

Seminarski rad u okviru kursa

Istraživanje podataka 1

Matematički fakultet

Klasifikacija skupa podataka **Online News Popularity**

Lea Petković 163/2016

mi16163@alas.matf.bg.ac.rs

15. avgust 2019. godine

1. Opšte o podacima

Podaci koji su korišćeni u ovom seminarskom radu predstavljaju skup podataka o člancima koje je *Mashable*¹ objavio u periodu od dve godine. Skup podataka moguće je besplatno preuzeti na veb-sajtu [UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity](http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). Cilj je, metodom klasifikacije, predvideti broj objava nekog članka na društvenim mrežama, tj. predvideti popularnost datog članka.

Podaci su smešteni u tabelu koja sadrži 39797 instanci i opisani pomoću 61 atributa. Sledi lista i opis svih atributa:

url: URL članka	weekday_is_monday: da li je članak bio objavljen u ponedeljak?
timedelta: dani protekli između objavljivanja članka i akvizicije podataka	weekday_is_tuesday: da li je članak bio objavljen u utorak?
n_tokens_title: broj reči u naslovu	weekday_is_wednesday: da li je članak bio objavljen u sredu?
n_tokens_content: broj reči u sadržaju članka	weekday_is_thursday: da li je članak bio objavljen u četvrtak?
n_unique_tokens: stopa jedinstvenih reči u sadržaju	weekday_is_friday: da li je članak bio objavljen u petak?
n_non_stop_words: stopa non-stop reči u sadržaju	weekday_is_saturday: da li je članak bio objavljen u subotu?
n_non_stop_unique_tokens: stopa jedinstvenih non-stop reči u sadržaju	weekday_is_sunday: da li je članak bio objavljen u nedelju?
num_hrefs: broj veza	is_weekend: da li je članak objavljen za vikend?
num_self_hrefs: broj veza ka drugim člancima koje je objavio <i>Mashable</i>	LDA_00: blizina LDA temi 0
num_imgs: broj slika	LDA_01: blizina LDA temi 1
num_videos: broj video snimaka	LDA_02: blizina LDA temi 2
average_token_length: prosečna dužina reči u sadržaju	LDA_03: blizina LDA temi 3
num_keywords: broj ključnih reči u metapodacima	LDA_04: blizina LDA temi 4
data_channel_is_lifestyle: da li je tema kanala podataka stil života?	global_subjectivity: subjektivnost teksta
data_channel_is_entertainment: da li je tema kanala podataka zabava?	global_sentiment_polarity: polaritet osećanja u tekstu
data_channel_is_bus: da li je tema kanala podataka buznis?	global_rate_positive_words: stopa pozitivnih reči u sadržaju
data_channel_is_socmed: da li su tema kanala podataka društveni mediji?	global_rate_negative_words: stopa negativnih reči u sadržaju

¹ Mashable je digitalni medijski veb-sajt. Osnovao ga je Pit Kešmor 2005. godine (engl. Pete Cashmore)

data_channel_is_tech: da li je tema kanala podataka tehnologija?	rate_positive_words: stopa pozitivnih reči među neneutralnim tokenima
data_channel_is_world: da li je tema kanala podataka svet?	rate_negative_words: stopa negativnih reči među neneutralnim tokenima
kw_min_min: najgora ključna reč (min. objava)	avg_positive_polarity: prosečni polaritet pozitivnih reči
kw_max_min: najgora ključna reč (max. objava)	min_positive_polarity: minimalni polaritet pozitivnih reči
kw_avg_min: najgora ključna reč (prosečno objava)	max_positive_polarity: maksimalni polaritet pozitivnih reči
kw_min_max: najbolja ključna reč (min. objava)	avg_negative_polarity: prosečni polaritet negativnih reči
kw_max_max: najbolja ključna reč (max. objava)	min_negative_polarity: minimalni polaritet pozitivnih reči
kw_avg_max: najbolja ključna reč (prosečno objava)	max_negative_polarity: maksimalni polaritet pozitivnih reči
kw_min_avg: prosečna ključna reč (min. objava)	title_subjectivity: subjektivnost naslova
kw_max_avg: prosečna ključna reč (max. objava)	title_sentiment_polarity: polaritet naslova
kw_avg_avg: prosečna ključna reč (prosečno objava)	abs_title_subjectivity: apsolutni nivo subjektivnosti
self_reference_min_shares: minimalni br. objava referenciranih članaka na <i>Mashable-u</i>	abs_title_sentiment_polarity: apsolutni nivo polariteta
self_reference_max_shares: maksimalni br. objava referenciranih članaka na <i>Mashable-u</i>	shares: broj objava (ciljni atribut)
self_reference_avg_shares: prosečan br. objava referenciranih članaka na <i>Mashable-u</i>	



