ANALIZA SKUPA PODATAKA „DATA MINING AMAZON REVIEWS DATASET”

Nikola Veselinović
200/2015

ANALIZA SKUPA PODATAKA

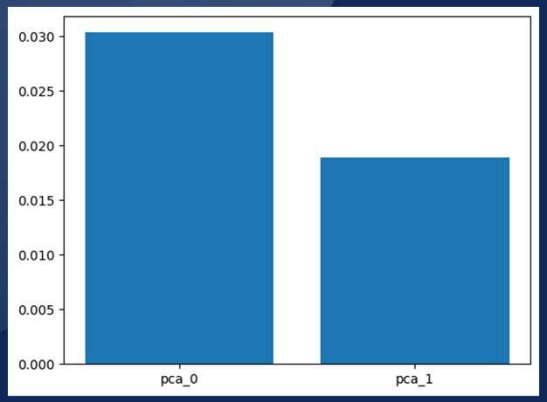
Karakteristike skupa podataka	Više-parametarski, tekst, teorija domena	**Karakteristike atributa**	Celobrojni
Predviđen zadatak	Klasifikacija	**Izostavljene vrednosti**	Nema
Broj instanci	1500	**Broj parametara**	10000

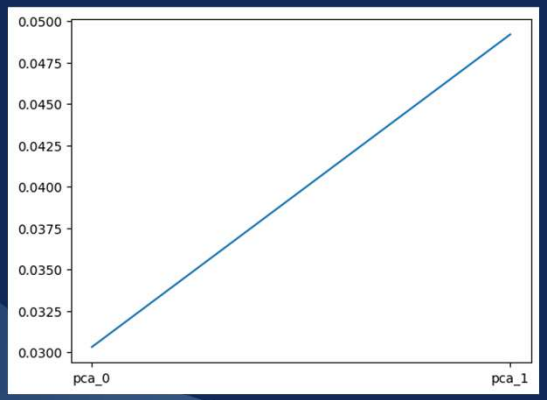
	the numeric	and numeric	a numeric	of numeric	to numeric	is numeric	I numeric	in numeric	that numeric	it numeric	...	ra_ numeric	le_to numeric	bra numeric	uch_a numeric	ave_a numeric
0	5	3	4	4	1	5	1	3	2	4	...	0	0	0	1	0
1	12	3	6	2	3	4	2	0	1	3	...	0	0	5	0	1
2	3	2	2	4	4	2	2	2	3	1	...	0	0	6	0	0
3	18	4	6	5	4	2	1	0	4	3	...	0	0	0	0	0
4	13	4	7	5	4	5	0	1	0	4	...	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	15	11	5	9	10	0	5	6	1	3	...	0	0	0	0	0
1496	12	7	7	5	5	3	3	2	1	5	...	0	0	0	0	0
1497	8	10	2	4	2	0	2	3	1	1	...	0	0	0	0	0
1498	11	12	10	7	8	4	4	7	2	2	...	0	0	0	0	0
1499	8	7	5	6	5	4	4	11	6	0	...	0	0	0	0	0

1500 rows × 10001 columns

- Skup se sastoji od 1001 atributa koji su podeljeni u 2 vrste (numeričke i tekstualne)
- Atribut „class“ je jedini tekstualni i predstavlja ima Autora komentara
- Ostali atributi zajedno sa svojim vrednostima predstavljaju sadržaj komentara
- Ne postoje nedostajuće vrednosti
- Jednak broj komentara po autoru

PRETPROCESIRANJE





- Uklanjanje duplikata atributa
- Rešavanje problema nepostojećih vrednosti
- Normalizacija podataka
- Podela na test i trening skupove

