

# Paramount TV Shows and Movies

---

## Skup Podataka

Ovaj skup podataka se sastoji od dve datoteke:

1. titles.csv
2. credits.csv

# titles.csv

Ovaj skup podataka sadrži 2825 jedinstvenih naslova Paramountovih filmova i serija, kao i 15 kolona (atributa) koje nam ga bolje opisuju:

- **id: ID naslova na JustWatch platformi**
- title: Naslov filma, odnosno serije
- **show\_type: Indikator da li je u pitanju film ili serija - moguće vrednosti su SHOW i MOVIE**
- description: Kratak opis naslova
- **release\_year: Godina premijere filma, odnosno serije**
- age\_certification: Naznaka o preporučenoj starosnoj dobi gledaoca
- **runtime: Dužina epizode serije, odnosno dužina filma**
- **genres: Lista žanrova**
- production\_countries: Lista zemalja u kojima je rađeno na filmu
- **seasons: Broj sezona (ukoliko je u pitanju SHOW tip)**
- imdb\_id: ID naslova na IMDBu
- **imdb\_score: Ocena na IMDBu**
- imdb\_votes: Glasovi na IMDBu
- tmdb\_popularity: Popularnost na TMDBu
- tmdb\_score: Ocena na TMDBu

# credits.csv

Ovaj skup podataka sadrži 39842 glumaca i režisera koji su radili na gorepomenutim naslovima, kao i 5 atributa koji nam ih bolje opisuju:

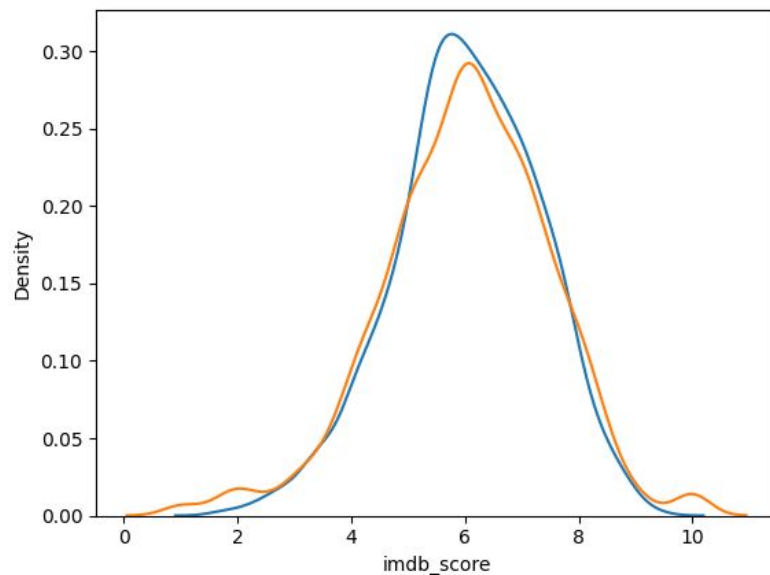
- person\_ID: ID osobe na JustWatch platformi
- **id: ID naslova na JustWatch platformi (preko ovoga se mogu povezati sa Titles datotekom)**
- name: Ime glumca, odnosno režisera
- character\_name: Ime lika kojeg glumac glumi u naslovu
- **role: Indikator da li je u pitanju glumac ili režiser (moguće vrednosti su ACTOR i DIRECTOR)**

# Pretprocesiranje

---

## Pretprocesiranje

- Kako svi filmovi imaju NaN za broj sezona, postavili smo taj broj na 0.0
- Izbacili smo kolonu “Age Certification”, jer nas nije interesovala za dalji rad, a imala je i mnogo nedostajucih vrednosti
- Izbacili smo sve instance koje nemaju “imdb\_id”, kao i “description” - nije ih bilo puno, pa nam nije pravilo preveliku razliku.
- Dodelili smo tmdb\_score odgovarajucim instancama bez imdb\_score-a, nakon utvrđivanje linearne zavisnosti i visoke korelacije:



# Klasifikacija

---

# Klasifikacija

Kako bismo predstavili KNN algoritam i algoritam stabala odlučivanja, odlučili smo da klasifikujemo glumce i režisere u jednu od tri klase:

- Visoko Rangirani (High Rated) - glumci/režiseri čija je prosečna IMDB ocena između 7 i 10
- Srednje Rangirani (Medium Rated) - glumci/režiseri čija je prosečna IMDB ocena između 4 i 7
- Nisko Rangirani (Low Rated) - glumci/režiseri čija je prosečna IMDB ocena između 0 i 4

Prosečna IMDB ocena je za svakoga izračunata na osnovu svih filmova u kojima su glumili/režirali.



# KNN

Za izgradnju modela, izabrali smo prosečnu godinu premijere filmova na kojim je određeni glumac/režiser radio, prosečan broj sezona, kao i dužinu trajanja filma/epizode serije.

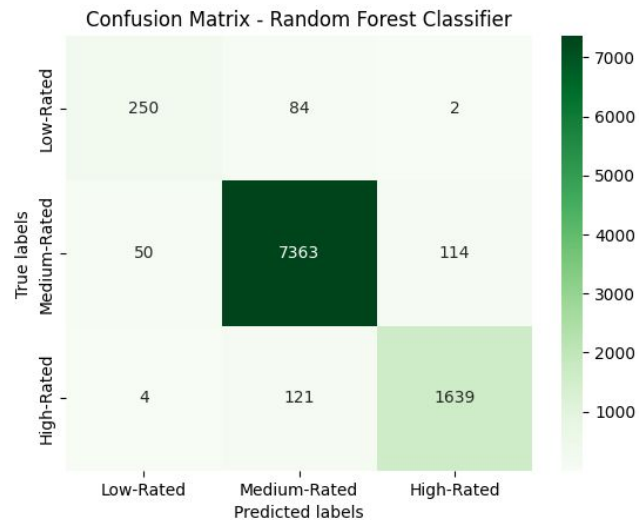
Cilj nam je da na osnovu ovih atributa, naš KNN model uspešno klasifikuje određenog glumca ili režisera kao visoko, srednje ili nisko rangiranog.

Nakon odvajanja karakteristika koje ćemo iskoristiti za treniranje i testiranje modela, skalirali smo ih koristeći MinMaxScaler.

# Ansambli - Random forest Classifier

Classification report for model RandomForestClassifier on test data

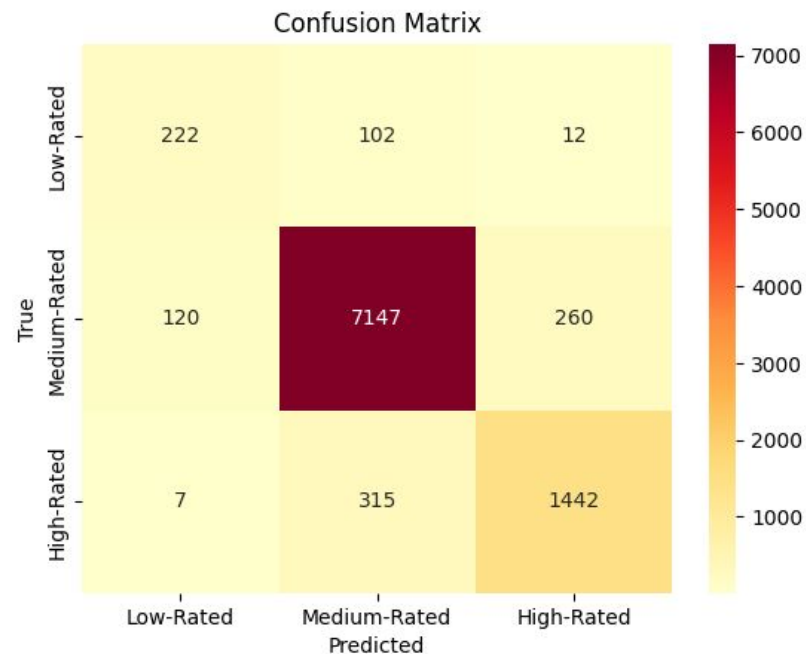
	precision	recall	f1-score	support
High-Rated	0.93	0.93	0.93	1764
Low-Rated	0.82	0.74	0.78	336
Medium-Rated	0.97	0.98	0.98	7527
accuracy			0.96	9627
macro avg	0.91	0.88	0.90	9627
weighted avg	0.96	0.96	0.96	9627



