HIGGS Dataset

Istraživanje podataka 1

Filip Ogrenjac

Matematički fakultet Univerziteta u Beogradu

Sadržaj

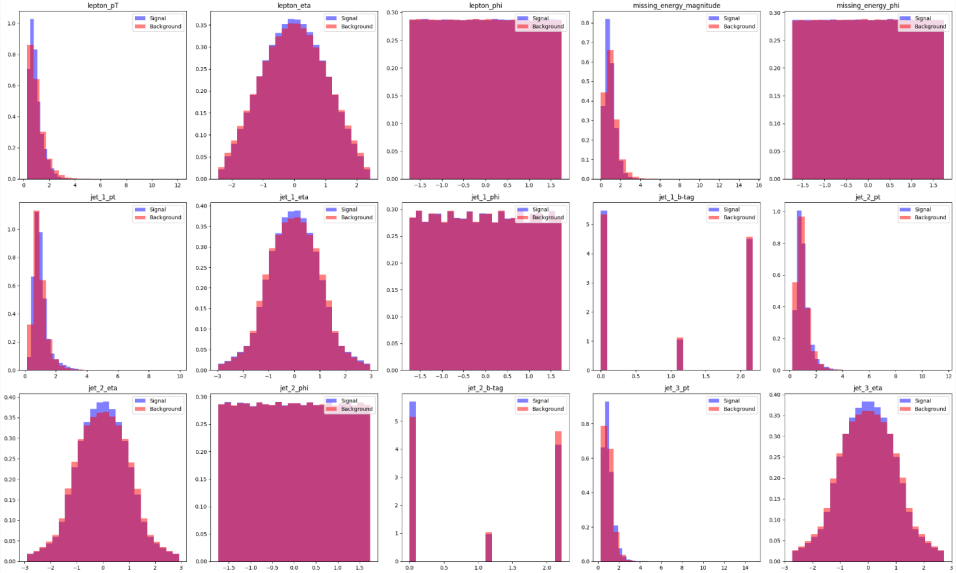
Analiza skupa podataka

- *Svaki uzorak - 28 osobina + target*
- *21 osobina - izmerene kinematičke osobine*
- *Preostalih 7 osobina - funkcije prvih 21 osobina*
- *"HIGGS dataset" - kolekcija od 11 miliona uzoraka*
- *Klasifikacija procesa u dve kategorije:*
  - *Signalni proces koji proizvodi Higsov bozon*
  - *Pozadinski proces koji ne proizvodi Higsov bozon*

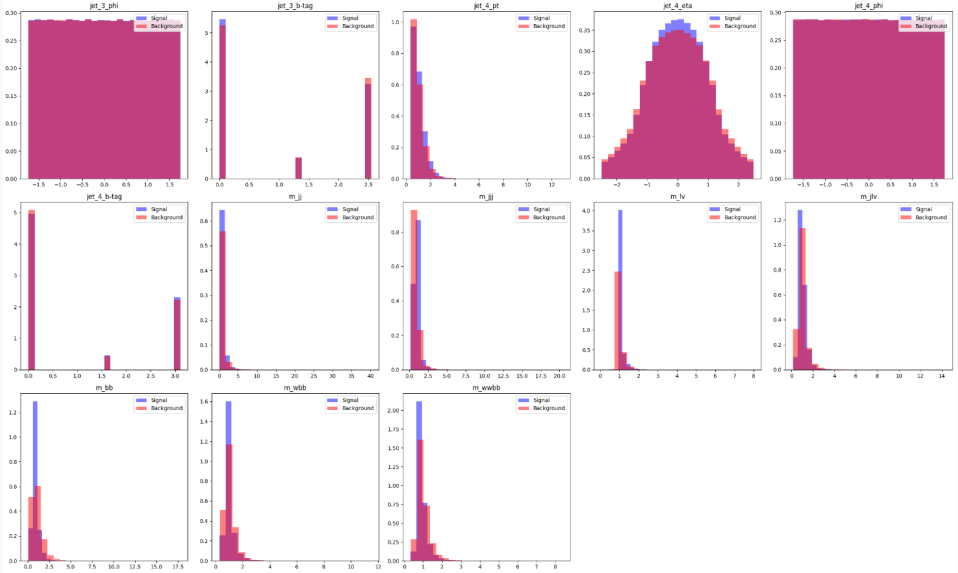
target	count	percentage
1.0	5829123	53.0%
0.0	5170877	47.0%

