvih 90 PCA komponenti (PCA Dataset)
  - Nadsemplovan dataset (Normalized SMOTE dataset)
  - Nadsemplovan dataset sa prvih 90 PCA komponenti (PCA+SMOTE dataset)
- Parametrizacija - Grid Search
- Rezultat - **F1 skor** *(otežinjen)*, tačnost
- Analiza rezultata - Matrica konfuzije

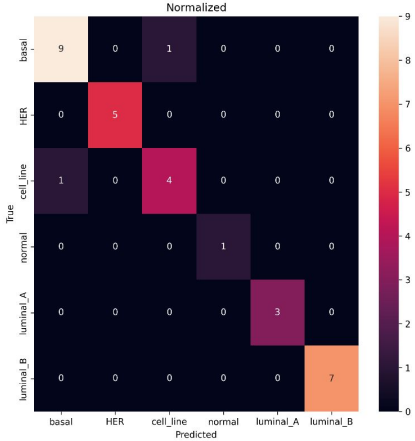


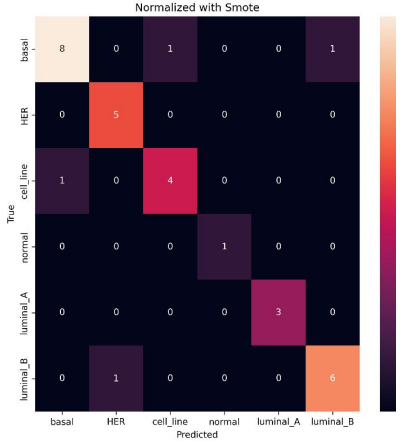
Rezultati

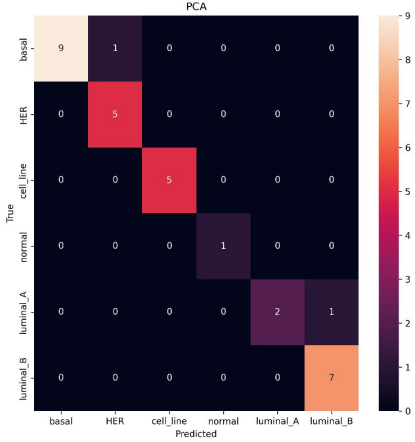
- XGBoost *(Normalized)*
  - Parametri: lr=0.1, max depth=6, estimarors=100
  - F1 skor - 93.55%
- SVM *(Normalized)*
  - Parametri: C=0.1, gamma=1, kernel=linear
  - F1 skor - 90.58%
- KNN *(Normalized with SMOTE)*
  - Parametri: n_neighbors=2, p=1, weights=uniform
  - F1 skor: 84.40%
- Ansambl:
  - F1 skor: 90.58%

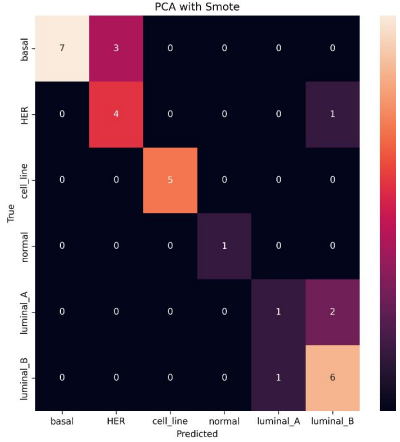


XGBoost

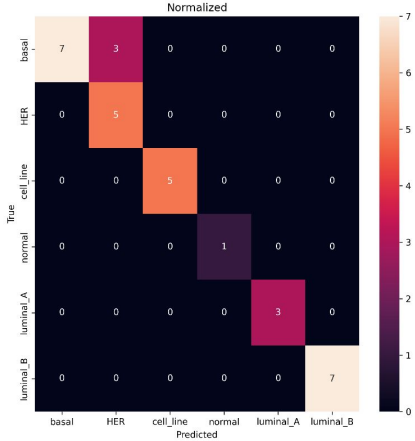

Normalized
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
9 0 1 0 0 0
0 5 0 0 0 0
1 0 4 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7
0 1 2 3 4 5 6 7 8 9

