Klasterovanje FIFA19

Seminarski rad na kursu Istraživanje podataka 1

Nikola Janković

Matematički fakultet, Univerzitet u Beogradu

26. avgust 2019

Sadržaj

Motivacija

- Da li postoji povezanost između pozicije i kvaliteta fudbalera u stvarnom svetu sa procenama koje su napravili kreatori igre FIFA19

Motivacija

- Da li postoji povezanost između pozicije i kvaliteta fudbalera u stvarnom svetu sa procenama koje su napravili kreatori igre FIFA19
- Popularnost ove oblasti.

Motivacija

- Da li postoji povezanost između pozicije i kvaliteta fudbalera u stvarnom svetu sa procenama koje su napravili kreatori igre FIFA19
- Popularnost ove oblasti.
- Autorova lična satisfakcija.

Skup podataka

- 18000 slogova
- 89 atributa

Skup podataka

- 18000 slogova
- 89 atributa

ID	Name	Age	Photo	Nationality	Flag	Overall	Pc
158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94
20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94
190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93
193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93
192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92

Skup podataka

	Age	Overall	Potential	Special	Int. Reput.	Weak Foot	Skill Moves
mean	25.12	62.24	71.31	1597.81	1.11	2.95	2.36
std	4.67	6.9	6.13	272.59	0.39	0.66	0.75
min	16	46	48	731	1	1	1
25%	21	62	67	1457	1	3	2
50%	25	66	71	1635	1	3	2
75%	28	71	75	1787	1	3	3
max	45	94	95	2346	5	5	5