Balansiranost klasa

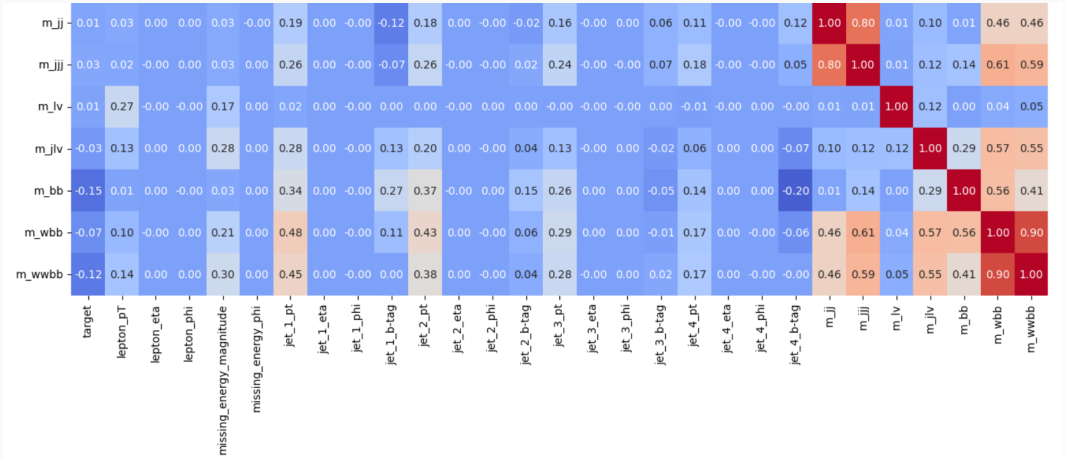
Analiza skupa podataka - Normalizovani histogrami



Analiza skupa podataka - Normalizovani histogrami



Analiza skupa podataka



Matrica korelacije poslednjih 7 atributa

Pretprocesiranje

```
1 print('Number of null values:\n')
2 df.isna().sum()
```

```
Number of null values:


target                      0
lepton_pT                   0
lepton_eta                  0
lepton_phi                  0
missing_energy_magnitude    0
missing_energy_phi          0
jet_1_pt                    0
jet_1_eta                   0
jet_1_phi                   0
jet_1_b-tag                 0
jet_2_pt                    0
jet_2_eta                   0
jet_2_phi                   0
jet_2_b-tag                 0
jet_3_pt                    0
jet_3_eta                   0
jet_3_phi                   0
jet_3_b-tag                 0
jet_4_pt                    0
jet_4_eta                   0
jet_4_phi                   0
jet_4_b-tag                 0
m_jj                        0
m_jjj                       0
m_lv                        0
m_jlv                       0
m_bb                        0
m_wbb                       0
m_wwbb                      0
```

