

# Klasifikacija sajtova analizom URL-ova

Marko Veljković 1096/2019

# Početni skup

- ❖ Sadrži 2 kolone 'url' koja sadrži url sajta i 'label' koja predstavlja kategoriju sajta i može imati 2 vrednosti: 'bad' ukoliko je sajt maliciozan i 'good' ukoliko je sajt bezbedan

creditgratosse.blogspot.com/2008/01/connnectoi_03.html	bad
aijcs.blogspot.com/2005/03/colourful-life-of-aij.html	bad
tudu-free.blogspot.com/2008/02/jogos-java-aplicativos.html#footer-wrap2	bad
floridarentfinders.com/uploads/ws/css/www.paypal.com/cgi-bin/websrcmd=_login-run/update.php	bad
paypollar.com.p12.hostingprod.com/wbsc.php	bad
01453.com/	good
015fb31.netsolhost.com/bosstweed.html	good
02bee66.netsolhost.com/lincolnhomepage/	good
02ec0a3.netsolhost.com/getperson.php?personID=14920&tree=ncshawfamily	good
032255.com/	good

- ❖ Ispitivanjem je utvrđeno da skup sadrži ponovljene redove i oni su uklonjeni iz skupa podataka

```
Original data length: 420464  
Data without duplicates length: 411248
```

- ❖ Skup ne sadrži null ni NaN vrednosti

## ❖ O URL-ovima

- ❖ Protokol (shema)
- ❖ Sub-domén
- ❖ Domen
- ❖ TLD
- ❖ Duboki URL



## ❖ Statički podaci

- ❖ Leksički podaci (dužina URL-a, broj specijalnih karaktera, postojanje IP adrese)
- ❖ Meta podaci (broj dana od registracije sajta, broj dana od poslednjeg ažuriranje sajta)

## ❖ Dinamički podaci

- ❖ Struktura stranice, linkovi, broj reči na stranici, prosečna dužina reči...

# Obrada URL-ova i dohvatanje meta podataka

## ❖ Leksički podaci

- ❖ Dužina URL-a; Da li se TLD nalazi u listi poznatih TLD-ova; Da li se TLD nalazi u listi sumnjivih TLD-ova; Da li sadrži IP adresu;
- ❖ Dužina dubokog URL-a; Broj „gen“ graničnika; Broj „sub“ graničnika; Broj specijalnih karaktera; Broj ne rezervisanih karaktera;
- ❖ Broj sub-domena; Da li sadrži http; Broj sumnjivih reči; Broj % znakova; Broj cifara; Odnos broja cifara i dužine URL-a;
- ❖ Da li sadži karakter = nakon karaktera ?; Da li sadrži ne standardni port

## ❖ Fункције за добијање мета података

- ❖ whois python biblioteka
- ❖ Podaci podeljeni u 12 (maksimalan broj jezgara) listi jednakih dužina
- ❖ Dohvatani meta podaci:
  - ❖ Broj dana proteklih od registracije sajta
  - ❖ Broj dana proteklih od zadnjeg ažuriranja
  - ❖ Broj dana do isteka važeња URL-a