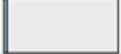
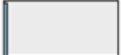

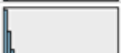



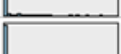
2. O klasifikaciji

Problem klasifikacije predstavlja problem učenja strukture skupa podataka koji je već podeljen u grupe koje nazivamo kategorijama ili klasama. Učenje ovih kategorija se obično postiže pomoću modela. Model se koristi za procenu oznake klase nekog podatka čije oznake nisu poznate. Dakle, jedan od ulaznih parametara problema klasifikacije jeste skup podataka koji je već podeljen u klase – trening skup. Nepoznati podaci, odnosno podaci koje je potrebno klasifikovati čine test skup. Klasifikacija pripada nadgledanom učenju zato što na osnovu trening skupa uči strukturu grupe.

Neki od algoritama klasifikacije, koji su takođe primenjeni nad skupom podataka Online News Popularity, su: C5.0, CR&T (CART), KNN, SVM ...

3. Analiza i pretprocesiranje podataka

Analiza i pretprocesiranje skupa vršeni su u IBM SPSS modeleru, učitavanjem čvora *Var*, a potom i drugih odgovarajućih čvorova. Analizirani su tipovi atributa, kao i njihove statistike:

Field	Sample Graph	Measurement	Min	Max	Sum	Mean	Std. Dev	Variance	Valid
n_tokens_title		Continuous	2.000	23.000	412248.000	10.399	2.114	4.469	39644
n_tokens_content		Continuous	0.000	8474.000	21666030.000	546.515	471.108	221942.284	39644
n_unique_tokens		Continuous	0.000	701.000	21733.464	0.548	3.521	12.395	39644
n_non_stop_words		Continuous	0.000	1042.000	39504.000	0.996	5.231	27.366	39644
n_non_stop_unique_tokens		Continuous	0.000	650.000	27321.669	0.689	3.265	10.659	39644
num_hrefs		Continuous	0.000	304.000	431473.000	10.884	11.332	128.415	39644
num_self_hrefs		Continuous	0.000	116.000	130573.000	3.294	3.855	14.862	39644
num_imgs		Continuous	0.000	128.000	180148.000	4.544	8.309	69.047	39644
num_videos		Continuous	0.000	91.000	49550.000	1.250	4.108	16.874	39644
shares		Continuous	1	843300	134606452	3395.380	11626.951	135185983.712	39644

Slika 3. 1. Tipovi i statistike nekih atributa

Imajući u vidu broj atributa, njihove opise, kao i to da, kao što je već napomenuto, želimo da klasifikujemo podatke prema popularnosti članka, bilo je jasno da nam nisu potrebni svi atributi. Takođe, tipovi (skoro) svih atributa su numerički, uključujući i ciljni atribut. Iz te činjenice uočeno je da je diskretizacija ciljnog atributa bila neophodna kako bi klasifikacija postala moguća. Postavilo se pitanje: *na koji način je najbolje diskretizovati potrebne podatke?* Prva ideja bila je da se interval podeli na jednake podintervale, ali na osnovu analize vrednosti ciljnog atributa *shares*, izveden je zaključak da je to najbolje učiniti podelom na podintervale tako da svaka klasa sadrži jednak broj instanci. U suprotnom bi skoro svi podaci pripadali malom broju od ukupnog broja klasa. Ovo je učinjeno primenom čvora *Binning*, a novonastali (diskretizovani) atribut, korišćenjem čvorova *Type* i *Filter*, postavljen je kao novi cilj.