Broj nedostajućih vrednosti za svaki atribut

	lower	min	num_lower	upper	max	num_upper	percentage
lepton_pT	-0.377456	0.274697	0	2.204436	12.098914	301057	3.91%
lepton_eta	-2.955563	-2.434976	0	2.954480	2.434868	0	0.00%
lepton_phi	-3.485763	-1.742508	0	3.485936	1.743236	0	0.00%
missing_energy_magnitude	-0.497921	0.000394	0	2.367797	12.843856	227758	2.96%
missing_energy_phi	-3.485751	-1.743944	0	3.485697	1.743257	0	0.00%
jet_1_pt	-0.058903	0.137502	0	1.908820	9.940391	381233	4.95%
jet_1_eta	-2.750389	-2.969725	24164	2.751328	2.969674	24157	0.63%
jet_1_phi	-3.472709	-1.741237	0	3.472926	1.741454	0	0.00%
jet_1_b-tag	-3.259614	0.000000	0	5.432690	2.173076	0	0.00%
jet_2_pt	-0.161345	0.188981	0	2.019807	10.860058	308401	4.01%
jet_2_eta	-2.780497	-2.913090	17098	2.779646	2.913210	17198	0.45%
jet_2_phi	-3.480257	-1.742372	0	3.479950	1.743175	0	0.00%
jet_2_b-tag	-3.322308	0.000000	0	5.537180	2.214872	0	0.00%
jet_3_pt	-0.205565	0.263608	0	2.078215	11.155643	247327	3.21%
jet_3_eta	-2.799752	-2.729663	0	2.800098	2.730009	0	0.00%
jet_3_phi	-3.484928	-1.742069	0	3.485188	1.742884	0	0.00%
jet_3_b-tag	-3.822337	0.000000	0	6.370561	2.548224	0	0.00%
jet_4_pt	-0.286977	0.365354	0	2.125675	12.882567	259522	3.37%
jet_4_eta	-2.854546	-2.497265	0	2.855290	2.498009	0	0.00%
jet_4_phi	-3.486105	-1.742691	0	3.486232	1.743372	0	0.00%
jet_4_b-tag	-4.652942	0.000000	0	7.754903	3.101961	0	0.00%
m_jj	0.439426	0.075070	198187	1.375836	40.192368	876420	13.96%
m_jjj	0.490419	0.234753	13670	1.439329	20.372782	562017	7.48%
m_lv	0.933704	0.083049	20800	1.072493	7.992739	1510177	19.88%
m_jlv	0.205605	0.157473	5	1.704215	14.262439	393746	5.11%
m_bb	-0.023144	0.047862	0	1.835426	17.762852	466826	6.06%
m_wbb	0.337827	0.295112	16	1.622123	11.496522	482902	6.27%
m_wwbb	0.337072	0.347443	0	1.492587	8.374498	461756	6.00%

Pregled informacija o autlajerima

Stabla odlučivanja

```
param_grid_dt = {"max_depth": [4, 6, 8, 12],
                 "criterion": ["gini", "entropy"]}
param_grid_rf = {"n_estimators": [100, 150],
                 "max_depth": [4, 6, 8, 12],
                 "criterion": ["gini", "entropy"]}
```

Parametri koji se prosleđuju GridSearch-u

```
print("Best parameters for Decision Tree: ", grid_search_dt.best_params_)
print("Best score for Decision Tree: ", grid_search_dt.best_score_)
```

```
Best parameters for Decision Tree:  {'criterion': 'entropy', 'max_depth': 12}
Best score for Decision Tree:  0.7026714285714286
```

Najbolji parametri za DecisionTree

```
print("Best parameters for Random Forest: ", grid_search_rf.best_params_)
print("Best score for Random Forest: ", grid_search_rf.best_score_)
```

```
Best parameters for Random Forest:  {'criterion': 'gini', 'max_depth': 12, 'n_estimators': 150}
Best score for Random Forest:  0.7185207792207792
```

Najbolji parametri za RandomForest

Stabla odlučivanja

```
Train set
-------------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.71      0.70      0.70    361961
         1.0       0.74      0.75      0.74    408039

    accuracy                           0.73    770000
   macro avg       0.72      0.72      0.72    770000
weighted avg       0.73      0.73      0.73    770000

-------------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------------
[[251796 110165]
 [101038 307001]]
```

Izveštaj o klasifikaciji na train skupu za DecisionTree

```
Test set
-------------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.69      0.67      0.68   1551263
         1.0       0.72      0.73      0.72   1748737

    accuracy                           0.70   3300000
   macro avg       0.70      0.70      0.70   3300000
weighted avg       0.70      0.70      0.70   3300000

-------------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------------
[[1044445  506818]
 [ 468398 1280339]]
```

