05 Domaci Asocijativna analiza

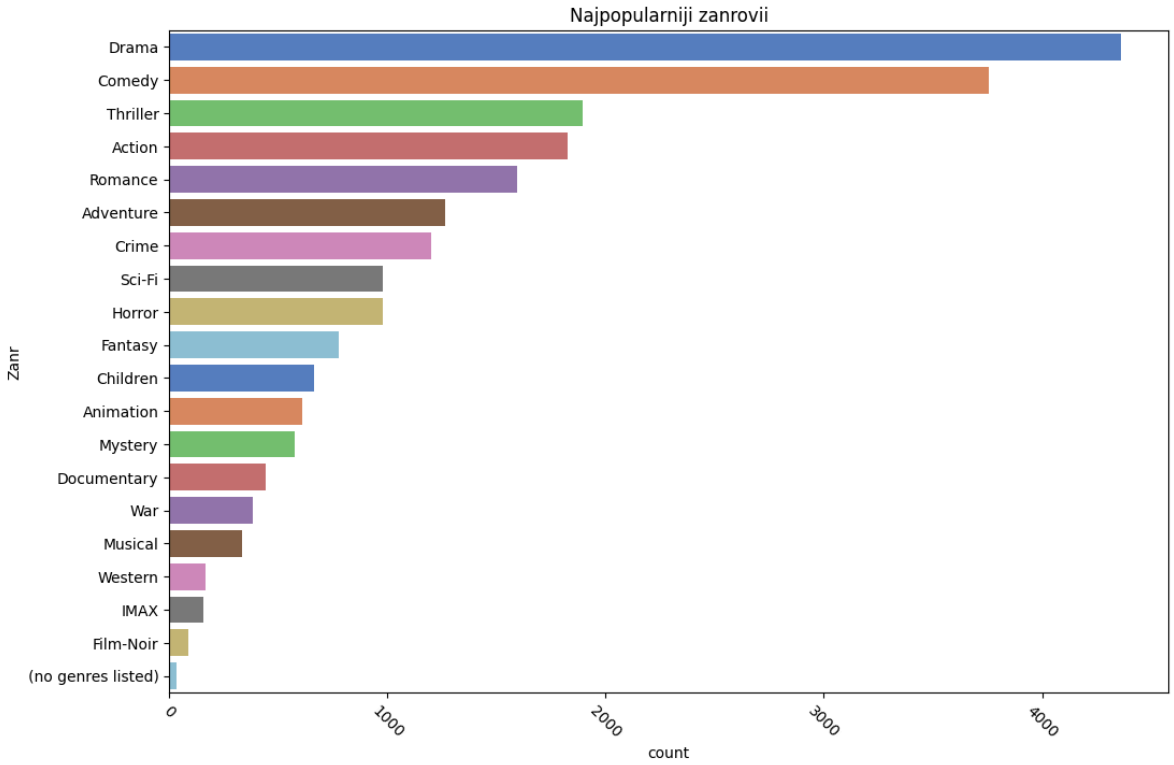
Predmet: Tehnike i metode analize podataka

Student 1636 Milica Jovanovic

Za izradu ovog domaceg zadatka, koriscen je Kaggle Movie Lens Dataset (https://www.kaggle.com/datasets/aigamer/movie-lens-dataset). Dataset sadrzi ukupno 9742 filma, sa 100836 ocena od strane korisnika i 3683 tagova.