Daljom analizom je uočeno da nema pojavljivanja nedostajućih vrednosti, te samim tim da njihova obrada nije neophodna. Takođe, nekoliko atributa je promenjen tip: iz numeričkog (na intervalu $[0.0, 1.0]$) u binarni atribut.

Elementi van granice i ekstremne vrednosti su zadržani u skupu podataka, zato što su skoro svi modeli primenjeni nad podacima bez njih loše klasifikovali podatke.

4. Primena algoritama

Pre primene klasifikacije nad podacima, skup podataka podeljen je na dva skupa: a) trening skup; b) test skup. Trening skup iznosi 70% svih instanci, dok test skup čini preostalih 30% svih instanci.

4.1. C5.0 algoritam

Prvi primenjeni algoritam za klasifikaciju bio je C5.0. Napravljena su dva modela primenom ovog algoritma.

U prvom slučaju, broj atributa u podacima je redukovano upotrebom rotacije osa (PCA). PCA algoritmom je početnih 60 atributa svedeno na 25 nezavisnih atributa koji imaju najveći uticaj na ciljni atribut. Kasnije, upotrebom čvora *Type* na čvor PCA algoritma, napomenuto je da će se u daljoj klasifikaciji koristiti samo izdvojeni atributi. U drugom slučaju, model je koristio svih 60 atributa. Takođe, radi što boljeg klasifikovanja podataka korišćena je opcija *use boosting*.

Na dobijene modele je primenjen čvor *Analysis*, čime su dobijene matrica konfuzije i statistike rada klasifikatora nad (trening i test) podacima.

Results for output field shares_diskretizovano10

Comparing \$C- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing	Training
Correct	6,867 57.77%	16,020 57.71%
Wrong	5,019 42.23%	11,738 42.29%
Total	11,886	27,758

Coincidence Matrix for \$C- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing		1	10	2	3	4	5	6	7	8	9
1		3,575	0	514	184	78	54	22	12	8	1
10		6	1	2	1	4	3	2	1	2	0
2		921	2	1,574	220	111	53	23	6	8	1
3		472	0	330	767	101	52	25	7	5	0
4		289	0	176	126	460	25	15	4	5	3
5		175	1	120	82	57	254	8	5	5	2
6		90	0	73	45	36	24	132	3	1	1
7		77	0	65	38	28	9	15	54	3	2
8		35	0	30	25	18	12	5	3	42	1
9		18	0	11	5	5	2	6	1	3	8

'Partition' = Training		1	10	2	3	4	5	6	7	8	9
1		8,346	1	1,085	387	205	134	71	26	20	9
10		14	4	10	8	7	3	7	0	0	0
2		2,188	0	3,565	525	267	131	60	31	22	4
3		1,117	1	778	1,866	226	96	46	21	12	4
4		651	1	436	295	1,071	86	45	13	12	5
5		374	0	285	185	140	614	34	12	8	1
6		226	0	177	127	88	50	320	18	6	2
7		178	0	111	63	62	46	33	127	6	1
8		86	0	61	43	32	22	11	15	86	1
9		62	0	42	14	25	6	13	6	6	21

Slika 4.1. Statistike modela C5.0 algoritma nad svim atributima

Results for output field shares_diskretizovano10

Comparing \$C- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing	Training
Correct	8,588 72.75%	20,259 72.77%
Wrong	3,216 27.25%	7,581 27.23%
Total	11,804	27,840