Skup podataka

- *Value*

Skup podataka

- *Value*
- *International reputation*

Skup podataka

- *Value*
- *International reputation*
- *Loaned From*

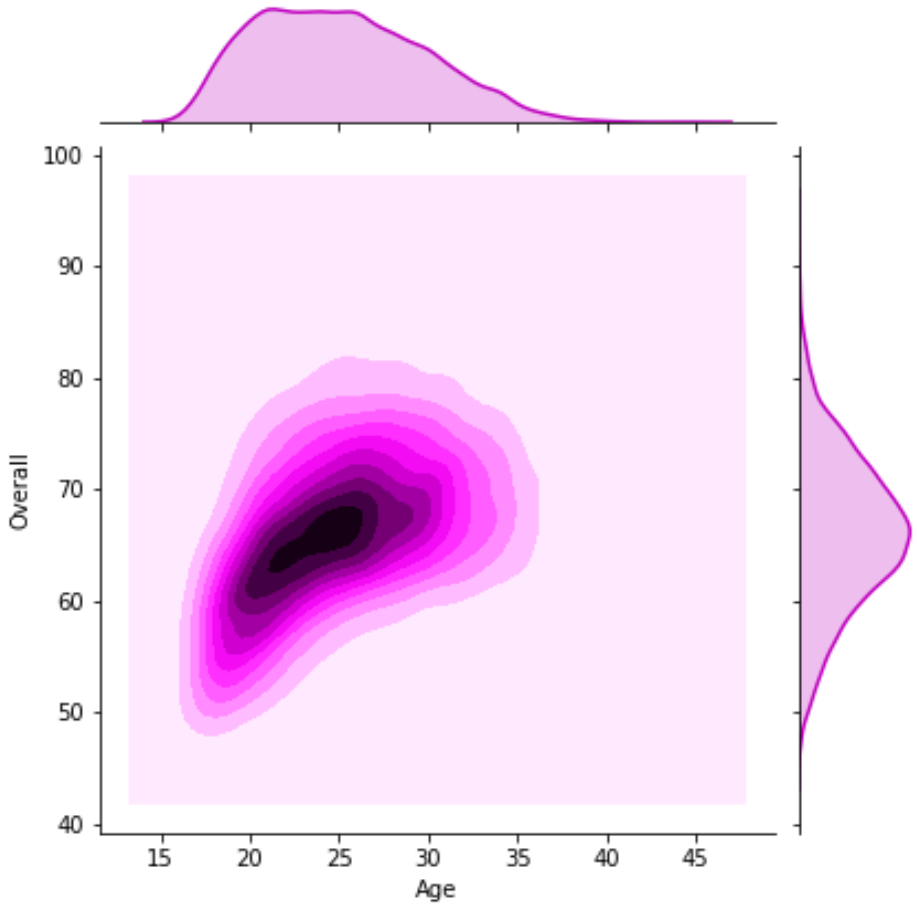
Skup podataka

- *Value*
- *International reputation*
- *Loaned From*
- *LS, ST, RS, ..., RB*

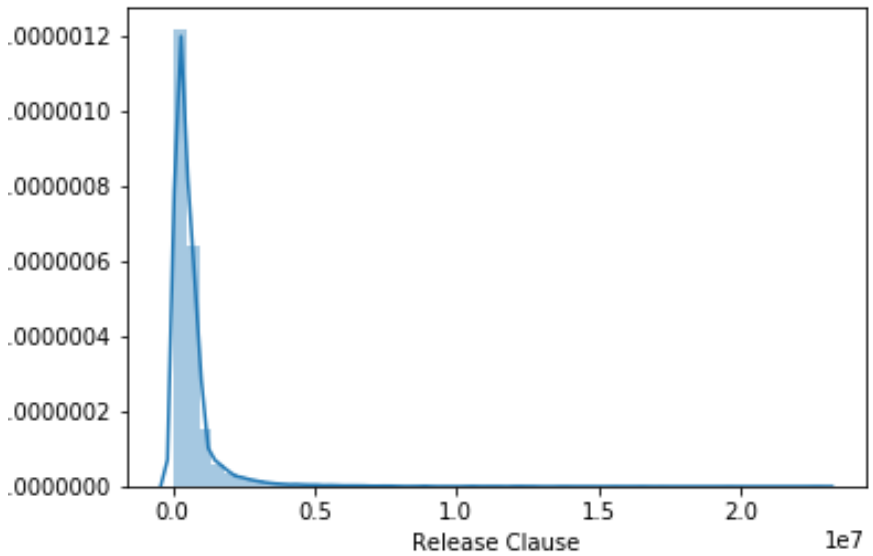
Skup podataka

- *Value*
- *International reputation*
- *Loaned From*
- *LS, ST, RS, ..., RB*
- *Release Clause*