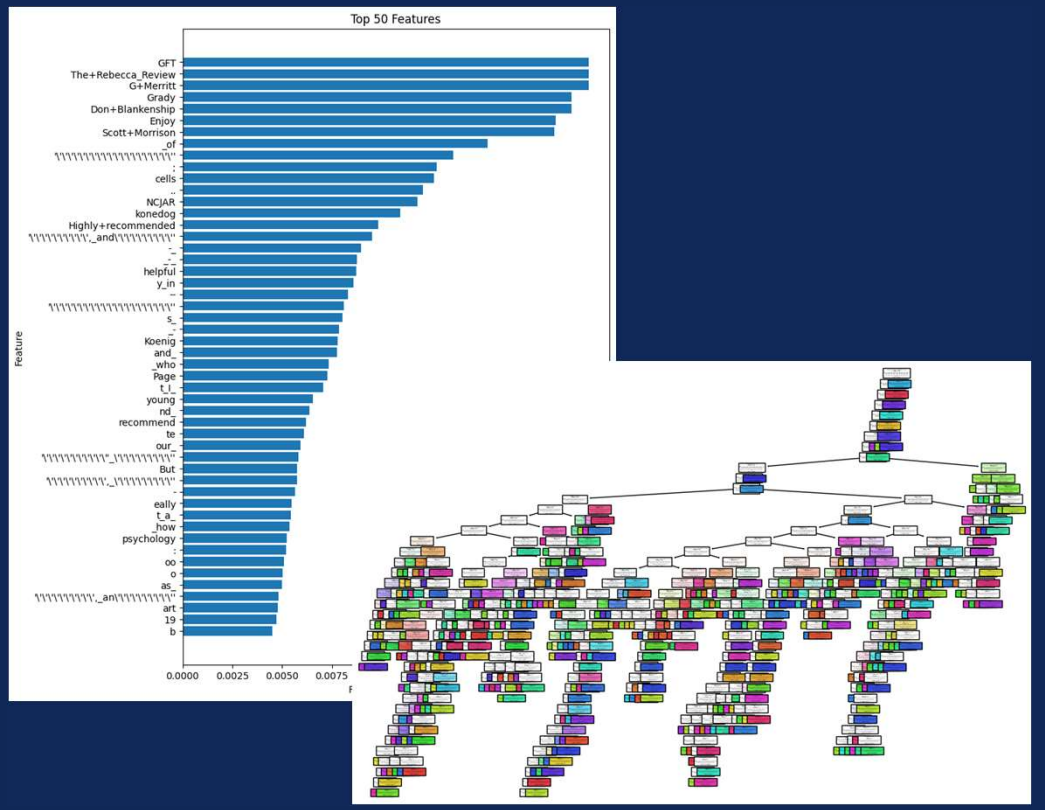
- Zamena tekstualnih vrednosti celobrojnim
- Smanjenje dimenzije skupa podataka

- Izmena vrednosti skupa podataka tako da budu tipa bool

KLASIFIKACIJA
STABLO ODLUČIVANJA (DECISION TREE)

```
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 1.0
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 1  0 23 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 0.9783333333333334
```

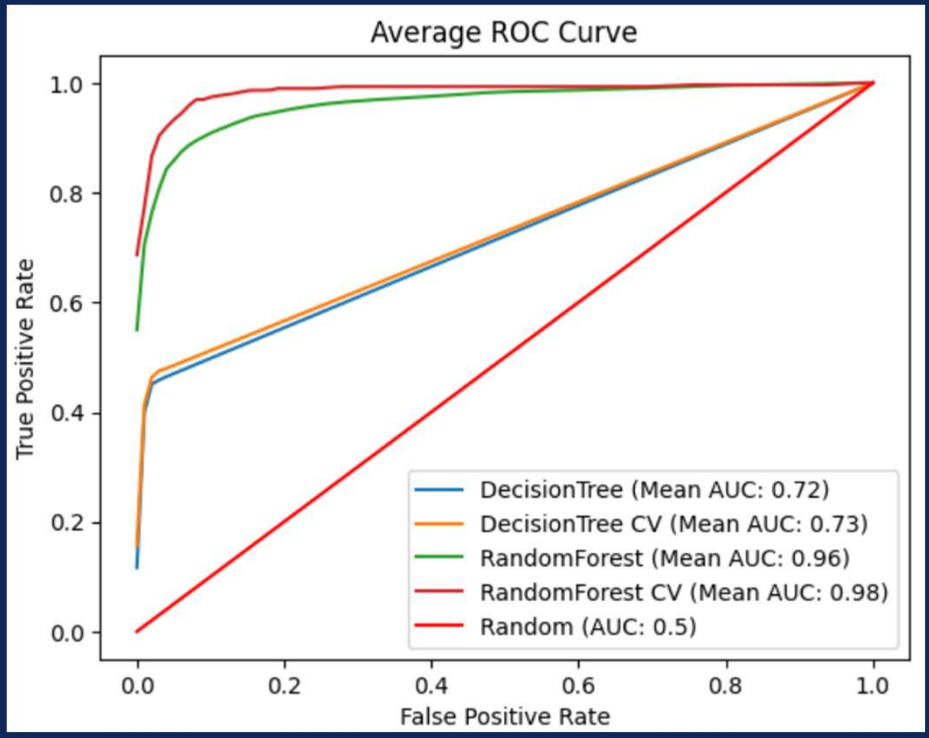
```
Test data:
Confusion matrix:
 [[1 0 0 ... 0 0 0]
 [0 3 0 ... 2 0 0]
 [0 0 2 ... 0 0 0]
 ...
 [0 0 1 ... 1 0 0]
 [0 0 0 ... 0 2 0]
 [0 0 0 ... 0 0 1]]
Accuracy score:
 0.45
Test data:
Confusion matrix:
 [[3 0 0 ... 0 0 0]
 [0 3 0 ... 1 0 1]
 [0 0 2 ... 0 0 0]
 ...
 [0 0 1 ... 1 0 0]
 [0 0 0 ... 0 2 0]
 [0 0 0 ... 0 0 1]]
Accuracy score:
 0.46
```