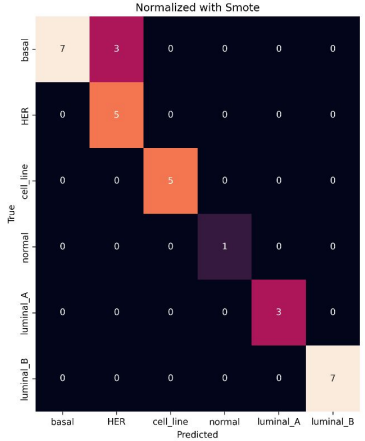

Normalized with Smote
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
8 0 1 0 0 1
0 5 0 0 0 0
1 0 4 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 1 0 0 0 6
0 1 2 3 4 5 6 7 8

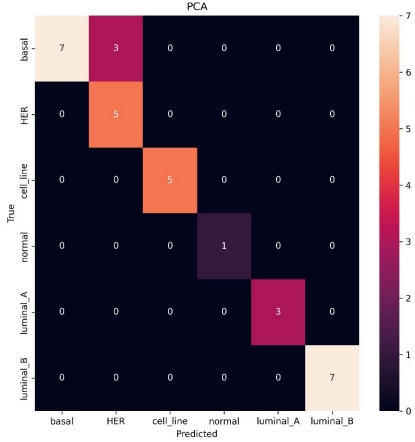

PCA
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
9 1 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 2 1
0 0 0 0 0 7
0 1 2 3 4 5 6 7 8 9

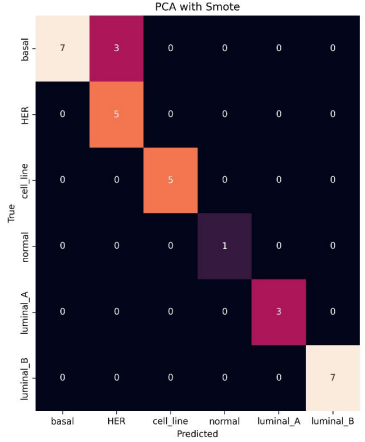

PCA with Smote
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
7 3 0 0 0 0
0 4 0 0 0 1
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 1 2
0 0 0 0 1 6
0 1 2 3 4 5 6 7 8

SVM

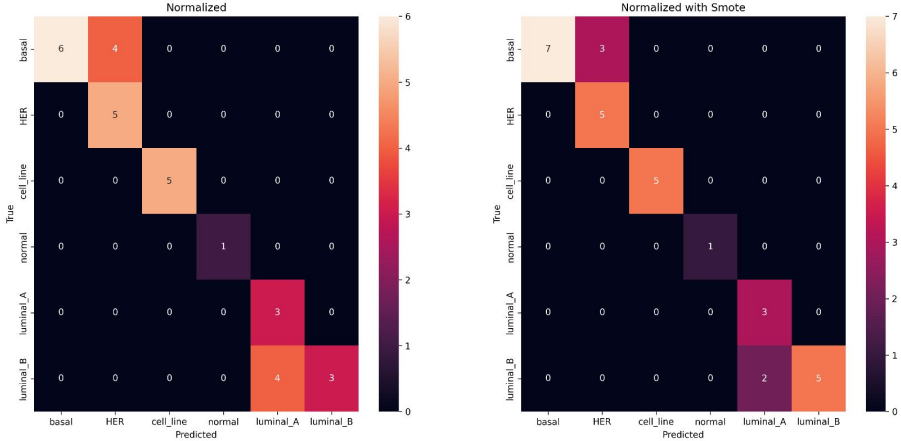

Normalized
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7
0 1 2 3 4 5 6 7


