

Progetto di Fondamenti di Analisi dati

STUDENTE : ANDREA FILIPPO SALEMI

CORSO DI LAUREA: INFORMATICA LM-18

DATASET USATO NETFLIX MOVIES AND TV
SHOW ([QUI](#))

Obiettivi

- Lo scopo principale è quello di analizzare questo dataset, ovvero un dataset in cui sono presenti dei dati di un catalogo di netflix fino alla metà del 2021. Il catalogo contiene diversi campi tra cui nome del titolo, il tipo di show , il genere e l'anno di uscita e così via. Nel dataset si sono proposte 5 domande che verranno analizzate nelle prossime slide, lo scopo è anche quello di usare le numerose tecniche durante il corso per lo più:
- Regressione lineare (tendenze temporali) e logistica (classificazioni delle serie e dei film)
- Pearson per le correlazioni;
- Clustering: k-means, PCA per ridurre la dimensionalità;
- Analisi sulla stima di densità utilizzando la kde ;
- Analisi predittiva usando: Naive Bayes e la knn;
- Valutazione tramite: K-fold cross-validation, metriche (Accuracy, Precision, Recall, F1, AUC).

Dataset e Preprocessing

Inizialmente il dataset ha 8807 record (film e serie TV, aggiornato al 2021).

- Variabili principali: *titolo, tipo, regista, cast, paese, anno, rating, durata, generi*

- **Pulizia dei dati (rimuovendo tutti**

- Rimossi duplicati e valori nulli passando da 8807 a 5332 record finali

• Parsing date in formato standard

• **Trasformazioni**

• **Normalizzazione & Standardizzazione** (release year, durata)

• **Encoding:**

• One-hot per generi

• Dummy per rating

Prima
della
One-Hot

```
<class 'pandas.core.frame.DataFrame'>
Index: 5332 entries, 7 to 8806
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  -
 0   show_id         5332 non-null   object  
 1   type            5332 non-null   object  
 2   title           5332 non-null   object  
 3   director        5332 non-null   object  
 4   cast            5332 non-null   object  
 5   country         5332 non-null   object  
 6   date_added      5328 non-null   datetime64[ns]
 7   release_year    5332 non-null   int64   
 8   rating          5332 non-null   object  
 9   duration        5332 non-null   object  
10   listed_in       5332 non-null   object  
11   description     5332 non-null   object  
dtypes: datetime64[ns](1), int64(1), object(10)
memory usage: 541.5+ KB
```

Dopo
La one-
hot

Codifica completata. Esempio delle nuove colonne:

	listed_in	rating_NC-17	rating_NR \
7	Dramas, Independent Movies, International Movies	False	False
8	British TV Shows, Reality TV	False	False
9	Comedies, Dramas	False	False
12	Dramas, International Movies	False	False
24	Comedies, International Movies, Romantic Movies	False	False

	rating_PG	rating_PG-13	rating_R	rating_TV-14	rating_TV-G \
7	False	False	False	False	False
8	False	False	False	True	False
9	False	True	False	False	False
12	False	False	False	False	False
24	False	False	False	True	False

	rating_TV-MA	rating_TV-PG	rating_TV-Y	rating_TV-Y7	rating_TV-Y7-FV \
7	True	False	False	False	False
8	False	False	False	False	False
9	False	False	False	False	False
12	True	False	False	False	False
24	False	False	False	False	False

	rating_UR
7	False
8	False
9	False
12	False
24	False

Analisi Descrittiva

Distribuzione anni di uscita

- Maggior parte dei contenuti pubblicati dopo il 2015
- Forte crescita recente del catalogo

Durata dei film

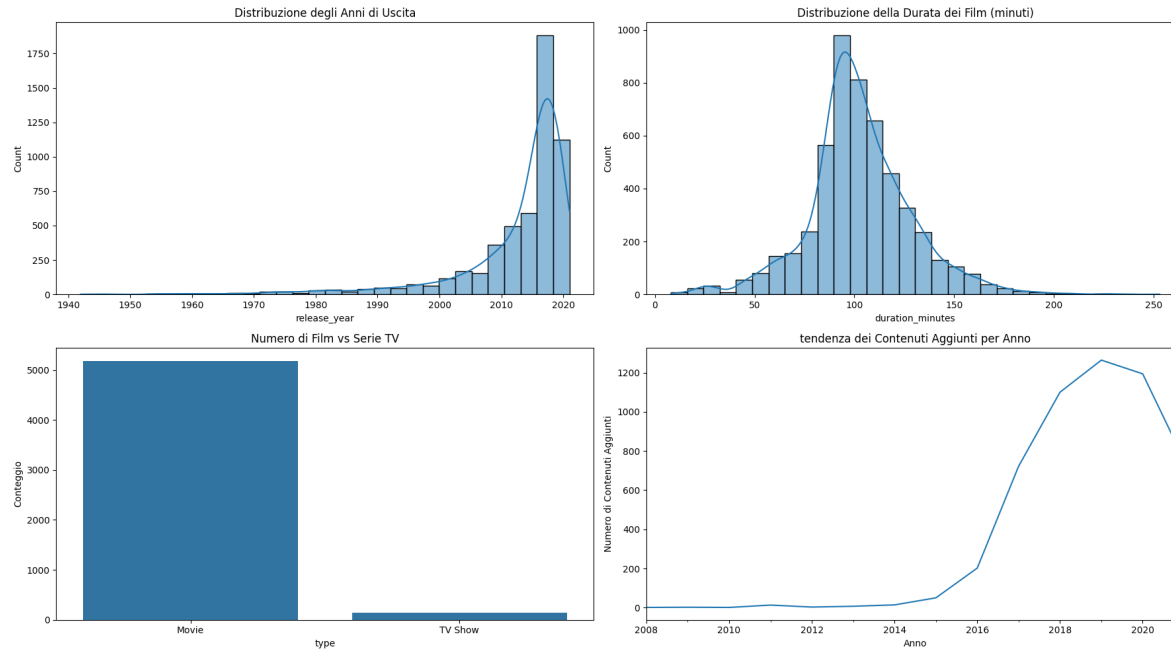
- Distribuzione ~ normale tra 80–120 min
- Pochi film molto brevi o molto lunghi