## ❖ Opis krajnjih podataka

```
print(df.describe())
```

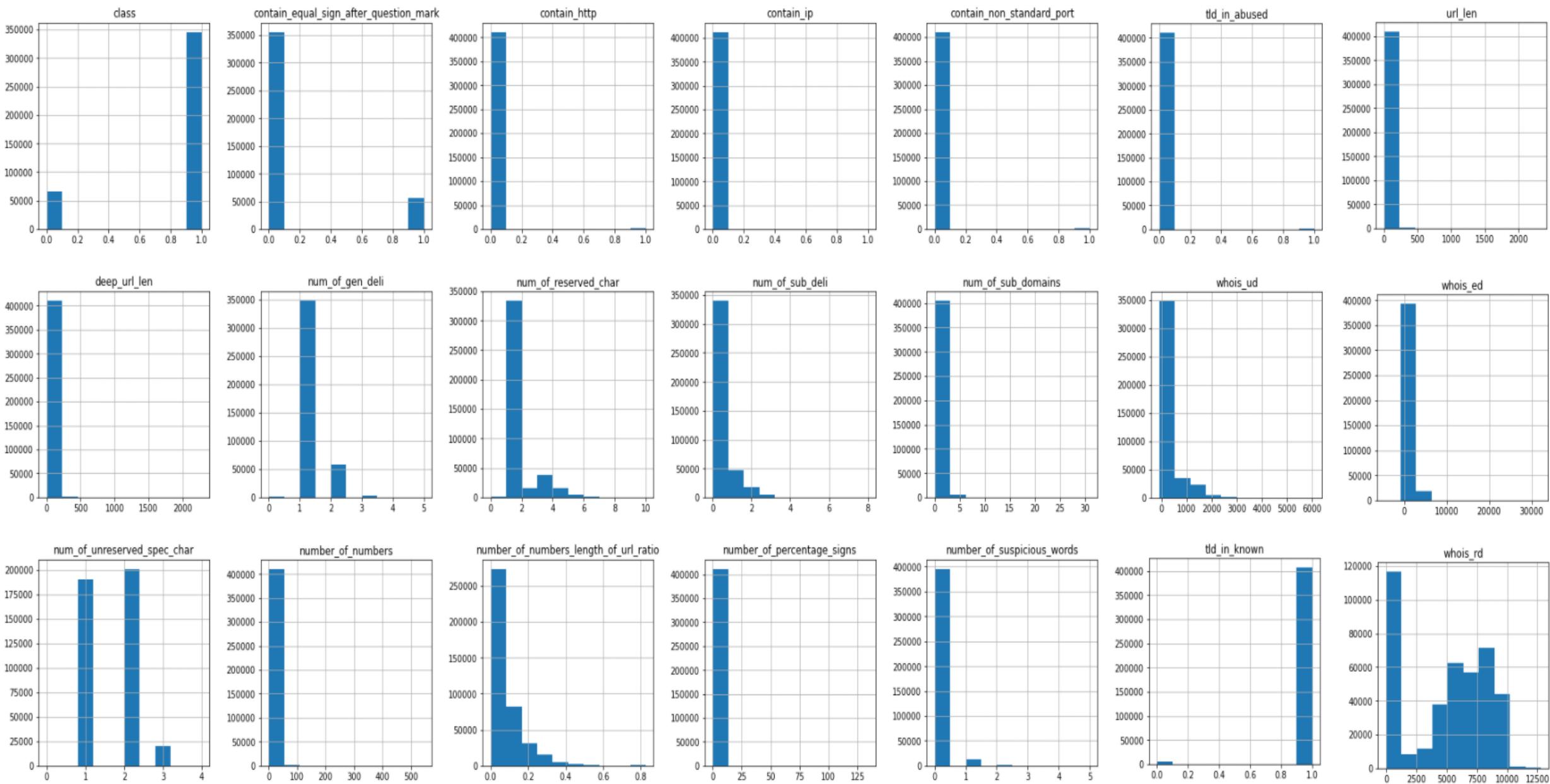
	url_len	tld_in_known	tld_in_abused	contain_ip	deep_url_len	num_of_gen_deli	num_of_sub_deli	num_of_reserved_char
count	411263.000000	411263.000000	411263.000000	411263.000000	411263.000000	411263.000000	411263.000000	411263.000000
mean	48.428648	0.988757	0.001734	0.000109	32.546478	1.154633	0.244892	1.399525
std	35.063820	0.105437	0.041601	0.010460	34.587544	0.390959	0.597133	0.926967
min	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	29.000000	1.000000	0.000000	0.000000	14.000000	1.000000	0.000000	1.000000
50%	41.000000	1.000000	0.000000	0.000000	26.000000	1.000000	0.000000	1.000000
75%	59.000000	1.000000	0.000000	0.000000	42.000000	1.000000	0.000000	1.000000
max	2307.000000	1.000000	1.000000	1.000000	2282.000000	5.000000	8.000000	10.000000

