Klasifikacija komentara(podataka) o ženskoj odeći

Seminarski rad u okviru kursa
Istraživanje podataka

Matematički fakultet
Miloš Miković
2.9.2019

Uvod:

U daljem tekstu prikazana je klasifikacija koja je vršena nad skupom podatak Women's E-Commerce Clothing Reviews preuzetog sa sajta kaggle.
(https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews)

Seminarski je organizovan u 3 datoteke:

1. "klasifikacijaDaLiJeKomentarRecomendedIliNe": u kojoj je vršena klasifikacija korišćenjem algoritama MNB I SVM. Od atributa iz skupa podataka korišćeni su samo komentari a konstruisali smo model koji predviđa da li korisnik preporucuje ili ne dati artikal na osnovu komentara koji je ostavio

2. "klasifikacijaRejtingaArtikla": u kojoj koristimo podatke iz samo skupa podataka, obrađujemo ih, uzimamo neke bitne osobine koje mogu da nam pomognu za klasifikaciju, grupišemo podatke po id-iju artikla I korišćenjem tih podataka klasifikujemo rejting datog artikla na osnovu prethodno obrađenih podataka (rejting u opsegu od 1 do 5).Korišćeni algoritmi SVM I KNN.

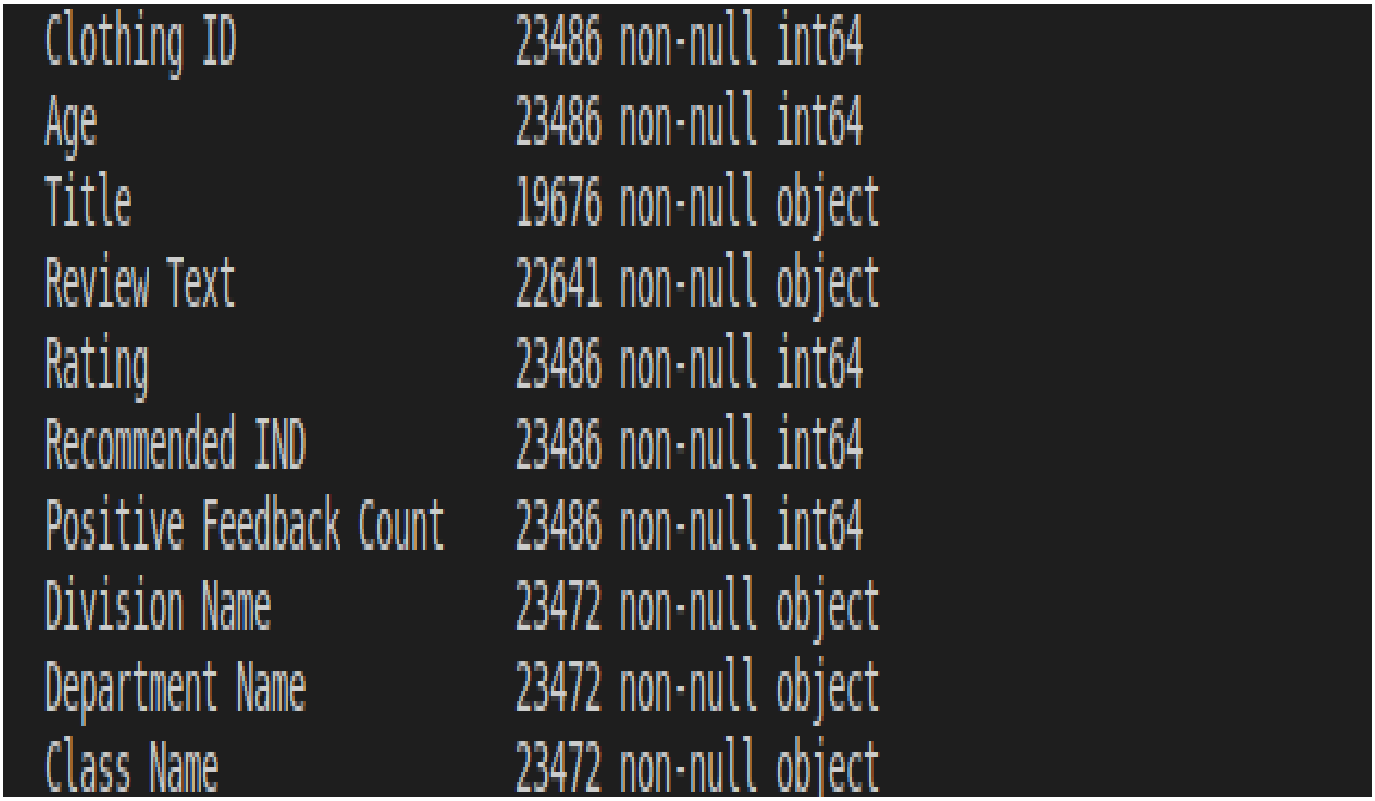
3. "KlasifikacijaRecomendedIliNeObradjeniPodaci": u kojoj koristimo podatke koje smo obradili u prethodnoj datoteci ali sada da predvidimo da li je na osnovu njih artikal preporucljiv ili nije, uzimajući u obzir da arikli sa ocenom većom od 3 jesu preporučljivi(intuitivno odabrano). Korišćeni algoritmi SVM I KNN.
(rejting u opsegu od 1 do 5)