Film vs Serie TV

- I film dominano il catalogo
- Le serie TV crescono solo negli ultimi anni

Trend contenuti aggiunti

- Crescita marcata dal 2016 al 2019
- Leggera diminuzione negli anni successivi





Analisi Inferenziale

ANOVA (durata film per rating)

- Differenze significative tra i rating ($p < 0.05$)
- Es.: TV-14 \approx 114 min, TV-MA \approx 97 min

Correlazione (anno di uscita – durata)

- Correlazione negativa significativa ($r \approx -0.21$)
- I film più recenti tendono a durare meno

Chi-quadro (genere vs rating)

- Associazione significativa ($p < 0.001$)
- Alcuni rating più frequenti in certi generi (es. Dramas, Comedies)

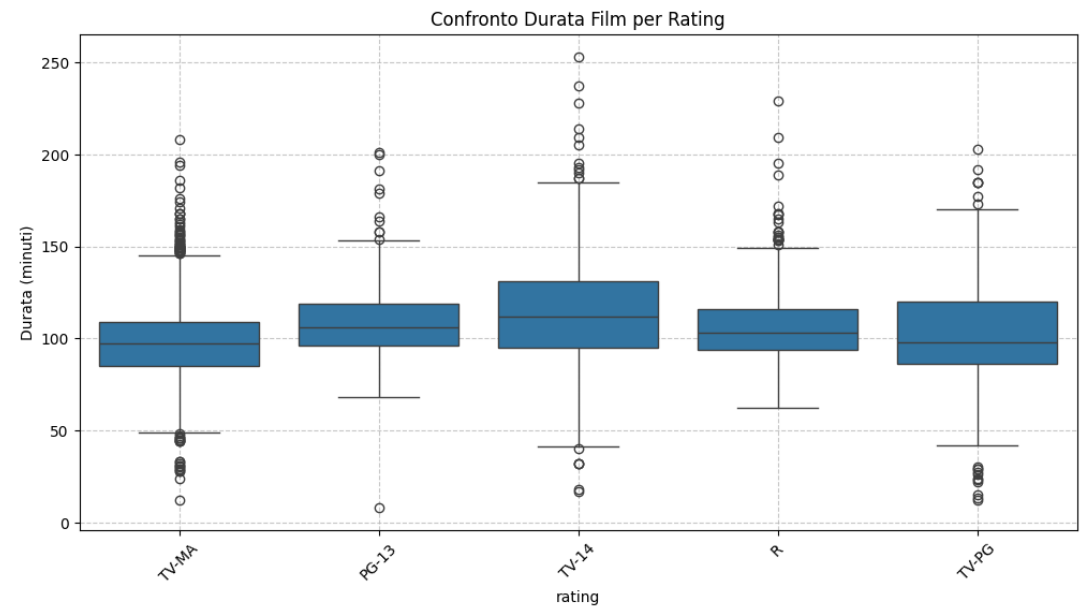
ANOVA (durata film per decade)

- Differenze significative tra decenni
- Film anni '90-2000 più lunghi in media rispetto a quelli 2010+

1) Per durata film tra i top 5 rating

Analisi inferenziale

1. **ANOVA** per durata film tra i top 5 rating:
F-statistic: 89.5722, p-value: 0.0000000000
Durata media per TV-MA: 97.41 minuti
Durata media per TV-14: 113.90 minuti
Durata media per R: 106.68 minuti
Durata media per PG-13: 108.90 minuti
Durata media per TV-PG: 100.58 minuti