	num_of_unreserved_spec_char	num_of_sub_domains	number_of_suspicious_words	number_of_percentage_signs	number_of_numbers
count	411263.000000	411263.000000	411263.000000	411263.000000	411263.000000
mean	1.586391	1.448409	0.048913	0.088267	3.947547
std	0.584327	0.904155	0.260331	1.172438	8.149198
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	0.000000	0.000000	0.000000
50%	2.000000	1.000000	0.000000	0.000000	1.000000
75%	2.000000	2.000000	0.000000	0.000000	6.000000
max	4.000000	31.000000	5.000000	134.000000	545.000000

	number_of_numbers_length_of_url_ratio	contain_equal_sign_after_question_mark	contain_non_standard_port
count	411263.000000	411263.000000	411263.000000
mean	0.068881	0.137265	0.004469
std	0.099882	0.344127	0.066702
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.021277	0.000000	0.000000
75%	0.109091	0.000000	0.000000
max	0.827586	1.000000	1.000000

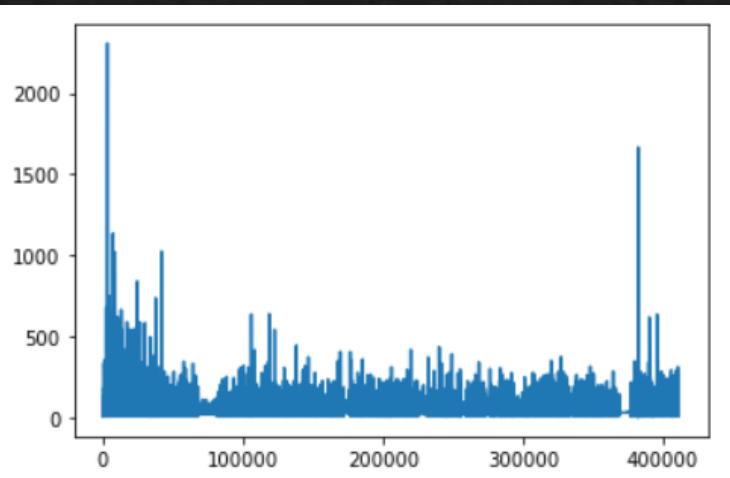
	whois_rd	whois_ed	whois_ud
count	411263.000000	411263.000000	411263.000000
mean	4911.572685	504.615064	288.278061
std	3472.109566	760.493166	425.042535
min	-41.000000	-4549.000000	-101.000000
25%	80.000000	7.000000	-1.000000
50%	5646.000000	221.000000	150.000000
75%	7910.000000	584.000000	337.000000
max	12801.000000	31969.000000	6093.000000