KLASIFIKACIJA
SLUČAJNA ŠUMA (RANDOM FOREST)

```
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 1.0
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 1.0
```

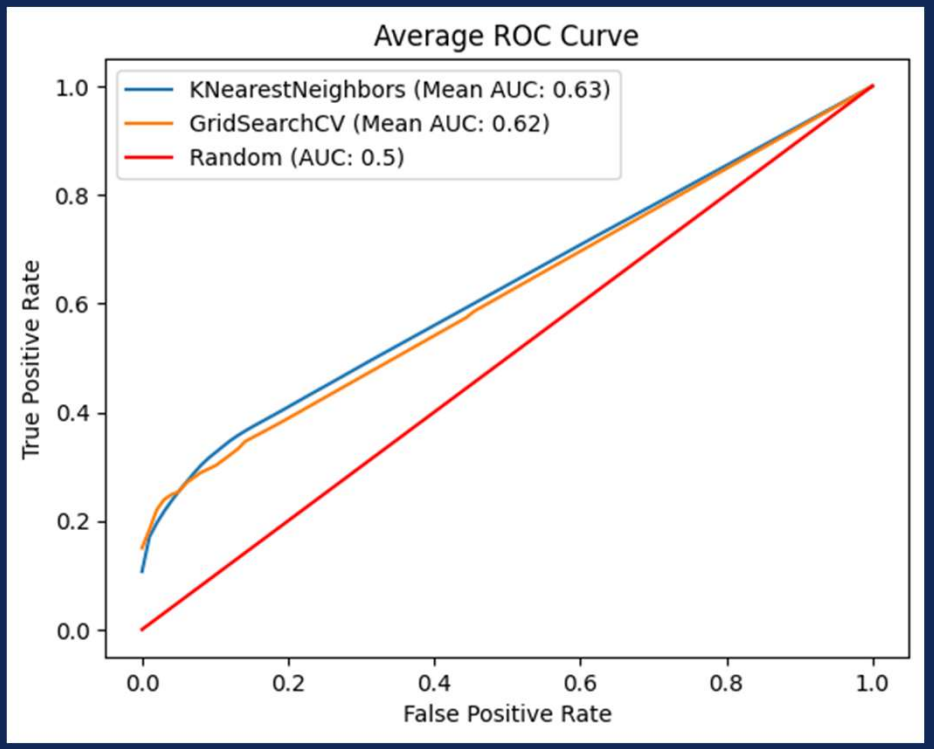
```
Test data:
Confusion matrix:
 [[6 0 0 ... 0 0 0]
 [0 6 0 ... 0 0 0]
 [0 0 6 ... 0 0 0]
 ...
 [0 0 1 ... 3 0 0]
 [0 1 0 ... 0 3 1]
 [0 0 0 ... 0 0 5]]
Accuracy score:
 0.7033333333333334
Test data:
Confusion matrix:
 [[3 0 0 ... 0 0 0]
 [0 6 0 ... 0 0 0]
 [0 0 5 ... 0 0 0]
 ...
 [0 0 0 ... 5 0 0]
 [0 0 0 ... 0 5 0]
 [0 0 0 ... 0 0 5]]
Accuracy score:
 0.8233333333333334
```