Podaci:

Clothing ID: id artikla, kategorički atribut (integer)

- **Age:** broj godina kupca (integer, veci od 0)
- **Title:** naziv(naslov) komentara (string)
- **Review Text:** tekst komentara (string)
- **Rating:** ocena artikla koju je ostavio kupac (integer od 1(worst) do 5(best))
- **Recommended IND:** označava da li kupac preporučuje ili ne preporučuje artikal (integer 0 ili 1)
- **Positive Feedback Count:** broj koji označava koliko je drugih kupaca reklo da je odgovor ovog korisnika koristan (integer veci ili jednak 0)
- **Division Name:** kategorija u koju produkt spada (kategorički atribut)
- **Department Name:** kategorija u koju produkt spada (kategorički atribut)
- **Class Name:** ime klase u koju produkt spada (kategorički atribut)

Slika pokazuje podatke iz skupa:naziv, broj podataka, null-not null, tip



Atribut id, prvi atribut iz skupa, odmah je izbačen iz skupa jer predstavlja redni broj review-a korisnika a to ni na koji način neće uticati na tok klasifikacije.

Preprocesiranje I obrada podataka za klasifikaciju

(još o samoj vizuelizaciji, raspodelama itd. U samo odbrani projekta)

Najpre ćemo ukloniti null vrednosti iz skupa podataka(u klasifikaciji da li je komentar preporučljiv ili nije, uklonićemo samo one null vrednosti koje se tiču atributa review text, u ostala 2 dokumenta uklonićemo sve null vrednosti)

Slika pokazuje broj null vrednosti pre uklanjanja null vrednosti za atribut Review Text

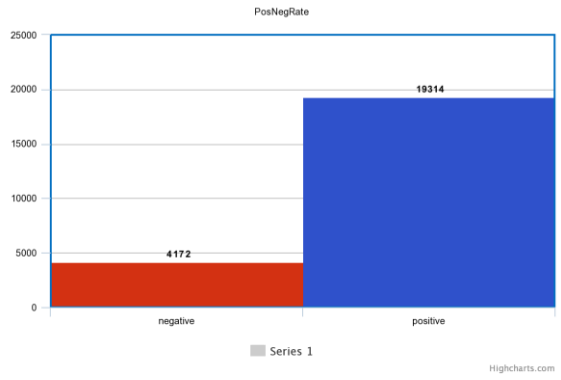
```
BROJ NULL VREDNOSTI:
Clothing ID                    0
Age                            0
Title                       3810
Review Text                  845
Rating                         0
Recommended IND                0
Positive Feedback Count        0
Division Name                 14
Department Name               14
Class Name                    14
```

Slika pokazuje broj null vrednosti posle **uklanjanja null vrednosti za atribut Review Text**