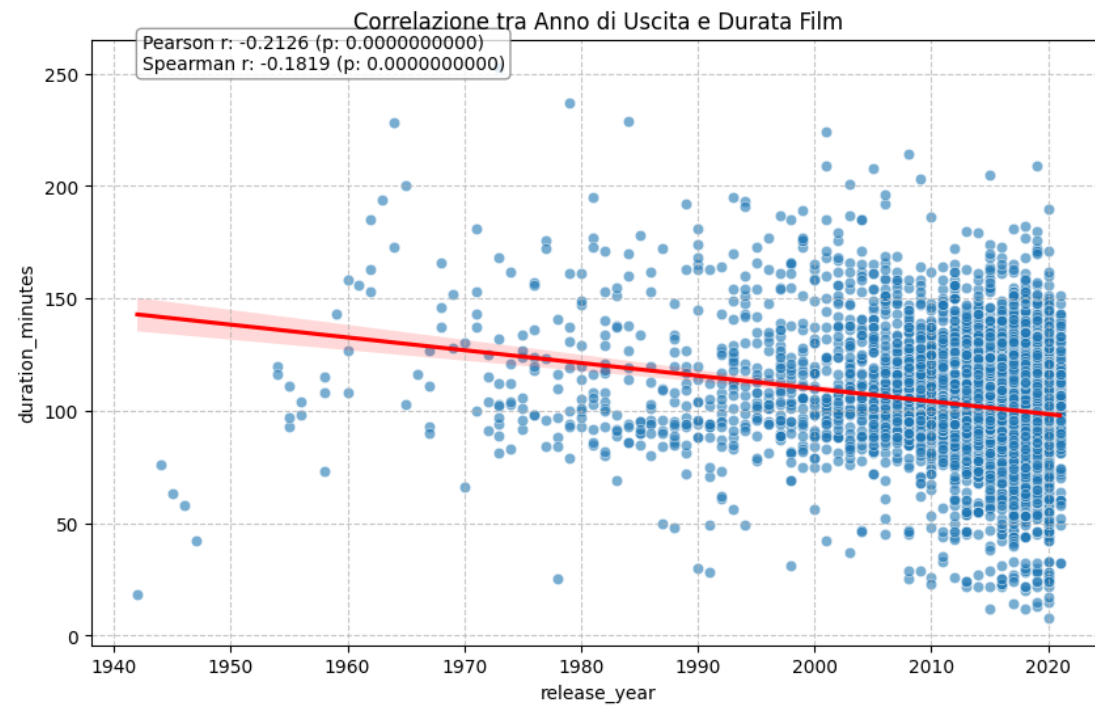


2) Correlazione tra anno di uscita e durata film

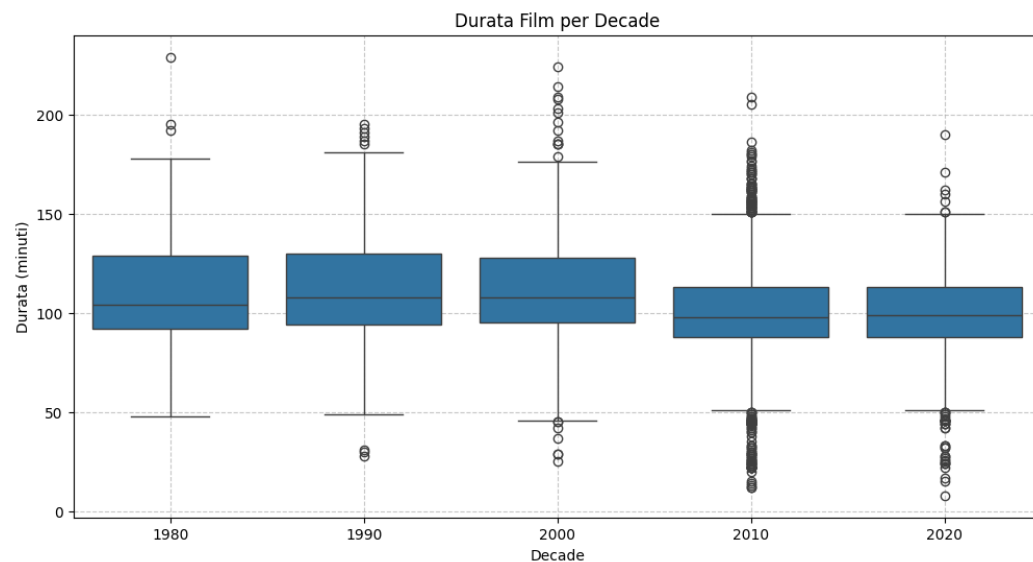
2. Correlazione tra Anno di Uscita e Durata Film:

Correlazione Pearson: -0.2126 , p-value: 0.0000000000

Correlazione Spearman: -0.1819 , p-value: 0.0000000000



Test Chi-quadro Generi vs Rating (Film) e ANOVA durata film per decade:



3. Test Chi-quadro Generi vs Rating (Film):

Dramas: Chi-square = 219.8250, p-value = 0.000000

Comedies: Chi-square = 172.3763, p-value = 0.000000

International Movies: Chi-square = 1543.7043, p-value = 0.000000

4. ANOVA durata film per decade:

F-statistic: 61.6206, p-value: 0.0000000000

Durata media anni 2010s: 99.96 minuti

Durata media anni 2000s: 112.91 minuti

Durata media anni 2020s: 97.80 minuti


Durata media anni 1990s: 114.64 minuti

Durata media anni 1980s: 113.43 minuti



Interpretazione generale dell'analisi

- Esistono differenze reali nelle durate in base a rating e periodo storico.
- Il rating e il genere influenzano in modo significativo le caratteristiche dei film.

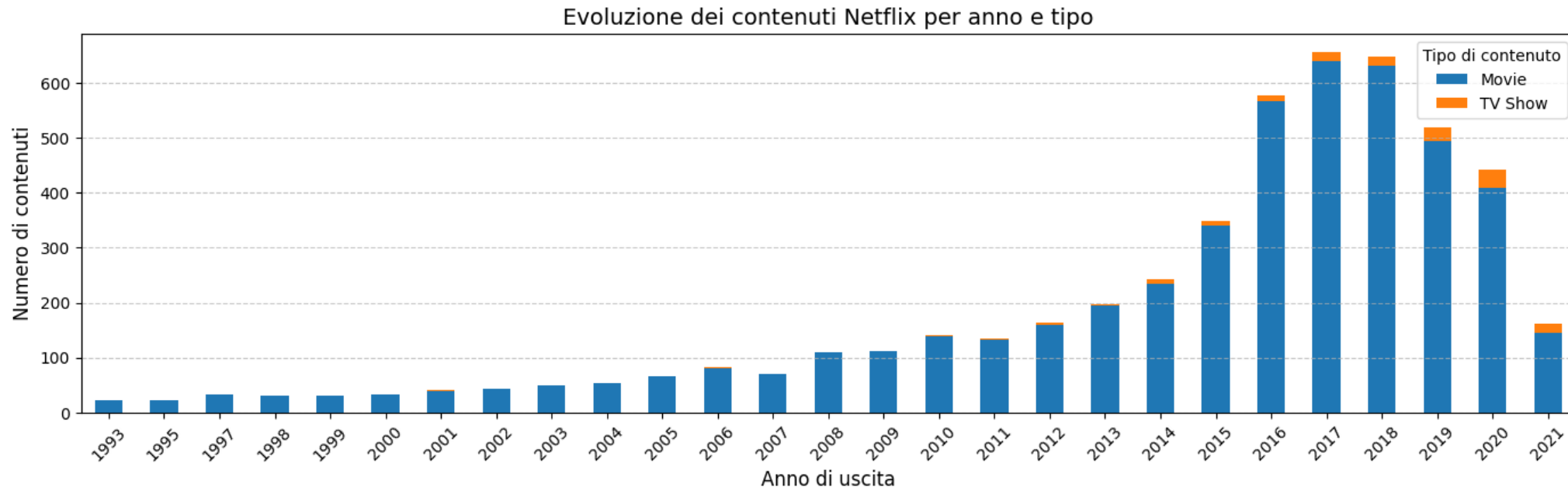


1) Com'è cambiata la produzione dei contenuti durante gli anni? (se col tempo sono state fatte più serie TV o più film)

- **Approccio adottato:**
- **Conteggio annuale** di Film e Serie TV (trend nel tempo)
- **Percentuali** : composizione del catalogo per anno
- **Regressione lineare** : stima dei trend di crescita
- **Test statistici (linregress)** : verifica della significatività
- **Confronto pre-2010 vs post-2010** : cambiamenti strategici
- **CAGR (tasso di crescita annuo composto)** : misura della crescita media

1) Com'è cambiata la produzione dei contenuti durante gli anni? (se col tempo sono state fatte più serie TV o più film)