Izveštaj o klasifikaciji na test skupu za DecisionTree

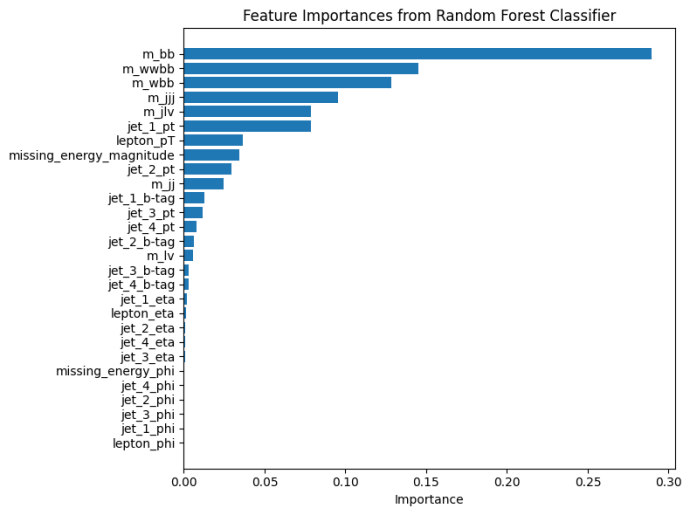
```
Train set
-------------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.74      0.70      0.72    361961
         1.0       0.74      0.78      0.76    408039

    accuracy                           0.74    770000
   macro avg       0.74      0.74      0.74    770000
weighted avg       0.74      0.74      0.74    770000

-------------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------------
[[252279 109682]
 [ 89192 318847]]
```

Izveštaj o klasifikaciji na train skupu za RandomForest

```
Test set
-------------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.71      0.67      0.69   1551263
         1.0       0.72      0.76      0.74   1748737

    accuracy                           0.72   3300000
   macro avg       0.72      0.72      0.72   3300000
weighted avg       0.72      0.72      0.72   3300000

-------------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------------
[[1044794  506469]
 [ 422190 1326547]]
```

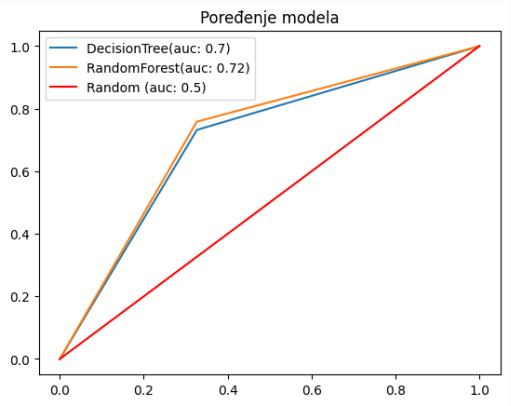
Izveštaj o klasifikaciji na test skupu za RandomForest

Stabla odlučivanja

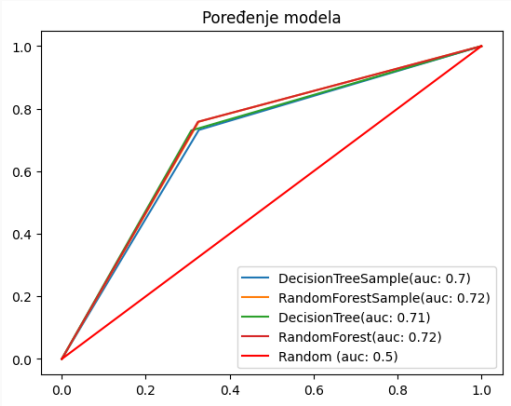


Važnost osobina kod RandomForest modela

Poređenje DecisionTree i RandomForest modela



ROC kriva - uzorak 10%



ROC kriva - uzorak 10% i ceo skup

K najbližih suseda

```
knn_params = {
    'n_neighbors': [3, 5, 7],
    'weights': ['uniform', 'distance'],
    'p': [1, 2]
}
```

Parametri koji se prosleđuju GridSearch-u

```
knn_grid_no_outliers.best_estimator_
```

```
KNeighborsClassifier(n_neighbors=7, p=1, weights='distance')
```

```
knn_grid_no_outliers.best_score_
```

```
0.6509490681095937
```

Najbolji parametri i skor za GridSearch bez autlajera

```
knn_grid_replaced.best_estimator_
```

```
KNeighborsClassifier(n_neighbors=7, p=1, weights='distance')
```

```
knn_grid_replaced.best_score_
```

```
0.6481818231230805
```

Najbolji parametri i skor za GridSearch sa zamenjenim autlajerima

K najbližih suseda

```
Train set
-------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     15558
         1.0       1.00      1.00      1.00     19371

    accuracy                           1.00     34929
   macro avg       1.00      1.00      1.00     34929
weighted avg       1.00      1.00      1.00     34929

-------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------
[[15558     0]
 [    0 19371]]
```