# KNN - rezultati

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
High-Rated	0.84	0.82	0.83	1764
Low-Rated	0.64	0.66	0.65	336
Medium-Rated	0.94	0.95	0.95	7527
accuracy			0.92	9627
macro avg	0.81	0.81	0.81	9627
weighted avg	0.92	0.92	0.92	9627

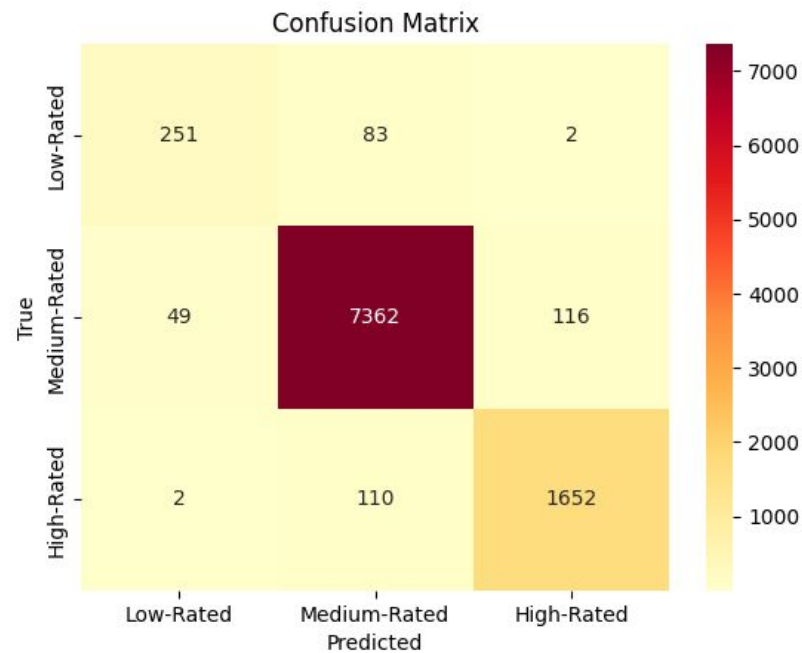


# KNN - optimizacije

## GridSearchCV

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
High-Rated	0.93	0.94	0.93	1764
Low-Rated	0.83	0.75	0.79	336
Medium-Rated	0.97	0.98	0.98	7527
accuracy			0.96	9627
macro avg	0.91	0.89	0.90	9627
weighted avg	0.96	0.96	0.96	9627