## ❖ Atributi

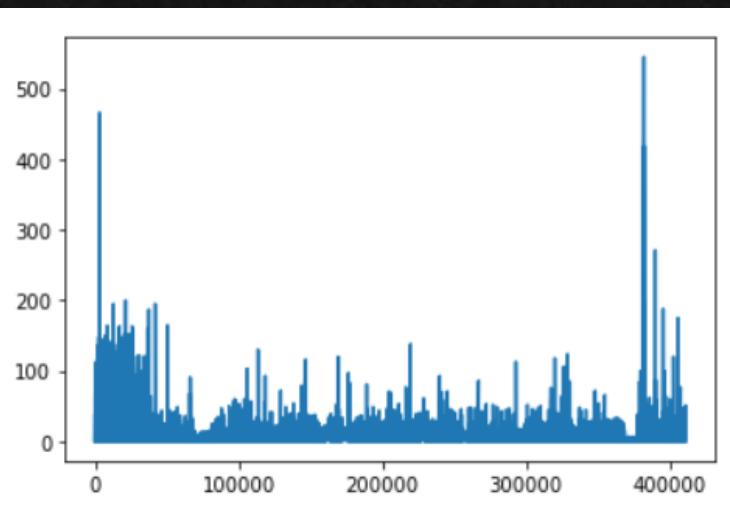


## ❖ Grafički prikaz atributa

### ❖ Dužina URL-ova

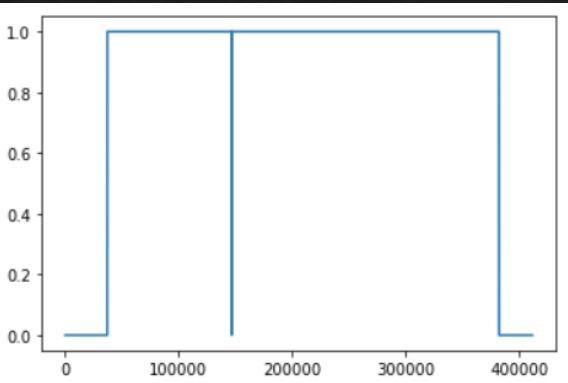


### ❖ Broj cifara u URL-u



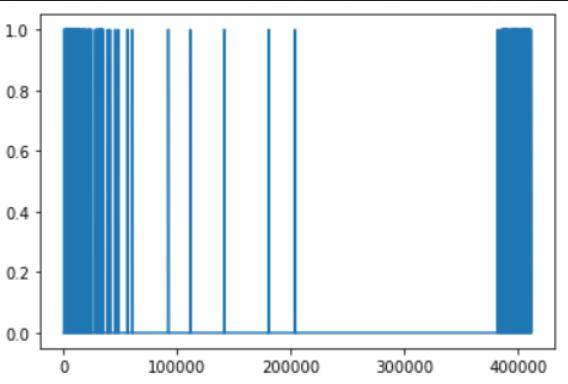
## ❖ Pripadnost klasi i TLD-ovi

❖ Klasa



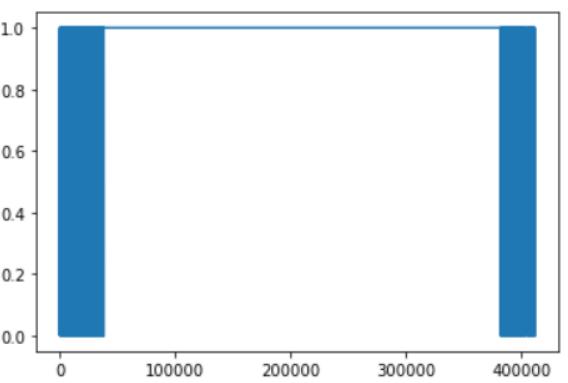
❖ Sumnjiv

TLD