Normalized with Smote
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7

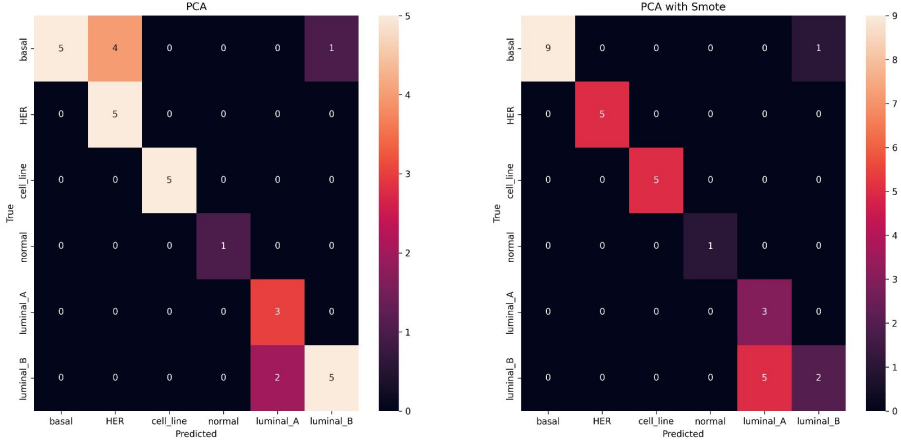

PCA
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7
0 1 2 3 4 5 6 7


PCA with Smote
True
basal
HER
cell_line
normal
luminal_A
luminal_B
Predicted
basal HER cell_line normal luminal_A luminal_B
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7

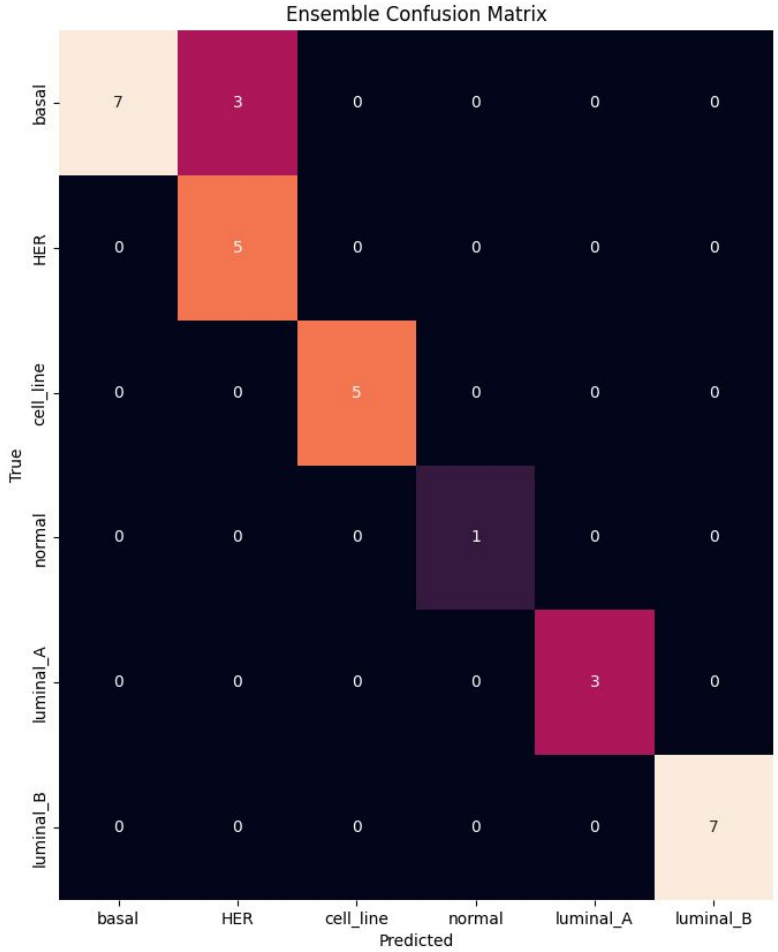
KNN


Normalized
basal
HER
cell_line
normal
luminal_A
luminal_B
True
Predicted
6 4 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 4 3

Normalized with Smote
basal
HER
cell_line
normal
luminal_A
luminal_B
True
Predicted
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 2 5