Train skup za kNN (GridSearch best) bez autlajera

```
Test set
-------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.64      0.54      0.59   1551263
         1.0       0.64      0.73      0.68   1748737

    accuracy                           0.64   3300000
   macro avg       0.64      0.63      0.63   3300000
weighted avg       0.64      0.64      0.64   3300000

-------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------
[[ 842954  708309]
 [ 478460 1270277]]
```

Test skup za kNN (GridSearch best) bez autlajera

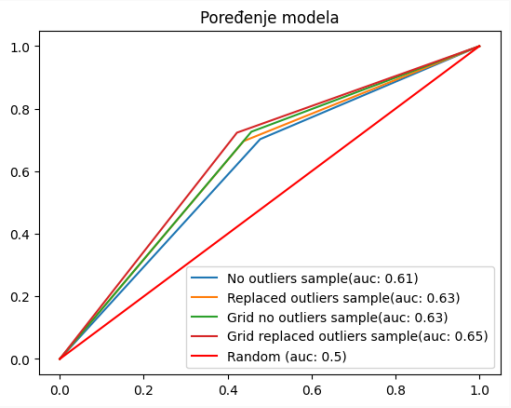
```
Train set
-------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     36196
         1.0       1.00      1.00      1.00     40804

    accuracy                           1.00     77000
   macro avg       1.00      1.00      1.00     77000
weighted avg       1.00      1.00      1.00     77000

-------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------
[[36196     0]
 [    0 40804]]
```

Train skup za kNN (GridSearch best) zamenjeni autlajeri

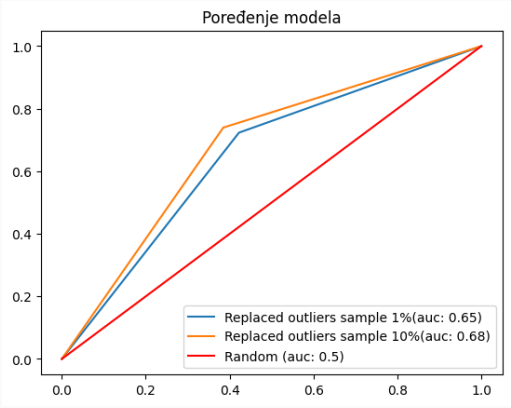
```
Test set
-------------------------------------------------------------------------
Classification report:
-------------------------------------------------------------------------
              precision    recall  f1-score   support

         0.0       0.65      0.58      0.61   1551263
         1.0       0.66      0.72      0.69   1748737

    accuracy                           0.65   3300000
   macro avg       0.65      0.65      0.65   3300000
weighted avg       0.65      0.65      0.65   3300000

-------------------------------------------------------------------------
Confusion matrix:
-------------------------------------------------------------------------
[[ 896265  654998]
 [ 483690 1265047]]
```

Test skup za kNN (GridSearch best) zamenjeni autlajeri

Poređenje KNeighbors modela



ROC kriva - uzorak 1%



ROC kriva - uzorak 1% i 10%

Analiza glavnih komponenti (PCA)

```
explained_variance_ratio = pca.explained_variance_ratio_
explained_variance_ratio
```

```
array([0.13467162, 0.06706966, 0.06485308])
```

```
sum(explained_variance_ratio)
```

```
0.26659435448558816
```

Ukupna varijansa korišćenjem svih atributa za PCA na 3 dimenzije

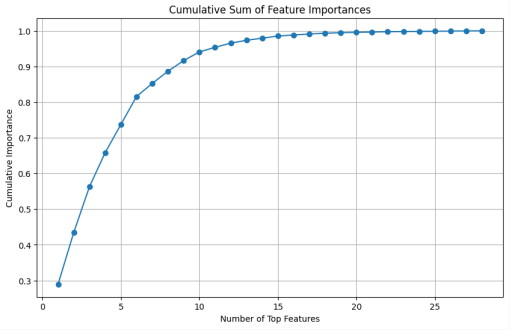
```
explained_variance_ratio = pca.explained_variance_ratio_
explained_variance_ratio
```

```
array([0.5698167 , 0.19029378, 0.14908575])
```

```
sum(explained_variance_ratio)
```

