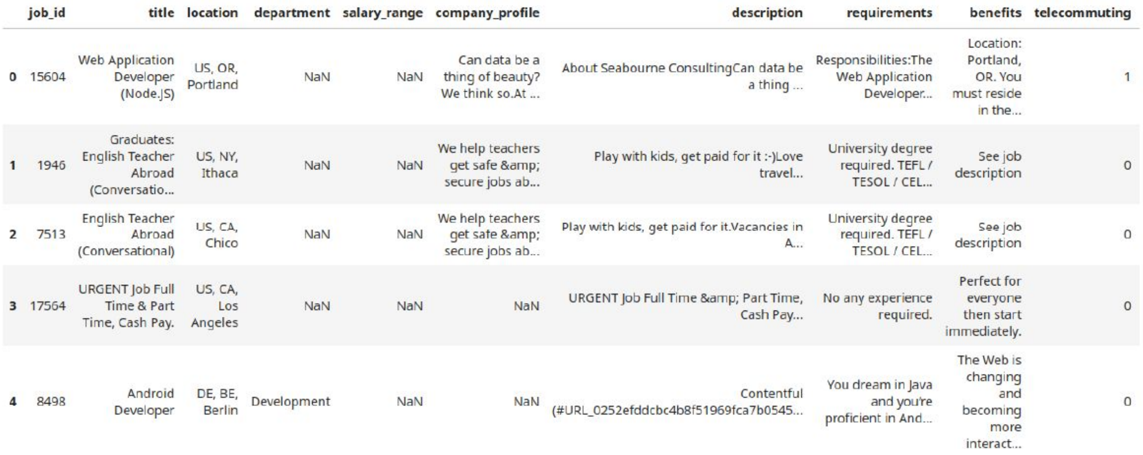
Detekcija lažnih oglasa za posao

Istraživanje Podataka 1

Uvod

- NLP skup podataka - *Kaggle - Real or Fake Job Postings*

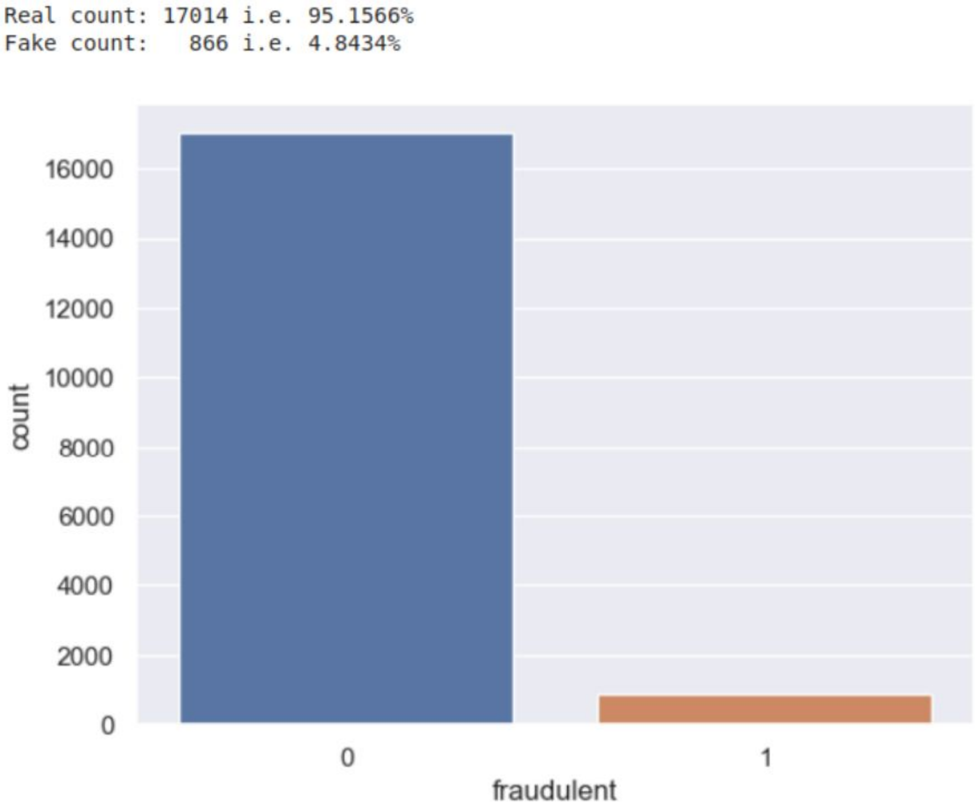
17880 instanci
14 tekstualnih kolona
4 binarna atributa