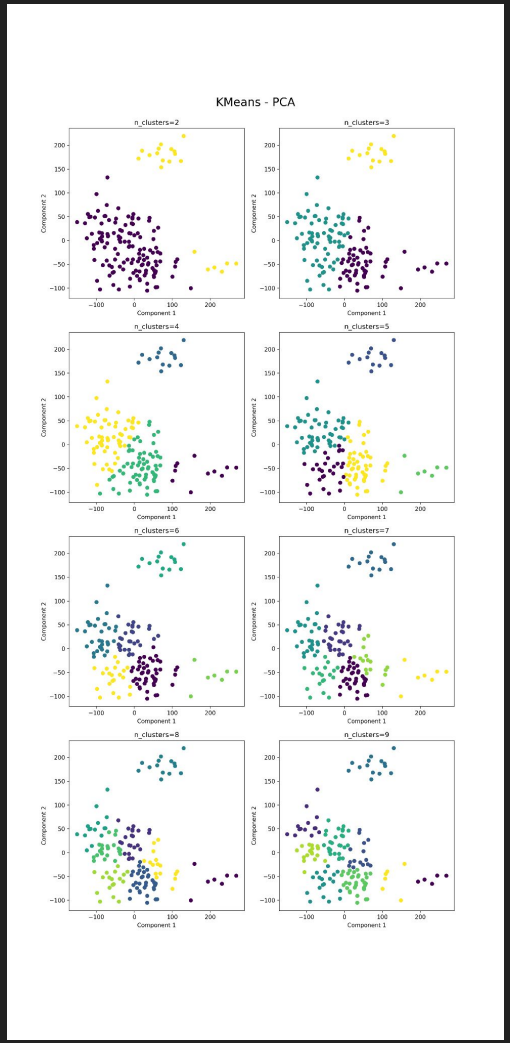

PCA
basal
HER
cell_line
normal
luminal_A
luminal_B
True
Predicted
5 4 0 0 0 1
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 2 5

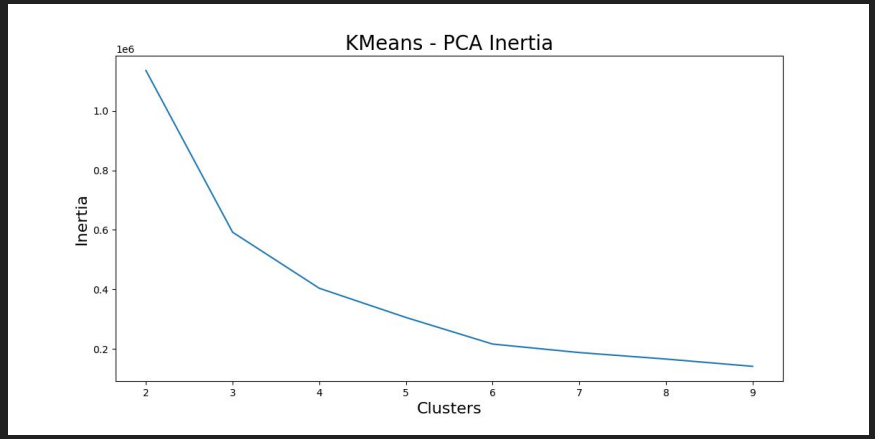
PCA with Smote
basal
HER
cell_line
normal
luminal_A
luminal_B
True
Predicted
9 0 0 0 0 1
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 5 2

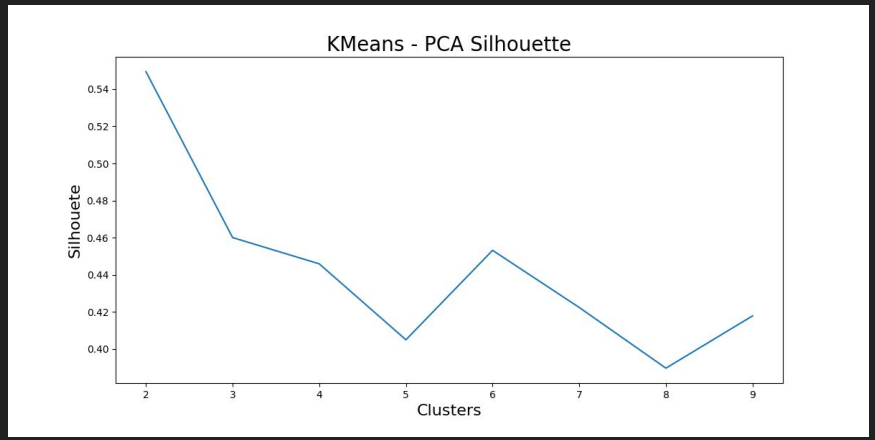

Ensemble Confusion Matrix
basal
HER
cell_line
normal
luminal_A
luminal_B
True
Predicted
7 3 0 0 0 0
0 5 0 0 0 0
0 0 5 0 0 0
0 0 0 1 0 0
0 0 0 0 3 0
0 0 0 0 0 7