- Fino al 2010 : produzione bassa, quasi solo film
 - Dal 2015 : crescita esplosiva, picco 2017–2018
 - Serie TV in aumento, ma i film restano prevalenti
 - Dopo il 2018 : leggera diminuzione (es. pandemia)
- Fino al 2010 : produzione bassa, quasi solo film
Dal 2015 → crescita esplosiva, picco 2017–2018
Serie TV in aumento, ma i film restano prevalenti
Dopo il 2018 : leggera diminuzione (es. pandemia)



1) Com'è cambiata la produzione dei contenuti durante gli anni? (se col tempo sono state fatte più serie TV o più film)

Totale (viola)

- Forte crescita fino al 2018 : +19,6 contenuti/anno
- Lieve calo negli anni successivi

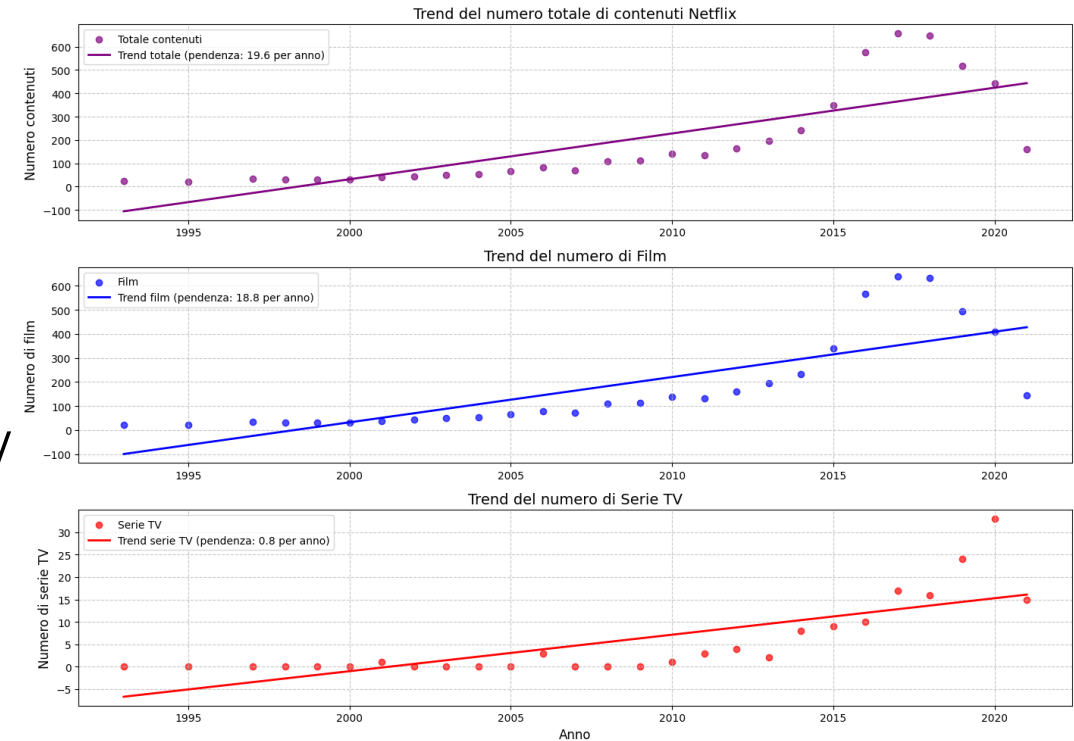
Film (blu)

- Pendenza $\approx +18,8$: crescita trainata quasi solo dai film
- Picco nel 2017-2018

Serie TV (rosso)

- Pendenza $\approx +0,8$: crescita lenta ma costante
- Più rilevanti solo negli ultimi anni

La crescita del catalogo è trainata dai **film**, ma le **serie TV** stanno diventando progressivamente più presenti.





Evoluzione dei contenuti Netflix

- Evoluzione dei contenuti Netflix
- Crescita continua di **Film** e **Serie TV** negli anni
- **Serie TV**: incremento più marcato e trend significativo (assoluto e percentuale)
- **Film**: crescita più lenta e stabile
- **Confronto pre-2010 vs post-2010** → cambio di strategia, maggiore focus sulle Serie TV
- **CAGR** → crescita media annua più alta per le Serie TV rispetto ai Film

In conclusione, Netflix ha reso le Serie TV un pilastro centrale del catalogo, soprattutto nell'ultimo decennio.

2) Una relazione tra tipo di contenuto (Film e SerieTV) e i loro rating.

L'obiettivo di questo quesito è di verificare se il **rating** dipende dal tipo di contenuto (Film / Serie TV).

Il metodo Metodo usato:

Creazione di **tabella di contingenza (crosstab)** usando :

```
content_rating_table = pd.crosstab(df['type'],  
                                   df['rating'])
```