	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting
0	15604	Web Application Developer (Node.JS)	US, OR, Portland	NaN	NaN	Can data be a thing of beauty? We think so.At ...	About Seabourne ConsultingCan data be a thing ...	Responsibilities:The Web Application Developer...	Location: Portland, OR. You must reside in the...	1
1	1946	Graduates: English Teacher Abroad (Conversatio...	US, NY, Ithaca	NaN	NaN	We help teachers get safe &amp; secure jobs ab...	Play with kids, get paid for it :-)Love travel...	University degree required. TEFL / TESOL / CEL...	See job description	0
2	7513	English Teacher Abroad (Conversational)	US, CA, Chico	NaN	NaN	We help teachers get safe &amp; secure jobs ab...	Play with kids, get paid for it.Vacancies in A...	University degree required. TEFL / TESOL / CEL...	See job description	0
3	17564	URGENT Job Full Time & Part Time, Cash Pay.	US, CA, Los Angeles	NaN	NaN	NaN	URGENT Job Full Time &amp; Part Time, Cash Pay...	No any experience required.	Perfect for everyone then start immediately.	0
4	8498	Android Developer	DE, BE, Berlin	Development	NaN	NaN	Contentful (#URL_0252efddcbc4b8f51969fca7b0545...	You dream in Java and you're proficient in And...	The Web is changing and becoming more interact...	0

Eksplorativna analiza & pretprocesiranje

Nebalansiranost

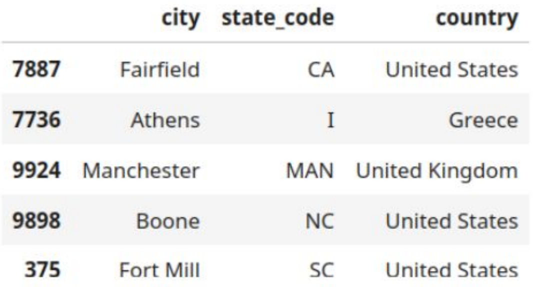


```
Real count: 17014 i.e. 95.1566%
Fake count:   866 i.e. 4.8434%
```

- Razmatrane opcije
  - Uzorkovanje
    - *oversampling*
    - *undersampling*
  - **Izbor modela**

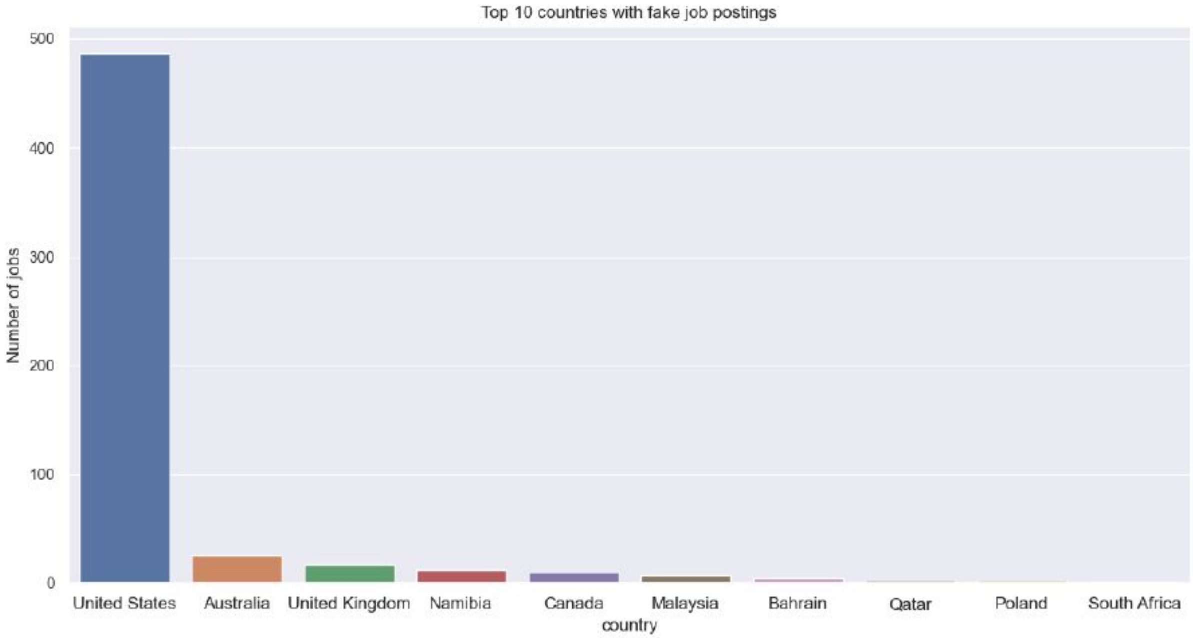
Lokacija

- Dodatni skup podataka - *country-list*



	city	state_code	country
7887	Fairfield	CA	United States
7736	Athens	I	Greece
9924	Manchester	MAN	United Kingdom
9898	Boone	NC	United States
375	Fort Mill	SC	United States