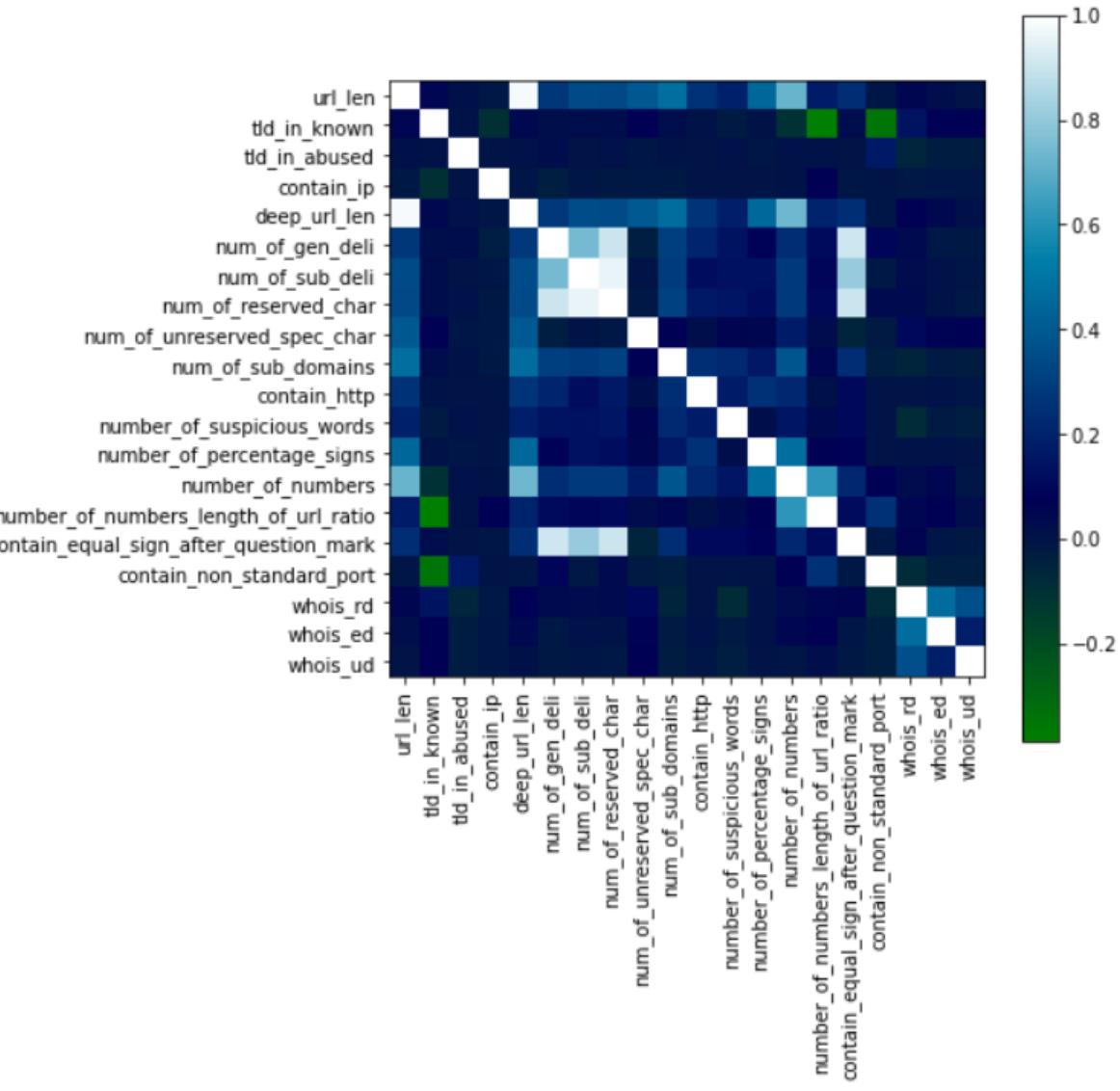


❖ Poznat

TLD



## ❖ Korelacija



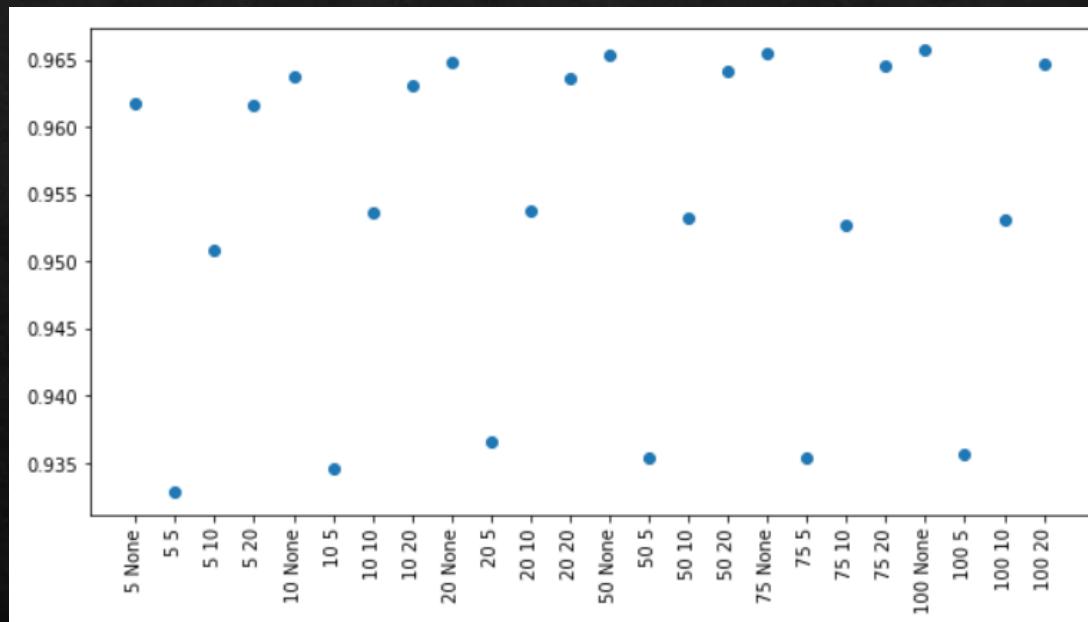
Prema dobijenim podacima, vidimo da su dužina URL-a i dužina, kao i broj „sub“ graničnika i broj rezervisanih karaktera visoko korelisani, redom 0.98 i 0.96, zbog čega izbacujemo atribute dužina dubokog URL-a i broj rezervisanih karaktera.