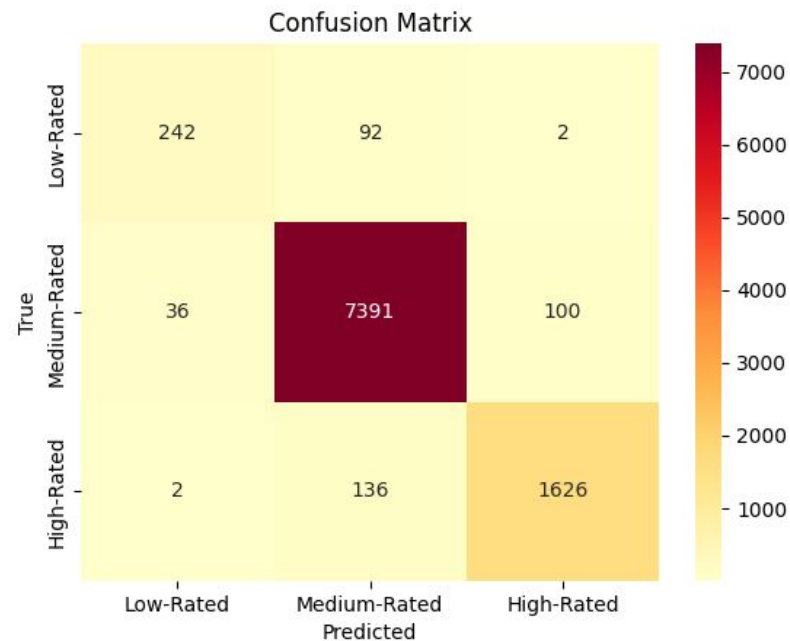


# KNN - optimizacije

## BaggingClassifier

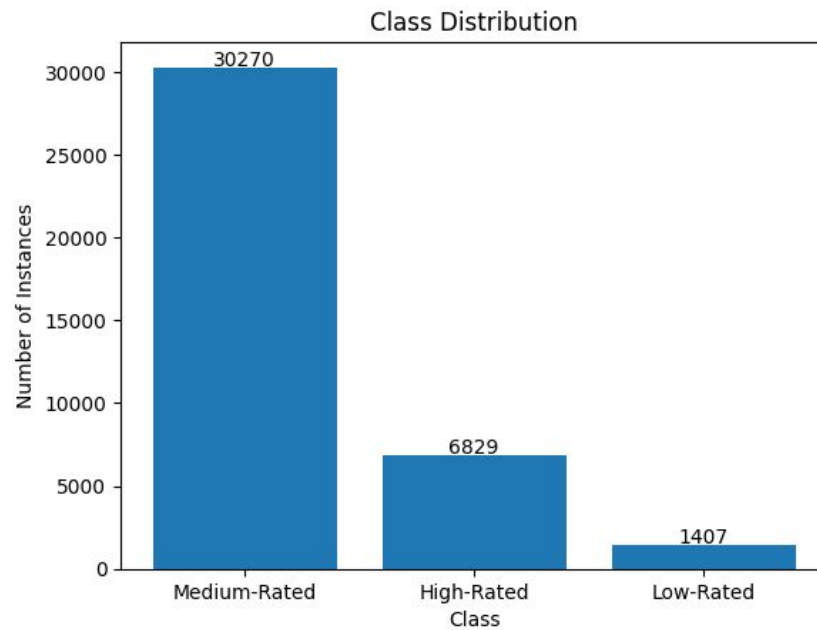
Classification report for model BaggingClassifier on test data

	precision	recall	f1-score	support
High-Rated	0.94	0.92	0.93	1764
Low-Rated	0.86	0.72	0.79	336
Medium-Rated	0.97	0.98	0.98	7527
accuracy			0.96	9627
macro avg	0.93	0.87	0.90	9627
weighted avg	0.96	0.96	0.96	9627



## Stabla odlučivanja

Nebalansirane klase:



# Stablo odlucivanja bez balansiranja klasa

Classification report for model DecisionTreeClassifier on test data

	precision	recall	f1-score	support
High-Rated	0.92	0.94	0.93	1764
Low-Rated	0.80	0.75	0.77	336
Medium-Rated	0.97	0.97	0.97	7527
accuracy			0.96	9627
macro avg	0.90	0.89	0.89	9627
weighted avg	0.96	0.96	0.96	9627

