

Državni Univerzitet u Novom Pazaru

Studijski Program Softversko Inženjerstvo

###### Mašinsko učenje

*“Klasifikacija I analiza utisaka Diznilend posetioca”*

*- seminarski rad -*

|  |  |
| --- | --- |
| Studenti:  Mentori: | Vahid Uglić  Muhamed Sadović    Doc. Dr. Ulfeta Marovac  Mast. Inž. Aldina Avdić |
|  |  |

Novi Pazar, 2023

SADRŽAJ

[1. Uvod 2](#_Toc72536346)

[2. Osnovni pojmovi 2](#_Toc72536347)

[3. Opis skupa podataka 3](#_Toc72536348)

[4. Opis metoda za klasifikaciju i metrika za evaluaciju 3](#_Toc72536349)

[5. Prikaz rezultata 6](#_Toc72536350)

[6. Analiza rezultata 9](#_Toc72536351)

[7. Zaključak 9](#_Toc72536352)

[8. Literatura 10](#_Toc72536353)

# Uvod

Zabavni park “DisneyLand” je jedan od najvećih tematskih parkova u svetu. Kao jedna globalna vodeća kompanija, poboljšanje iskustva njihovih korisnika se može rešiti korišćenjem mašinskog učenja u klasifikaciji recenzija njihovih posetilaca. Za rešenje ovog problema ćemo koristiti NLP , tekst mining kao i predstavljanje rečenica (komentara posetilaca) preko vektora reči. Recenzije ćemo razvrstati na ocene od 1-5 pomoću metoda za klasifikaciju i metrika za evaluaciju.

# Osnovni pojmovi

Tekst-mining je postupak ispitivanja velikih skupova tekstova kako bi se otkrile nove informacije ili pomoglo u odgovoru na određena istraživačka pitanja.

Tekst-mining identifikuje činjenice, veze i tvrdnje koje bi inače ostale zakopane u masu tekstualnih velikih podataka. Jednom izvučene, ove informacije se pretvaraju u strukturirani oblik koji se može dalje analizirati. Tekst mining koristi razne metodologije za obradu teksta,i to jednu od najvažnijih „Obrada prirodnog teksta“ (NLP).

NLP ili “Razumevanje prirodnog jezika” pomaže mašinama da „čitaju“ tekst (ili neki drugi unos kao što je govor) simulirajući ljudsku sposobnost razumevanja prirodnog jezika kao što su engleski, španski ili kineski. Obrada prirodnog jezika uključuje i razumevanje prirodnog jezika i stvaranje prirodnog jezika, što simulira ljudsku sposobnost stvaranja teksta na prirodnom jeziku, npr. da sumiraju informacije ili učestvuju u dijalogu.

Prestavljanje rečenice preko vektora reči pretvara atribute niza u skup numeričkih atributa koji predstavljaju informacije o pojavljivanju reči iz teksta koji se nalazi u nizovima. Rečnik se određuje od prve serije filtriranih podataka (obično trening podaci).

TF-IDF je statistička mera koja procenjuje koliko je neka reč relevantna za dokument u zbirci dokumenata. To se postiže množenjem dve metrike: koliko se puta reč pojavljuje u dokumentu i obrnuta učestalost reči u skupu dokumenata. Izračunava se primenom formule:

CountVectorizer je jedna od najčešće korišćenih tehnika za pretvaranje teksta u numerički oblik kako bi se mogao koristiti u algoritmima mašinskog učenja. Ova tehnika se koristi u polju obrade prirodnog jezika.

CountVectorizer izračunava broj pojavljivanja svake reči u datom tekstu i pretvara ih u numerički vektor.

SMOTE (Synthetic Minority Over-sampling Technique) je tehnika koja se koristi za balansiranje klase u problemima mašinskog učenja, posebno u situacijama kada postoji neravnoteža između broja instanci različitih klasa u skupu podataka. Neravnotežnost klase se javlja kada jedna klasa ima znatno manje uzoraka od drugih klasa, što može dovesti do niskih performansi modela u prepoznavanju manje zastupljenih klasa.

# Opis skupa podataka

Skup podataka uključuje 40000+ recenzija za 3 filijale Disneilanda - Pariz, Kalifornija i Hong Kong, koje su posetioci postavili na Trip Advisor.