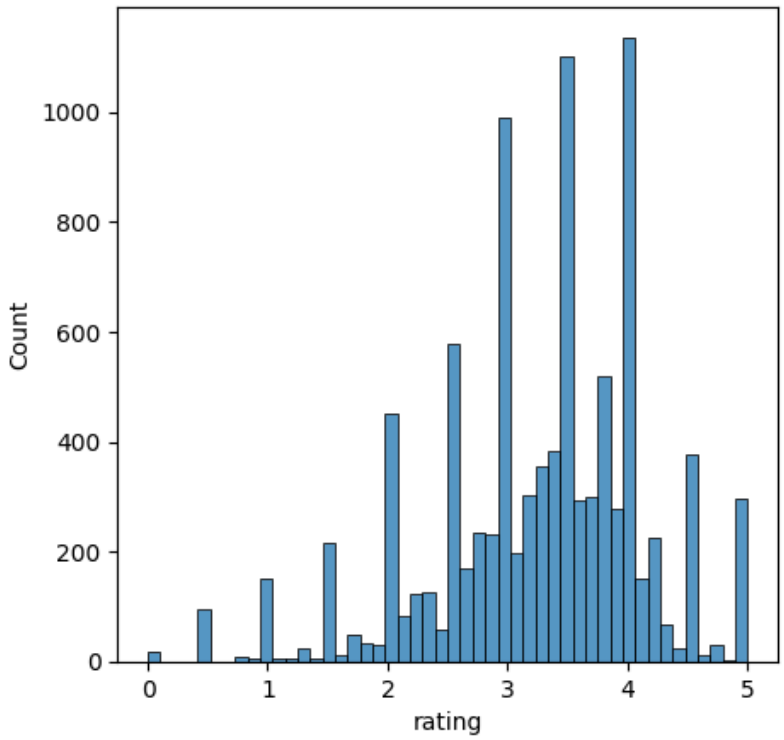
- **userId (ID korisnika):** Jedinstveni ID dodeljen svakom korisniku.
- **movieId (ID filma):** Jedinstveni ID dodeljen svakom filmu. U dataset su uključeni samo filmovi koji imaju barem jednu ocenu ili tag.
- **rating (ocena) (ratings.csv):** Ocene se vrše na skali od 5 zvezdica, sa poluzvezdama (0.5 zvezdica - 5.0 zvezdica).
- **genres (žanrovi):** Žanrovi su navedeni kao lista koja je odvojena vertikalnom crtom (pipe-separated) i izabrani su iz sledećih kategorija: *Action*
- *Adventure*
- *Animation*
- *Children's*
- *Comedy*
- *Crime*
- *Documentary*
- *Drama*
- *Fantasy*
- *Film-Noir*
- *Horror*
- *Musical*
- *Mystery*
- *Romance*
- *Sci-Fi*
- *Thriller*
- *War*
- *Western*
- *(no genres listed)*

Vizuelizacija najpopularnijih zanrova:



Najvise filmova su deame ili komedije, odmah zatim trileri i akcije.

Raspodela ocena:



Pre generisanja asocijativnih pravila, algoritmi zahtevaju okvir podataka u kojem su sve transakcije kodirane u obliku "one-hot encoding" za sve stavke sto je I odradjeno.

```
transactions=[]
for item in data['movieId'].unique():
    genres = data[data['movieId'] == item]['genres'].tolist()
    flat_genres = [genre for sublist in genres for genre in sublist]  # Spajanje svih u jednu listu
    unique_genres = list(set(flat_genres))  # Pretvaranje u jedinstvenu listu
    transactions.append(unique_genres)

transactions[0:10]
```

✓ 3.5s

```
['Animation', 'Comedy', 'Fantasy', 'Adventure', 'Children'],
['Fantasy', 'Adventure', 'Children'],
['Romance', 'Comedy'],
['Romance', 'Comedy', 'Drama'],
['Comedy'],
['Action', 'Thriller', 'Crime'],
['Romance', 'Comedy'],
['Adventure', 'Children'],
['Action'],
['Action', 'Adventure', 'Thriller']]
```

```
te = TransactionEncoder()
encodedData = te.fit(transactions).transform(transactions)
df = pd.DataFrame(encodedData, columns=te.columns_)
df.head()
```

✓ 0.0s

Zatim se prelazi na primenu algoritma:

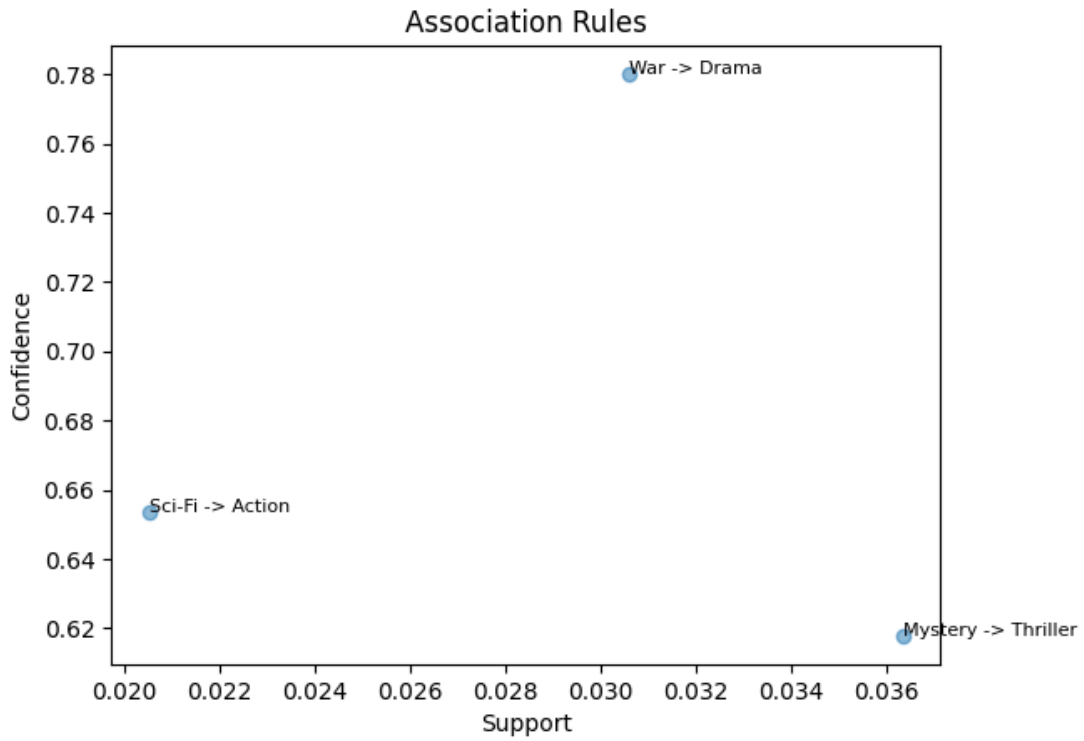
1. Apriori

Potrebno je da naci najbolje vrednosti za min_support i min_confidence

```
✓ 0.1s
Za min_support = 0.02 i min_confidence = 0.5 prosecni lift je 1.8877621110735707.
Za min_support = 0.02 i min_confidence = 0.6 prosecni lift je 2.801204807151955.
Za min_support = 0.02 i min_confidence = 0.7 prosecni lift je 1.742669136599872.
Za min_support = 0.02 i min_confidence = 0.8 prosecni lift je nan.
Za min_support = 0.02 i min_confidence = 0.9 prosecni lift je nan.
Za min_support = 0.03 i min_confidence = 0.5 prosecni lift je 1.639758552812204.
Za min_support = 0.03 i min_confidence = 0.6 prosecni lift je 2.4601988241201895.
Za min_support = 0.03 i min_confidence = 0.7 prosecni lift je 1.742669136599872.
Za min_support = 0.03 i min_confidence = 0.8 prosecni lift je nan.
Za min_support = 0.03 i min_confidence = 0.9 prosecni lift je nan.
Najbolji parametri su: min_support = 0.02, i min_confidence = 0.6.
```

Sa ovim parametrima se primenjuje algoritam:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
0	War	Drama	0.039212	0.447649	0.030589	0.780105	1.742669	1.0	0.013036	2.511880	0.443560	0.067042	0.601892	0.424219
1	Mystery	Thriller	0.058817	0.194416	0.036338	0.617801	3.177729	1.0	0.024902	2.107761	0.728137	0.167534	0.525563	0.402354
2	Sci-Fi	Action	0.031410	0.187641	0.020530	0.653595	3.483217	1.0	0.014636	2.345111	0.736028	0.103413	0.573581	0.381502



Analiza pravila:

Pravilo: `War -> Drama`

- **Support:** 0.030589 → 3.06% filmova su istovremeno `War` i `Drama`.
- **Confidence:** 0.780105 → 78% `War` filmova su takođe `Drama`.
- **Lift:** 1.742669 → `Drama` je 1.74 puta verovatniji kod `War` filmova nego u slučajnoj distribuciji.
- **Zhang's Metric:** 0.443560 → Srednja povezanost između `War` i `Drama`.
- **Jaccard:** 0.067042 → Mali procenat zajedničkog pojavljivanja u odnosu na ukupna pojavljivanja.