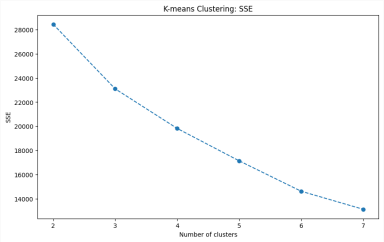
```
0.9091962345452064
```

Ukupna varijansa korišćenjem najvažnijih 5 atributa za PCA na 3 dimenzije

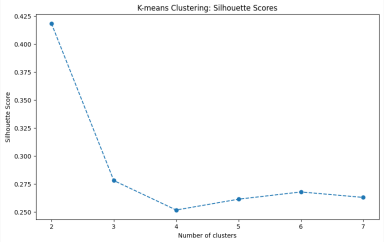


Kumulativna suma značaja atributa za RandomForest model

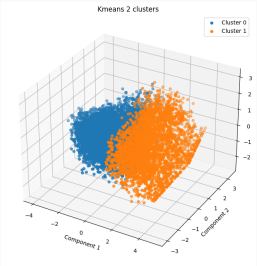
K-sredina



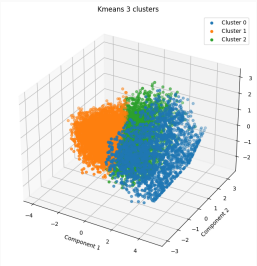
SSE u zavisnosti od broja k



Silhouette score u zavisnosti od broja k

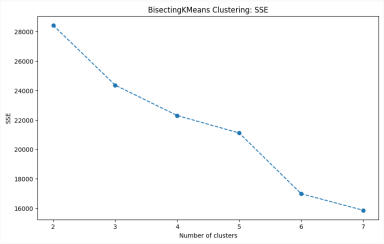


K-means: $k = 2$

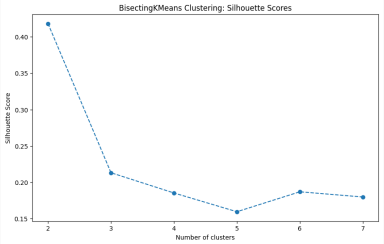


K-means: $k = 3$

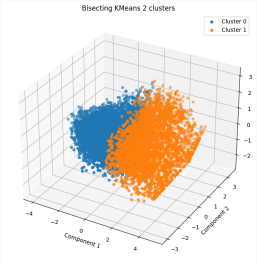
K-sredina sa bisekcijom



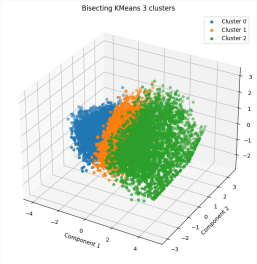
SSE u zavisnosti od broja k



Silhouette score u zavisnosti od broja k

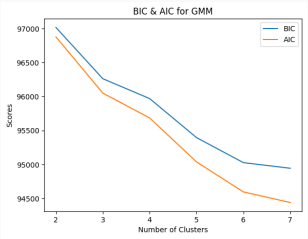


Bisecting K-means: $k = 2$

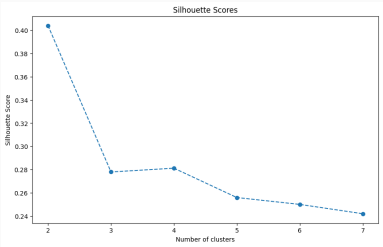


Bisecting K-means: $k = 3$

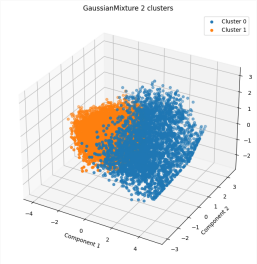
Gausov model mešavine



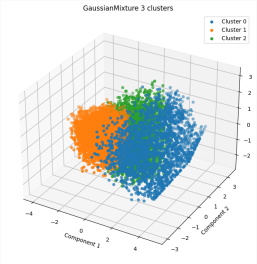
BIC i AIC score u zavisnosti od broja k



Silhouette score u zavisnosti od broja k



GMM: $k = 2$



GMM: $k = 3$

Poređenje modela za klasterovanje

```
Adjusted Rand Index: 0.017671976200457913
Normalized Mutual Information: 0.011367572951140239
```