# Podela skupova i odabir meta podataka za modele

- ❖ Podela podataka na trening, validacione i test skupove
  - ❖ Veličina test skupa 30%
  - ❖ Veličina validacionog skupa: 20% trening skupa
- ❖ Skaliranje podataka
- ❖ Validacija modela
  - ❖ Za nasumične šume 2 meta podatka
    - ❖ Broj esimitora (5, 10, 20, 50, 75, 100)
    - ❖ Maksimalna dubina stabla (Neograničena, 5, 10, 20)
  - ❖ Za k najbližih suseda 1 meta podatak
    - ❖ Broj suseda (1, 2, 3, 5, 7, 10, 20)

## ❖ Rezultati validacije

Nasumična šuma



Najbolje rezultate dobijamo za 75 stabala i  
bez ograničavanja dubine

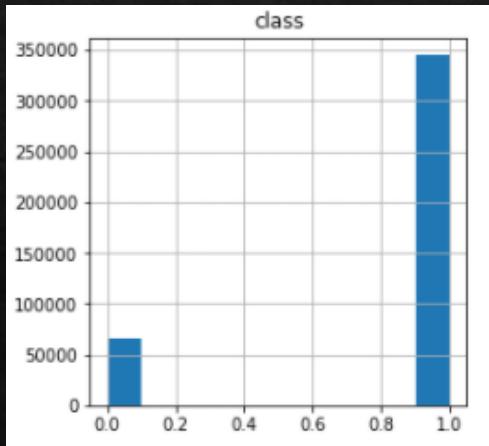
K najbližih suseda

k	f1 score
1	0.9492091986752916
2	0.9378218577150486
3	0.9492039423805912
5	0.9504145700490965
7	0.9497377089178968
10	0.9478814381903078
20	0.9468790828159656

Najbolji rezultat daje model sa 5 suseda

# Nebalansirani skupovi

- ❖ Odnos



- ❖ RUS

- ❖ Prvo smo izdvojili podatke za smanjeno uzorkovanje, odnos 1:2

- ❖ SMOTE

- ❖ Drugi način uzorkovanja je bio uvećano uzorkovanje, pri jednakom broju instanci iz obe klase

# Klasifikacija

## ❖ Modeli

- ❖ Stablo odlučivanja sa Gini greškom
- ❖ Stablo odlučivanja sa greškom entropije
- ❖ Nasumična šuma (75 stabala, bez ograničenja dubine)
- ❖ K najbližih suseda (5 suseda)

## ❖ Rezultati klasifikacije pomoću drveta odlučivanja

-----Drvo odlucivanja: NormalGiniNone -----

Rezultat trening skupa: 0.982

Rezultat test skupa: 0.925

Matrica kofuzije trening skupa:

[[ 35073 2147]

[ 1958 191129]]

Matrica kofuzije test skupa:

[[15362 4577]

[ 4628 98812]]

Metrika:

	precision	recall	f1-score	support
0	0.77	0.77	0.77	19939
1	0.96	0.96	0.96	103440
accuracy			0.93	123379
macro avg	0.86	0.86	0.86	123379
weighted avg	0.93	0.93	0.93	123379

Vreme izvrsavanja: 2.386

-----Drvo odlucivanja: RatioGiniNone -----

Rezultat trening skupa: 0.978

Rezultat test skupa: 0.904

Matrica kofuzije trening skupa:

[[36095 1125]

[ 1363 73077]]

Matrica kofuzije test skupa:

[[16779 3160]

[ 8732 94708]]

Metrika:

	precision	recall	f1-score	support
0	0.66	0.84	0.74	19939
1	0.97	0.92	0.94	103440
accuracy			0.90	123379
macro avg	0.81	0.88	0.84	123379
weighted avg	0.92	0.90	0.91	123379

Vreme izvrsavanja: 1.033

-----Drvo odlucivanja: SmoteGiniNone -----

Rezultat trening skupa: 0.980

Rezultat test skupa: 0.913

Matrica kofuzije trening skupa:

[[188958 4129]

[ 3460 189627]]

Matrica kofuzije test skupa:

[[15983 3956]

[ 6805 96635]]

Metrika:

	precision	recall	f1-score	support
0	0.70	0.80	0.75	19939
1	0.96	0.93	0.95	103440
accuracy			0.91	123379
macro avg	0.83	0.87	0.85	123379
weighted avg	0.92	0.91	0.92	123379

Vreme izvrsavanja: 4.340

-----Drvo odlucivanja: NormalEntropyNone -----

Rezultat trening skupa: 0.982

Rezultat test skupa: 0.927

Matrica kofuzije trening skupa:

[[ 35027 2106]

[ 2095 191079]]

Matrica kofuzije test skupa:

[[15632 4453]

[ 4531 98763]]

Metrika:

	precision	recall	f1-score	support
0	0.78	0.78	0.78	20085
1	0.96	0.96	0.96	103294
accuracy			0.93	123379
macro avg	0.87	0.87	0.87	123379
weighted avg	0.93	0.93	0.93	123379

Vreme izvrsavanja: 2.071

-----Drvo odlucivanja: RatioEntropyNone -----

Rezultat trening skupa: 0.978

Rezultat test skupa: 0.905

Matrica kofuzije trening skupa:

[[36067 1066]

[ 1424 72842]]

Matrica kofuzije test skupa:

[[16992 3093]

[ 8637 94657]]

Metrika:

	precision	recall	f1-score	support
0	0.66	0.85	0.74	20085
1	0.97	0.92	0.94	103294
accuracy			0.90	123379
macro avg	0.82	0.88	0.84	123379
weighted avg	0.92	0.90	0.91	123379

Vreme izvrsavanja: 1.147

-----Drvo odlucivanja: SmoteEntropyNone -----

Rezultat trening skupa: 0.980

Rezultat test skupa: 0.916

Matrica kofuzije trening skupa:

[[189261 3913]

[ 3682 189492]]

Matrica kofuzije test skupa:

[[16224 3861]

[ 6543 96751]]

Metrika:

	precision	recall	f1-score	support
0	0.71	0.81	0.76	20085
1	0.96	0.94	0.95	103294
accuracy			0.92	123379
macro avg	0.84	0.87	0.85	123379
weighted avg	0.92	0.92	0.92	123379

Vreme izvrsavanja: 4.197

-----Nasumicna suma: Norma75 -----

Rezultat trening skupa: 0.982

Rezultat test skupa: 0.941

Matrica kofuzije trening skupa:

[[ 34667 2553]

[ 1558 191529]]

Matrica kofuzije test skupa:

[[ 15359 4580]

[ 2742 100698]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.85	0.77	0.81	19939
---	------	------	------	-------

1	0.96	0.97	0.96	103440
---	------	------	------	--------

accuracy			0.94	123379
----------	--	--	------	--------

macro avg	0.90	0.87	0.89	123379
-----------	------	------	------	--------

weighted avg	0.94	0.94	0.94	123379
--------------	------	------	------	--------

Vreme izvrsavanja: 34.473

-----Nasumicna suma: Ratio75 -----

Rezultat trening skupa: 0.978

Rezultat test skupa: 0.928

Matrica kofuzije trening skupa:

[[ 35832 1388]

[ 1104 73336]]

Matrica kofuzije test skupa:

[[ 16967 2972]

[ 5945 97495]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.74	0.85	0.79	19939
---	------	------	------	-------

1	0.97	0.94	0.96	103440
---	------	------	------	--------

accuracy			0.93	123379
----------	--	--	------	--------

macro avg	0.86	0.90	0.87	123379
-----------	------	------	------	--------

weighted avg	0.93	0.93	0.93	123379
--------------	------	------	------	--------

Vreme izvrsavanja: 16.195

-----Nasumicna suma: Smote75 -----

Rezultat trening skupa: 0.980

Rezultat test skupa: 0.933

Matrica kofuzije trening skupa:

[[ 188715 4372]

[ 3222 189865]]

Matrica kofuzije test skupa:

[[ 16632 3307]