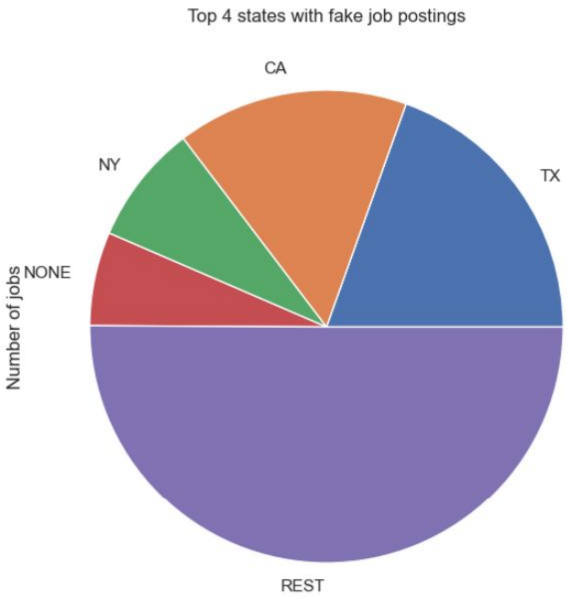
slučajan uzorak transformisanog skupa podataka



Dominacija SAD nije problem
(u ovom radu)

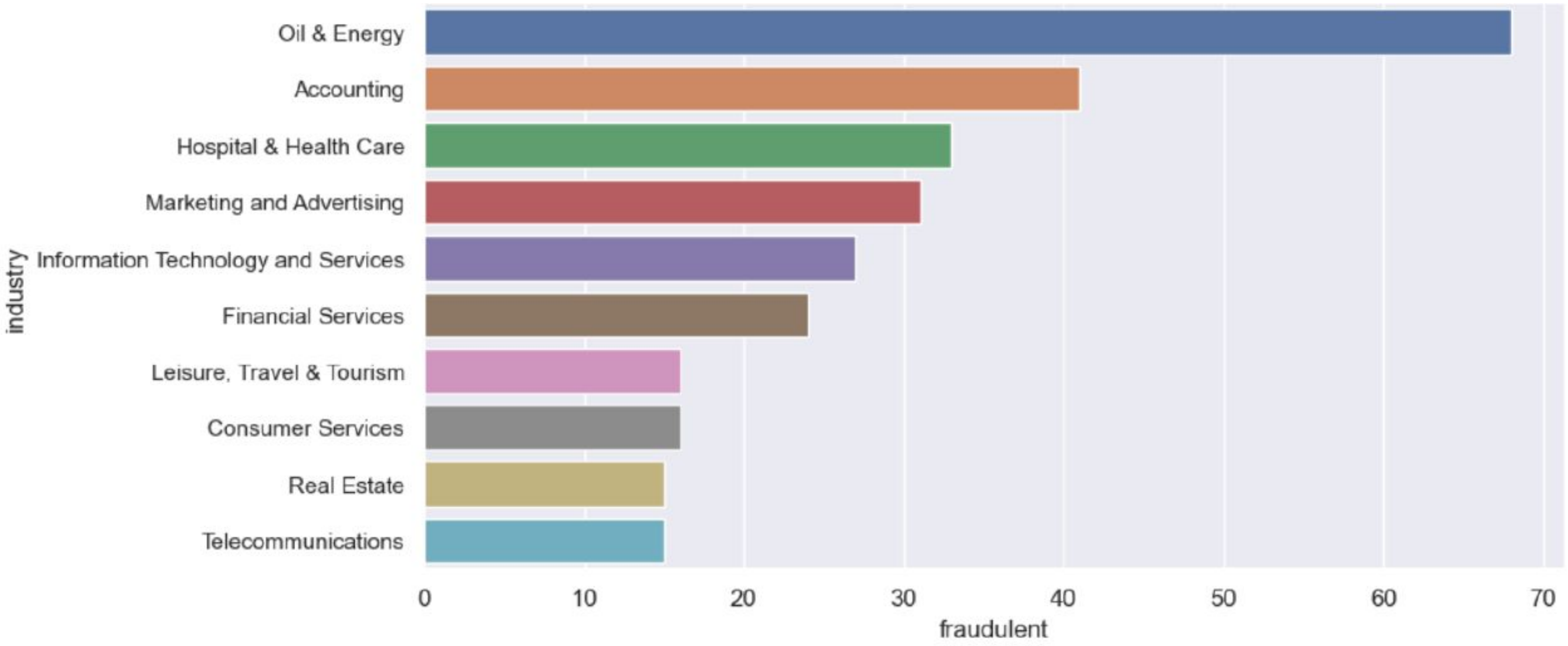
TF-IDF

Lokacija

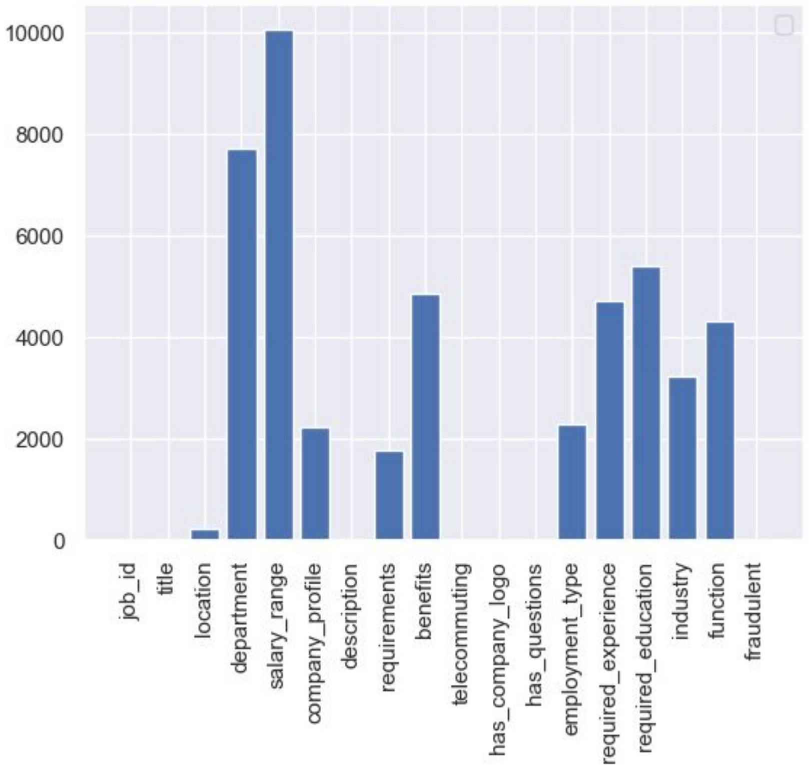


Najčešće države sa lažnim oglasima

Industrija


Oil & Energy
Accounting
Hospital & Health Care
Marketing and Advertising
Information Technology and Services
Financial Services
Leisure, Travel & Tourism
Consumer Services
Real Estate
Telecommunications
industry
0
10
20
30
40