Coincidence Matrix for \$C- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing		1	10	2	3	4	5	6	7	8	9
1		4,065	0	270	84	31	13	4	0	1	0
10		8	7	2	2	2	1	0	0	0	0
2		780	0	1,950	87	43	22	8	4	0	0
3		413	0	183	1,057	39	18	8	3	0	0
4		273	0	113	46	640	6	2	5	0	0
5		154	0	68	32	16	420	3	2	1	0
6		120	0	50	25	17	4	226	0	0	0
7		68	0	29	8	10	2	3	146	0	0
8		41	0	24	12	6	3	0	1	62	0
9		27	0	7	5	4	1	1	0	1	15

'Partition' = Training		1	10	2	3	4	5	6	7	8	9
1		9,408	0	565	175	64	29	15	6	1	1
10		18	10	9	9	4	1	0	1	1	0
2		1,851	0	4,597	201	104	42	13	6	2	2
3		989	0	441	2,644	70	42	12	7	0	0
4		617	1	267	134	1,576	29	6	2	1	0
5		371	0	164	89	41	987	8	2	3	1
6		255	0	116	51	23	15	516	0	1	0
7		204	0	75	38	21	5	3	306	0	0
8		113	0	55	23	17	10	4	0	157	0
9		72	0	25	24	6	4	1	3	0	58

Slika 4.2. Statistike modela C5.0 algoritma nad atributima koji su rezultat rada PCA algoritma

U slučaju oba klasifikatora primećujemo da između klasifikacije trening i test skupa skoro da ne postoji razlika, što je dobro jer ne dolazi do problema prilagođavanja trening podacima. Klasifikator nad redukovanim atributima se pokazao kao bolji od klasifikatora nad svim atributima.

4.2. CART algoritam

Naredni algoritam korišćen za klasifikaciju podataka je CR&T (CART) algoritam. Kao i prethodni (C5.0), algoritam je primenjen nad svim atributima i nad redukovanim skupom atributa. Dubina stabla koja je korišćena je 15. Takođe, prevencija od prilagođavanja je smanjena i postavljena na 15%, umesto podrazumevanih 30%.

Results for output field shares_diskretizovano10

Comparing SR- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing		Training	
Correct	4,819	40.59%	11,147	40.14%
Wrong	7,054	59.41%	16,624	59.86%
Total	11,873		27,771	

Coincidence Matrix for SR- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing	1	2	3	4
1	3,735	642	42	2
10	26	2	2	0
2	1,888	967	70	7
3	1,014	662	104	10
4	622	409	59	13
5	370	263	42	9
6	252	141	26	0
7	179	81	11	1
8	102	32	10	0
9	60	15	3	0

'Partition' = Training	1	2	3	4
1	8,728	1,492	88	3
10	39	5	1	0
2	4,416	2,185	171	8
3	2,364	1,556	204	12
4	1,459	961	165	30
5	970	571	125	12
6	621	334	40	5
7	441	181	20	4
8	259	114	10	1
9	125	46	4	1

Slika 4.3. Statistike modela CART algoritma nad svim atributima

Results for output field shares_diskretizovano10

Comparing SR- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing		Training	
Correct	4,767	39.59%	11,053	40.04%
Wrong	7,274	60.41%	16,550	59.96%
Total	12,041		27,603	

Coincidence Matrix for SR- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing	1	2	3	4
1	3,967	385	55	11
10	15	4	0	0
2	2,180	653	86	26
3	1,229	420	119	31
4	760	290	67	28
5	516	179	39	15
6	289	112	23	10
7	218	51	8	3
8	124	43	9	4
9	63	7	2	0

'Partition' = Training	1	2	3	4
1	9,189	983	126	16
10	44	6	6	0
2	4,990	1,535	197	45
3	2,782	1,029	272	44
4	1,658	683	175	57
5	1,048	428	105	32
6	699	220	51	15
7	462	140	24	12
8	253	76	13	6
9	137	36	3	6

Slika 4.4. Statistike modela CART algoritma nad atributima koji su rezultat rada PCA algoritma

Na osnovu dobijenih statistika uočeno je da je razlika između dobijenih modela zanemarljiva, kao i to da, imajući u vidu klasifikatore dobijene algoritmom C5.0, CART klasifikatori slabije klasifikuju podatke.