Opis kolona:

* Review\_ID: jedinstveni ID koji se daje svakoj recenziji
* Rating: u rasponu od 1 (nezadovoljan) do 5 (zadovoljan)
* Year\_Month: datum kada je recenzent posetio tematski park
* Reviewer\_Location: zemlja porekla posetioca
* Review\_Text: komentari posetioca
* Disneyland\_Branch: lokacija Disney-Land parka

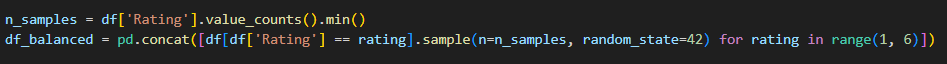
# Opis metoda za klasifikaciju i metrika za evaluaciju

**Priprema podataka**

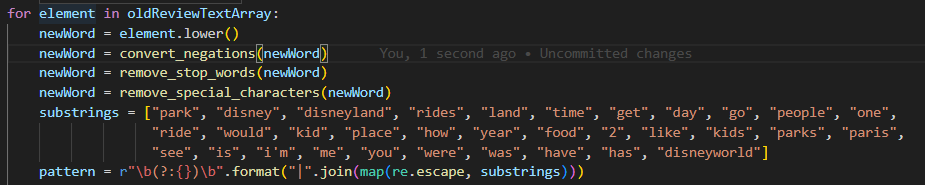
U okviru istraživačkog rada, izvršili smo uvoz skupa podataka iz datoteke "UpdatedDisneyLand.csv" koristeći programski jezik Python i biblioteku pandas. Prilikom uvoza podataka, primenili smo odgovarajuće kodiranje ('latin-1') kako bi obezbedili tačno i precizno čitanje podataka.



Zbog razlike u količini podataka u datasetu za različite ocene ,izvršili smo filtriranje odnosno izdvojili smo podjednak broj podataka za svaku ocenu. I na taj nacin smo izbalansirali taj uzorak.



Podatke smo preprocesirali kako bi se sve reči pretvorile u mala slova, uklonile negacije i uklonile stop reči(the, is, and, ...). To smo rešili pomocu koda:



Takodje smo ovime rešili i problem specijalnih karaktera.

Nakon preprocesiranja podataka, u nekim klasifikatorima smo koristili CountVectorizer a u nekim TF-IDFVectorizer kako bi dodatno obradili podatke.

Nakon obrade podataka, podatke dalje delimo u trening i test skupove u odnosu 80:20 pomocu koda



**Logistička regresija**

Logistička regresija predstavlja model binarne klasifikacije.

U logističkoj regresiji, koristimo:

Da procenimo verovatnoću pripadanja klasi.

**Naivni-Bajes**

Modelovanje raspodele ciljne promenljive *y* pri datim vrednostima promenljive *x*. Najčešće se primenjuje na problem klasifikacije. Naivni Bajesov klasifikator se zasniva na primeni Bajesove formule:

Za dato *x*, od svih ishoda *y* bira se onaj sa maksimalnom verovatnoćom .

Za ocenu verovatnoća i , potrebno je eksponencijalno mnogo podataka u odnosu na broj atributa.

Uvodi dve “naivne” pretpostavke nad atributima:

* svi atributi su a priori podjednako važni
* svi atributi su statistički nezavisni (vrednost jednog atributa nam ne govori ništa o vrednosti drugog atributa)

**J48**

Stablo odlučivanja je tehnika klasifikacije koja se sastoji od tri komponente korenskog čvora, grane (ivice ili veze) i čvora lista. Root predstavlja uslov ispitivanja za različite atribute, grana predstavlja sve moguće ishode koji mogu biti tamo u testu, a čvorovi listova sadrže oznaku klase kojoj pripada. Korijenski čvor je na početku drveta koji se naziva i vrhom drveta.

J48 klasifikator

Algoritam za generisanje stabla odluka koje generiše C4.5. Poznat je i kao statistički klasifikator.

**SVM - Support Vector Machines**

SVM je razvijen za numeričke ulazne promenljive, iako će automatski pretvoriti nominalne vrednosti u numeričke vrednosti. Ulazni podaci se takođe normalizuju pre upotrebe.