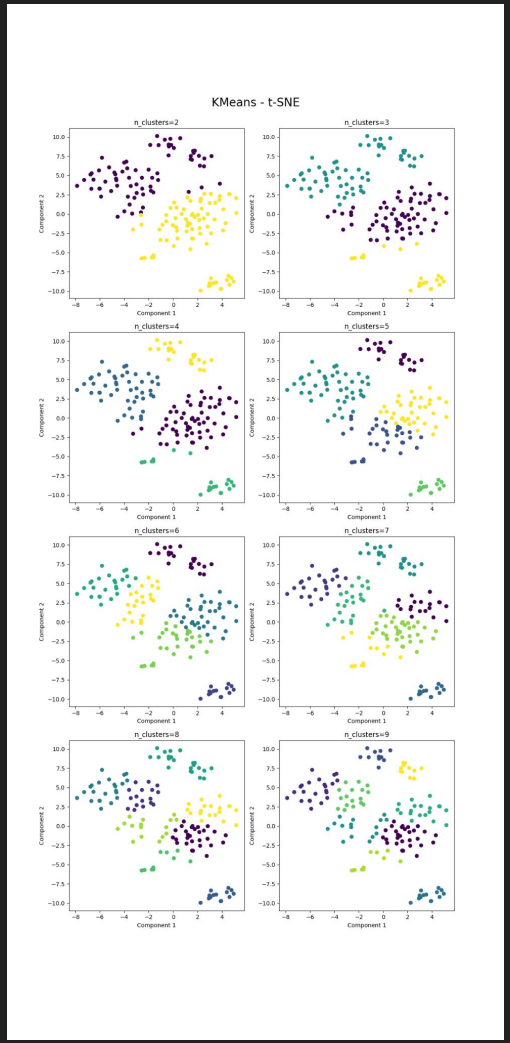
Klasterovanje

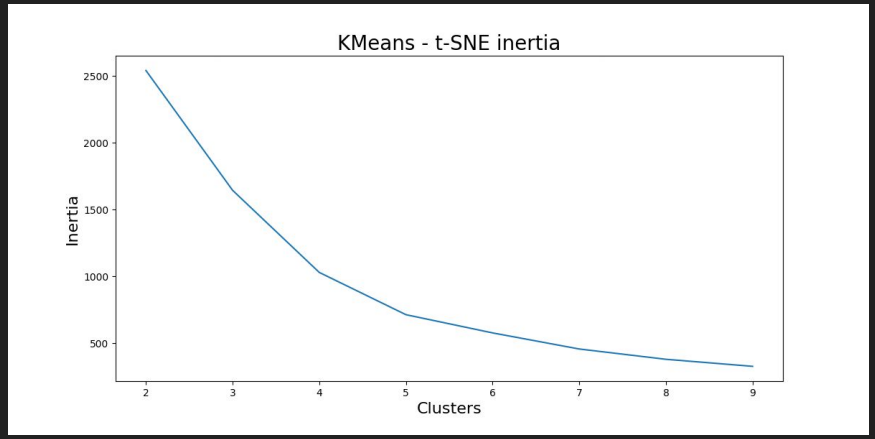
- KMeans
  - Broj klastera - [2, 9]
  - Inertia
  - Skor siluete
- Gausova mešavina
  - Broj klastera - [2, 9]
  - Kovarijantni tip?
  - BIC skor
  - Skor siluete
- Korišćeni su PCA i t-SNE
- Rezultati nezadovoljavajući po metrikama kvaliteta