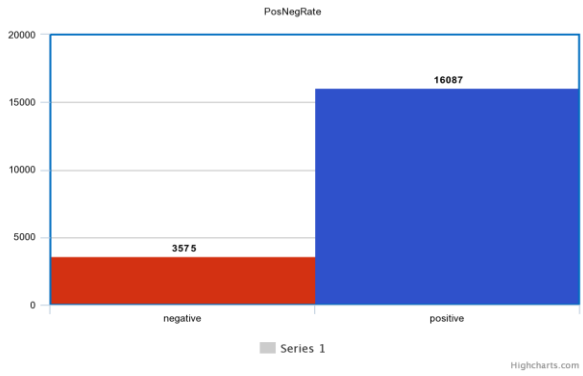
```
BROJ NULL VREDNOSTI POSLE CISCENJA:
Clothing ID                    0
Age                            0
Title                       2966
Review Text                    0
Rating                         0
Recommended IND                0
Positive Feedback Count        0
Division Name                 13
Department Name               13
```

Uporedimo broj pozitivnih I negativnih komentara pre I posle izbacivanja null vrednosti

Pre



Posle



Histogram pokazuje da se odnos nije promenio.Izbalansiranost odnosa pozitivnih I negativnih komentara bila bi poželjnija svakako, I doprinela bi boljem procesu klasifikacije.

Sada treba odabrati podatke od značaja za sam proces klasifikacije, u klasifikaciji komentara uzeti su atributi **Review Text** I **Recommended IND**(prvi dokument)**,** dok su za ostale dve klasifikacije (2. I 3. dokument) korišćeni sledeći atributi: **Clothing ID, Age, Rating, Positive Feedback Count, Review Text** kao I atributi dobijeni obradom navedenih atributa.

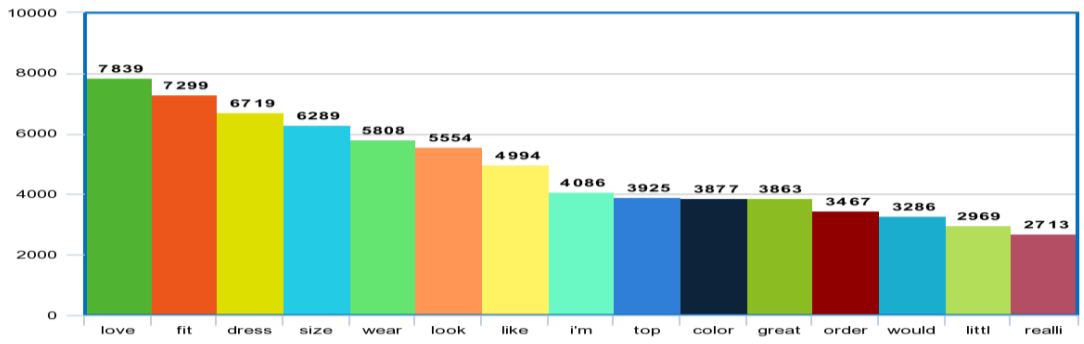
U sve 3 klasifikacije obrada **Review Text-a** tekla je na manje više isti način: najpre tekst čistimo od veznika I reči dužine 1, uklanjamo brojeve I beline, prebacujemo tekst u mala slova, I uzimamo samo korene reči.