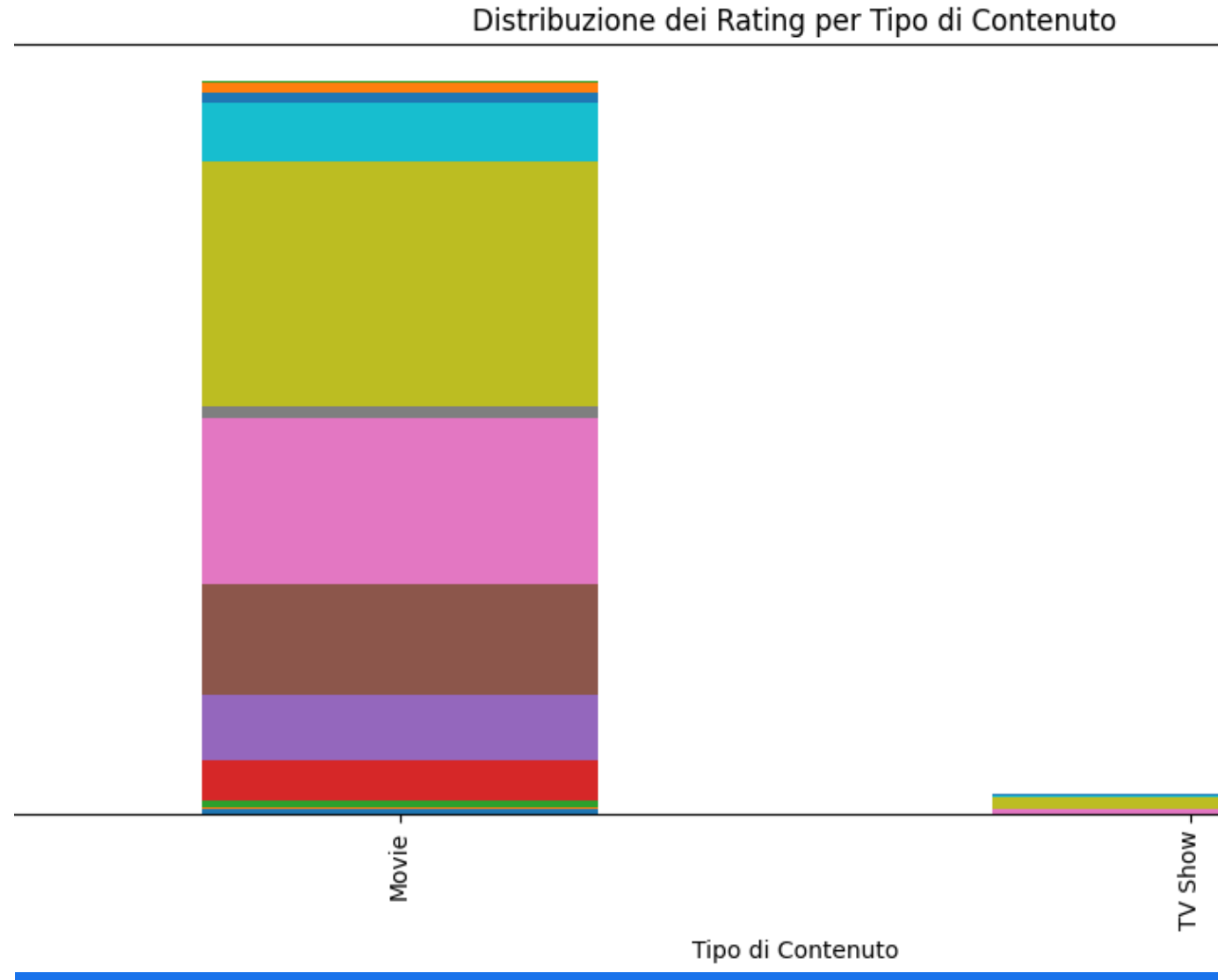
- Applicazione del **Test Chi-Quadro (X^2)**:

```
chi2, p_value, dof, expected =  
stats.chi2_contingency(content_rating_table)
```

- Calcolo valori attesi e gradi di libertà.

L' **Interpretazione** ci permette di capire se alcuni **rating** sono più comuni nei Film o nelle Serie TV, identificando le differenze **statisticamente significative**

- **Risposta al quesito dai risultati :**
Come si può notare, a prima vista : I TV Show sono decisamente di meno questo è probabilmente dovuto alla possibilità per netflix di poter pubblicare più film rispetto a una Serie TV poiché si diladono nel tempo.

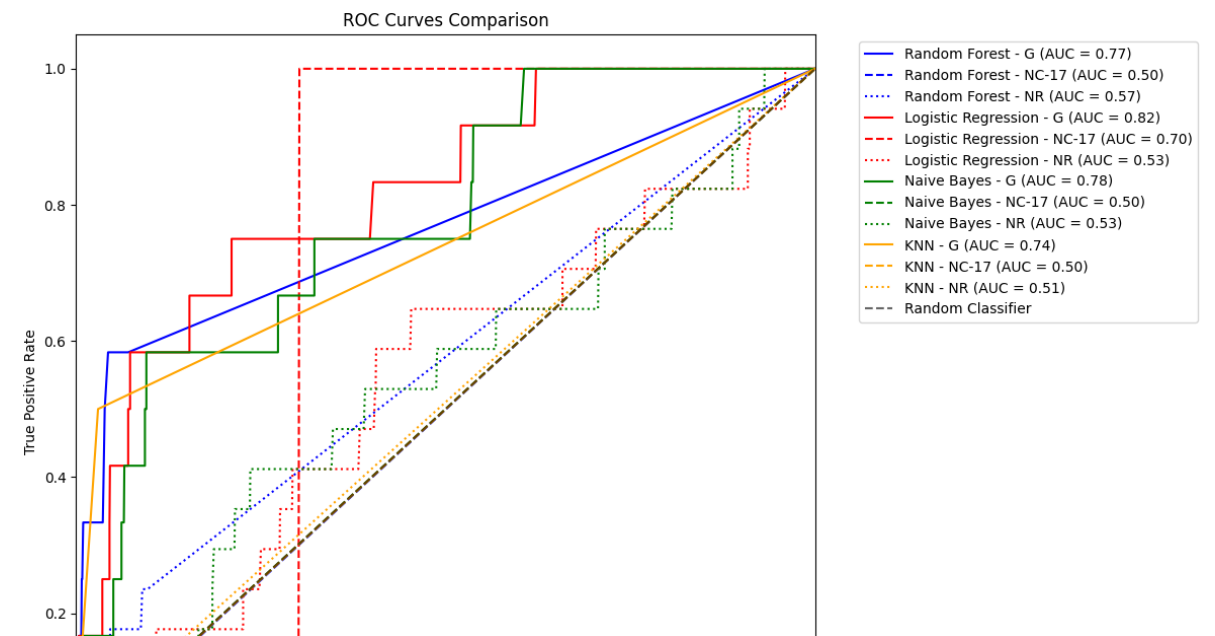
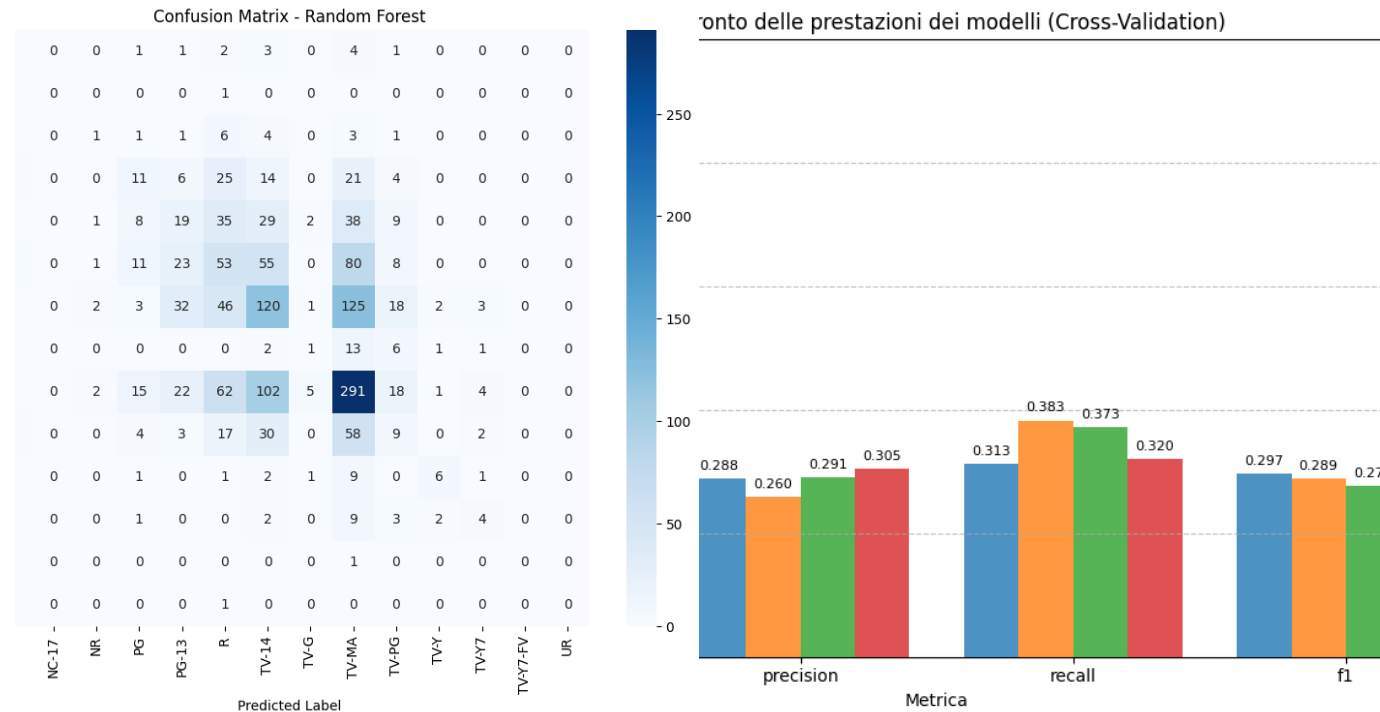