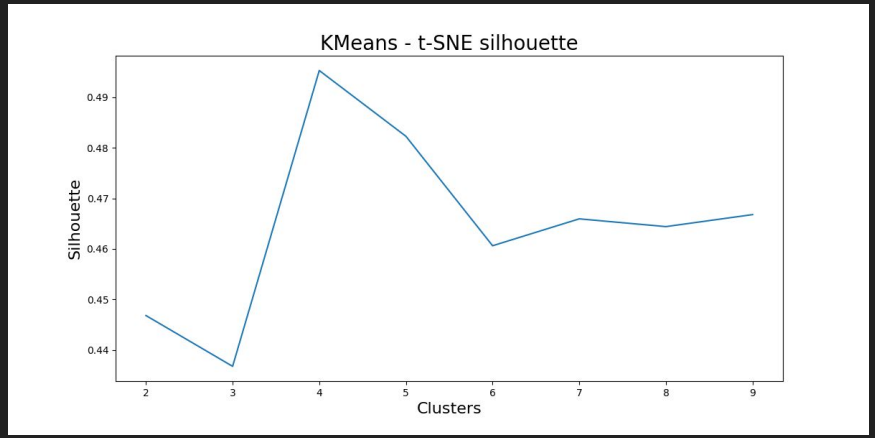

KMeans - PCA
n_clusters=2
n_clusters=3
n_clusters=4
n_clusters=5
n_clusters=6
n_clusters=7
n_clusters=8
n_clusters=9
Component 1
Component 2


KMeans - PCA Inertia
1e6
Inertia
Clusters


KMeans - PCA Silhouette
Silhouete
Clusters


KMeans - t-SNE
n_clusters=2
n_clusters=3
n_clusters=4
n_clusters=5
n_clusters=6
n_clusters=7
n_clusters=8
n_clusters=9
Component 1
Component 2