Slika preprocesiranog teksta

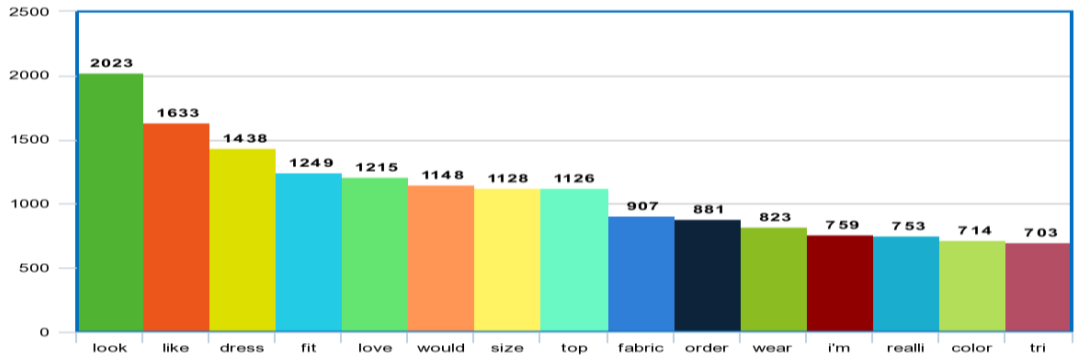
```
i', 'casual', 'wear.', 'nice', 'mute', 'print', 'good', 'qualiti', 'terri', 'material.', 'recommend,', 'size', 'want', 'fit', 'feel.']
['love', 'way', 'pant', 'look', 'pictures,', 'great', 'quality,', 'style', 'realli', 'me.', 'gave', 'try,', "that'", 'counts!']
['saw', 'shirt', 'retail', 'websit', 'new', 'it.', 'arrived,', 'perfectli', 'draped,', 'light', 'airy,', 'incred', 'soft', 'comfortable.', 'perfect', '
throw', 'yoga', 'pant', 'errand', 'dress', 'bit', 'casual', 'night', 'friends.', 'would', 'buy', 'sever', 'differ', 'color', 'offered.', '/.']
['great', 'qualiti', 'extrem', 'flattering.', 'bonu', 'sale.', 'felt', 'like', 'dress', 'godsend.', 'five', 'month', 'pregnant', 'need', 'dress', 'made
', 'feel', 'comfort', 'frumpy.', 'think', "i'll", 'get', 'lot', 'mileag', 'it,', "there'", 'enough', 'stretch', 'get', 'next', 'month', '(hopefully!)',
 'wear', 'too.', 'also', 'dress', 'down,', 'appropri', 'mani', 'occasions.', 'believ', 'run', 'true', 'size', '(i', 'size', 'up,', 'xs', 's,', 'obviou
', 'reasons,']
['yes,', 'great', 'dress!', 'sure', 'onlin', 'color', 'combination.', 'think', 'would', 'prefer', 'gray', 'color', 'sold', 'out.', 'receiv', 'good', 'r
eview', 'onlin', 'thought', 'worth', 'risk', 'sale', 'price.', 'alway', 'hunt', 'great', 'dress', 'great', 'price', '(who', "isn't?!).", 'receiv', 'tri
', 'on,', 'oh', 'wow!', 'love', 'it.', 'flattering.', 'pretti', 'dress.', 'think', 'wear', 'time.', 'actual', 'think']
['cute', 'dress', 'me.', 'waist', 'high', 'sleev', 'tight.', 'mayb', 'differ', 'bodi', 'type', 'dress', 'would', 'perfect.', 'return', 'it.']
['bottom', 'cute', 'defiantli', 'cheeky!', 'would', 'recommend', 'size', 'want', 'coverage.']
["i'm", 'impress', 'beauti', 'color', 'combin', 'embroideri', 'disappoint', 'rayon', 'fabric', 'used,', 'especi', 'price', 'point.', 'sleeveless', 'she
er', 'blue', 'outer', 'dress', 'flowi', 'swing', 'silhouett', 'retain', 'mani', 'wrinkles.', 'thick', 'shoulder', 'upper', 'arms,', 'armhol', 'cut', 'a
```

Sada ćemo prikazati frekvencije reči u pozitivnim I negativnim komentarima (prvih 15):

Pozitivni



Negativni



U 2. I 3. klasifikaciji najpre sve podatke grupišemo po **Clothing ID-u** jer želimo da ispitujemo rejting I preporučljivost svakog artikla ponaosob.Pravimo novi data frame u koji sada postavljamo atribute:

ID- Clothing ID

Avg_Rating- prosečan rejting za artikal sa id-em ID

Num_Reviwes-broj review-a za artikal sa id-em ID

AVG_cnt-prosečan broj reči u komentaru (mislim da može da bude korisna svar vodeći se logikom da će za dobar komentar najčešće biti ispisan duži tekst, mada možda grešim)

Pos_Negative_Proba- kako smo pri prvoj klasifikaciji klasifikovali da li je komentar preporučljiv ili nije sada koristimo taj klasifikator u predikciji datog komentara za trenutni artikal, I uzimamo razliku verovatnoće da komentar bude recommended nasuprot da nije recommended I zbir ovih vrednosti delimo sa brojem komentara za dati artikal (pokazatelj da li skup komentara za dati artikal više naginje ka tome da bude recommended ili not recommended)

avg_feedback- koliko je prosek **Positive Feedback Count** po artiklu

Avg_age-prosek godina kupaca koji su ostavili reviw za artikal sa id-em ID

Na slici vidimo obrađene podatke koje ćemo koristiti u klasifikaciji(mali deo podataka)