K-means: $k = 2$ - *ARI i NMI*

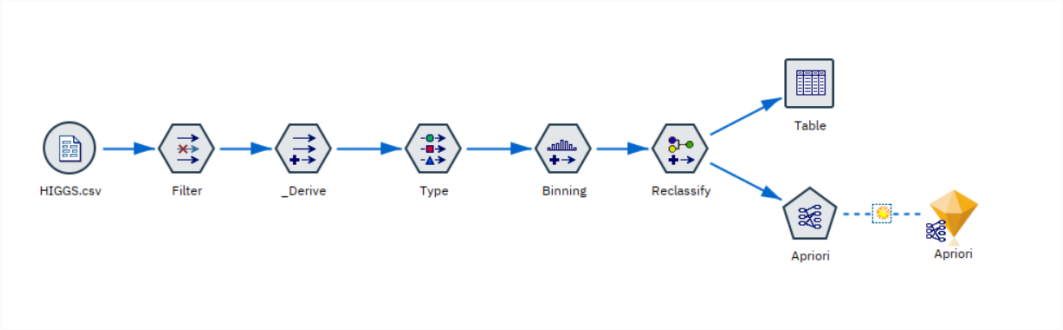
```
Adjusted Rand Index: 0.017623281588269646
Normalized Mutual Information: 0.011330090937800304
```

Bisecting K-means: $k = 2$ - *ARI i NMI*

```
Adjusted Rand Index: 0.015533695769814688
Normalized Mutual Information: 0.009679067376613715
```

GMM: $k = 2$ - *ARI i NMI*

Apriori algoritam



Tok operacija

Apriori algoritam

Consequent	Antecedent	Support %	Confidence %	Lift
target_Derive = Not b...	m_jlv_LMH = Medium m_wwbb_LMH = High m_jjj_LMH = Medium	2.822	86.435	1.839
target_Derive = Not b...	m_jlv_LMH = Medium m_wwbb_LMH = High m_jjj_LMH = Low	1.963	83.641	1.779
target_Derive = Not b...	m_wwbb_LMH = High m_jjj_LMH = Medium m_jlv_LMH = Low	1.41	82.628	1.758
target_Derive = Not b...	m_jlv_LMH = Medium m_wbb_LMH = High m_jjj_LMH = Medium	2.698	81.694	1.738
target_Derive = Not b...	m_jlv_LMH = Medium m_bb_LMH = High m_wwbb_LMH = High	5.02	80.145	1.705
target_Derive = Not b...	m_bb_LMH = High m_wwbb_LMH = High m_jjj_LMH = Medium	5.103	79.616	1.694
target_Derive = Boson	m_wbb_LMH = Mediu... m_wwbb_LMH = Low m_bb_LMH = Medium	3.683	88.944	1.678
target_Derive = Not b...	m_jlv_LMH = Medium m_wwbb_LMH = High m_jj_LMH = Low	2.298	78.797	1.676
target_Derive = Not b...	m_jlv_LMH = Medium m_bb_LMH = High m_jjj_LMH = Medium	4.583	77.935	1.658

Prvih nekoliko pravila sortiranih po liftu

- Analysis
  - Number of Rules: 1,097
  - Number of Valid Transactions: 11,000,000
  - Minimum Support: 1.004%
  - Maximum Support: 33.334%
  - Minimum Confidence: 50.001%
  - Maximum Confidence: 88.944%
  - Minimum Lift: 0.944%
  - Maximum Lift: 1.839%
  - Minimum Deployability: 0.236%
  - Maximum Deployability 16.587%
  - Minimum Rule Support: 0.629%
  - Maximum Rule Support: 23.432%
- Fields
- Build Settings
  - Use partitioned data: false
  - Maximum number of antecedents: 3
  - Minimum antecedent support (%): 1.0
  - Minimum rule confidence (%): 50.0
  - Optimize: Speed
  - Only true values for flags: false
  - Transactional data: false
- Training Summary

Informacije o algoritmu

Zaključak

- *"HIGGS dataset" - kompleksan skup podataka, što odražava složenost povezanu sa fizikom elementarnih čestica*
- *Metode su pružile uvid u podatke ali proces je također istakao potrebu za strpljenjem u istraživanju podataka*
- *Disciplina koja zahteva upornost i stalno usavršavanje metoda kroz težak niz pokušaja i učenja na greškama*