Statistike

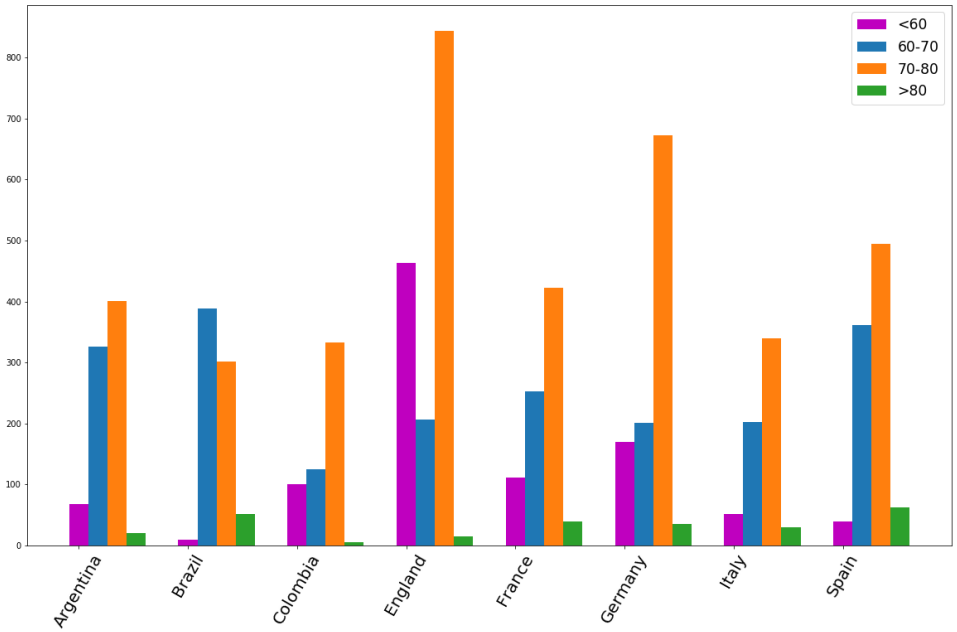


Statistike

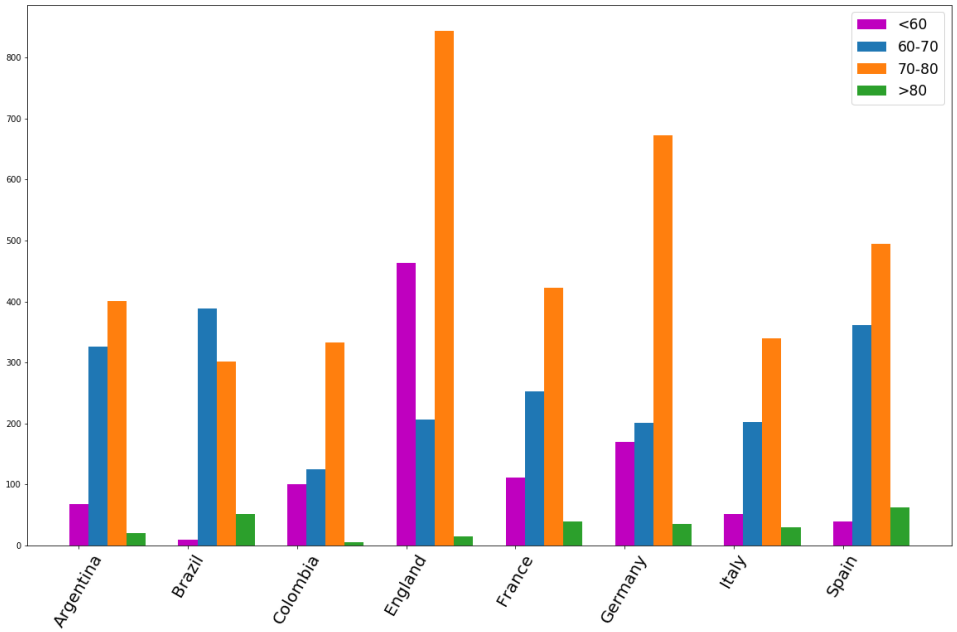


Slika: Raspodela izlazne klauze

Statistike


<60
60-70
70-80
>80
0
100
200
300
400
500
600
700
800
Argentina
Brazil
Colombia
England
France
Germany
Italy
Spain

Statistike



Zašto?

Pretprocesiranje

- ID, Name, Photo, Club, Club Logo, Flag, Jersey Number, Loaned From, Work Rate, Real Face, Joined, Body Type

Pretprocesiranje

- ID, Name, Photo, Club, Club Logo, Flag, Jersey Number, Loaned From, Work Rate, Real Face, Joined, Body Type
- LS, ST, RS, ..., RB

Pretprocesiranje

- ID, Name, Photo, Club, Club Logo, Flag, Jersey Number, Loaned From, Work Rate, Real Face, Joined, Body Type
- LS, ST, RS, ..., RB
- Wage, Value, Release Clause