KMeans - t-SNE inertia
Inertia
Clusters


KMeans - t-SNE silhouette
Silhouette
Clusters

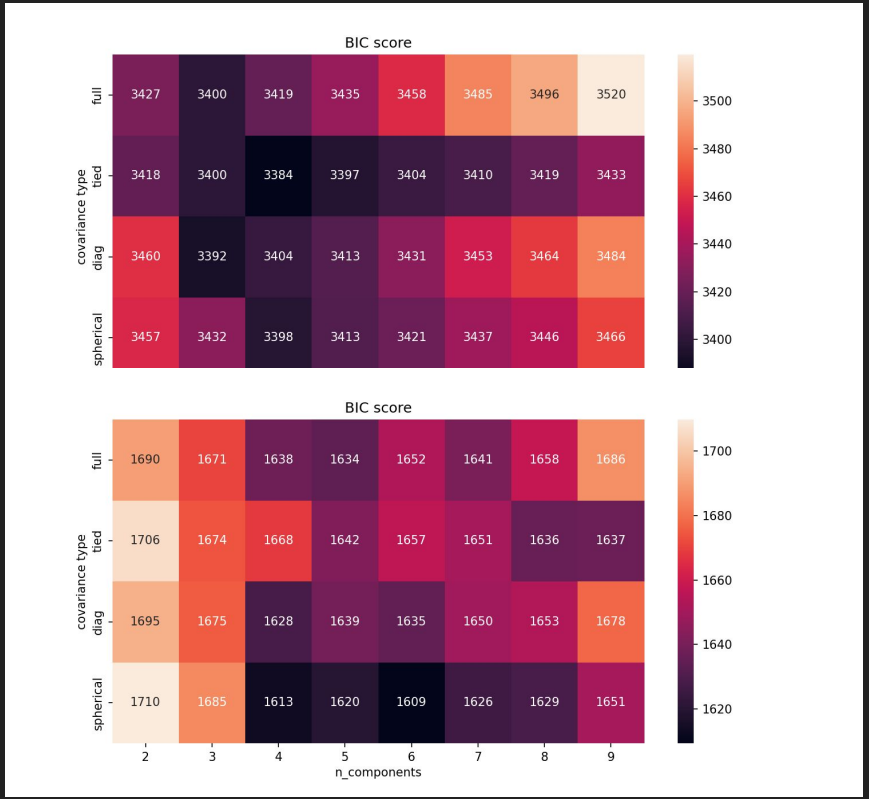
Gausova mešavina

- Kovarijantni tipovi
  - Full
  - Tied
  - Diag
  - Spherical

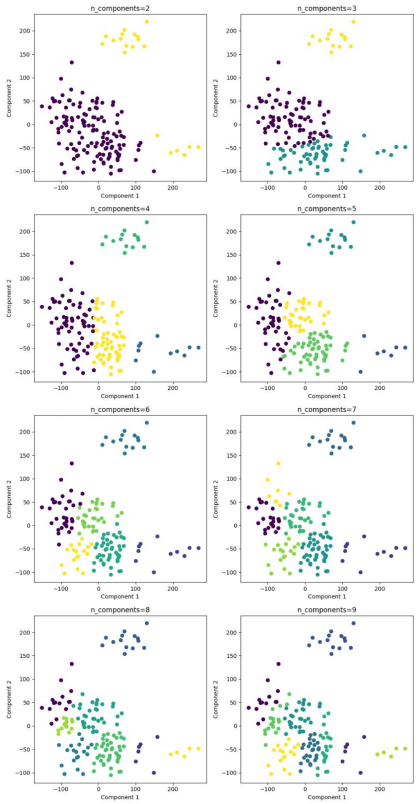
PCA

- Kako odrediti najboljeg?
  - Niži BIC - bolji kvalitet
  - Visok skor siluete - bolji kvalitet
  - PCA vs t-SNE