	ID	Avg Rating	Num Reviews	AVG cnt	Pos Negative Proba	Avg Feedback	Avg age
0	1077	4.0	251	329.669323	0.767052	2.135458	41.968127
1	1049	4.0	25	359.720000	0.859120	3.120000	44.440000
2	847	4.0	4	358.250000	0.901973	2.000000	42.000000
3	1080	4.0	241	342.116183	0.778027	3.062241	40.792531
4	858	4.0	18	370.444444	0.694176	2.111111	44.500000
5	1095	4.0	287	369.362369	0.742063	3.714286	41.128920
6	767	5.0	1	377.000000	0.781479	0.000000	44.000000
7	1065	4.0	16	352.000000	0.784321	3.000000	40.687500
8	853	4.0	6	348.666667	0.618518	3.166667	39.166667
9	1120	4.0	2	330.500000	0.970794	0.000000	49.000000
10	697	4.0	2	303.000000	0.625352	0.000000	35.000000
11	949	4.0	69	315.000000	0.829165	2.086957	46.782609
12	1003	4.0	21	327.904762	0.849482	1.619048	39.809524
13	684	5.0	1	192.000000	0.570737	2.000000	53.000000
14	4	5.0	1	445.000000	0.920498	0.000000	28.000000
15	1060	4.0	81	342.925926	0.803180	3.530864	43.123457
16	910	4.0	4	241.000000	0.829370	2.750000	52.000000
17	89	4.0	1	499.000000	0.785428	1.000000	67.000000
18	862	4.0	658	293.607903	0.777421	2.583587	43.542553
19	368	2.0	2	383.000000	0.481071	0.500000	44.000000
20	1078	4.0	871	328.477612	0.778484	2.859931	42.719862
21	845	4.0	4	398.750000	0.367699	7.750000	47.250000
22	822	4.0	29	330.620690	0.794416	2.241379	44.448276
23	850	4.0	280	320.457143	0.800940	3.103571	43.764286
24	1082	4.0	119	336.352941	0.750703	3.495798	44.756303
25	836	4.0	172	322.988372	0.763463	3.970930	43.424419
26	1081	4.0	487	336.971253	0.794412	3.203285	42.293634
27	1072	4.0	166	337.367470	0.736666	3.379518	43.879518
28	1133	3.0	6	364.833333	0.855663	2.166667	48.833333
29	861	4.0	203	267.275862	0.794073	2.640394	43.645320

Čak I na ovom malom delu podataka, vidimo da su mahom dobro ocenjeni artikli, ali I da postoje oni sa malim brojem review-a.

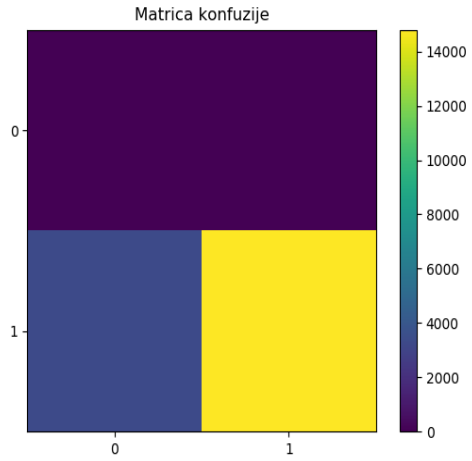
Klasifikacija:

Klasifikacija komentara (recomended ili not recommended)

Podatke **Review Text-a** koji su prošlli kroz fazu preprocesiranja teksta sada koristimo u klasifikaciji.

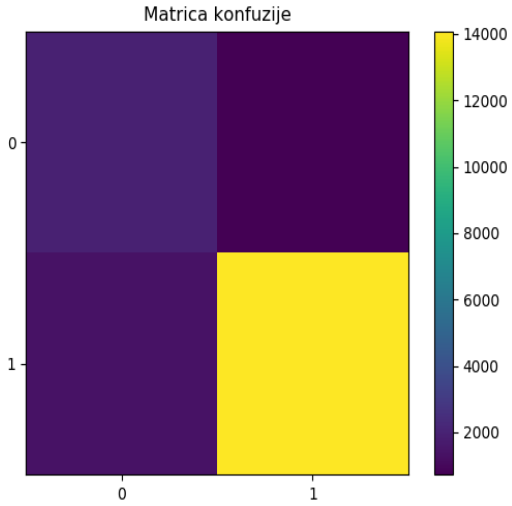
Multinomial Naive Bayes

```
Multinomial Naive Bayes
Accuracy:  0.818
Confusion Matrix:
 [[    4     0]
 [ 3302 14806]]
```