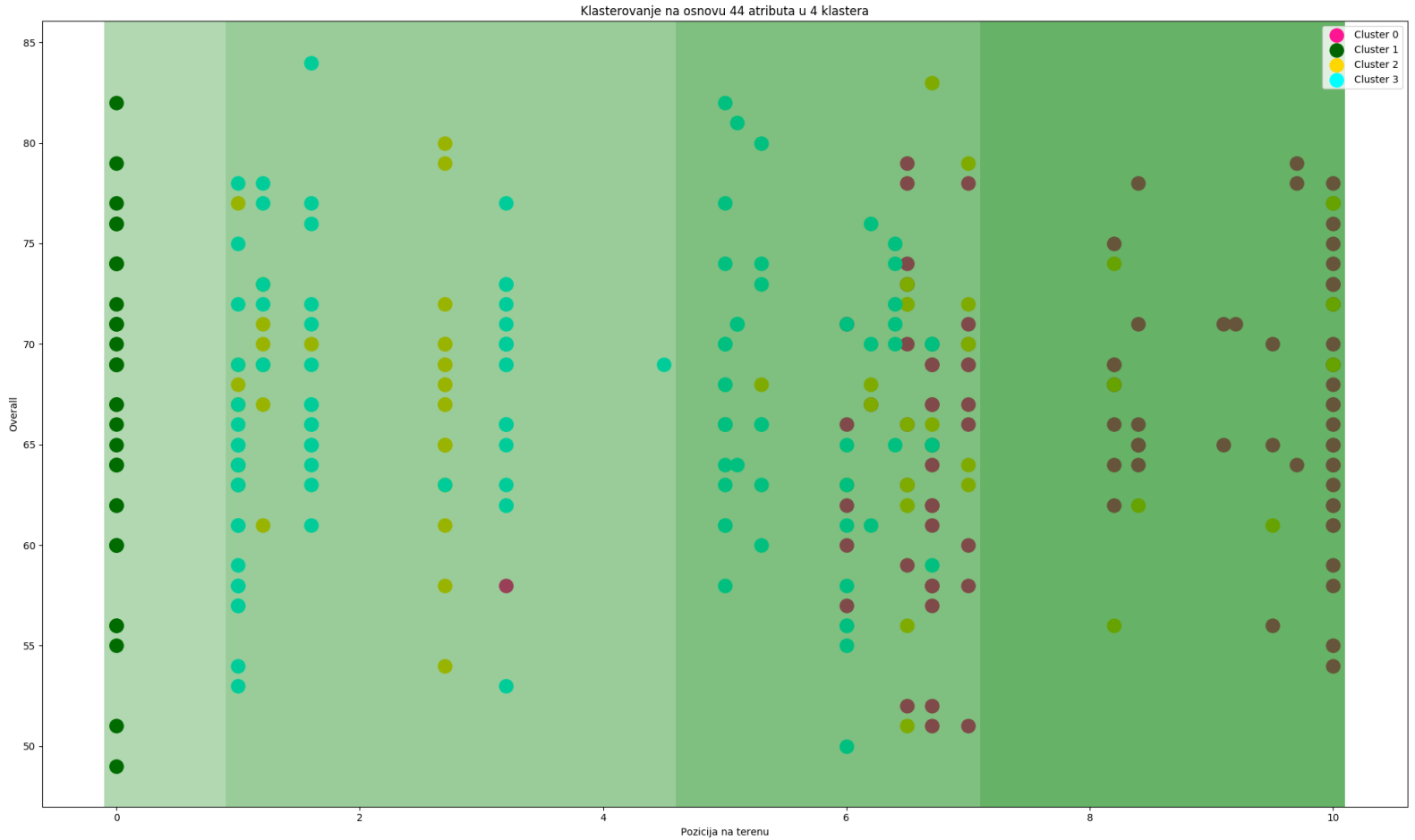
Pretprocesiranje

- ID, Name, Photo, Club, Club Logo, Flag, Jersey Number, Loaned From, Work Rate, Real Face, Joined, Body Type
- LS, ST, RS, ..., RB
- Wage, Value, Release Clause
- Country

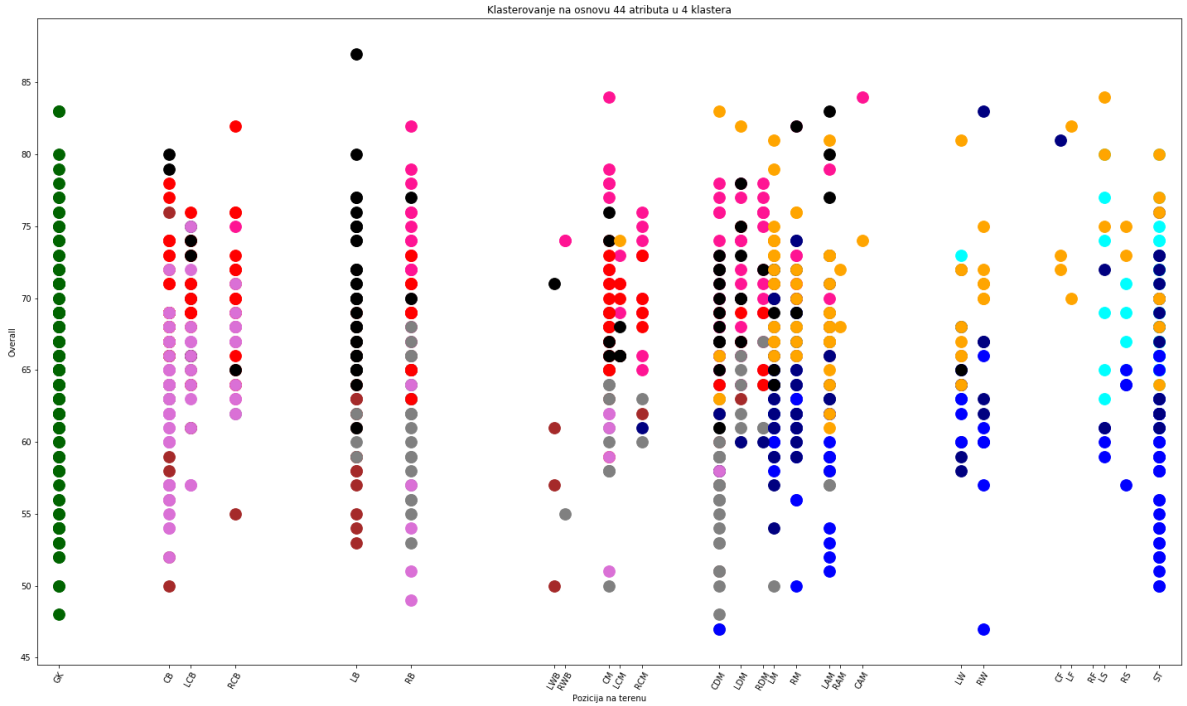
Primena algoritama

- Na kursu
  - K-means
  - DBSCAN
  - Self Organizing Map
  - Hijerahijsko klasterovanje
- Dodatno
  - Meanshift
  - BIRCH