SVM rade pronalaženjem linije koja najbolje razdvaja podatke u dve grupe. To se radi pomoću procesa optimizacije koji uzima u obzir samo one instance podataka u skupu podataka o obuci koje su najbliže liniji koja najbolje razdvaja klase. Primeri se nazivaju vektori podrške, pa otuda i naziv tehnike. SMO se odnosi na specifični efikasni algoritam optimizacije koji se koristi unutar SVM implementacije, što znači Sekvencijalna minimalna optimizacija.

**KNN(iBk)**

Algoritam K-najbližih suseda, često skraćen k-nn, pristup je klasifikaciji podataka koji procenjuje koliko je verovatno da će tačka podataka biti član jedne ili druge grupe u zavisnosti od toga u kojoj su grupi najbliže tačke podataka .

K-najbliži sused je primer algoritma „lenji učenik“, što znači da ne gradi model pomoću skupa treninga dok se ne izvrši upit za skup podataka.

**Metrike ze evaluaciju**

**Metrike za evaluaciju u klasifikaciji:**

Matrica zabune je matrica u kojoj je predstavljena raspodela podataka u zavisnosti od njihove stvarne klase i one u koju ih je model svrstao.

**Performanse binarne klasifikacije:**

Accuracy

Pri radu sa nebalansiranim podacima koristimo metriku F-measure koja predstavlja harmonijsku sredinu preciznosti (eng.precision) i odziv (eng.recall).

Kod preciznosti i odziva kao pozitivna klasa uzima se ona koja je malobrojnija.

# Prikaz rezultata

**Trening set rez.**

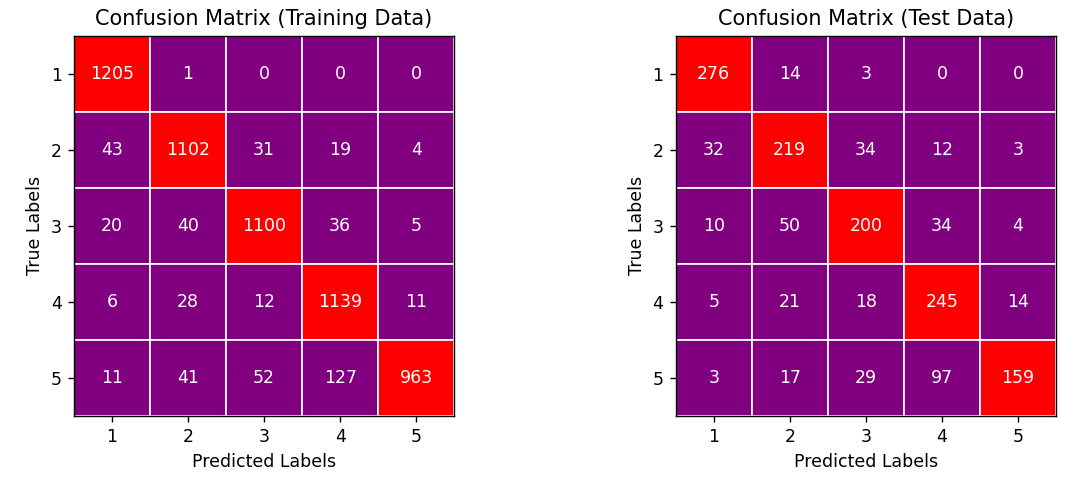
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metoda | Accuracy | Precision | Recall | F-Meassure |
| Logistic Reg. | **85,128%** | 0,953 | 0,9 | 0,88 |
| Naive Bayes | **84,32 %** | 0,84 | 0,844 | 0,815 |
| KNN | **62,48%** | 0,682 | 0,614 | 0,586 |
| J48 | **88,35%** | 0,914 | 0,897 | 0,865 |
| SVM (SMO) | **97,949 %** | 0,966 | 0,911 | 0,938 |

