



Istraživanje podataka

**SKUP PODATAKA
AVILA BIBLIJA**

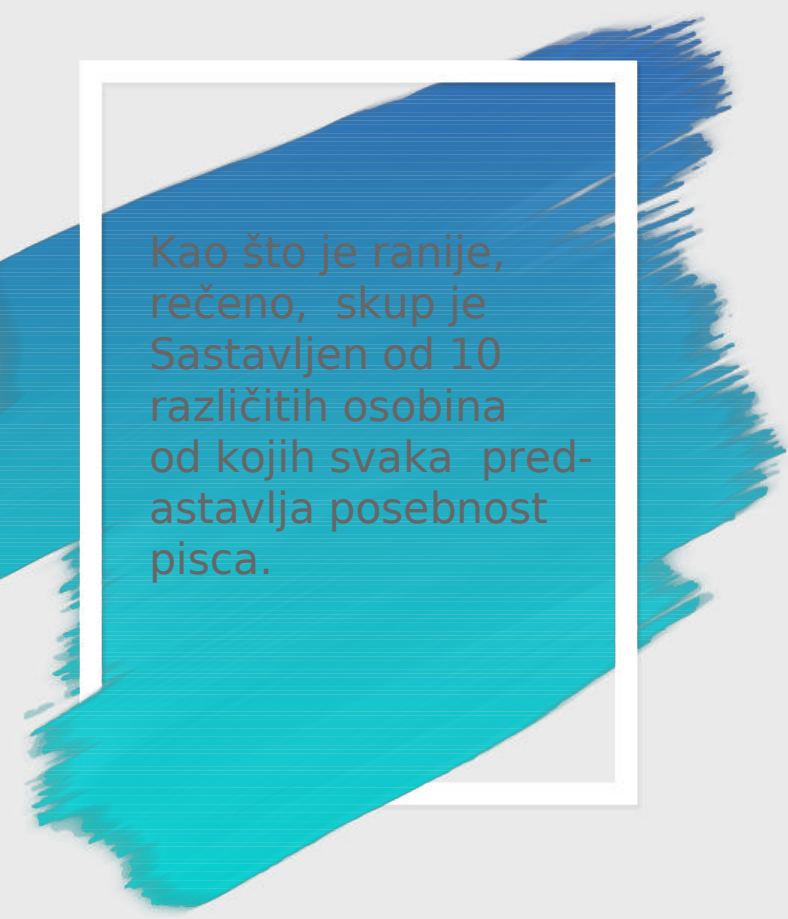
Lazar Dačković



AVILA SKUP PODATAKA

- **Avila** skup podataka je napravljen na osnovu 800 slika "Avila Biblije", velike Latinske kopije napisane tokom 12 veka.
- Paleografska analiza rukopisa je prepoznala 12 pisaca. Broj stranica koje je napisao svaki pisac nije jednak.
- Svaka instanca ima 10 atributa.
- Zadatak predviđanja se sastoji od prepoznavanja jednog od 12 pisaca (označenih sa: A, B, C, D, E, F, G, H, I, W, X, Y).
- Podaci su normalizovani korišćenjem metode Z-normalizacija, I podeljeni u 2 skupa podataka: trening skup broji 10430 uzoraka, I test skup broji 10437 uzoraka.

VIZUELIZACIJA PODATAKA



Kao što je ranije,
rečeno, skup je
Sastavljen od 10
različitih osobina
od kojih svaka pred-
stavlja posebnost
pisca.

Osobine su:

F1 : Među kolonsko rastojanje

F2: Gornja margina

F3 : Donja margina

F4: Eksploatacija

F5: Broj redova

F6 : Modularni odnos

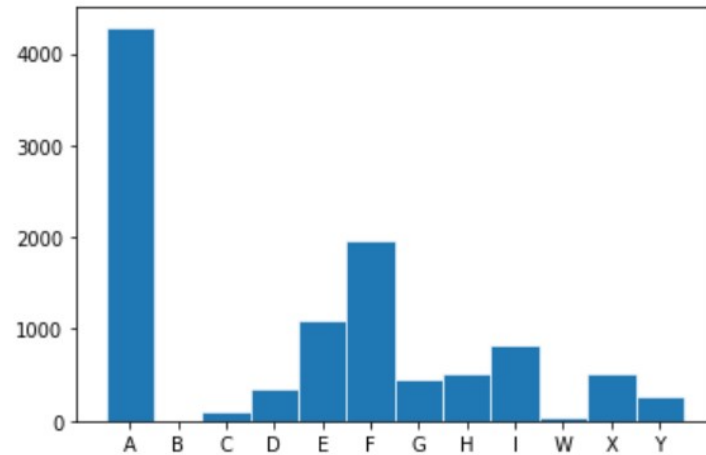
F7: Međuredno rastojanje

F8 : Težina

F9: Broj vrha

F10 : Modularni odnos/međuredno rastojanje

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Copyist
0	0.130292	0.870736	-3.210528	0.062493	0.261718	1.436060	1.465940	0.636203	0.282354	0.515587	A
1	-0.116585	0.069915	0.068476	-0.783147	0.261718	0.439463	-0.081827	-0.888236	-0.123005	0.582939	A
2	0.031541	0.297600	-3.210528	-0.583590	-0.721442	-0.307984	0.710932	1.051693	0.594169	-0.533994	A
3	0.229043	0.807926	-0.052442	0.082634	0.261718	0.148790	0.635431	0.051062	0.032902	-0.086652	F
4	0.117948	-0.220579	-3.210528	-1.623238	0.261718	-0.349509	0.257927	-0.385979	-0.247731	-0.331310	A
...
10424	0.080916	0.588093	0.015130	0.002250	0.261718	-0.557133	0.371178	0.932346	0.282354	-0.580141	F
10425	0.253730	-0.338346	0.352988	-1.154243	0.172340	-0.557133	0.257927	0.348428	0.032902	-0.527134	F
10426	0.229043	-0.000745	0.171611	-0.002793	0.261718	0.688613	0.295677	-1.088486	-0.590727	0.580142	A
10427	-0.301743	0.352558	0.288973	1.638181	0.261718	0.688613	0.069175	0.502761	0.625350	0.718969	E
10428	-0.104241	-1.037102	0.388552	-1.099311	0.172340	-0.307984	0.786433	-1.337547	0.999528	-0.551063	X



- Izgled podataka trening skupa

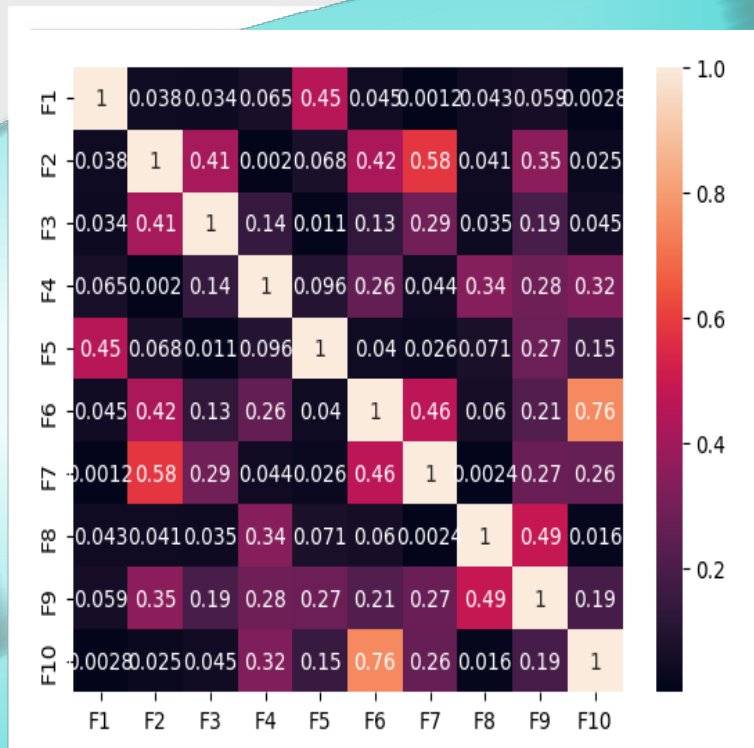
Distribucija klasa

□ Matrica korelacije nam pomaže da
Uočimo neke pravilnosti među atributima

□ Ovo je bilo predvidivo da F10, F7 i F6 budu
u visokoj korelaciji jer je $F10 = F6/F7$.

Ali takođe smo otkrili i neke druge
iznenađujuće korelacije npr. F2 i F7 koje
odgovaraju gornjoj margini i među linijskoj
udaljenosti.