K-means


Klasterovanje na osnovu 44 atributa u 4 klastera
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Overall
Pozicija na terenu
85
80
75
70
65
60
55
50
0
2
4
6
8
10

K-means

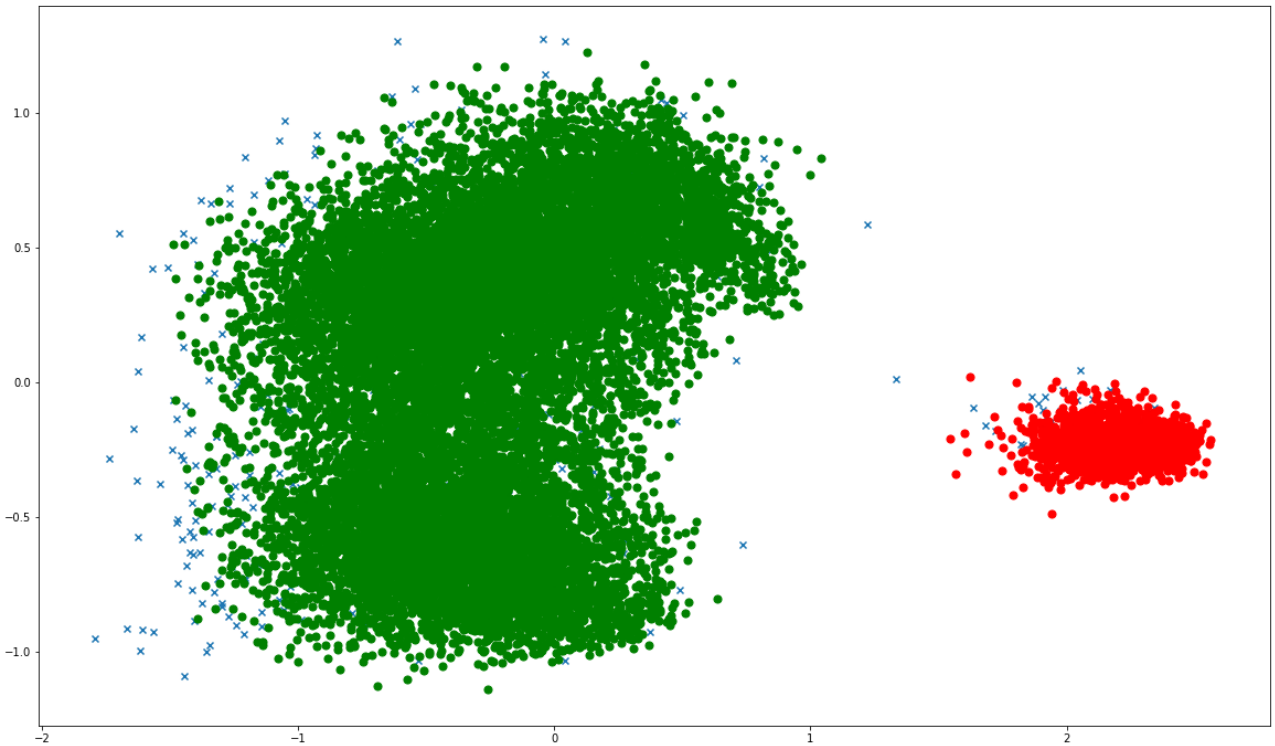


Senka koeficijent dobijen ovakvim klasterovanjem je 0.176189407.

DBSCAN

EPS	MIN_SAMPLE	SENKA KOEF.	EPS	MIN_SAMPLE	SENKA KOEF.
0.2	15	0.1869359	0.3	15	0.2102658
0.2	17	-0.167595	0.3	17	0.0920832
0.2	19	-0.217335	0.3	19	0.2000161
0.2	22	-0.119365	0.3	22	0.1222602
0.2	25	-0.110591	0.3	25	0.1196769
0.25	15	0.02033613	**0.35**	**15**	**0.2683946**
0.25	17	-0.0788248	0.35	17	0.2625584
0.25	19	-0.0136990	0.35	19	0.2588889
0.25	22	0.08744345	0.35	22	0.2474575
0.25	25	-0.0008812	0.35	25	0.1936369
0.28	15	0.1888508			
0.28	17	0.1003213			
0.28	19	0.1004942			
0.28	22	0.0512825			
0.28	25	0.1040002			