50
60
70
fraudulent

Nedostajući podaci & elementi van granica



1. Mala slova
2. Interpunkcijski znakovi
3. Tokenizacija
4. Lematizacija
5. Stop reči
6. Predugačke i prekratke reči

Klasifikacija

Komplementarni Naivni Bajes

- Dizajniran kao nadogradnja Multinomijalnog Naivnog Bajesa sa boljom podrškom za **disbalansirane skupove podataka**
  - Originalni radi - *Rennie et al.*

- Unakrsna validacija za α

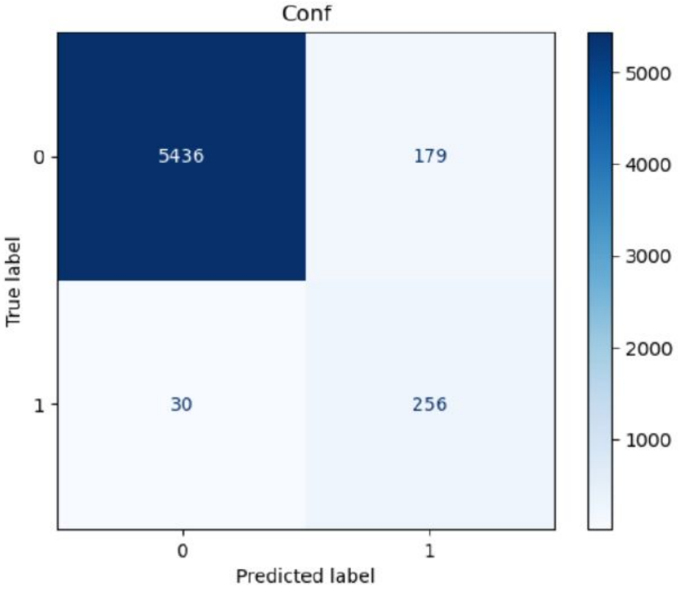
$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