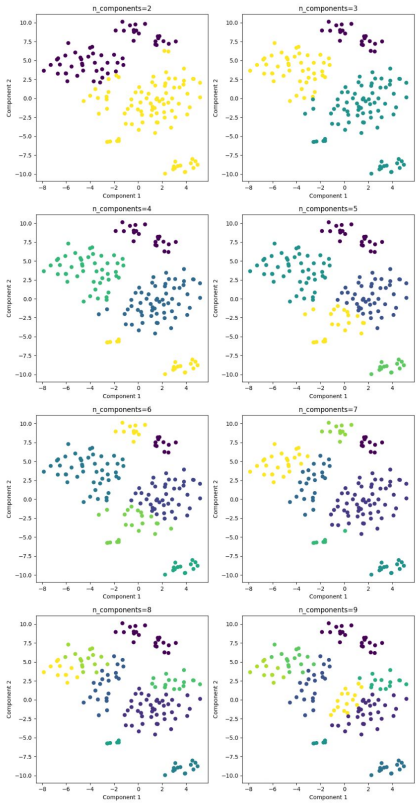
t-SNE



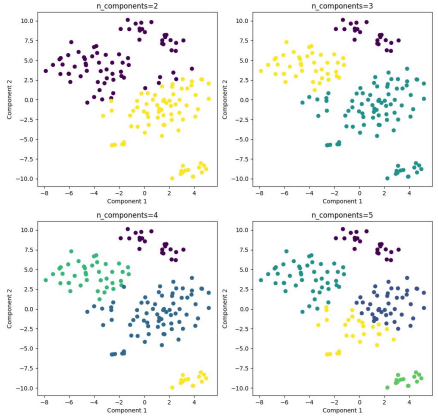
Gaussian Mixture (full) - PCA



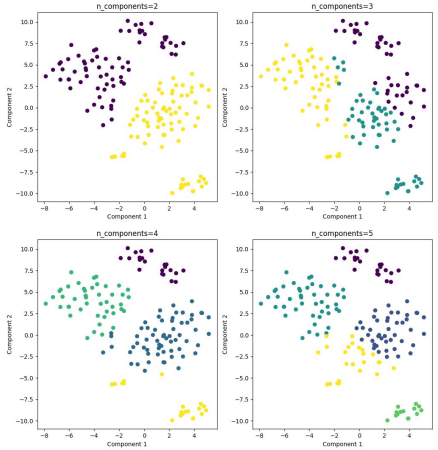
Gaussian Mixture (full) - t-SNE



Gaussian Mixture (spherical) - t-SNE



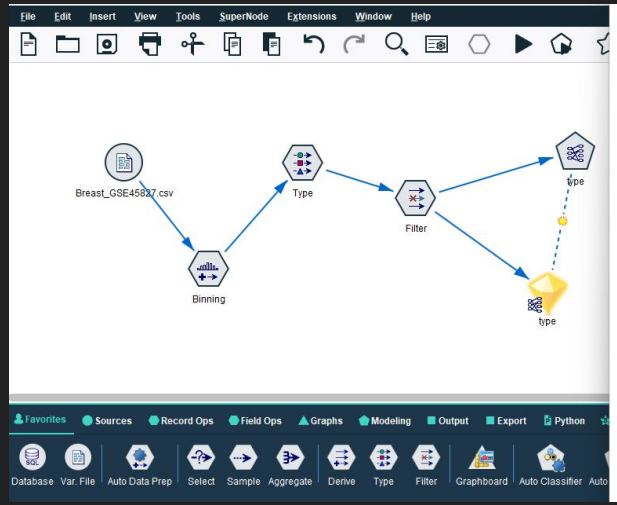
Gaussian Mixture (tied) - t-SNE

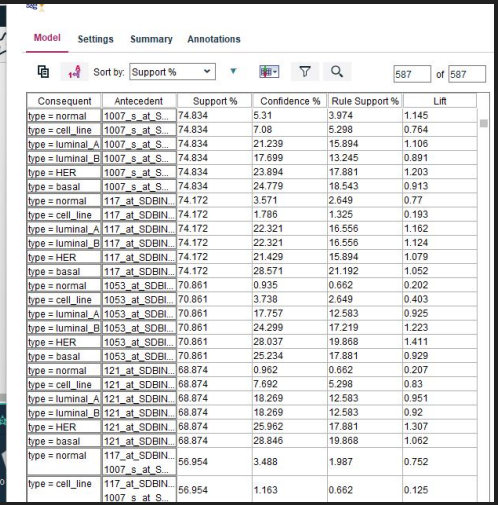


Pravila pridruživanja

- Apriori algoritam
- Zahtevaju se kategoričke klase
  - Odraditi pridruživanje na par atributa
  - Razdvojiti kontinualne atribute u kategoričke
  - Binning - po +2 STD

- Rezultat
  - Mali lift - Velika podrška
  - Veliki lift - Mala podrška
  - Ne nalazi dobra pravila





Consequent	Antecedent	Support %	Confidence %	Rule Support %	Lift
type = normal	1007_s_at_S...	74.834	5.31	3.974	1.145
type = cell_line	1007_s_at_S...	74.834	7.08	5.298	0.764
type = luminal_A	1007_s_at_S...	74.834	21.239	15.894	1.106
type = luminal_B	1007_s_at_S...	74.834	17.699	13.245	0.891
type = HER	1007_s_at_S...	74.834	23.894	17.881	1.203
type = basal	1007_s_at_S...	74.834	24.779	18.543	0.913
type = normal	117_at_SDBIN...	74.172	3.571	2.649	0.77
type = cell_line	117_at_SDBIN...	74.172	1.786	1.325	0.193
type = luminal_A	117_at_SDBIN...	74.172	22.321	16.556	1.162
type = luminal_B	117_at_SDBIN...	74.172	22.321	16.556	1.124
type = HER	117_at_SDBIN...	74.172	21.429	15.894	1.079
type = basal	117_at_SDBIN...	74.172	28.571	21.192	1.052
type = normal	1053_at_SDBI...	70.861	0.935	0.662	0.202
type = cell_line	1053_at_SDBI...	70.861	3.738	2.649	0.403
type = luminal_A	1053_at_SDBI...	70.861	17.757	12.583	0.925
type = luminal_B	1053_at_SDBI...	70.861	24.299	17.219	1.223
type = HER	1053_at_SDBI...	70.861	28.037	19.868	1.411
type = basal	1053_at_SDBI...	70.861	25.234	17.881	0.929
type = normal	121_at_SDBIN...	68.874	0.962	0.662	0.207
type = cell_line	121_at_SDBIN...	68.874	7.692	5.298	0.83
type = luminal_A	121_at_SDBIN...	68.874	18.269	12.583	0.951
type = luminal_B	121_at_SDBIN...	68.874	18.269	12.583	0.92
type = HER	121_at_SDBIN...	68.874	25.962	17.881	1.307
type = basal	121_at_SDBIN...	68.874	28.846	19.868	1.062
type = normal	117_at_SDBIN... 1007_s_at_S...	56.954	3.488	1.987	0.752
type = cell_line	117_at_SDBIN... 1007_s_at_S...	56.954	1.163	0.662	0.125

Hvala na pažnji

Branko Grbić 2/2020
mi20002@alas.matf.bg.ac.rs
GitHub link