4.3. Algoritam k najbližih suseda (KNN)

Još jedan primenjeni algoritam je k najbližih suseda. Za broj razmatranih suseda, odnosno k, uzete su vrednosti od 3 do 5. Kao mera rastojanja iskorišćeno je Euklidsko rastojanje. Algoritam je primenjen nad svim atributima i nad atributima dobijenim primenom PCA algoritma. Rezultat rada dat je na narednim slikama:

Results for output field shares_diskretizovano10

Comparing \$KNN- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing		Training	
Correct	5,622	46.77%	13,077	47.34%
Wrong	6,399	53.23%	14,546	52.66%
Total	12,021		27,623	

Coincidence Matrix for \$KNN- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing

	1	10	2	3	4	5	6	7	8	9
1	3,660	0	511	139	59	27	7	5	3	1
10	14	0	3	2	2	1	1	0	0	0
2	1,443	0	1,232	166	101	28	13	4	2	0
3	781	1	432	413	83	26	7	4	1	0
4	496	0	262	142	195	15	11	3	1	1
5	348	0	174	82	54	85	9	4	0	0
6	218	0	105	48	24	21	16	3	3	1
7	162	0	64	34	8	4	1	13	1	0
8	89	0	20	22	12	1	4	0	6	0
9	44	0	23	11	6	2	0	0	1	2

'Partition' = Training

	1	2	3	4	5	6	7	8	9
1	8,550	1,155	365	131	73	28	13	4	1
10	34	7	5	3	2	0	1	0	0
2	3,231	2,746	433	191	84	18	17	2	1
3	1,847	1,012	1,036	172	73	24	10	3	1
4	1,138	580	310	459	66	23	7	8	1
5	738	363	181	107	188	20	4	4	1
6	485	225	114	62	31	57	5	3	0
7	327	140	62	42	19	8	33	0	0
8	189	97	43	17	10	8	5	5	0
9	84	39	17	12	6	1	3	0	3

Slika 4.5. Statistike modela KNN algoritma nad svim atributima

Results for output field shares_diskretizovano10

Comparing \$KNN- shares_diskretizovano10 with shares_diskretizovano10

'Partition'	Testing		Training	
Correct	5,534	46.19%	13,008	47.02%
Wrong	6,448	53.81%	14,654	52.98%
Total	11,982		27,662	

Coincidence Matrix for \$KNN- shares_diskretizovano10 (rows show actuals)

'Partition' = Testing

	1	10	2	3	4	5	6	7	8	9
1	3,617	0	559	147	58	29	14	7	3	0
10	10	0	4	3	0	0	1	0	0	0
2	1,451	0	1,196	171	72	29	11	4	2	2
3	829	0	418	419	74	26	6	3	3	0
4	524	0	271	128	159	26	9	2	2	0
5	321	0	164	85	50	81	9	1	0	0
6	222	0	103	57	25	15	31	1	1	0
7	151	0	53	23	17	10	5	19	0	0
8	80	0	42	21	7	4	1	1	12	1
9	50	1	19	3	3	3	1	0	0	0

'Partition' = Training

	1	10	2	3	4	5	6	7	8	9
1	8,433	0	1,251	335	155	67	31	22	4	0
10	33	0	12	6	1	2	2	0	1	0
2	3,253	0	2,852	384	167	63	27	14	10	4
3	1,859	1	1,024	1,017	157	50	21	13	4	2
4	1,156	0	637	310	412	44	21	11	4	2
5	772	0	396	194	90	176	15	4	3	1
6	480	0	230	111	52	24	59	4	3	1
7	348	0	152	67	29	5	4	33	1	1
8	183	0	79	42	27	6	3	1	18	0
9	94	0	47	13	8	2	2	0	0	0

Slika 4.6. Statistike modela KNN algoritma nad atributima koji su rezultat rada PCA algoritma

Dobijeni modeli daju vrlo slične rezultate te zaključujemo da nije važno koji ćemo koristiti. Podaci su bolje klasifikovani u odnosu na klasifikaciju podataka CART modela, ali još uvek lošiji od C5.0 klasifikatora.