SVM

```
Linear SVC
Accuracy:  0.884
Confusion Matrix:
 [[ 1926   722]
 [ 1380 14084]]
```



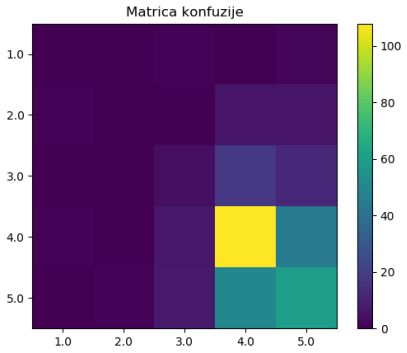
Svm je nesto bolje uradio klasifikaciju od MNB, model pravi greske pri svrstavanju komentara u not recommended, mislim da je to posledica broja not recommended review-a koji je dosta manji. Dalje upravo ova mana skupa će uticati I na ocenu rejtinga u daljem tekstu, jer manji broj not recommended komentara povući će I manji broj loših ocena po artiklu.

Klasifikacija rejtinga artikla:

Kao što smo prikazali na jednoj od prethodnih slika, za klasifikaciju rejtinka koristićemo sledeće atribute **ID, Avg_Rating, Num_Reviwes, AVG_cnt, Pos_Negative_Proba, avg_feedback, Avg_age.**Najpre treba napomenuti da ukupno postoji 1095 jedinstvenih artikala, I da prethodo prikazana tabela sadrži samo prvih 29 artikala.Veliki problem u klasifikaciji rejtinga pravi broj review-a ostavljenih po artiklu.Ako bi uzeli svih 1095 artikala I analizirali broj review-a dobili bi podatak da postoji veliki broj artikala koji imaju manje od 5 review-a, čak 799 artikala što će nam praviti problem u klasifikaciji.Ako bi uzeli sve podatke iz prethodno navedene tabele dobili bi sledeće rezultate:

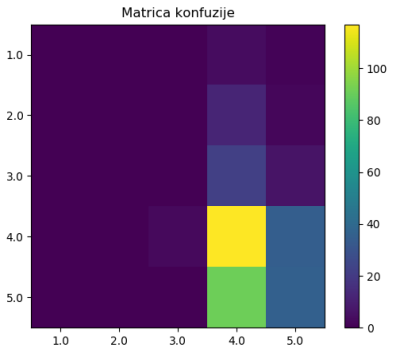
KNN rezultati

```
PRECIZNOST
0.5227963525835866
[[  0   0   1   0   2]
 [  1   0   0   6   6]
 [  0   0   4  18  12]
 [  1   0   7 108  44]
 [  0   1   8  50  60]]
```



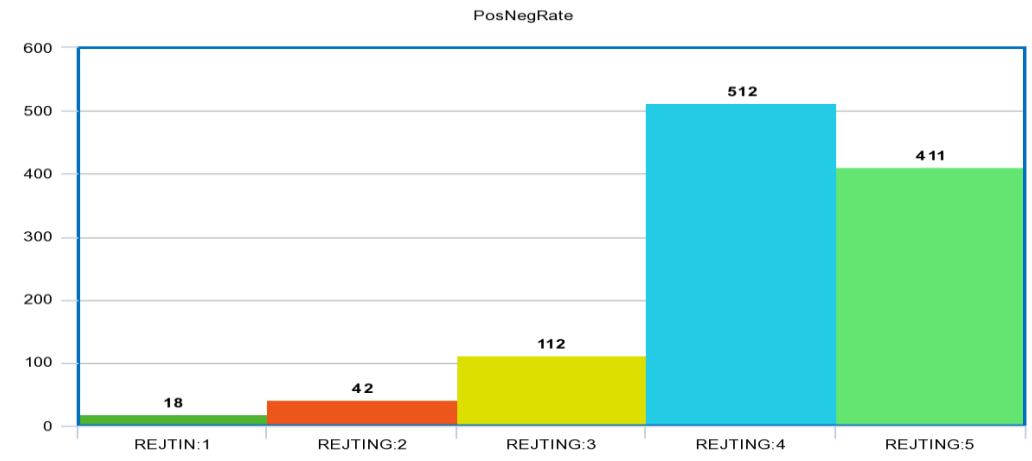
SVM rezultati

```
0.46504559270516715
[[  0   0   0   4   1]
 [  0   0   0  12   2]
 [  0   0   0  22   6]
 [  0   0   3 117  35]
 [  0   0   0  91  36]]
```