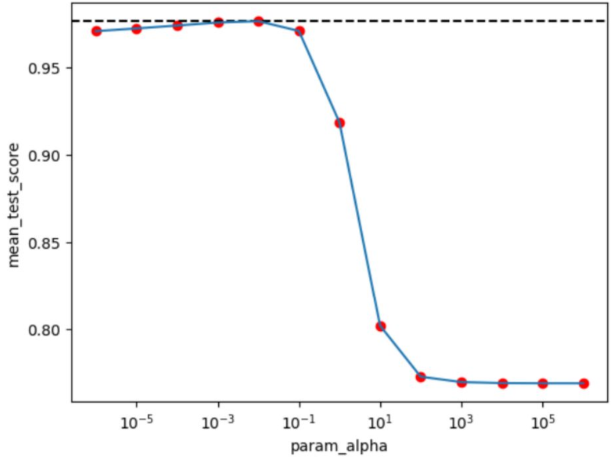
$$w_{ci} = \log \hat{\theta}_{ci}$$

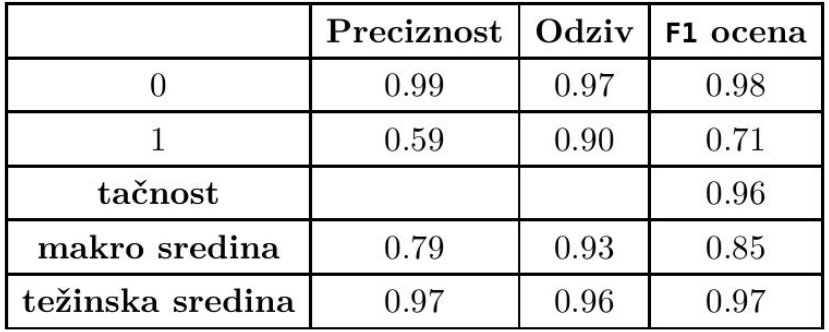
$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{cj}|}$$

$$\hat{c} = \arg\min_c \sum_i t_i w_{ci}$$

Komplementarni Naivni Bajes



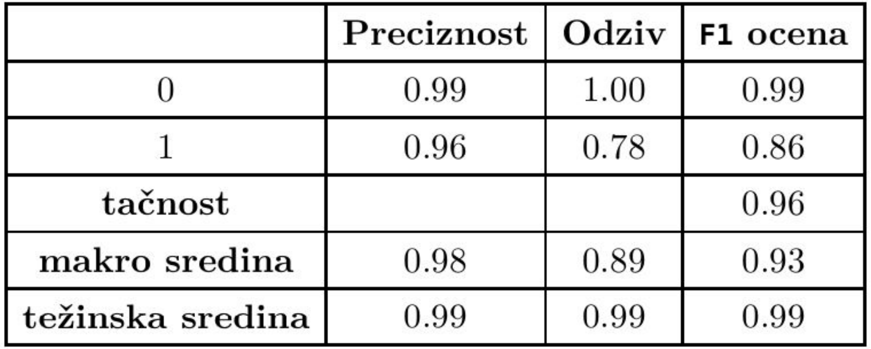




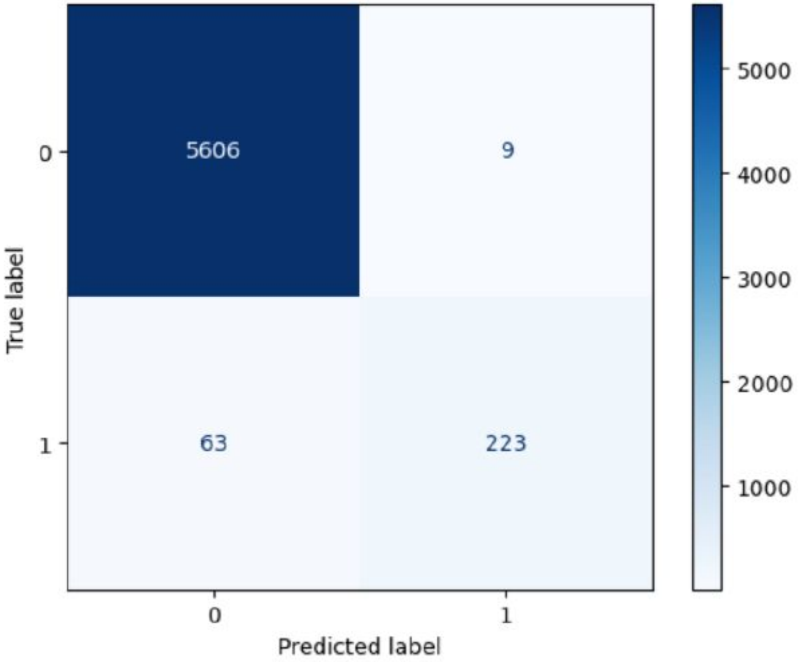
	Preciznost	Odziv	F1 ocena
0	0.99	0.97	0.98
1	0.59	0.90	0.71
tačnost			0.96
makro sredina	0.79	0.93	0.85
težinska sredina	0.97	0.96	0.97

