Препознавање слова

– Истраживање података 1 –

Аутор:
Лазар Васовић, 99/2016
Математички факултет, 2019

Замисао

- Генерализација (уопштавање)
- Индуктивно закључивање
  (од посебног ка општем)
- Модел законитости у подацима
- Класификација → препознавање
  врсте објекта
- Учење/тренирање и примена наученог
- Реализација горњег на скупу слова

Скуп података

- Велика слова енглеске абецеде
- Црно-беле правоугаоне растерске слике
- 20 фонтова, случајно изобличавање
- 20.000 инстанци, 26 категорија
- Слике трансформисане у 16 нумеричких улазних атрибута → статистичке особине расподеле пиксела: димензије, средње вредности и одступања, корелације...

	slovo	x_kutija	y_kutija	širina	visina	broj_piksela	x_mean	y_mean	x2_var
1	T	2	8	3	5	1	8	13	0
2	I	5	12	3	7	2	10	5	5
3	D	4	11	6	8	6	10	6	2
4	N	7	11	6	6	3	5	9	4
5	G	2	1	3	1	1	8	6	6
6	S	4	11	5	8	3	8	8	6
7	B	4	2	5	4	4	8	7	6
8	A	1	1	3	2	1	8	2	2
9	J	2	2	4	4	2	10	6	2
10	M	11	15	13	9	7	13	2	6

	slovo	y2_var	xy_kor	x2y_mean	xy2_mean	x_ivice	xivy_kor	y_ivice	yivx_kor
1	T	6	6	10	8	0	8	0	8
2	I	4	13	3	9	2	8	4	10
3	D	6	10	3	7	3	7	3	9
4	N	6	4	4	10	6	10	2	8
5	G	6	6	5	9	1	7	5	10
6	S	9	5	6	6	0	8	9	7
7	B	6	7	6	6	2	8	7	10
8	A	2	8	2	8	1	6	2	7
9	J	6	12	4	8	1	6	1	7
10	M	2	12	1	9	8	1	1	8