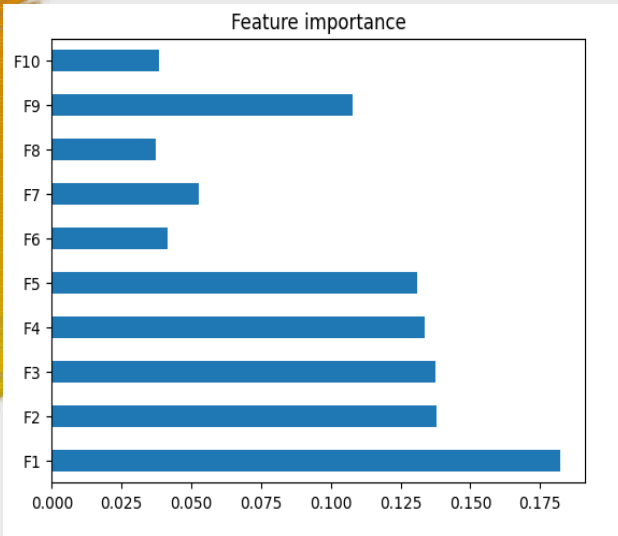
F1 i F5 koje odgovaraju međustubnoj
udaljenosti i broju reda.



KLASIFIKACIJA

Metoda slučajnih šuma
Konstruiše više manjih
stabala i pravi njihovu
najbolju kombinaciju.

Pravi model koji je 98%
precizan



Važnost atributa u modelu

True labels	A	4263	0	0	0	3	13	2	3	2	0	0	0
	B	0	5	0	0	0	0	0	0	0	0	0	0
	C	3	0	99	0	1	0	0	0	0	0	0	0
	D	0	0	0	344	9	0	0	0	0	0	0	0
	E	17	0	0	0	1072	2	0	0	0	0	4	0
	F	56	0	0	0	3	1900	2	1	0	0	0	0
	G	8	0	0	0	5	1	433	0	0	0	0	0
	H	17	0	3	0	1	0	0	499	0	0	0	0
	I	3	0	0	0	0	0	0	0	829	0	0	0
	W	0	0	0	0	0	0	0	0	0	44	0	0
X	6	0	0	0	0	8	0	0	0	0	0	508	0
Y	0	0	0	0	0	1	0	0	0	0	0	2	264
	A	B	C	D	E	F	G	H	I	W	X	Y	
Predicted labels													

Matrica konfuzije

Knn model, vodi se izrekom "S kim si, takav si". Onih instanci kojih ima najviše u okolini njima si verovatno najslićniji

Elemtni van granica mogu drastično da utiču na model. Knn ako izvršimo odsecanje radi sa 76% tačnosti.

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
A	0.79	0.80	0.80	4286
B	1.00	1.00	1.00	5
C	0.73	0.54	0.62	103
D	0.72	0.61	0.66	353
E	0.76	0.70	0.73	1095
F	0.63	0.68	0.65	1962
G	0.61	0.60	0.61	447
H	0.62	0.62	0.62	520
I	0.98	0.96	0.97	832
W	0.72	0.77	0.75	44
X	0.90	0.85	0.87	522
Y	0.83	0.80	0.82	267
accuracy			0.76	10436
macro avg	0.77	0.75	0.76	10436
weighted avg	0.76	0.76	0.76	10436