4.4. Metod potpornih vektora (SVM)

Sledeći klasifikator je metod potpornih vektora (SVM). Klasifikator je primenjen nad svim atributima, a potom i nad redukovanim brojem atributa. Korišćen je čvor *SVM*, gde je za opciju *Mode* izabran *Expert* sa podrazumevanim podešavanjima.

Results for output field shares_diskretizovano10

Comparing SS- shares_diskretizovano10 with shares_diskretizovano10

Partition	Testing		Training	
Correct	5,589	47.22%	13,067	46.99%
Wrong	6,246	52.78%	14,742	53.01%
Total	11,835		27,809	

Coincidence Matrix for SS- shares_diskretizovano10 (rows show actuals)

Partition = Testing

	1	10	2	3	4	5	6	7	8	9
1	3,765	1	417	98	25	10	24	20	13	10
10	6	13	1	3	0	1	0	4	0	0
2	1,638	1	945	197	53	23	28	17	20	6
3	856	3	352	397	65	23	16	16	10	14
4	548	2	193	139	182	23	20	18	10	8
5	324	1	113	96	49	77	10	7	7	4
6	197	0	60	44	21	12	70	7	6	2
7	105	1	36	18	13	3	11	54	6	2
8	68	0	17	6	7	0	6	3	56	1
9	29	2	9	4	1	1	3	1	1	30

Partition = Training

	1	10	2	3	4	5	6	7	8	9
1	8,806	13	995	272	87	35	36	46	35	24
10	17	21	3	2	1	1	1	0	0	1
2	3,736	4	2,246	427	150	65	52	43	39	22
3	2,057	4	830	933	164	52	51	36	29	18
4	1,164	2	515	300	425	41	47	42	26	13
5	827	4	270	196	103	178	40	25	22	9
6	493	1	132	101	59	19	151	19	18	7
7	294	2	99	46	26	14	24	145	14	5
8	127	1	48	25	13	5	17	13	105	10
9	61	1	14	8	6	5	12	4	5	57

Slika 4.7. Statistike modela SVM algoritma nad svim atributima

Results for output field shares_diskretizovano10

Comparing SS- shares_diskretizovano10 with shares_diskretizovano10

Partition	Testing		Training	
Correct	4,784	40.17%	10,938	39.44%
Wrong	7,126	59.83%	16,796	60.56%
Total	11,910		27,734	

Coincidence Matrix for SS- shares_diskretizovano10 (rows show actuals)

Partition = Testing

	1	10	2	3	4	5	6	7	8	9
1	4,039	0	345	28	2	3	0	0	1	0
10	16	1	1	1	0	0	0	0	0	0
2	2,267	0	595	75	1	1	1	0	2	0
3	1,317	0	328	116	1	3	2	3	0	0
4	790	0	237	71	9	2	0	0	0	0
5	514	0	146	55	0	12	0	0	0	0
6	317	0	68	27	1	1	3	0	1	0
7	195	1	47	14	0	1	0	3	0	0
8	124	0	31	4	0	2	0	0	5	0
9	62	0	14	2	1	0	0	0	0	1

Partition = Training

	1	10	2	3	4	5	6	7	8	9
1	9,416	0	822	58	6	8	1	2	1	0
10	52	0	2	0	0	0	0	0	2	0
2	5,359	0	1,254	139	6	4	1	1	5	1
3	3,101	0	827	210	2	10	1	2	3	0
4	1,858	0	562	152	22	9	3	0	3	0
5	1,208	1	323	82	2	16	0	1	2	0
6	755	0	177	50	3	7	7	0	2	0
7	518	0	99	23	2	2	2	7	3	1
8	277	1	62	13	1	1	0	1	6	0
9	146	0	20	5	0	3	0	0	0	0

Slika 4.8. Statistike modela metoda potpornih vektora nad atributima koji su rezultat rada PCA algoritma

Iz dobijenih rezultata zaključujemo da je klasifikator nad svim atributima dao bolje rezultate od klasifikatora nad atributima koji su rezultat PCA algoritma. Prvi klasifikator daje slične rezultate kao klasifikator dobijen primenom algoritma k najbližih suseda, a drugi daje rezultate nalik na klasifikator dobijen CART algoritmom.