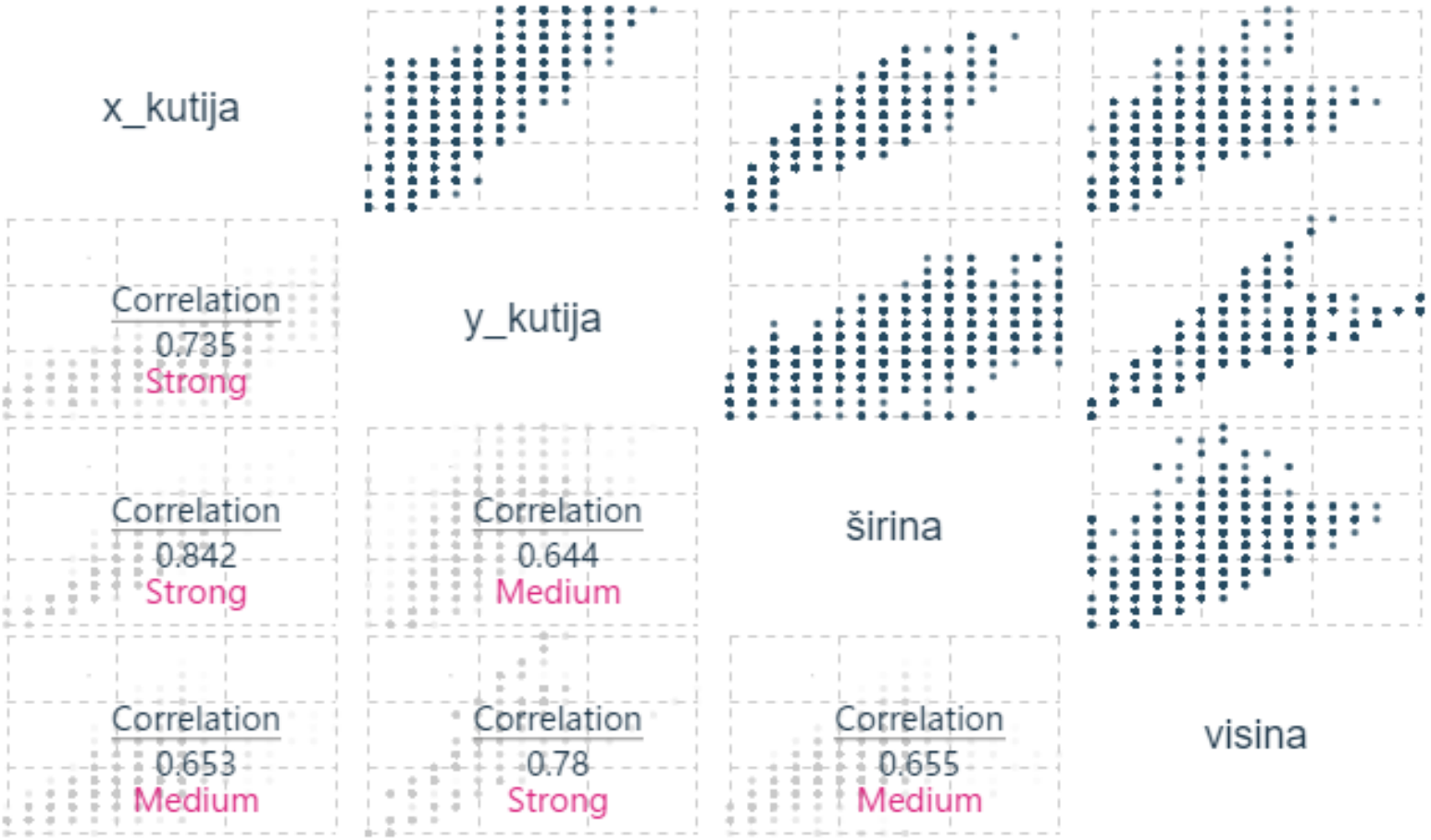
Скуп података

- Вишедимензиони подаци
- Сви улазни квантитативни, циљни именски
- Сви дискретни (16 односно 26 вредности)
- Сви нумерички у целобројном интервалу [0, 15]
- Нема недостајућих нити бланко вредности
- Нема некоректних нити дуплираних слогова

Анализа података

- Расподеле вредности атрибута махом се уклапају у познате расподеле: нормалну, експоненцијалну...
- Атрибути махом некорелисани у пару, са неколико битних изузетака
- Нема потребе за претпроцесирањем
- Надаље су слогови подразумевано подељени на тренинг и тест скуп у односу 70-30%

Dijagram raspršenih vrednosti atributa i njihovih korelacija




Pseudolinearna zavisnost širine kutije i x koordinate njenog centra
-prosečna širina obuhvatajućeg pravougaonika piskela i njegov centar-
širina
15
12
9
6
3
0
0
3
6
9
12
15
x_kutija

Дрвета одлучивања

- Хантов алгоритам → подела слогова према тестном атрибуту који максимизује одређени критеријум, прављење стабла категорисања
- C5.0 → ентропија, информациона добит, *n*-арно дрво, најмање две инстанце у листу
- Матрица конфузије није помогла

Results for output field slovo

Comparing $C-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	13,392	95.58%	5,180	86.49%
Wrong	619	4.42%	809	13.51%
Total	14,011		5,989	