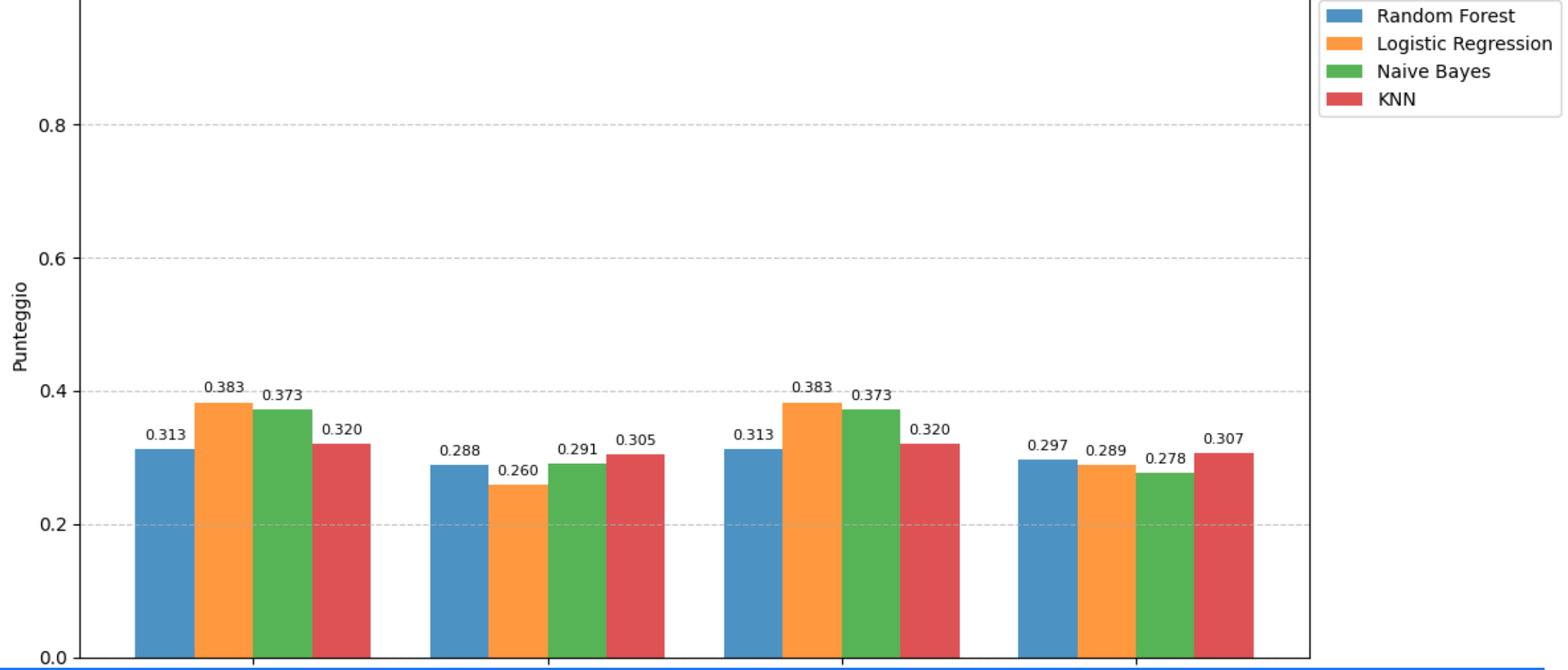
3) adesso si andrà a rispondere alla terza domanda che tratta della : Predizione fascia di rating delle serie TV e film?