Gradijentno Pojačavajuća Drveta Odlučivanja

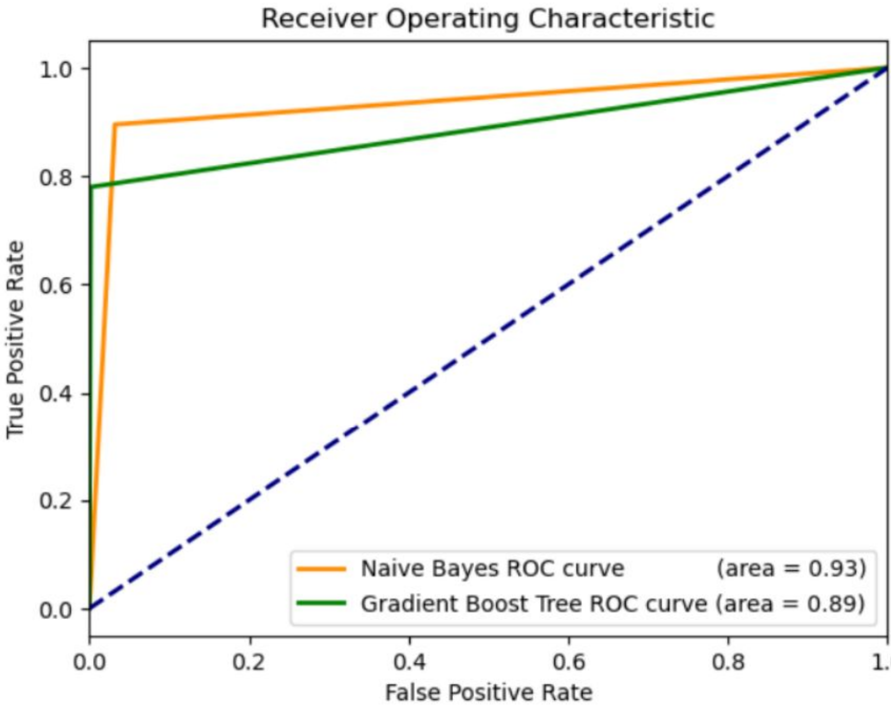
- Zasnovan na LightGBM
  - *Originalni rad - Ke et al.*
- Ansambl - pojačavanje
- Implementacija sa histogramima



	Preciznost	Odziv	F1 ocena
0	0.99	1.00	0.99
1	0.96	0.78	0.86
tačnost			0.96
makro sredina	0.98	0.89	0.93
težinska sredina	0.99	0.99	0.99