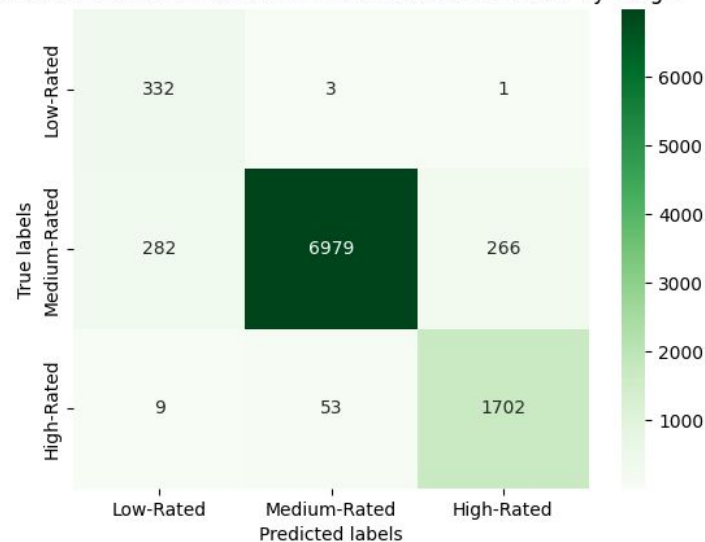


## Optimizacija Stabla Odlučivanja uz GridSearchCV - balansirani model

Classification report for model DecisionTreeClassifier on testing data

	precision	recall	f1-score	support
High-Rated	0.86	0.96	0.91	1764
Low-Rated	0.53	0.99	0.69	336
Medium-Rated	0.99	0.93	0.96	7527
accuracy			0.94	9627
macro avg	0.80	0.96	0.85	9627
weighted avg	0.95	0.94	0.94	9627

Confusion Matrix - GridSearchCV with classes balanced by weight

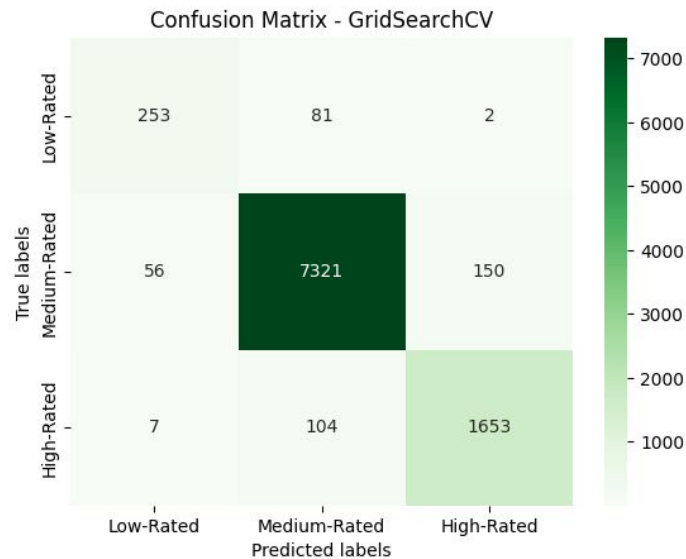




## Optimizacija Stabla Odlučivanja uz GridSearchCV - nebalansirani model

Classification report for model DecisionTreeClassifier on test data

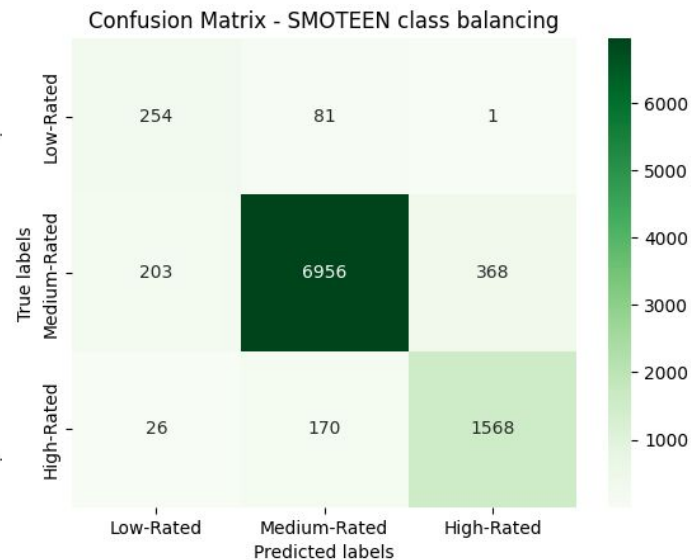
	precision	recall	f1-score	support
High-Rated	0.91	0.94	0.93	1764
Low-Rated	0.80	0.75	0.77	336
Medium-Rated	0.98	0.97	0.97	7527
accuracy			0.96	9627
macro avg	0.90	0.89	0.89	9627
weighted avg	0.96	0.96	0.96	9627