KLASIFIKACIJA
K NAJBLIŽIH SUSEDA (KNN)

```
Train data:                          Test data:
Confusion matrix:                    Confusion matrix:
 [[ 8  0  0 ...  0  0  0]            [[1 0 2 ... 0 0 0]
 [ 0 12  0 ...  0  0  0]             [0 3 0 ... 0 0 0]
 [ 0  0 20 ...  0  0  0]             [0 0 4 ... 0 0 0]
 ...                                 ...
 [ 0  0  2 ... 13  0  0]             [0 0 1 ... 0 0 0]
 [ 0  1 10 ...  0  1  0]             [0 1 2 ... 0 0 0]
 [ 0  0  5 ...  0  0  7]]            [0 0 1 ... 0 0 0]]
Accuracy score:                      Accuracy score:
 0.275                                0.11333333333333333
```

```
Train data:                          Test data:
Confusion matrix:                    Confusion matrix:
 [[24  0  0 ...  0  0  0]            [[0 0 0 ... 0 0 0]
 [ 0 24  0 ...  0  0  0]             [0 1 0 ... 0 0 0]
 [ 0  0 24 ...  0  0  0]             [0 0 6 ... 0 0 0]
 ...                                 ...
 [ 0  0  0 ... 24  0  0]             [0 0 2 ... 0 0 0]
 [ 0  0  0 ...  0 24  0]             [0 0 0 ... 2 0 0]
 [ 0  0  0 ...  0  0 24]]            [0 0 1 ... 0 0 2]]
Accuracy score:                      Accuracy score:
 1.0                                  0.19666666666666666
```



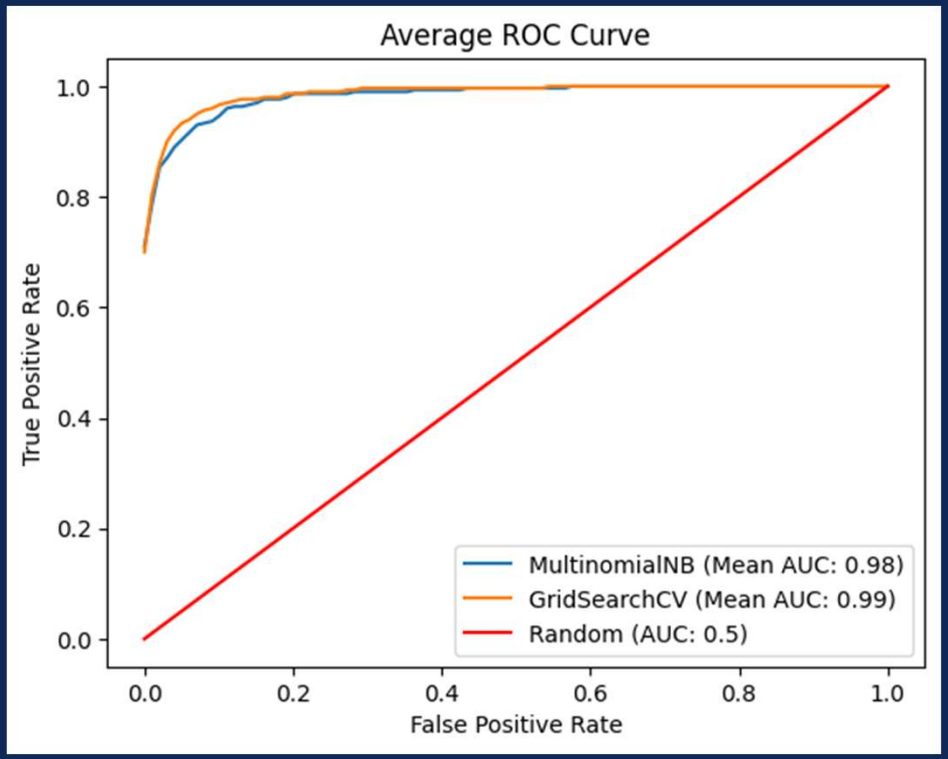
KLASIFIKACIJA
NAIVNI BAJES (MULTINOMIALNI)

```
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 0.9975
```