Classification report for model KNeighborsClassifier on test data

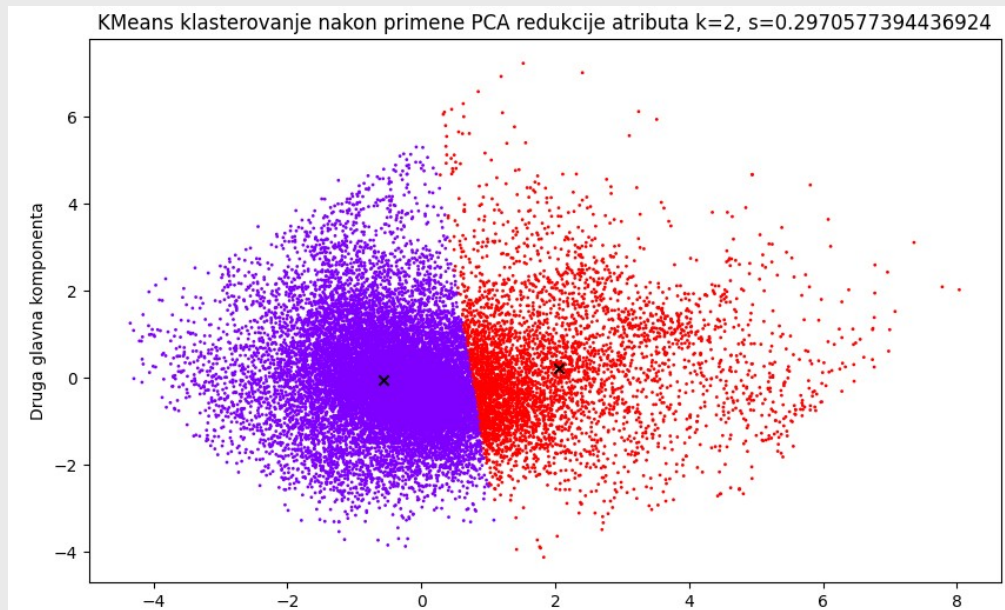
	precision	recall	f1-score	support
A	0.65	0.74	0.70	4286
B	0.00	0.00	0.00	5
C	0.22	0.32	0.26	103
D	0.35	0.22	0.27	353
E	0.48	0.54	0.51	1095
F	0.50	0.58	0.54	1962
G	0.51	0.50	0.51	447
H	0.40	0.57	0.47	520
I	0.00	0.00	0.00	832
W	0.25	0.05	0.08	44
X	0.69	0.68	0.69	522
Y	0.00	0.00	0.00	267
accuracy			0.57	10436
macro avg	0.34	0.35	0.33	10436
weighted avg	0.51	0.57	0.53	10436

Ako izvršimo odbacivanje radi sa 57% tačnosti

KLASTEROVANJE

K-means pretpostavlja da su klasteri sferičnog oblika i da su konveksni. Moramo mu reći koliko klastera želimo.

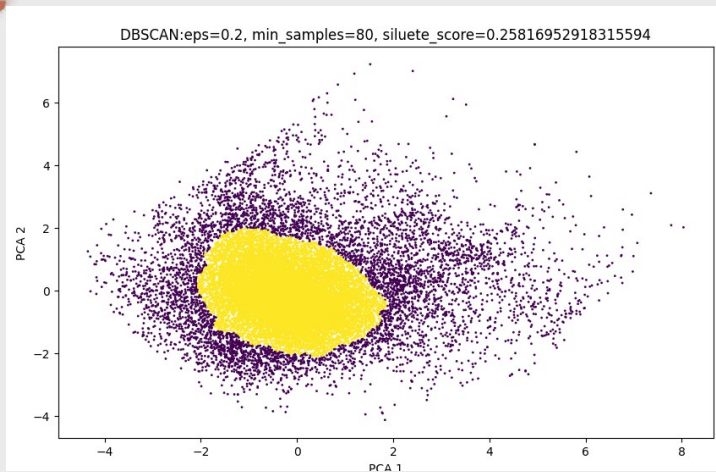
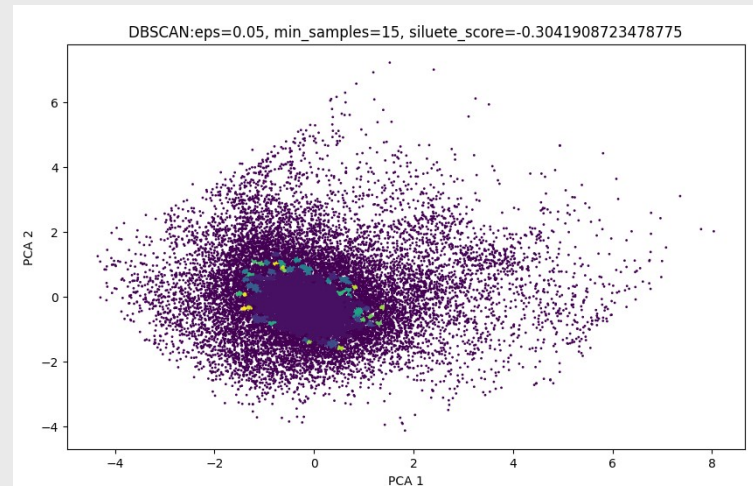
- Metriku koju koristimo za ocenu klasterovanja: Silueta



Zbog vizuelizacije je odrađena PCA redukcija atributa, sveli smo 10 atributa na 2

□ Metriku koju koristimo za ocenu klasterovanja:
Silueta

DBSCAN ne pretpostavlja unapred broj klastera.
Bitni parametri eps, minSamples.



The background of the slide is a large, abstract brushstroke in shades of orange and yellow, with a white rectangular border. The text is centered within this area.

HVALA NA PAŽNJI!