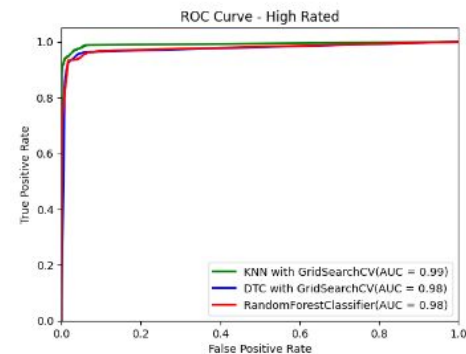
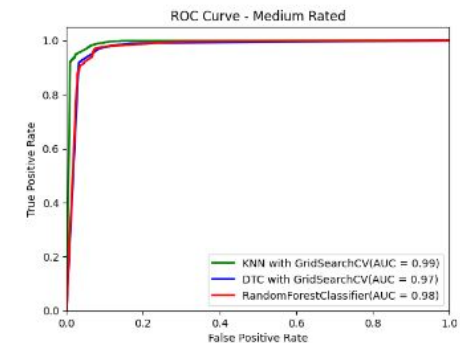
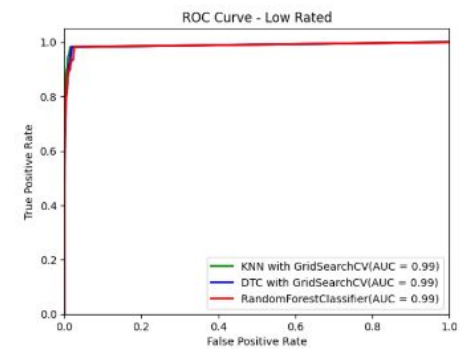
## Balansiranje klasa pomoću kombinacije oversamplinga i undersamplinga - SMOTEENN

Classification report for model DecisionTreeClassifier on test data

	precision	recall	f1-score	support
High-Rated	0.81	0.89	0.85	1764
Low-Rated	0.53	0.76	0.62	336
Medium-Rated	0.97	0.92	0.94	7527
accuracy			0.91	9627
macro avg	0.77	0.86	0.80	9627
weighted avg	0.92	0.91	0.92	9627



# Poređenje modela pomoću ROC krive



# Klasterovanje

---

# Algoritam K sredina

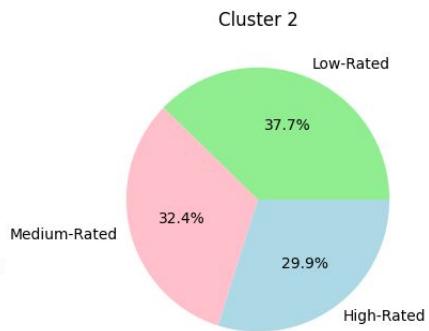
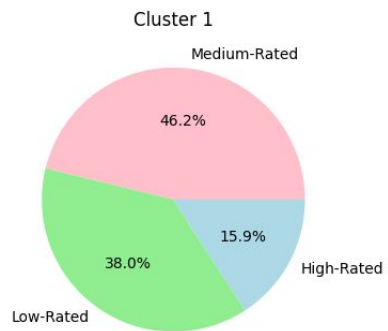
U ovom projektu, klasterovali smo naslove filmova, odnosno serija na osnovu njihovih ocena (Visoka, Srednja, Niska).