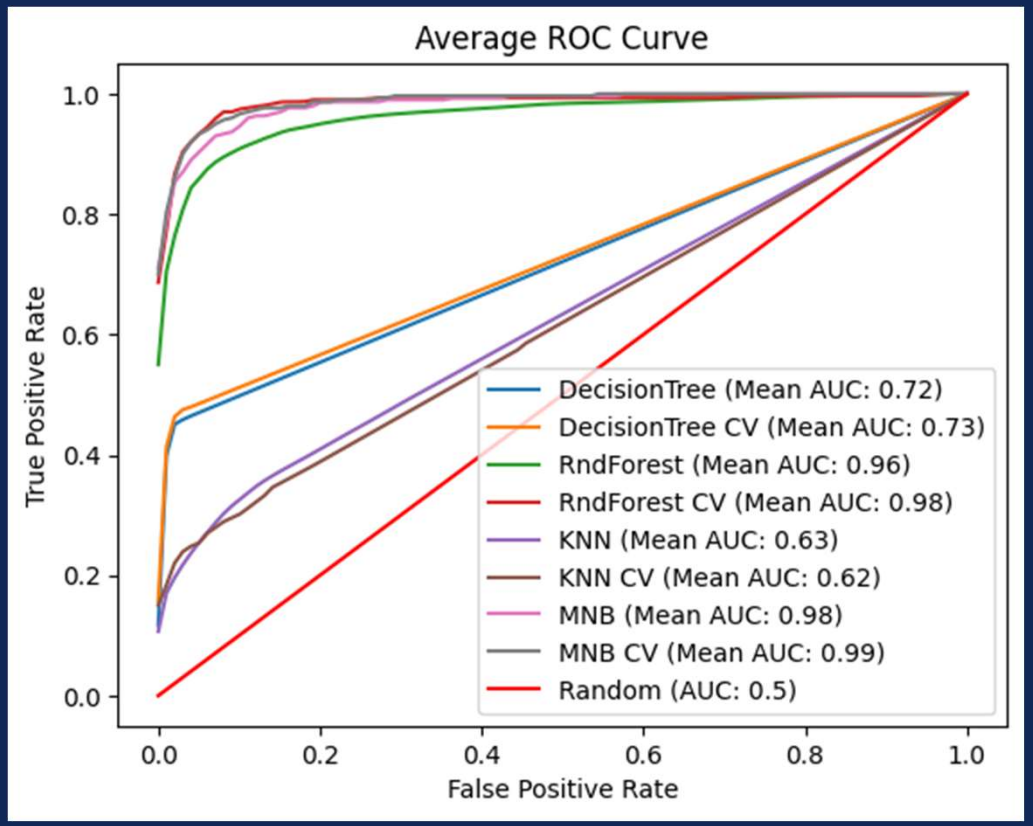
```
Test data:
Confusion matrix:
 [[3 0 0 ... 0 0 0]
 [0 4 0 ... 0 0 0]
 [0 0 5 ... 0 1 0]
 ...
 [0 0 0 ... 1 0 1]
 [0 0 0 ... 0 6 0]
 [0 0 0 ... 0 0 5]]
Accuracy score:
 0.71
```

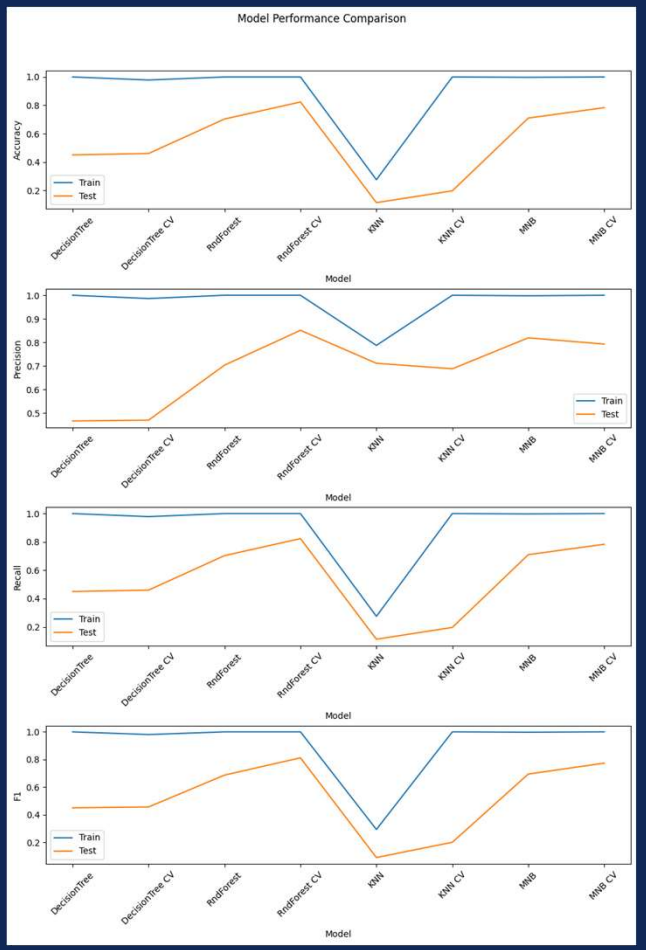
```
Train data:
Confusion matrix:
 [[24  0  0 ...  0  0  0]
 [ 0 24  0 ...  0  0  0]
 [ 0  0 24 ...  0  0  0]
 ...
 [ 0  0  0 ... 24  0  0]
 [ 0  0  0 ...  0 24  0]
 [ 0  0  0 ...  0  0 24]]
Accuracy score:
 1.0
```

```
Test data:
Confusion matrix:
 [[5 0 0 ... 0 0 0]
 [0 6 0 ... 0 0 0]
 [0 0 6 ... 0 0 0]
 ...
 [0 0 0 ... 5 0 0]
 [0 0 0 ... 0 5 0]
 [0 0 0 ... 0 0 6]]
Accuracy score:
 0.7833333333333333
```