**Test set rez.**

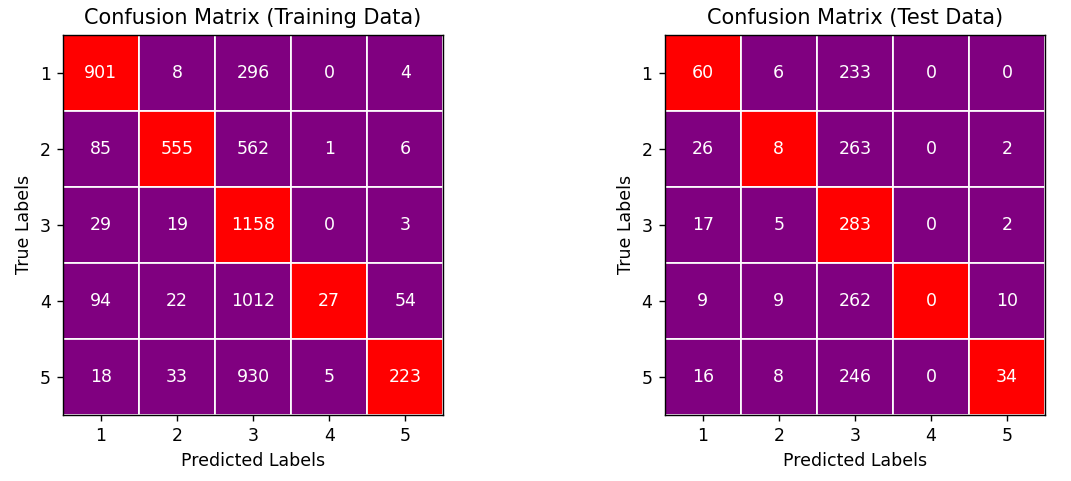
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metoda | Accuracy | Precision | Recall | F-Meassure |
| Logistic Reg. | **85 %** | 0,852 | 0,85 | 0.852 |
| Naive Bayes | **72 %** | 0,722 | 0,716 | 0,712 |
| KNN | **60 %** | 0,658 | 0,596 | 0,562 |
| J48 | **85 %** | 0,852 | 0,85 | 0,852 |
| SVM (SMO) | **83 %** | 0,836 | 0,83 | 0,832 |