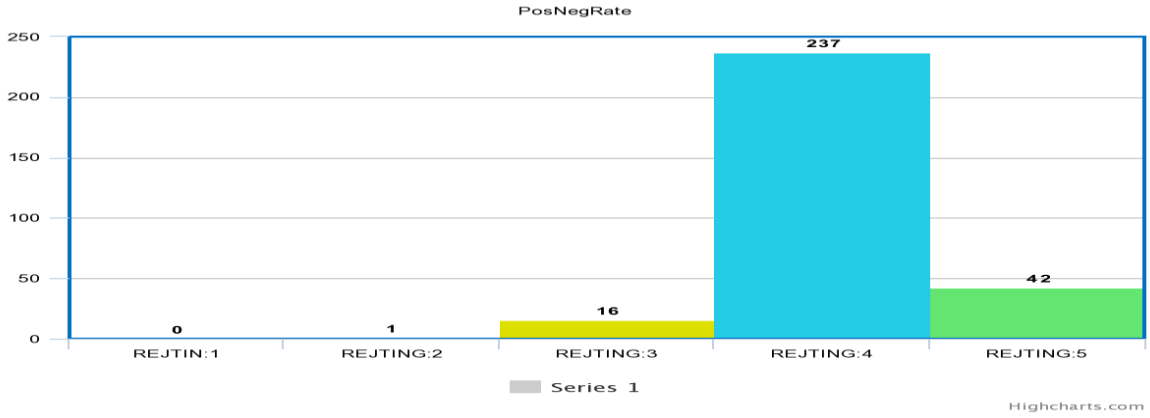
Zapažamo da modeli prave greške pri klasifikaciji rejtinga 4 I 5 (što je donekle očekivano) ali takođe zapažamo da broj onih artikala sa ocenom 1 I 2 ne igra skoro nikakvu ulogu u klasifikaciji I loše utiče na model, pre svega zato što ih ima jako malo.Takođe problem pravi broj review-a po artiklu što je napomenuto, upravo zbog nedovoljno informacija, model I ima lošu preciznost.(Diskutovanje o prilagođenosti modela podacima ostavljam za odbranu seminarskog)

Sada ćemo prikazati odnos artikala I njihove prosečne ocene, na celom skupu artikala.



Jasno se vidi da onih sa rejtingom 1 I 2, pa čak I 3 ima jako malo u odnosu na one sa rejtingom 4 I 5.

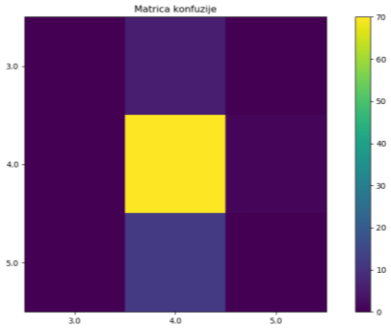
Sada ćemo pokušati da poboljšamo naš model koristeći samo one artikle koji imaju više od 5 review-a , čime dobijamo više informacija I bolji model.Problem će nam sada praviti broj onih koji imaju prosečnu ocenu 1 I 2 jer ih skoro neće ni biti, ali iz tog raloga pokušaćemo da napravimo model koji će bolje vršiti ocenjivanje artikala sa rejtingom 3, 4 I 5.
Prikažimo broj artikala sa prosečnim ocenama od 1 do 5 nakom čišćenja prvobitnog skupa od artikala koji imaju manje od 5 review-a.



Zapažamo da nema artikala sa ocenom 0 , dok sa ocenom 1 postoji samo 1 , što nas dalje navodi da klasifikaciju probamo da izvšimo samo za artikle sa ocenom 3, 4 I 5.Ipak gubimo dobar deo informacija ali konstruišemo bolji model.