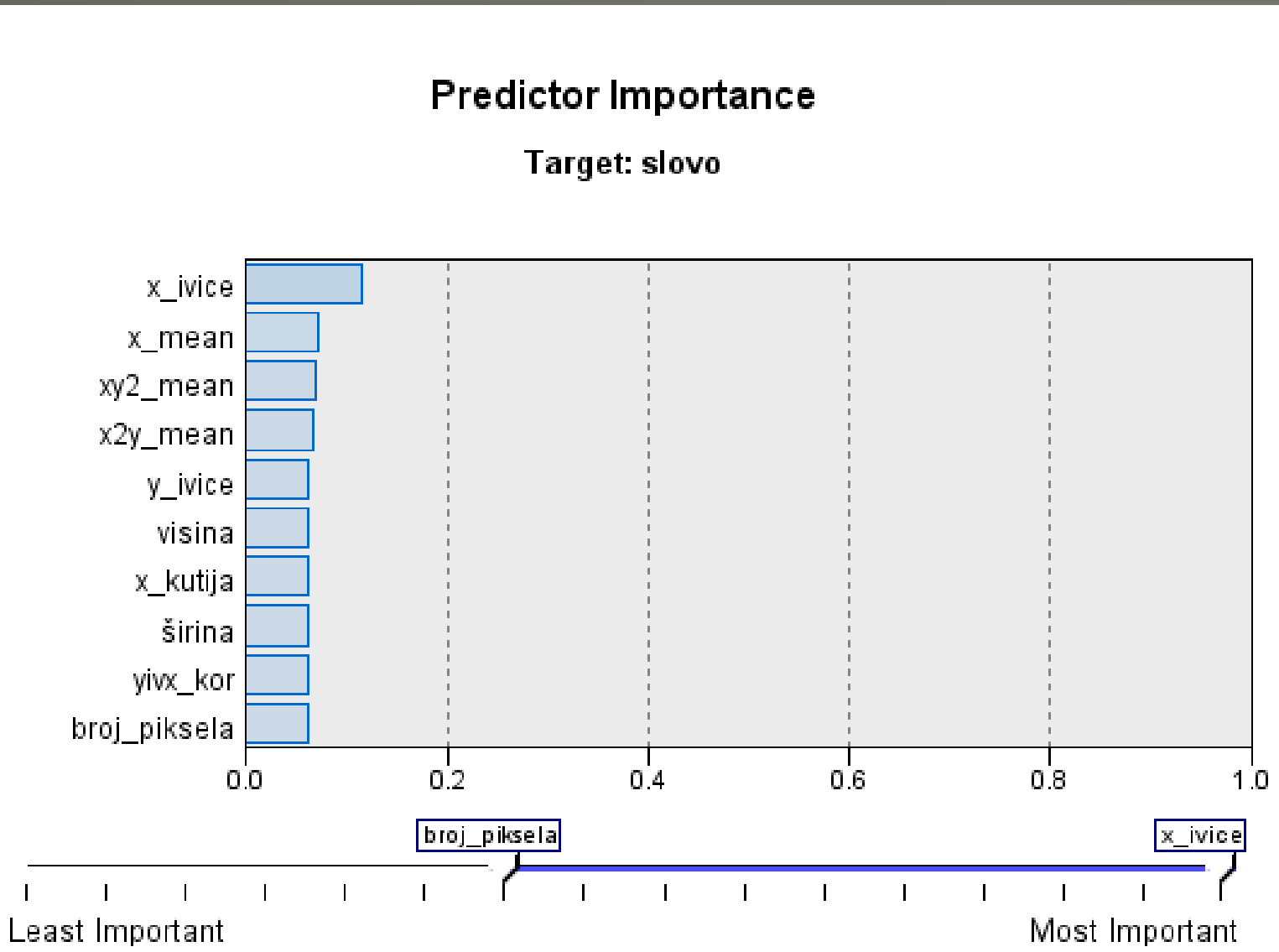
Дрвета одлучивања

- Али зато додатне погодности алгоритма јесу
- Појачавање (бустовање) → итеративно побољшавање класификатора
- Расејавање атрибута → анализа и искључивање оних небитних
- Унакрсна валидација → тестни блокови

Results for output field slovo

Comparing $C-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	13,982	99.79%	5,626	93.94%
Wrong	29	0.21%	363	6.06%
Total	14,011		5,989	


Predictor Importance
Target: slovo
x_ivice
x_mean
xy2_mean
x2y_mean
y_ivice
visina
x_kutija
širina
yivx_kor
broj_piksela
0.0
0.2
0.4
0.6
0.8
1.0
broj_piksela
x_ivice
Least Important
Most Important