DBSCAN



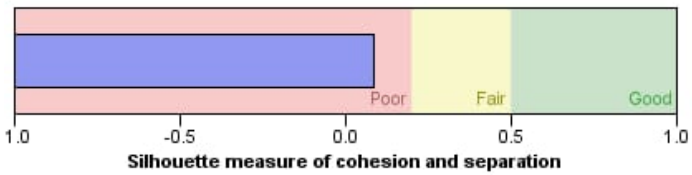
Slika: DBSCAN pca

SOM

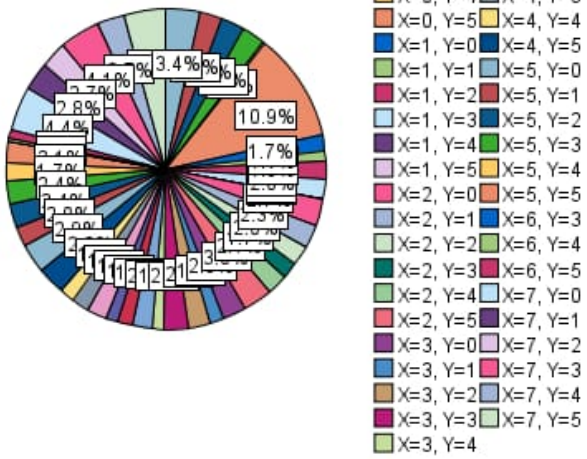
Model Summary

Algorithm	Kohonen
Inputs	50
Clusters	45

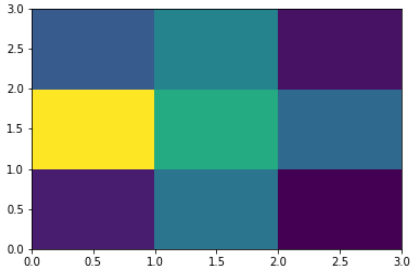
Cluster Quality

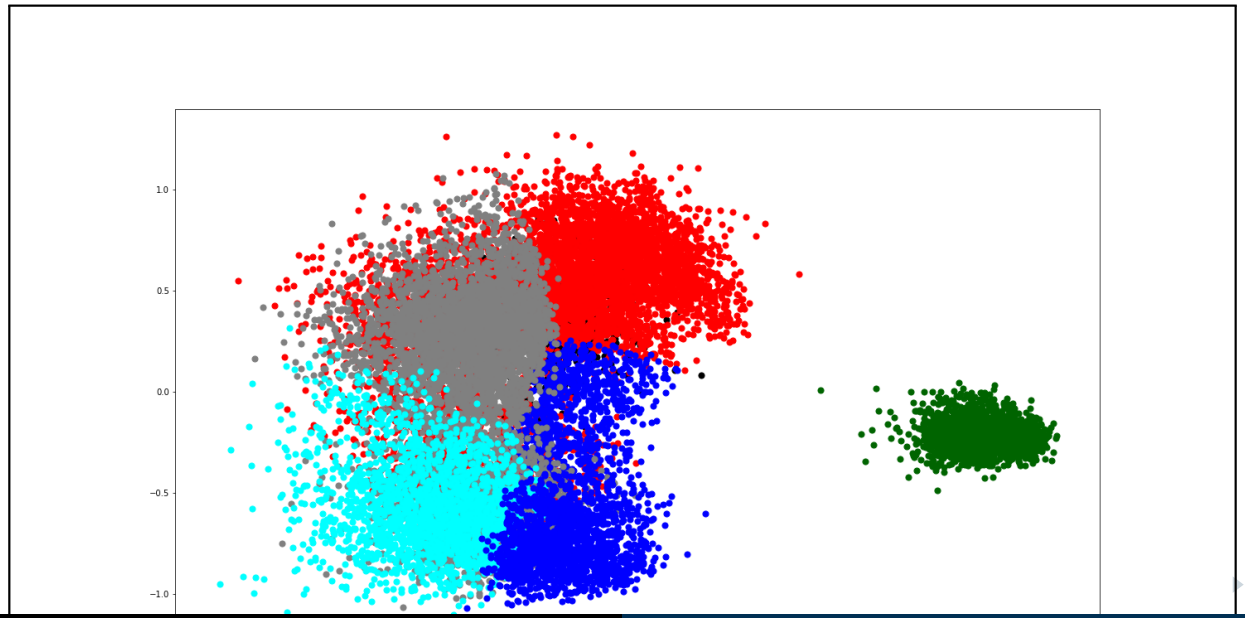


Cluster Sizes



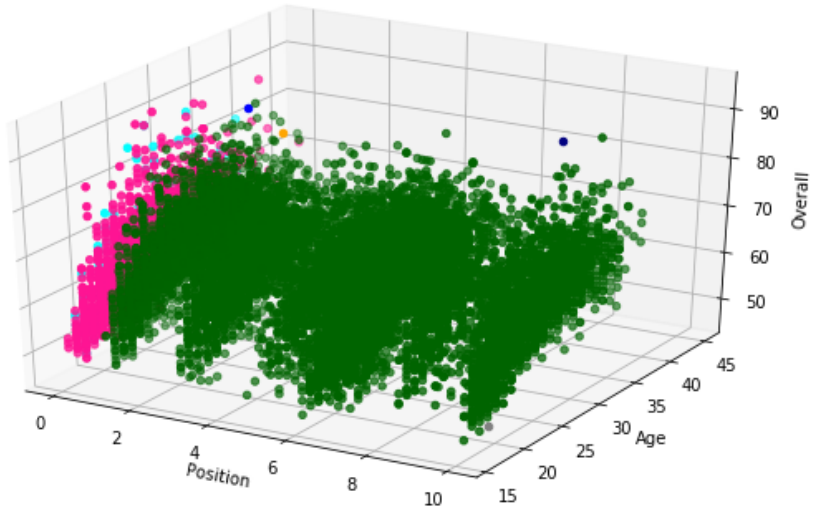