Atributi koje smo izabrali za treniranje modela su:

- 'runtime'
- 'release\_year'
- 'seasons'
- 'encoded\_type'
- 'encoded\_genre'

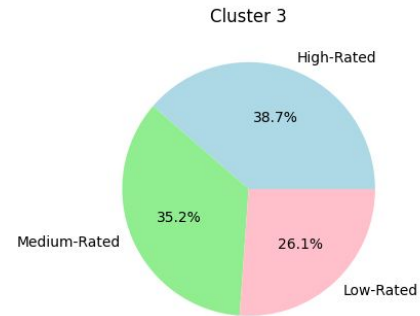
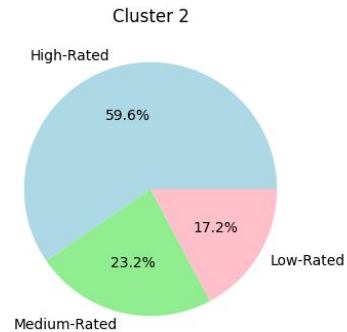
Ove atribute smo skalirali korišćenjem Standard Scalera.

# Algoritam K Sredina



# Algoritam Sakupljajućeg Klasterovanja

```
model = AgglomerativeClustering(n_clusters=3, linkage='complete', compute_distances=True)
```



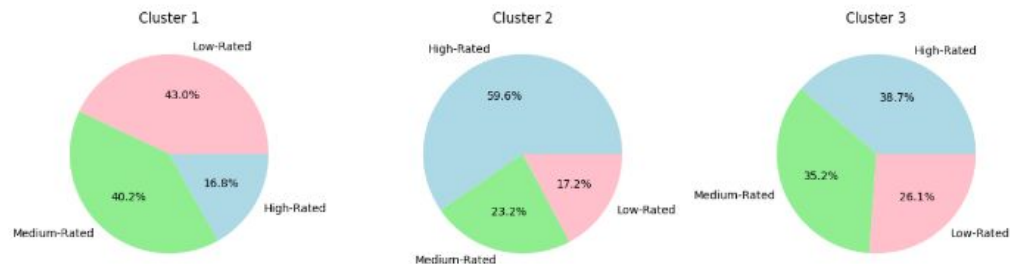
# Poređenje modela

K Means Clusters



	Low Rated	Medium Rated	High Rated
K Means	<b>37.7%</b>	32.4%	29.9%
Agglomerative	<b>43%</b>	40.2%	16.8%

Agglomerative Clusters



	Low Rated	Medium Rated	High Rated
K Means	38%	<b>46.2%</b>	15.9%
Agglomerative	26.1%	<b>35.2%</b>	38.7%

	Low Rated	Medium Rated	High Rated
K Means	17.1%	22.9%	<b>60%</b>
Agglomerative	17.2%	23.2%	<b>59.6%</b>















# Pravila Pridruživanja - SPSS

---

# Apriori Algoritam

Radi lakšeg rada sa podacima, veći deo pretprocesiranja odrađen je u Pythonu. Podaci koje smo odlučili da izdvojimo za ovaj algoritam su:

- release\_year
- imdb\_score
- encoded\_type
- encoded\_genre

Field	Measurement	Values	Missing	Check	Role
 release_year	 Continuous	[1912,2022]		None	 Both
 imdb_score	 Continuous	[1.7,9.4]		None	 Both
 encoded_type	 Flag	1/0		None	 Both
 encoded_ge...	 Nominal	0,1,2,3,4,5...		None	 Both

	release_year	imdb_score	encoded_type	encoded_genre
0	1926	8.2	0	1
1	1940	7.8	0	3
2	1945	7.3	0	16
3	1936	4.0	0	4
4	1916	7.7	0	9

# Apriori Algoritam

## Binovanje (Diskretizacija)

	release_year	imdb_score	encoded_type	encoded_genre	imdb_score_BIN	release_year_BIN
1	1926	8.200	0	1	3	1
2	1940	7.800	0	3	3	2
3	1945	7.300	0	16	3	2
4	1936	4.000	0	4	1	2
5	1916	7.700	0	9	3	1
6	1946	7.300	0	16	3	2
7	1925	7.500	0	10	3	1
8	1932	7.100	0	16	3	2
9	1928	7.800	0	6	3	1
10	1921	8.300	0	6	3	1