Дрвета одлучивања

- Одлично се показало и дрво направљено над редукованим скупом атрибута, из којег су избачени висококорелисани
- Честе грешке → „F“ и „P“, „H“ и „P“, „H“ и „R“, „I“ и „J“, „K“ и „X“, дакле, визуелно слична слова

Results for output field slovo

Comparing $C-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	13,926	99.39%	5,453	91.05%
Wrong	85	0.61%	536	8.95%
Total	14,011		5,989	

Дрвета одлучивања

- C&RT→ Гинијев индекс, бинарно дрво, ограничена дубина стабла
- CHAID, QUEST → подједнако лоши модели

Results for output field slovo

Comparing $R-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	7,448	53.16%	3,208	53.56%
Wrong	6,563	46.84%	2,781	46.44%
Total	14,011		5,989	

Дрвета одлучивања

- DecisionTreeClassifier→ Гинијев индекс, чисти или минимални листови, без одсецања

Izvestaj klasifikacije:

	precision	recall	f1-score	support
A	0.88	0.96	0.92	232
B	0.82	0.80	0.81	235
C	0.93	0.88	0.90	183
D	0.82	0.82	0.82	219
E	0.87	0.87	0.87	247
F	0.81	0.82	0.82	239
G	0.81	0.85	0.83	205
H	0.77	0.83	0.80	224
I	0.90	0.93	0.92	240
J	0.91	0.88	0.89	209
K	0.88	0.81	0.84	226
L	0.95	0.86	0.90	223
M	0.94	0.92	0.93	249
N	0.91	0.91	0.91	231
O	0.79	0.83	0.81	245
P	0.86	0.84	0.85	239
Q	0.87	0.80	0.83	253
R	0.79	0.84	0.81	227
S	0.81	0.86	0.83	224
T	0.91	0.87	0.89	244
U	0.91	0.92	0.92	259
V	0.90	0.90	0.90	249
W	0.94	0.93	0.93	222
X	0.90	0.92	0.91	219
Y	0.89	0.90	0.89	231
Z	0.91	0.88	0.89	226
accuracy			0.87	6000
macro avg	0.87	0.87	0.87	6000
weighted avg	0.87	0.87	0.87	6000

Test skup:
Matrica konfuzije:

	A	B	C	D	E	F	G	...	T	U	V	W	X	Y	Z
A	222	0	0	1	0	0	2	...	0	0	0	0	0	0	0
B	1	188	0	3	1	0	2	...	0	0	2	0	1	0	1
C	0	0	161	0	1	1	6	...	0	3	0	0	1	0	0
D	1	1	0	180	0	1	1	...	0	0	0	0	0	0	0
E	0	0	3	3	214	1	4	...	3	1	0	0	3	0	4
F	1	2	2	0	0	197	1	...	5	0	2	0	1	1	2
G	0	0	2	4	8	0	175	...	2	2	0	0	0	0	0
H	0	2	1	9	2	2	2	...	0	0	0	1	3	0	1
I	0	0	0	1	0	2	3	...	0	0	0	0	0	0	2
J	1	0	0	0	0	3	0	...	0	0	0	0	1	1	2
K	1	4	0	1	2	0	2	...	0	0	1	0	6	0	0
L	2	3	0	0	2	0	5	...	0	0	0	0	2	0	0
M	2	0	0	0	0	0	1	...	0	6	1	3	0	0	0
N	3	2	0	2	0	0	0	...	0	2	1	0	0	1	0
O	2	1	0	3	0	1	2	...	0	6	1	2	0	2	0
P	0	4	0	3	1	19	0	...	1	0	0	1	0	1	0
Q	1	2	0	5	4	1	2	...	0	1	2	0	0	4	4
R	3	4	2	2	3	0	3	...	0	0	2	0	1	0	0
S	6	6	3	0	1	1	0	...	1	0	0	0	2	0	0
T	0	1	0	0	0	6	2	...	212	0	2	0	1	12	3
U	0	1	0	1	0	1	2	...	0	238	1	3	0	1	0
V	1	5	0	0	0	3	0	...	2	0	223	4	0	3	0
W	1	0	0	0	0	2	0	...	0	1	5	207	0	1	0
X	1	2	0	1	2	0	1	...	1	0	0	0	202	0	1
Y	1	0	0	0	0	2	0	...	5	1	6	0	0	208	0
Z	2	0	0	1	4	0	1	...	2	0	0	0	1	0	198