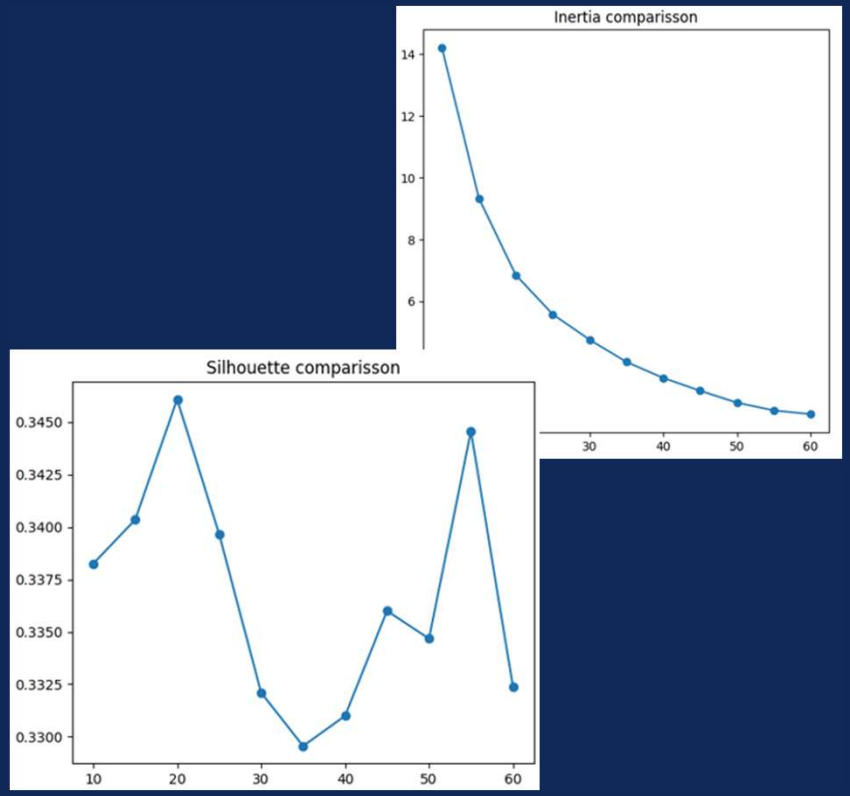


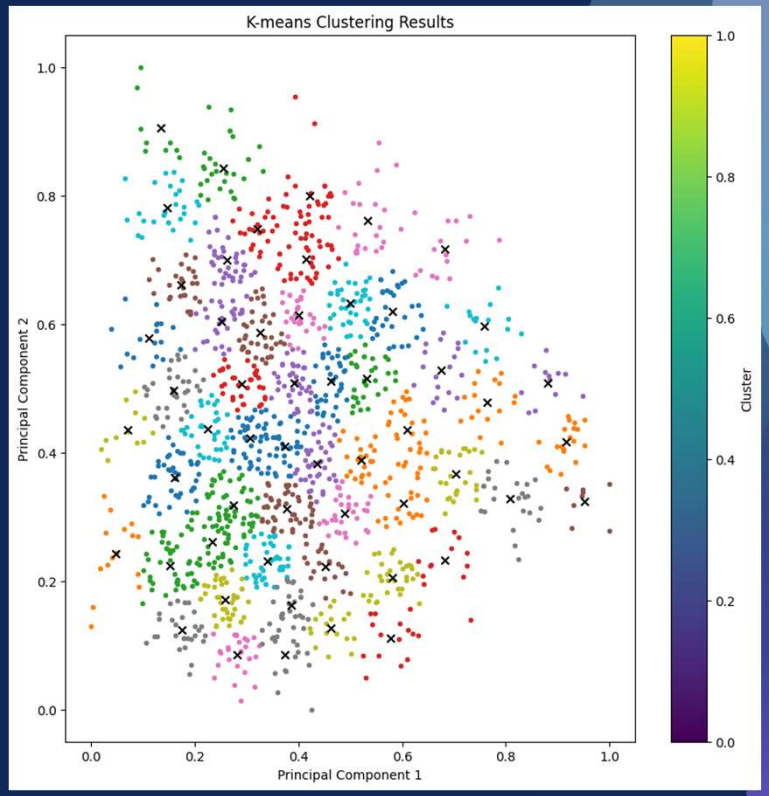
KLASIFIKACIJA