Pravilo: `Mystery -> Thriller`

- **Support:** 0.036338 → 3.63% filmova su istovremeno `Mystery` i `Thriller`.
- **Confidence:** 0.617801 → 61.78% `Mystery` filmova su takođe `Thriller`.
- **Lift:** 3.177729 → `Thriller` je 3.17 puta verovatniji kod `Mystery` filmova nego slučajno.
- **Zhang's Metric:** 0.728137 → Snažna povezanost između `Mystery` i `Thriller`.
- **Jaccard:** 0.167534 → Viši procenat zajedničkog pojavljivanja u odnosu na ukupna pojavljivanja.

Pravilo: `Sci-Fi -> Action`

- **Support:** 0.020530 → 2.05% filmova su istovremeno `Sci-Fi` i `Action`.
- **Confidence:** 0.653595 → 65.36% `Sci-Fi` filmova su takođe `Action`.
- **Lift:** 3.483217 → `Action` je 3.48 puta verovatniji kod `Sci-Fi` filmova nego slučajno.
- **Zhang's Metric:** 0.736028 → Vrlo snažna povezanost između `Sci-Fi` i `Action`.
- **Jaccard:** 0.103413 → Ograničeno zajedničko pojavljivanje u odnosu na ukupna pojavljivanja.

---

3. Zaključci:

- **Najsigurnije pravilo:** `War -> Drama`, jer ima najviši **confidence** (78%).
- **Najjača povezanost:** `Sci-Fi -> Action`, jer ima najviši **lift** (3.48) i solidan **Zhang's Metric** (0.736).
- **Najveća proporcionalna sličnost:** `Mystery -> Thriller`, jer ima najveći **Jaccard** (0.1675).

2. FP Growth

Prvo se nalaze najbolje vrednosti za min_support i min_confidence

```
✓ 7.4s

Za min_support = 0.02 i min_confidence = 0.5 prosecni lift je 1.887762111073571.
Za min_support = 0.02 i min_confidence = 0.6 prosecni lift je 2.801204807151955.
Za min_support = 0.02 i min_confidence = 0.7 prosecni lift je 1.742669136599872.
Za min_support = 0.02 i min_confidence = 0.8 prosecni lift je nan.
Za min_support = 0.02 i min_confidence = 0.9 prosecni lift je nan.
Za min_support = 0.03 i min_confidence = 0.5 prosecni lift je 1.639758552812204.
Za min_support = 0.03 i min_confidence = 0.6 prosecni lift je 2.4601988241201895.
Za min_support = 0.03 i min_confidence = 0.7 prosecni lift je 1.742669136599872.
Za min_support = 0.03 i min_confidence = 0.8 prosecni lift je nan.
Za min_support = 0.03 i min_confidence = 0.9 prosecni lift je nan.
Najbolji parametri su: min_support = 0.02, i min_confidence = 0.6.
```

Sa ovim parametrima se primenjuje algoritam:

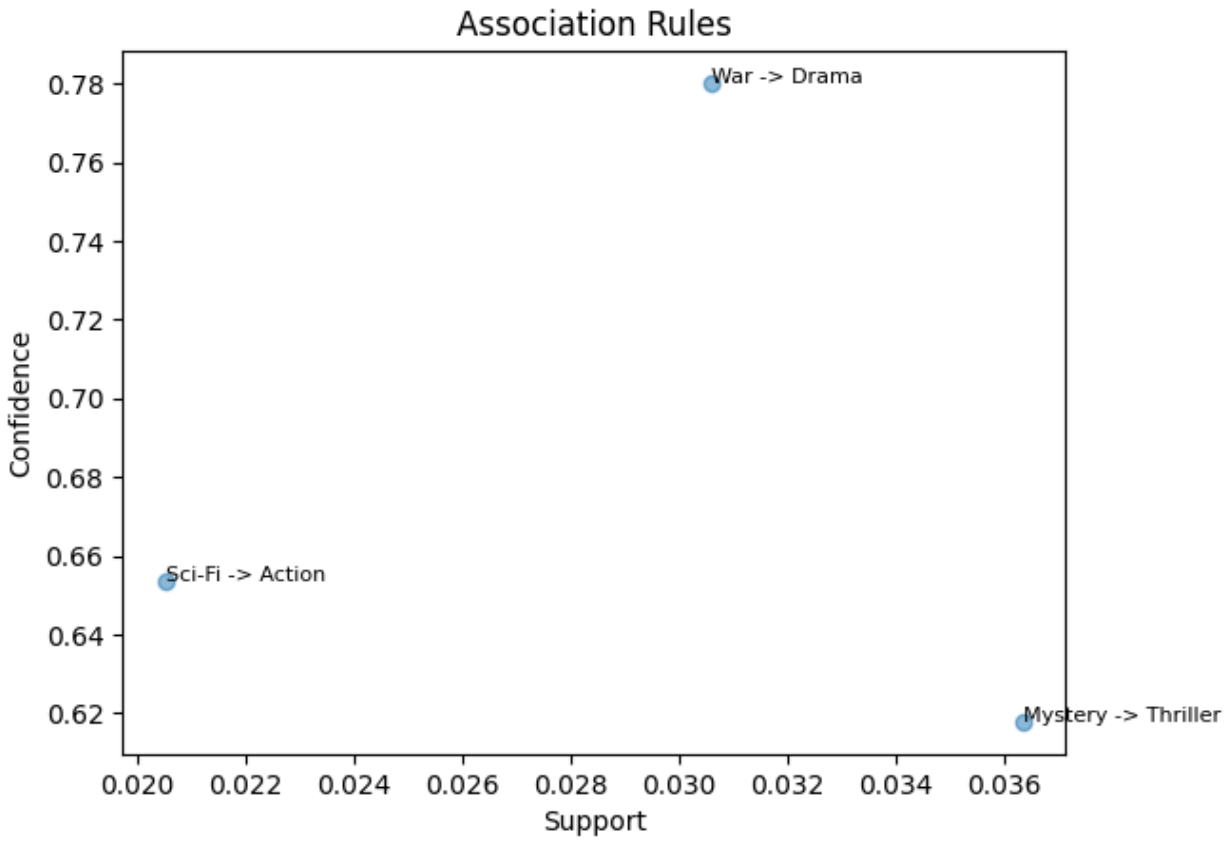


Tabela sa pravilima:

1. `Mystery -> Thriller`
    - **Support:** 0.036338 → Oko 3.63% svih podataka uključuje oba ova žanra zajedno.
    - **Confidence:** 0.617801 → Ako je film Mystery, postoji 61.78% šanse da je i Thriller.
    - **Lift:** 3.177729 → Mystery filmovi su 3.17 puta verovatniji da budu i Thrillery u poređenju sa slučajnom distribucijom.
    - **Zhang's Metric:** 0.728137 → Visoka vrednost znači jaku povezanost između ova dva žanra.
2. `Sci-Fi -> Action`
    - **Support:** 0.020530 → Oko 2.05% filmova pripada oba žanra.
    - **Confidence:** 0.653595 → Ako je film Sci-Fi, postoji 65.36% šanse da je i Action.
    - **Lift:** 3.483217 → Sci-Fi filmovi su 3.48 puta verovatniji da budu i Action u poređenju sa slučajnošću.
    - **Zhang's Metric:** 0.736028 → Ukazuje na snažnu vezu između Sci-Fi i Action žanra.
3. `War -> Drama`
    - **Support:** 0.030589 → Oko 3.06% filmova pripada oba žanra.
    - **Confidence:** 0.780105 → Ako je film War, postoji 78.01% šanse da je i Drama.
    - **Lift:** 1.742669 → War filmovi su 1.74 puta verovatniji da budu i Drama u poređenju sa slučajnom distribucijom.
    - **Zhang's Metric:** 0.443560 → Srednje jaka povezanost između War i Drama žanra.

Zaključci:

- *Mystery -> Thriller* ima najviši lift (3.17), što ukazuje na jaku zavisnost između ova dva žanra.
- *Sci-Fi -> Action* ima najviši lift (3.48), što znači da su Sci-Fi filmovi često Action filmovi.
- *War -> Drama* ima najviši confidence (78%), što ga čini najsigurnijim pravilom u smislu predviđanja.