[26 rows x 26 columns]

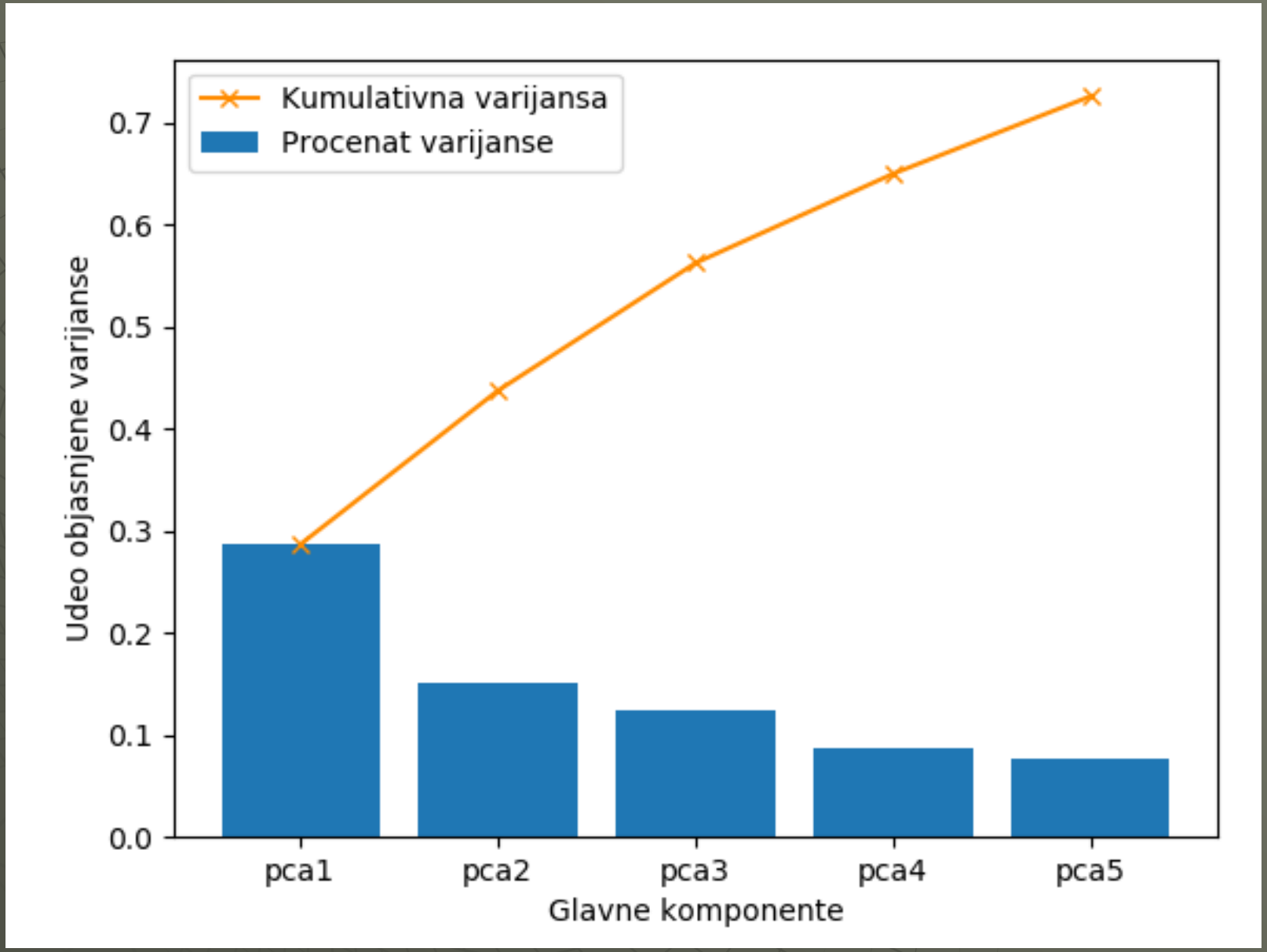
```
Vaznost prediktora:
x_kutija          0.010274
y_kutija          0.016941
sirina            0.010076
visina            0.014296
broj_piksela      0.017669
x_mean            0.048655
y_mean            0.055109
x2_var            0.068161
y2_var            0.119419
xy_kor            0.063463
x2y_mean          0.083342
xy2_mean          0.086849
x_ivice           0.137278
xivy_kor          0.099428
y_ivice           0.114702
yivx_kor          0.054338
dtype: float64
```