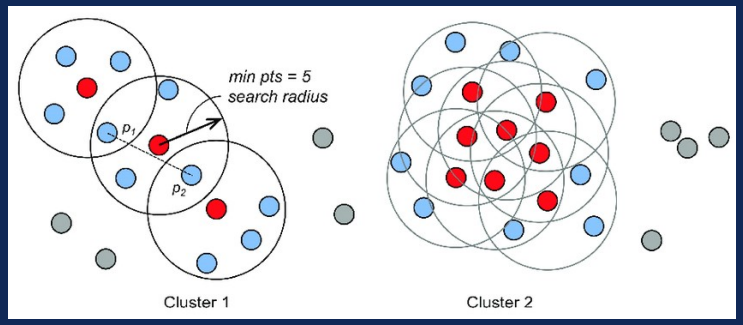


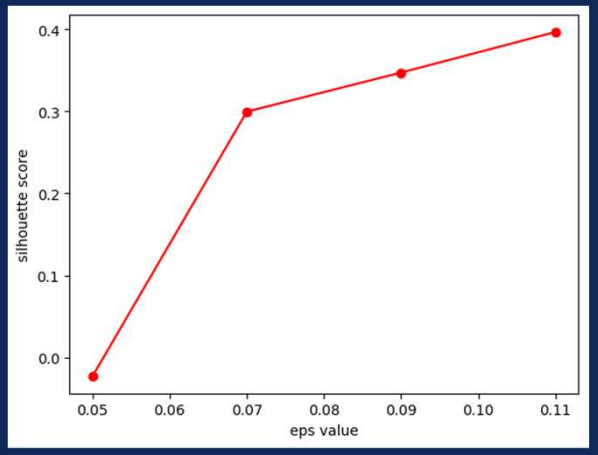
KLASTEROVANJE
K SREDINA (KMEANS)