SOM

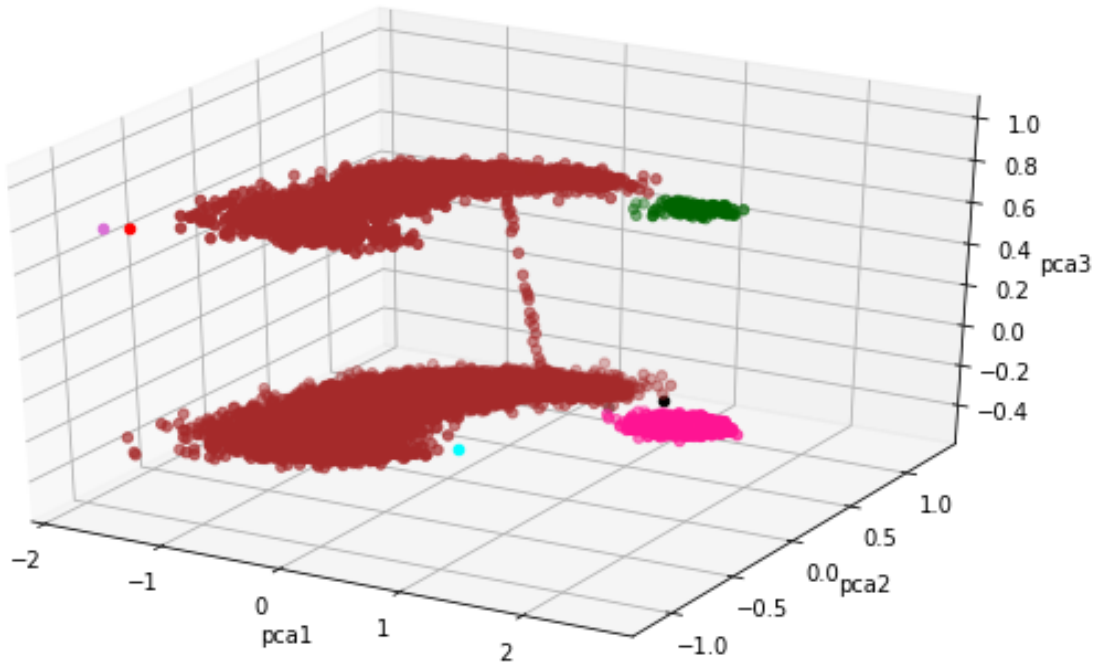




Hijerarhijsko klasterovanje

Primećeno je da najbolji senka koeficijent za 3-7 klastera daje *single* veza. Pa je za tu vezu isproban algoritam za 11 klastera




1.0
0.8
0.6
0.4
pca3
0.2
0.0
−0.2
−0.4
1.0
0.5
0.0
pca2
−0.5
−1.0
−2
−1
0
pca1
1
2