Неуронска мрежа и слично

- MLPClassifier → стохастички градијентни спуст, унакрсном валидацијом изабрана ReLU за активациону функцију, прец. 92% на тест
- Занимљивост→ нема већег мешања „F“ и „P“, али има „R“ и „B“, „E“ и „G“, као и „K“ и „X“
- Још неки алг. → анализа дискриминанти (70%), Бајесова мрежа (71%), логистичка регресија (78%), случајна дрвета (82%), SPSS вишеслојни перцептрон (86%), случајна шума (93-94% на тест скупу, пример дист. учења у ансамблу)

Редукција и потпорни вектори

- PCA → анализа главних компоненти, циљ да што мање димензија објасни што већи удео варијансе, фактори су лин. комб. атрибута, SPSS предлаже првих пет фактора (70% вар.), у Python-у тестирано за све комбинације
- SVM/SVC → класификација подржавајућим векторима, раздвајајућа хиперраван


Kumulativna varijansa
Procenat varijanse
Udeo objasnjene varijanse
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
pca1
pca2
pca3
pca4
pca5
Glavne komponente

Редукција и потпорни вектори

- Прва слика → SVM на пет фактора
- Друга слика → SVM на полазном скупу
- Без неважних предиктора прец. 72%

Results for output field slovo

Comparing $S-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	6,978	49.8%	2,956	49.36%
Wrong	7,033	50.2%	3,033	50.64%
Total	14,011		5,989	

Results for output field slovo

Comparing $S-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	12,134	86.6%	5,139	85.81%
Wrong	1,877	13.4%	850	14.19%
Total	14,011		5,989	

Редукција и потпорни вектори

- Python → бољи резултати
- SVM на пет фактора → 76%
- SVM на полазном скупу → 97%
- Кернел RBF (radial basis function)
- Мешање „F“ и „P“, „D“ и „H“

Лењи класификатори

- KNN ⟶ *k* најближих суседа, имплицитни модел, еуклидско растојање, k = {3, 4, 5}
- Друга слика ⟶ k = 1, најбољи лењи модел

Results for output field slovo

Comparing $KNN-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	13,731	98%	5,702	95.21%
Wrong	280	2%	287	4.79%
Total	14,011		5,989	

Results for output field slovo

Comparing $KNN-slovo with slovo

'Partition'	1_Training		2_Testing	
Correct	14,011	100%	5,712	95.37%
Wrong	0	0%	277	4.63%
Total	14,011		5,989	

Лењи класификатори

- Python → унакрсна валидација бира k = 4, еуклидско растојање, тежинске суседе, за алгоритам имплицитно *kd* дрво, прец. 96%
- Вероватносни класификатори → MNB (58%), GNB (67%), обичан наивни Бајес (76%)

Izvestaj klasifikacije:

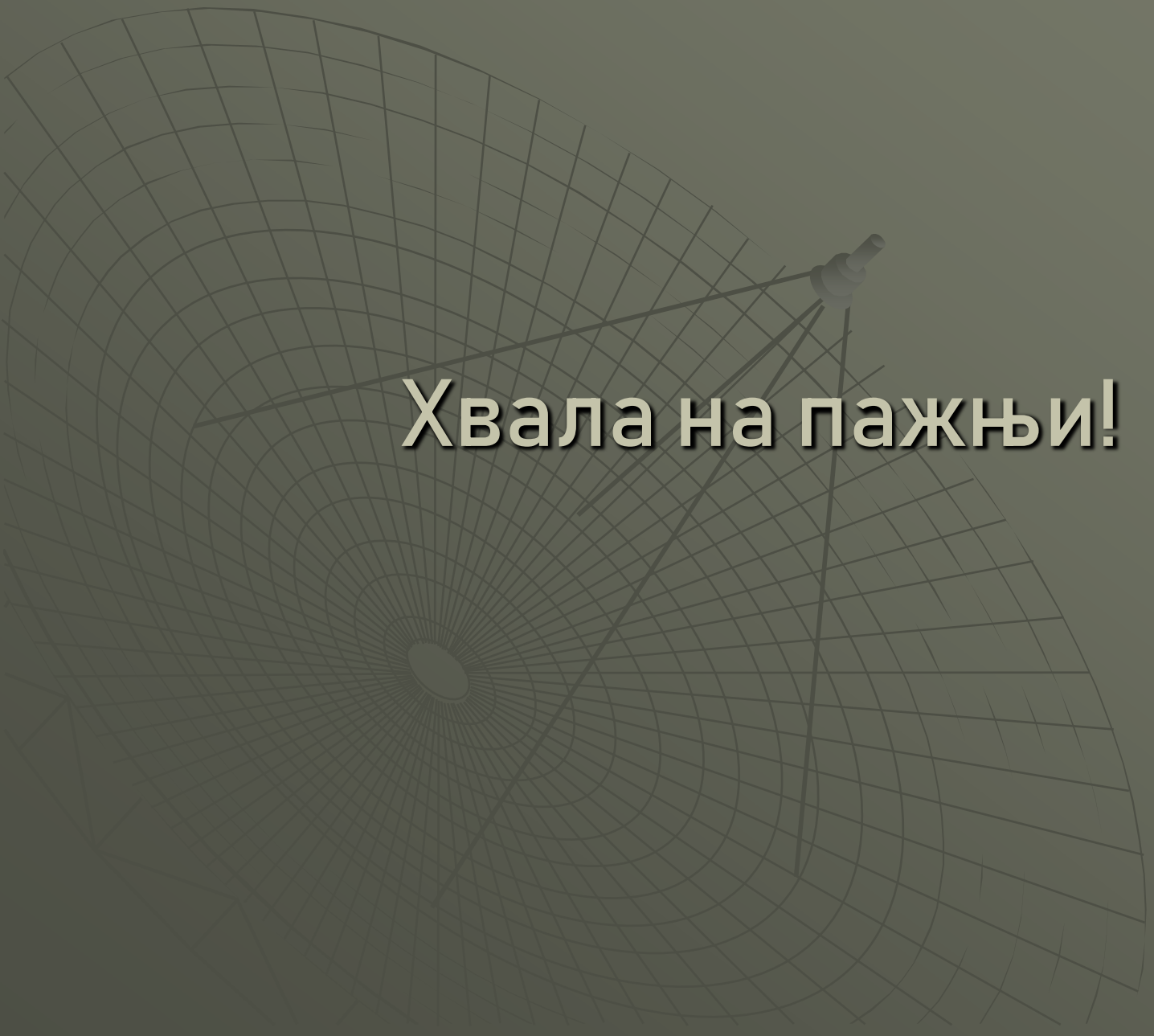
	precision	recall	f1-score	support
accuracy			0.76	6000
macro avg	0.77	0.76	0.76	6000
weighted avg	0.77	0.76	0.76	6000

Најбољи на факторима

- Најбоље се показали → SVM и KNN
- Учинак на факторима (*n* је број компоненти) добијеним из анализе главних компоненти

n	16	15	14	13	12	11	10	9	8
KNN	95	96	95	95	95	95	93	92	91
SVM	97	97	97	97	96	96	95	93	91

n	7	6	5	4	3	2	1
KNN	88	84	74	63	41	22	13
SVM	89	85	76	65	44	20	6



Хвала на пажњи!