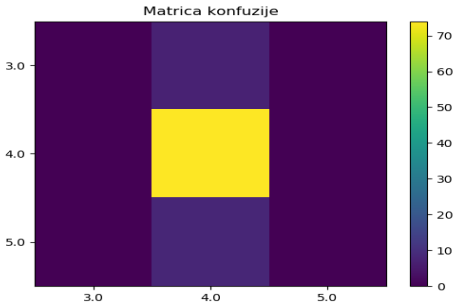
KNN rezultati:

```
PRECIZNOST
0.7865168539325843
[[ 0  6  0]
 [ 0 70  1]
 [ 0 12  0]]
```



SVM rezultati:

```
0.8314606741573034
[[ 0  7  0]
 [ 0 74  0]
 [ 0  8  0]]
```



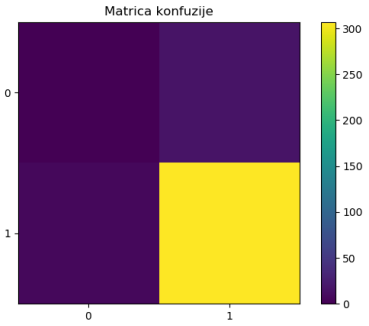
Klasifikacija artikala sa pros. ocenom 3, 4 I 5 daje mnogo bolje rezultate, što se vidi sa slika iznad.Preciznost znatno povećala a može se primetiti da model najbolje klasifikuje one sa ocenom 4 baš iz razloga što njih najviše I ima u skupu.SVM se pokazao nešto bolji u ovom slučaju.

Klasifikacija artikla (recommended ili not recommended)

Na kraju, pokušaćemo da klasifikujemo artikle po tome da li su sveobuhvatno prihvatljivi ili nisu.Koristićemo podatke kao u klasifikaciji rejtinga (prikazane u tabeli , u odelju preprocesiranje I obrada).Jedina izmena koju ćemo izvršiti je ta što ćemo pretpostaviti da za artikle sa rejtingom manjim od 3 važi da su not recommended dok ćemo za one sa rejtingom 4 I 5 reći da jesu recommended.U ovom slučaju nećemo izbacivati artikle sa brojem review-a manjim od 5, već koristimo ceo skup podataka iz tabele.Dobijamo sledeće rezultate.

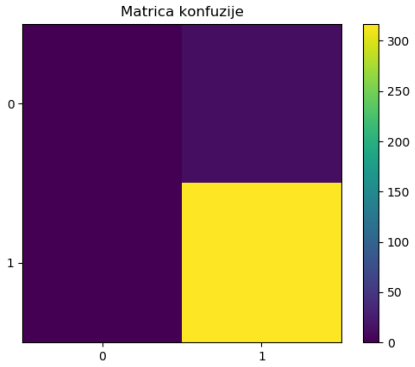
KN

```
0.9331306990881459
[[  0  16]
 [  6 307]]
```



SV

```
0.9635258358662614
[[  0  12]
 [  0 317]]
```



Oba modela imaju jako veliku preciznost, to je posledica, kao što je već rečeno velikog broja podataka koji su dobro ocenjeni, vrlo je moguće da je model preprilagodjen podacima (diskusija na odbrani projekta), sa druge strane možda je to posledica malo broja podataka koji su not recomended.Dodatna objašnjenja biće izložena u odbrani seminarskog kao I moguće ideje za dalje poboljšanje modela ako ih uočim.

Zaključak:

Moje mišljenje je da sve u svemu proces klasifikacije nije prošao loše, postoji manjak negativnih I loše ocenjenih artikala što pokazuje da su kupci ovih artikala mahom bili zadovoljni.Što se tiče klasifikacije komentara, ona je prošla zadovoljavajuće pogotovo svm-metodom koja je dala prilično realan model za ovaj skup podataka.Ocena rejtinga je nešto lošije prošla, zbog navedenih poteškoća I manjka review-a za veliki broj artikala, ali odbacivanjem artikala sa brojem review-a manjim od 5 dobijamo bolji model, ali sada sa manjim brojem podataka za trening I test fazu što nam daje model koji ima veću preciznost ali je diskutabilan za primenu u realnom vremenu.Na dalje to nas je dovelo do poslenja dva modela koja skoro sve artikle iz test skupa klasifikuju kao pozitivne, što nije dobro, ali verovatno je model preprilagodjen zbog količine pozitivnih komentara.Iako je preciznost jako velika poslednja dva modela zahtevaju dodatna unapredjenja.