[ 4960 98480]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.77	0.83	0.80	19939
---	------	------	------	-------

1	0.97	0.95	0.96	103440
---	------	------	------	--------

accuracy			0.93	123379
----------	--	--	------	--------

macro avg	0.87	0.89	0.88	123379
-----------	------	------	------	--------

weighted avg	0.94	0.93	0.93	123379
--------------	------	------	------	--------

Vreme izvrsavanja: 58.589

-----K najblizih suseda: Normal5 -----

Rezultat trening skupa: 0.945

Rezultat test skupa: 0.930

Matrica kofuzije trening skupa:

[[ 29043 8379]

[ 4299 188699]]

Matrica kofuzije test skupa:

[[ 14468 5524]

[ 3140 100308]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.82	0.72	0.77	19992
---	------	------	------	-------

1	0.95	0.97	0.96	103448
---	------	------	------	--------

accuracy			0.93	123440
----------	--	--	------	--------

macro avg	0.88	0.85	0.86	123440
-----------	------	------	------	--------

weighted avg	0.93	0.93	0.93	123440
--------------	------	------	------	--------

Vreme izvrsavanja: 2226.056

-----K najblizih suseda: Ratio5 -----

Rezultat trening skupa: 0.899

Rezultat test skupa: 0.881

Matrica kofuzije trening skupa:

[[ 33383 4039]

[ 7311 67533]]

Matrica kofuzije test skupa:

[[ 17059 2933]

[ 11809 91639]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.59	0.85	0.70	19992
---	------	------	------	-------

1	0.97	0.89	0.93	103448
---	------	------	------	--------

accuracy			0.88	123440
----------	--	--	------	--------

macro avg	0.78	0.87	0.81	123440
-----------	------	------	------	--------

weighted avg	0.91	0.88	0.89	123440
--------------	------	------	------	--------

Vreme izvrsavanja: 680.158

-----K najblizih suseda: Smote5 -----

Rezultat trening skupa: 0.943

Rezultat test skupa: 0.908

Matrica kofuzije trening skupa:

[[ 182391 10607]

[ 11437 181561]]

Matrica kofuzije test skupa:

[[ 16767 3225]

[ 8138 95310]]

Metrika:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.67	0.84	0.75	19992
---	------	------	------	-------

1	0.97	0.92	0.94	103448
---	------	------	------	--------

accuracy			0.91	123440
----------	--	--	------	--------

macro avg	0.82	0.88	0.85	123440
-----------	------	------	------	--------

weighted avg	0.92	0.91	0.91	123440
--------------	------	------	------	--------

Vreme izvrsavanja: 4383.985

## ❖ Upoređivanje modela

Modeli	Rezultati modeliranja
Stablo odlučivanja, Gini, početni podaci	0.982, 0.925
Stablo odlučivanja, Gini, RUS podaci	0.978, 0.904
Stablo odlučivanja, Gini, SMOTE podaci	0.980, 0.913
Stablo odlučivanja, Entropy, početni podaci	0.982, 0.927
Stablo odlučivanja, Entropy, RUS podaci	0.978, 0.905
Stablo odlučivanja, Entropy, SMOTE podaci	0.980, 0.916
Nasumična šuma, 75 drveta, bez ograničenja, početni podaci	0.982, 0.941
Nasumična šuma, 75 drveta, bez ograničenja, RUS podaci	0.978, 0.928
Nasumična šuma, 75 drveta, bez ograničenja, SMOTE podaci	0.980, 0.933
K najbližih suseda, sa 5 suseda, početni podaci	0.945, 0.930
K najbližih suseda, sa 5 suseda, RUS podaci	0.899, 0.881
K najbližih suseda, sa 5 suseda, SMOTE podaci	0.943, 0.908

# Reference:

- ❖ Di Ai, Autumn 2018, Classifying Spam using URLs
- ❖ DOYEN SAHOO, CHENGHAO LIU and STEVEN C.H. HOI, Aug 2019, Malicious URL Detection using Machine Learning: A Survey
- ❖ L. Manister, Jan 2005, Uniform Resource Identifier (URI): Generic Syntax (RFC 3986)

HVALA NA PAŽNJI!