4.5. Algoritam slučajne šume (engl. Random Decision Forest)

Algoritam slučajne šume zasnovan je na stablima odlučivanja i jedan je od poznatijih ansambl metoda u oblasti mašinskog učenja. Ansambl predstavlja skup modela koji zajedno čine jedan model.

Algoritam je primenjen u programskom jeziku pajton (engl. *Python*) nad skupom podataka koji je preprocesiran u IBM SPSS modeleru. Korišćeno je 3 stabla odlučivanja (parametar *n_estimators* je postavljen na 3) koja kasnije pri klasifikaciji donose odluku glasanjem. Na narednoj slici nalaze se statistike klasifikacije:

Matrica konfuzije:										
[3031	929	123	263	25	5	39	3	2	0]
[1671	769	176	230	23	9	35	1	0	0]
[560	320	83	102	11	5	29	2	3	0]
[902	497	119	210	12	3	33	2	0	0]
[212	127	31	41	6	3	4	1	0	1]
[140	69	23	29	3	2	8	1	0	0]
[351	182	72	73	10	5	14	2	0	0]
[82	41	10	20	0	1	3	1	0	0]
[43	19	4	2	5	2	1	0	0	0]
[11	6	1	4	0	0	1	0	0	0]]

Izveštaj klasifikacije				
	precision	recall	f1-score	support
0	0.43	0.69	0.53	4420
1	0.26	0.26	0.26	2914
2	0.13	0.07	0.09	1115
3	0.22	0.12	0.15	1778
4	0.06	0.01	0.02	426
5	0.06	0.01	0.01	275
6	0.08	0.02	0.03	709
7	0.08	0.01	0.01	158
8	0.00	0.00	0.00	76
9	0.00	0.00	0.00	23
accuracy			0.35	11894
macro avg	0.13	0.12	0.11	11894
weighted avg	0.28	0.35	0.30	11894

Slika 4.9. Matrica konfuzije i izveštaj klasifikacije *Random Decision Forest* klasifikatora

Na osnovu dobijenog izveštaja, preciznosti klasifikacije trening skupa od približno 0.856 i preciznosti klasifikacije test skupa od približno 0.346, zaključujemo da je došlo do problema prilagođavanja trening podacima što ovaj klasifikator nad datim podacima čini lošim.

5. Zaključak

Cilj istraživanja Online News Popularity skupa podataka, odnosno skupa podataka o člancima koje je objavio *Mashable*, bio je odrediti njihovu popularnost. Pošto je nad podacima bilo potrebno izvršiti klasifikaciju, bilo je neophodno diskretizovati ciljni atribut prilikom pretprocesiranja podataka. Primenjeno je pet algoritma, od toga četiri nad skupom podataka sa svim atributima i nad skupom podataka sa redukovanim brojem atributa. Najbolje rezultate dao je klasifikator dobijen algoritmom C5.0 nad redukovanim skupom podataka, sa preciznošću preko 70%. Ostali algoritmi (primenjeni korišćenjem IBM SPSS modelera) klasifikovali su podatke sa preciznošću od 40% do 50%, dok se u slučaju algoritma slučajne šume javlja problem preprilagođenosti trening podacima.

Ono što je zanimljivo jeste da između modela nad redukovanim skupom atributa i modela primenjenih nad svim atributima (sa izuzetkom modela dobijenog sa C5.0) nema značajne razlike u klasifikaciji, a metod potpornih vektora se čak pokazao boljim nad svim atributima skupa podataka. Ovo može biti posledica toga da svi atributi u maloj meri utiču na promenljivost ciljnog atributa.