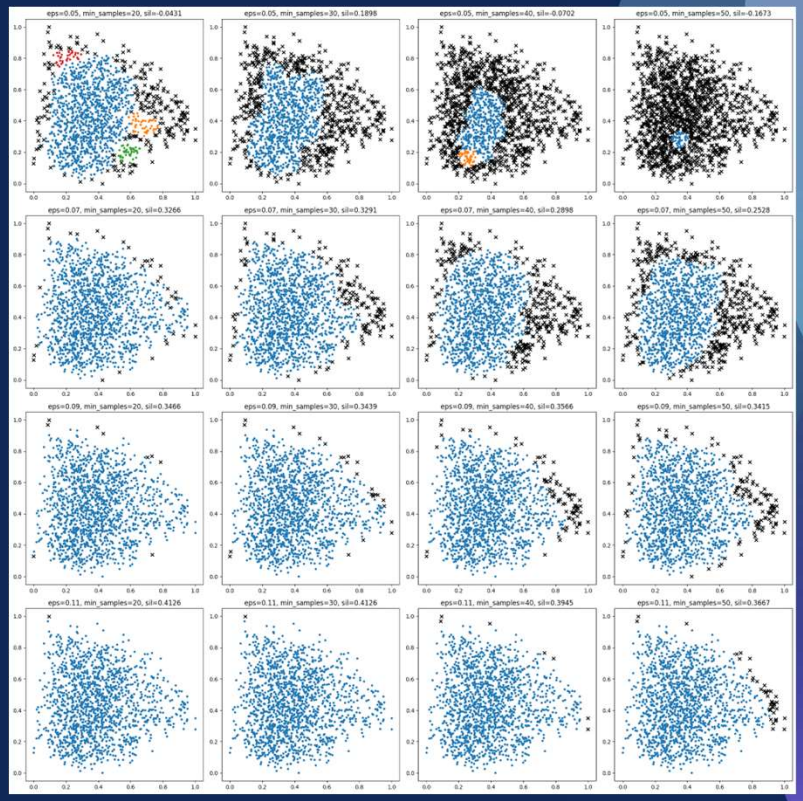




KLASTEROVANJE
DBSCAN

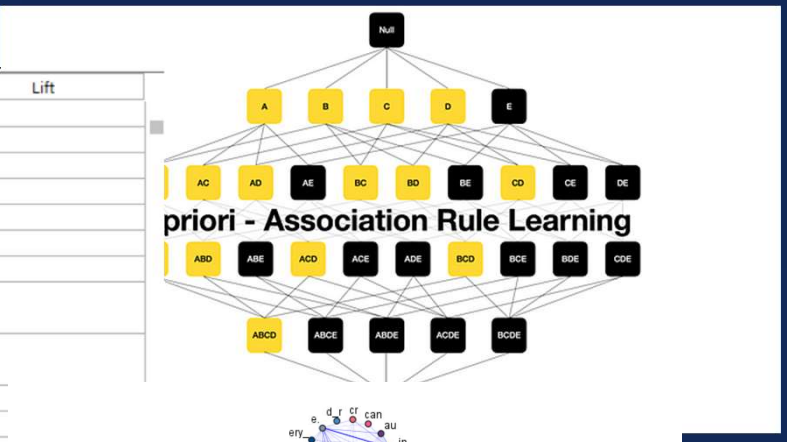


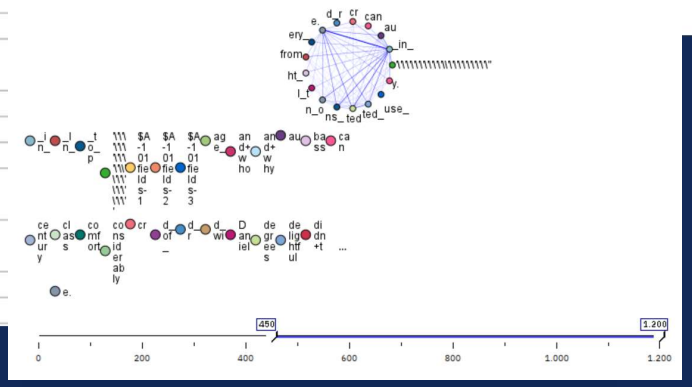




PRAVILA ASOCIJACIJE
APRIORI ALGORITAM

Consequent	Antecedent	Support %	Confidence %	Lift
e.	_in_	93,733	85,277	1,006
_in_	e.	84,733	94,335	1,006
ted_	ted	69,133	90,55	1,446
e.	ted	69,133	85,921	1,014
_in_	ted	69,133	94,311	1,006
e.	ns_	65,933	85,035	1,004
_in_	ns_	65,933	95,046	1,014
ted_	ted _in_	65,2	90,9	1,452
e.	ted _in_	65,2	86,401	1,02
e.	au	63,2	85,654	1,011
_in_	au	63,2	93,776	1,0
e.	ns_ _in_	62,667	85,638	1,011
ted	ted_	62,6	100,0	1,446
e.	ted_	62,6	85,729	1,012
_in_	ted_	62,6	94,675	1,01
e.	ted_ ted	62,6	85,729	1,012
_in_	ted_ ted	62,6	94,675	1,01
e.	n_o	60,2	86,489	1,021
_in_	n_o	60,2	95,238	1,016





ZAKLJUČAK

Svaka metoda istraživanja podatak ima svoju prednost i manu. Za ovaj konkretan skup podataka kao najbolje su se pokazale metode za klasifikaciju, primarno Multinomialni Naivni Bajes i Slučajna Šuma.

HVALA NA PAŽNJI

Nikola Veselinović

mi15200@alas.matf.bg.ac.rs

https://github.com/MATF-istrazivanje-podataka-1/2023_Data_Mining_Amazon_reviews_Dataset