- Nel seguente quesito si costruisce e si valuta un modello di classificazione prevedendo il rating dei titoli di Netflix usando delle caratteristiche come il tipo , la durata dei film o serie TV.
- per fare tutto questo verrà applicazione la normalizzazione (Standardizzazione), ci sarà una fase di training e in fine di predizione e valutazione per poter predire la fasce di rating e verranno visualizzate i risultati usando le Curve di ROC e le matrici di confusioni il quale rispettivamente serve per valutare tra le diverse classi di rating, mostrando il compromesso tra il tasso di veri positivi e il tasso di falsi positivi per ciascuna classe per capire, appunto sia efficace per ogni punto. più è vicino a 1 , migliore sarà la performance. Mentre, la matrice di confusione mostra ogni classe, mostrando quante volte ha predetto correttamente (valori in diagonale) mentre, viceversa, valori non correttamente non corretti (fuori dalla diagonale) questo fa in modo che ci siano più o meno fuori efficiente il modello di classificazione.
- Tecniche usate sono: Clustering (K-means), PCA (per ridurre la dimensionalità e visualizzare i cluster).

Plot:

Plot:





Confronto delle prestazioni dei modelli (Cross-Validation)

- Questo grafico a barre confronta le performance di quattro modelli di classificazione (Random Forest, Logistic Regression, Naive Bayes, KNN) nella predizione del rating dei contenuti Netflix. Le metriche mostrate sono Accuracy, Precision, Recall e F1-score, calcolate tramite validazione incrociata (cross-validation).
- Si può subito vedere quale modello ottiene i punteggi migliori su ciascuna metrica.
- Permette di identificare il modello più affidabile e robusto per il problema affrontato.

Curve di ROC confronto

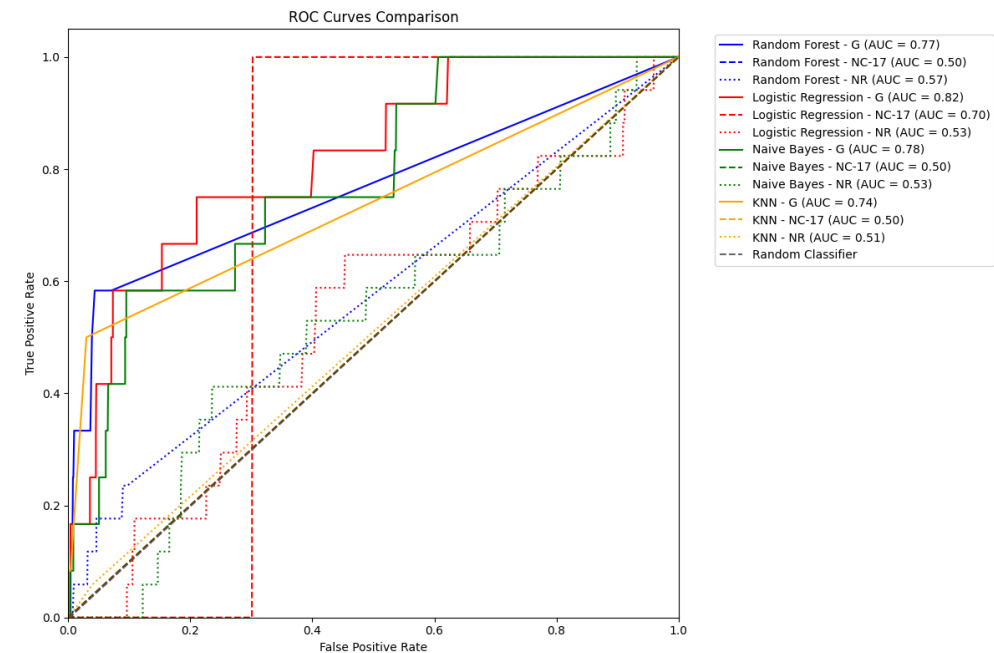
Il grafico mostra le curve ROC (Receiver Operating Characteristic) per i principali modelli e per alcune classi di rating.

Cosa mostra:

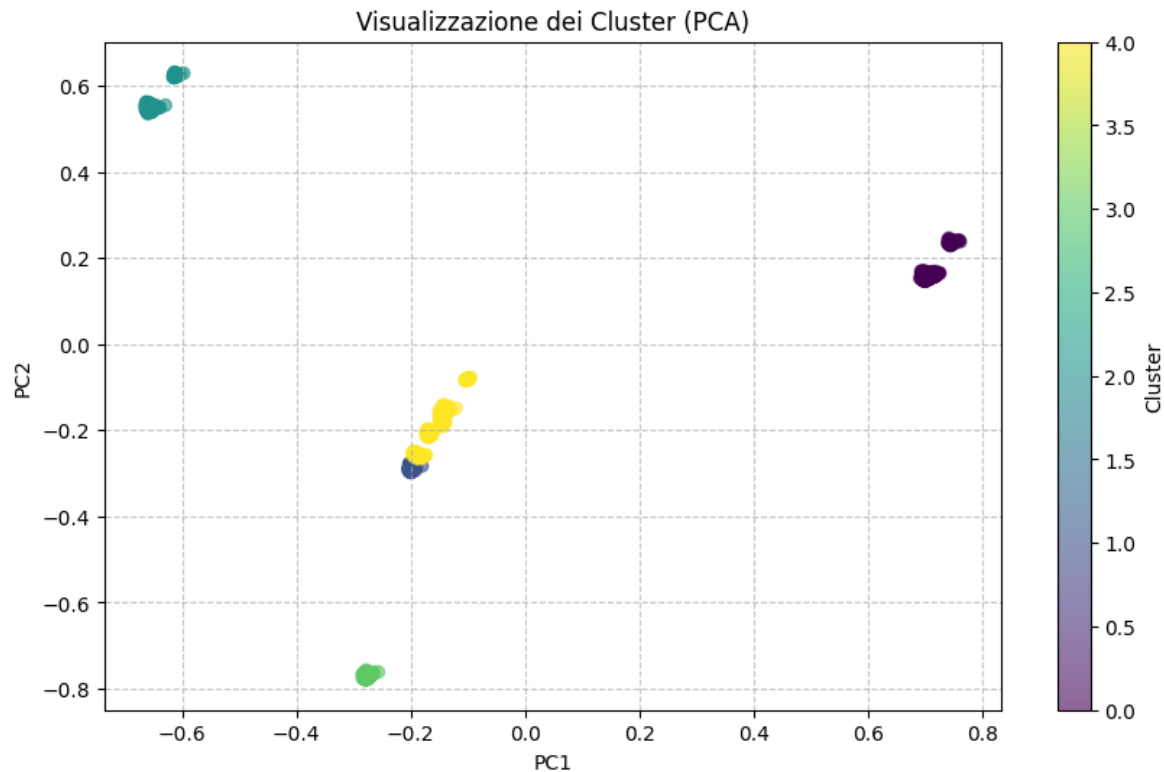
Ogni curva rappresenta la capacità del modello di distinguere tra una classe e le altre (one-vs-rest).