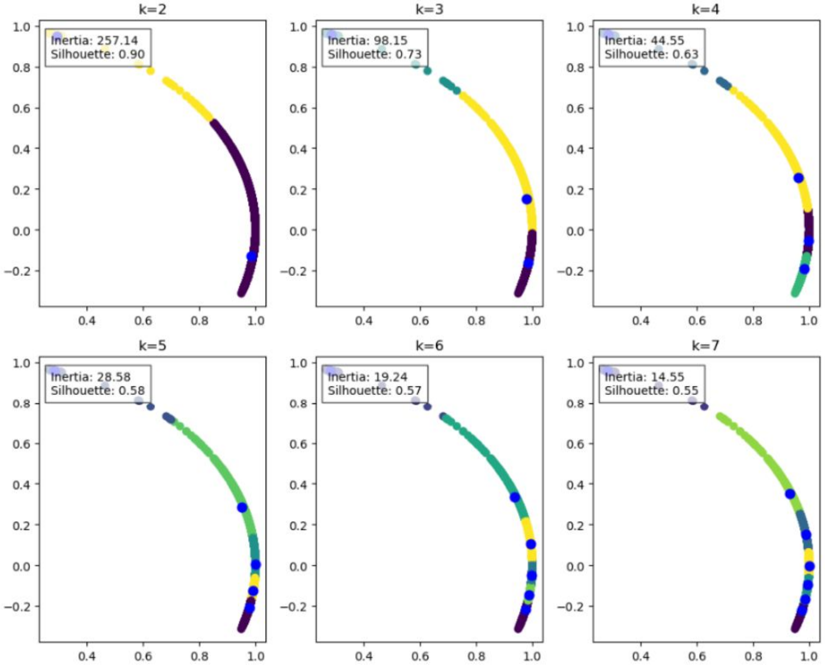
Poređenje

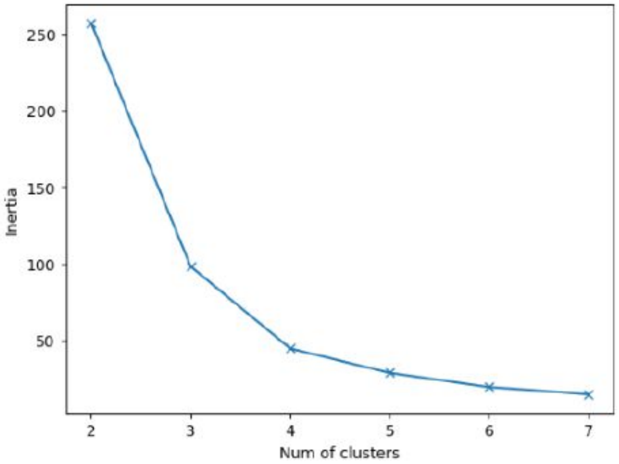

Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes ROC curve (area = 0.93)
Gradient Boost Tree ROC curve (area = 0.89)

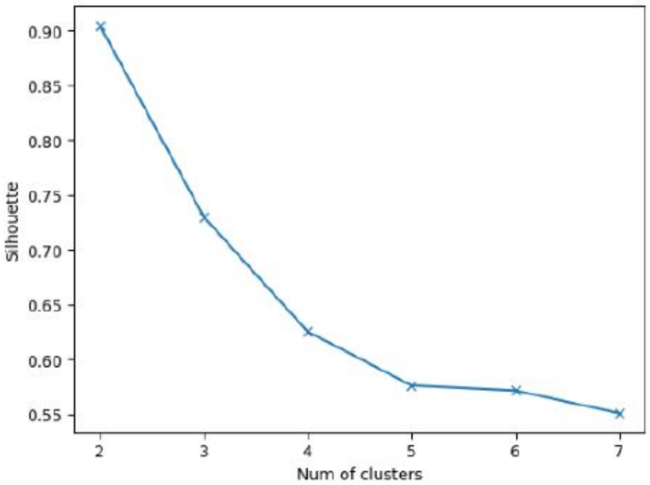
Klasterovanje

K sredina

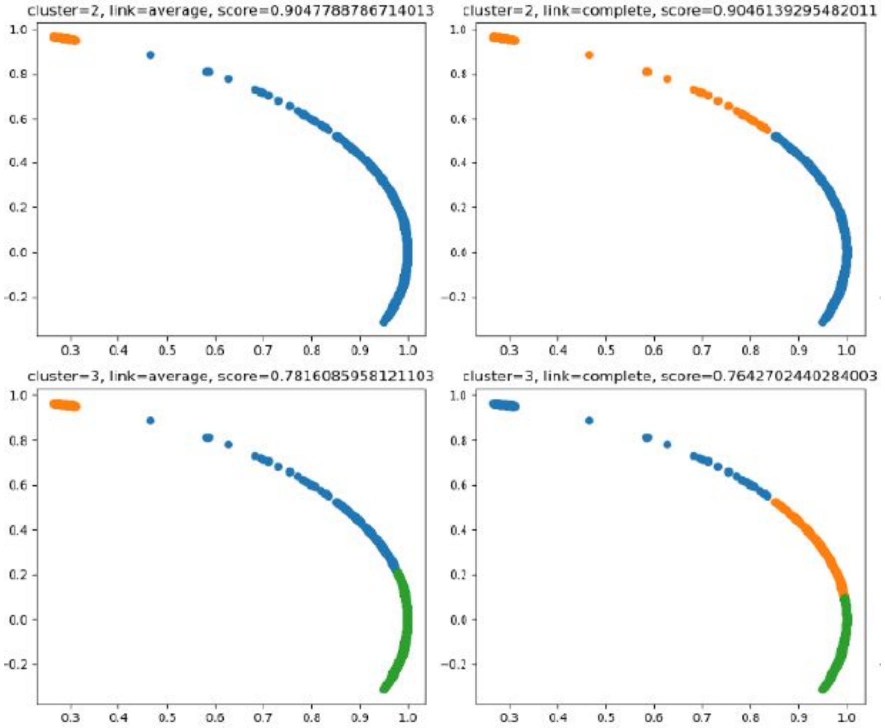
- Skrivena semantička analiza - LSA (eng. *Latent Semantic Analysis)*
- k-mean++