# Apriori Algoritam - balansiranje klasa





Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
release_year		Continuous	1912	2022	1989.415	31.588	-0.722	—	2637
imdb_score		Continuous	1.700	9.400	6.043	1.266	-0.268	—	2637
encoded_type		Flag	0	1	—	—	—	2	2637
encoded_genre		Nominal	1	18	—	—	—	18	2637
imdb_score_BIN		Nominal	1	3	—	—	—	3	2637
release_year_BIN		Nominal	1	4	—	—	—	4	2637

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
release_year		Continuous	1912	2022	1999.637	26.059	-1.458	—	6880
imdb_score		Continuous	1.700	9.400	5.629	1.813	-0.152	—	6880
encoded_type		Flag	0	1	—	—	—	2	6880
encoded_genre		Nominal	1	18	—	—	—	18	6880
imdb_score_BIN		Nominal	1	3	—	—	—	3	6880
release_year_BIN		Nominal	1	4	—	—	—	4	6880





\* Indicates a multimode result   \* Indicates a sampled result

# Apriori Algoritam

Consequents:

-  imdb\_score\_BIN
-  encoded\_type
-  encoded\_genre
-  release\_year\_BIN

Antecedents:

-  encoded\_type
-  encoded\_genre
-  release\_year\_BIN
-  imdb\_score\_BIN

Model name: ☒ Auto ☐ Custom

☒ Use partitioned data

Minimum antecedent support (%):

Minimum rule confidence (%):

Maximum number of antecedents:

☒ Only true values for flags

Optimize: ☒ Speed ☐ Memory

# Apriori Algoritam - Rezultati

1. encoded\_type -> encoded\_genre = 12 and release\_year\_BIN = 3

- Kada nam je tip naslova SHOW(serija), velika je verovatnća da će biti praćena sa encoded\_genre = 12 (REALITY) i release\_year\_BIN = 3 (Oko 1990-2000tih godina)

2. encoded\_type -> encoded\_genre = 12 and imdb\_score\_BIN = 1

- Kada nam je tip naslova serija, velika je verovatnoća da bude "reality" žanra i da mu IMDB ocena bude u najnižem rasponu.

3.encoded\_type -> encoded\_genre = 12

- Kada nam je tip naslova serija, velika je verovatnoća da bude "reality" žanra

4. encoded\_type -> encoded\_genre = 12 and release\_year\_BIN = 4

- Kada nam je tip naslova SHOW(serija), velika je verovatnća da će biti praćena sa encoded\_genre = 12 (REALITY) i release\_year\_BIN = 4 (poslednjih 20ak godina)

5. encoded\_type -> encoded\_genre = 2

- Kada nam je tip naslova serija, velika je verovatnoća da bude "animation" žanra

Zaključak je da je većina naslova koji su serije upada u "reality" žanr, dok im je ocena ili u najnižem ili najvišem rasponu, a sve su imale premijere u poslednjih tridesetak godina.

Consequent	Antecedent	Support %	Confidence %
encoded_type	encoded_genre = 12 release_year_BIN = 3	4.51	100.0
encoded_type	encoded_genre = 12 imdb_score_BIN = 1	5.188	100.0
encoded_type	encoded_genre = 12	9.845	99.85
encoded_type	encoded_genre = 12 release_year_BIN = 4	5.335	99.724
encoded_type	encoded_genre = 2	4.348	95.593

# Zaključak

Kada je u pitanju klasifikacija, algoritam koji je pokazao najbolje rezultate na ovom skupu je KNN algoritam sa GridSearchCV optimizacijom, sa tačnošću 96% i AUC-om 99.

Najbolje rezultate algoritama klasterovanja dao je algoritam K sredina, sa najboljim razdvajanjem visoko, srednje i nisko ocenjenih naslova.

Zaključak Apriori algoritma u IBM SPSS modeleru je da je većina naslova koji su serije upada u “reality” žanr, dok im je ocena ili u najnižem ili najvišem rasponu, a sve su imale premijere u poslednjih tridesetak godina.