**Matrice konfuzije**

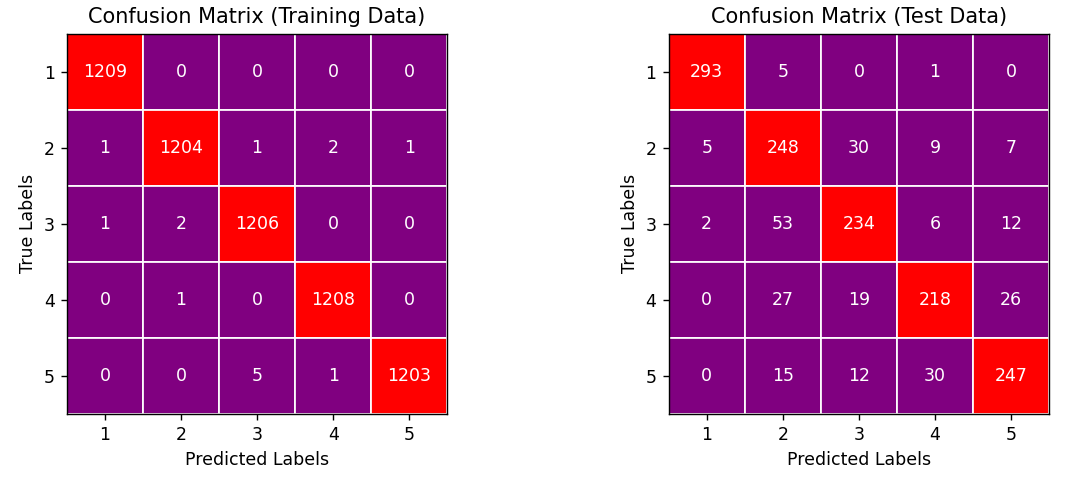
Naivni Bajes



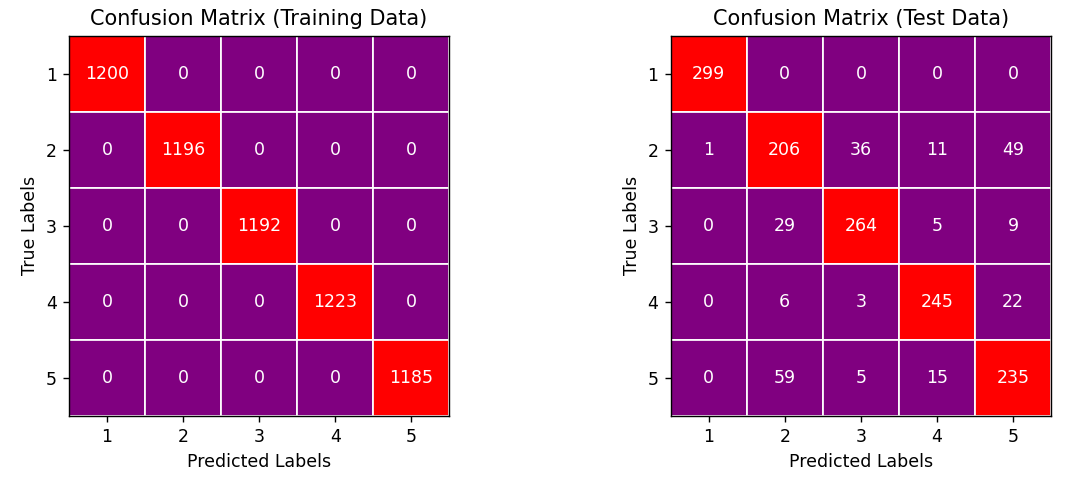
KNN



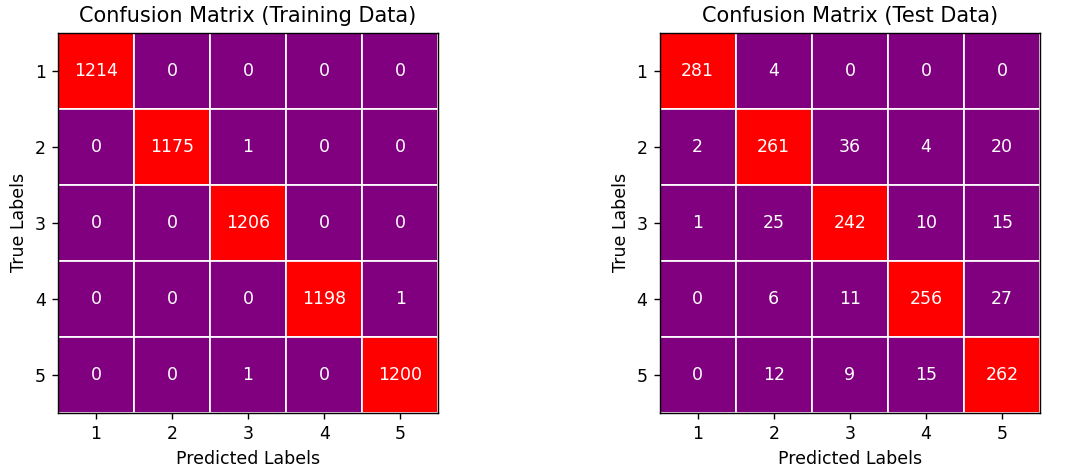
SVM



J48



Logisticka regresija



# Analiza rezultata

Na osnovu rezultata dobijenih primenjujući određene metode za klasifikaciju nad skupom podataka u kome je bilo potrebno klasifikovati recenzije Diznilenda po ocenama od 1-5, vidimo da su podaci približno isti. Kod Traning set-a najbolji rezultat smo dobili primenjujući Logističku regresiju i SVM, dok kod Test set-a najbolji rezultat su dali Logistička regresija i J48. KNN se pokazao kao najmanje pouzdan, i pri treniranju, kao i na test skupu.

# Zaključak

Jasno je da tačnost određenih modela zavisi od problema za koji se isti koristi, kao i od veličine skupa podataka. Svi algoritmi su uglavnom dali koliko-toliko zadovoljavajuće rezultate, a sledeći korak u našoj analizi bi bio proširivanje trening skupa ili možda dodavanje novih atributa. Sve u svemu, ovo je bilo jedno veoma značajno iskustvo, koje se može primeniti u kasnijim radovima iz oblasti mašinskog učenja.

# Literatura

* <http://moodle.np.ac.rs/course/view.php?id=564> – Predavanja, Vežbe
* <https://www.turing.com/kb/document-classification-using-naive-bayes>
* <https://realpython.com>
* https://www.youtube.com/