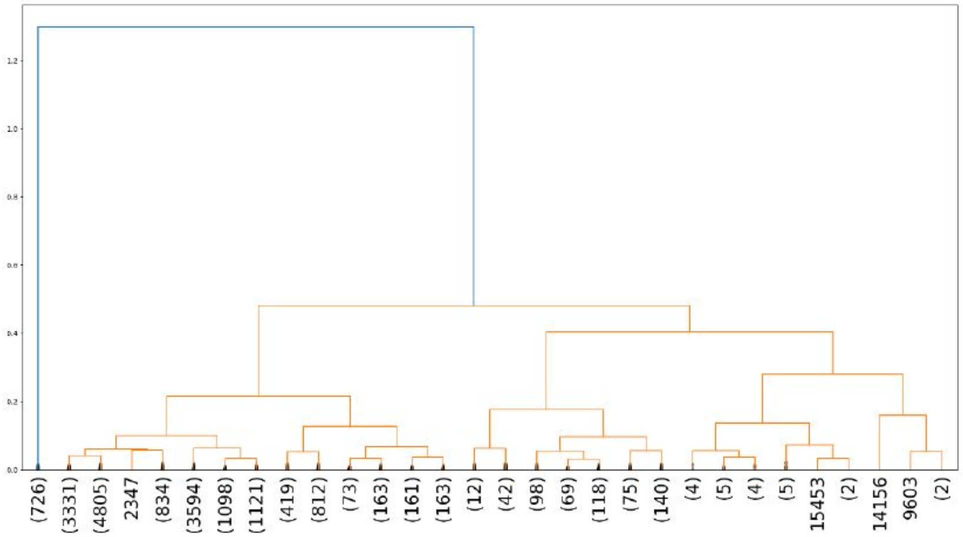






Hijerahijsko klasterovanje

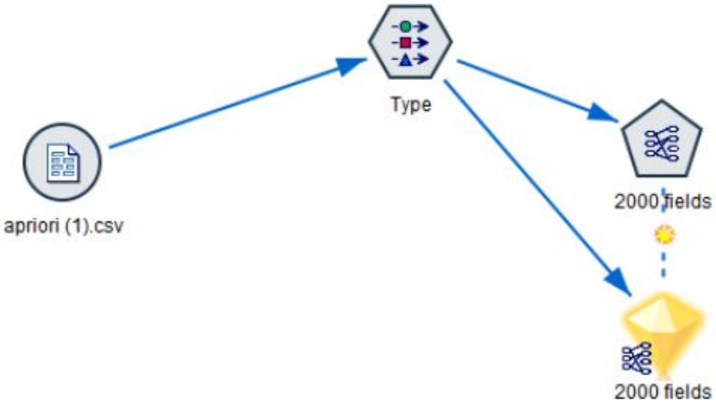


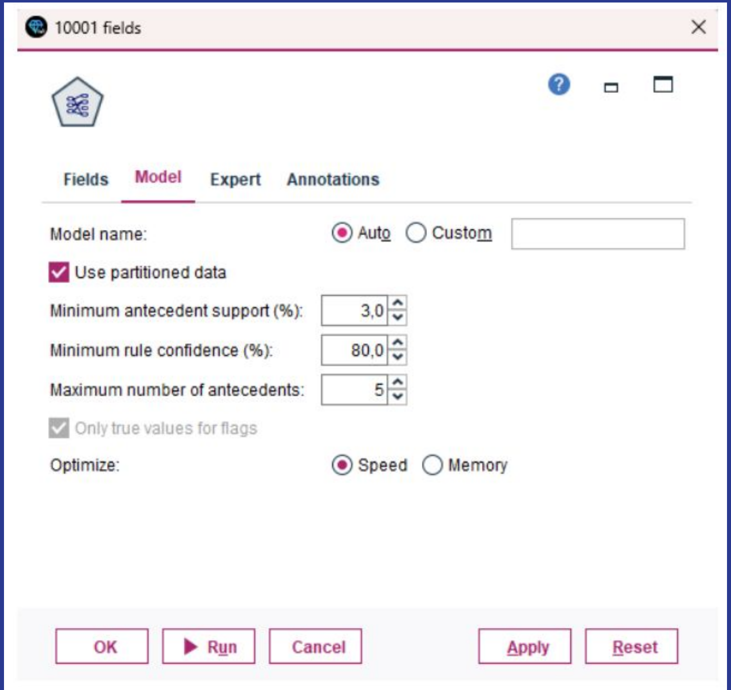


Pravila pridruživanja

Apriori

- Otkriveno 483943 pravila





Model name: ● Auto ○ Custom

☑ Use partitioned data

Minimum antecedent support (%): 3,0

Minimum rule confidence (%): 80,0

Maximum number of antecedents: 5

☑ Only true values for flags

Optimize: ● Speed ○ Memory

Zaključak