Meanshift

```
1  Input: bandwith, skup podataka
2  WHILE postoji objekat koji nije dodeljen nijednom klasteru
       DO :
3    izaberi jedan od nedodeljenih objekata i oznaci
4    da pripada novom klasteru
5    REPEAT:
6    azuriraj srednju tacku ( centroid ) u trenutnom klasteru ;
7    sve tacke koje se nalaze na razdaljini manjoj
8    od bandwidth oznaci ih da pripadaju trenutnom
9    klasteru;
10   UNTIL postoji promena na trenutnom klasteru
```

Meanshift

- Bandwith izabran 1.5 uz pomoć *sklearn.estimate_bandwith()*

Meanshift

- Bandwith izabran 1.5 uz pomoć *sklearn.estimate_bandwith()*
- Odličan senka koeficijent, ali samo 2 klastera.

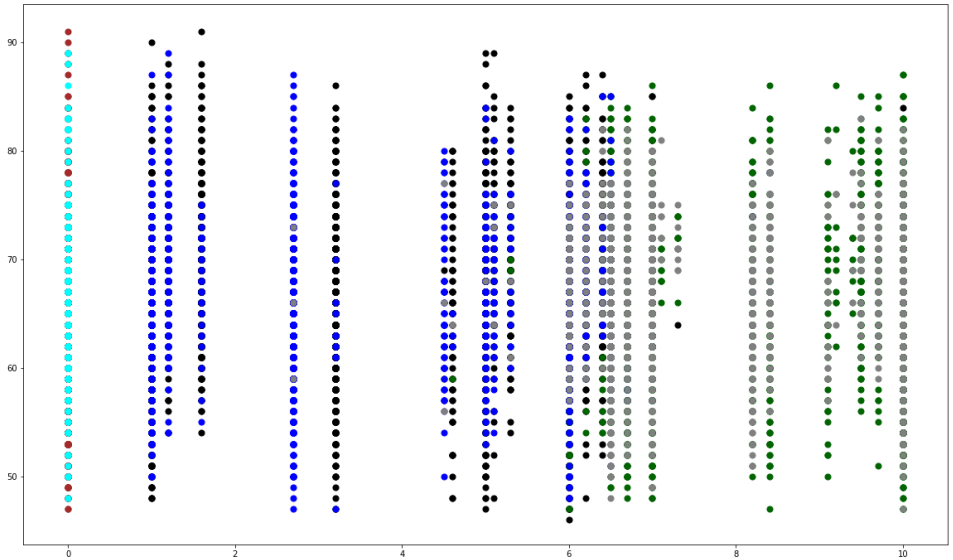
Meanshift

- Bandwith izabran 1.5 uz pomoć *sklearn.estimate_bandwith()*
- Odličan senka koeficijent, ali samo 2 klastera.
- Smanjen bandwith na 0.5.

Meanshift

- Bandwith izabran 1.5 uz pomoć *sklearn.estimate_bandwith()*
- Odličan senka koeficijent, ali samo 2 klastera.
- Smanjen bandwith na 0.5.
- Dobijena 4 klastera, senka koeficijent i dalje preko 0.5

Meanshift

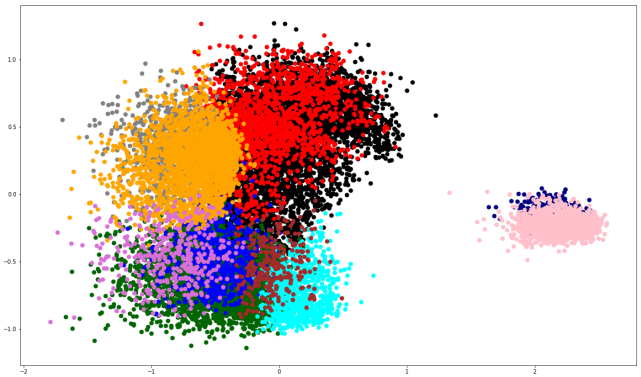


BIRCH

Definition

BIRCH - (balanced iterative reducing and clustering using hierarchies) Hibridni algoritam za klasteorvanje koji se zasniva na pravljenju CF(Clusters Features)-stabla.

Za k=11 senka koeficijent = 0.61



Zaključak

- Koji se algoritam najbolje pokazao?

Zaključak

- Koji se algoritam najbolje pokazao?
- Dalje mogućnosti sa ovim skupom?

Zaključak

- Koji se algoritam najbolje pokazao?
- Dalje mogućnosti sa ovim skupom?
- Pitanja?