Più la curva si avvicina all'angolo in alto a sinistra, migliore è la capacità predittiva (AUC più alto).

Il confronto tra modelli evidenzia quali sono più efficaci nel riconoscere le diverse fasce di rating.



È possibile identificare gruppi naturali di contenuti (cluster) basati su caratteristiche testuali e categorica?



Obiettivo:

- Raggruppare i titoli in **5 cluster** usando descrizione + tipo + rating
- Visualizzare i cluster nello spazio ridotto con **PCA**

Tecniche:

- **K-Means** : crea gruppi di contenuti simili senza etichette predefinite
- **PCA** : riduce la dimensionalità, facilita visualizzazione e interpretazione

Risultati Clustering (KMeans + PCA):

- **5 cluster distinti** di titoli Netflix, basati su descrizioni testuali + variabili categoriche (tipo, rating);
- **PCA** → i cluster risultano **ben separati** nello spazio ridotto
- Ogni cluster mostra **caratteristiche comuni** (generi, tipologie, rating prevalenti)

Capire la struttura del catalogo: Segmentare i contenuti per strategie di produzione/distribuzione e supportare sistemi di raccomandazione personalizzati



5) Distribuzione delle uscite delle serie TV per capire se ci sono dei picchi significativi?

Il principale obiettivo è identificare i picchi e i trend di crescita/diminuzione nelle uscite di Serie TV con l'applicazione dei seguenti metodi:

- **Analisi descrittiva** : media, deviazione standard, z-score per rilevare picchi
- **Visualizzazioni** : grafici a barre, linee, heatmap (uscite per anno, % Serie TV, distribuzione mensile)
- **Regressione lineare** : stima del trend delle uscite nel tempo

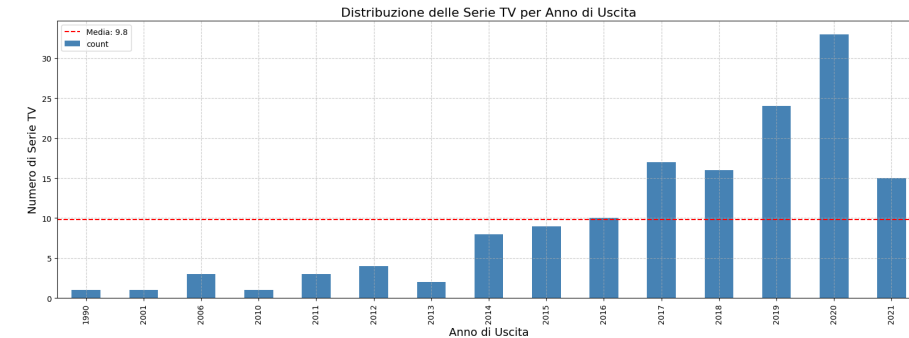
L'analisi mirata serve per capire l'evoluzione delle Serie TV e i momenti di crescita più significativi.

Plot:

1. Distribuzione delle Serie TV per Anno di Uscita (grafico in alto):

Mostra il numero di nuove Serie TV pubblicate ogni anno su Netflix.

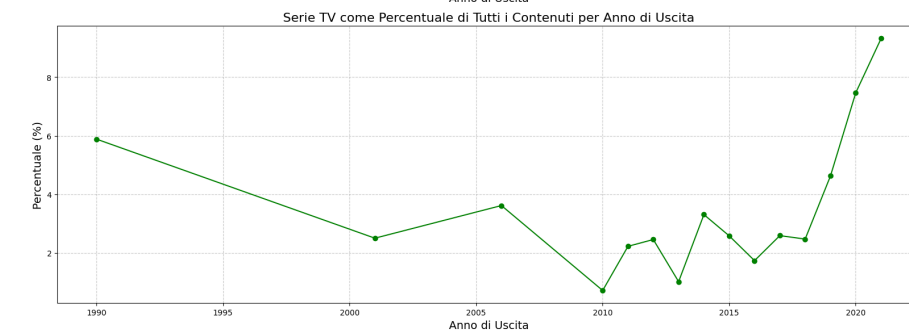
- Le barre blu rappresentano il conteggio annuale.
- La linea rossa tratteggiata indica la media storica.
- Si notano alcuni anni con valori nettamente superiori alla media: questi sono i “picchi” produttivi, che evidenziano strategie di espansione o cambiamenti di mercato.



2. Serie TV come Percentuale di Tutti i Contenuti (grafico centrale):

Mostra la quota percentuale delle Serie TV rispetto al totale dei contenuti pubblicati ogni anno.

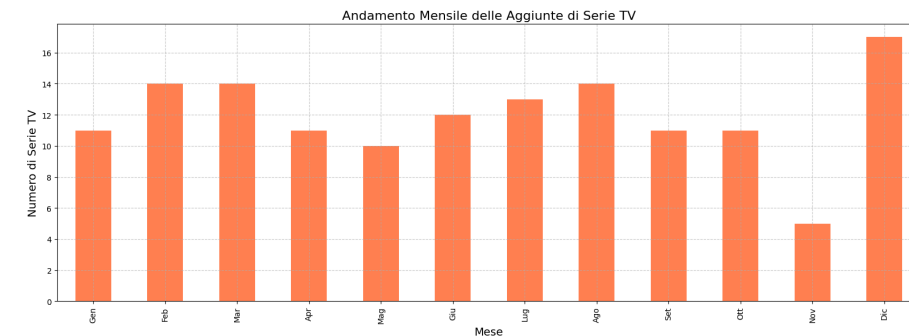
La linea verde evidenzia come la presenza delle Serie TV sia cresciuta nel tempo, soprattutto negli ultimi anni, diventando una parte sempre più rilevante del catalogo Netflix.



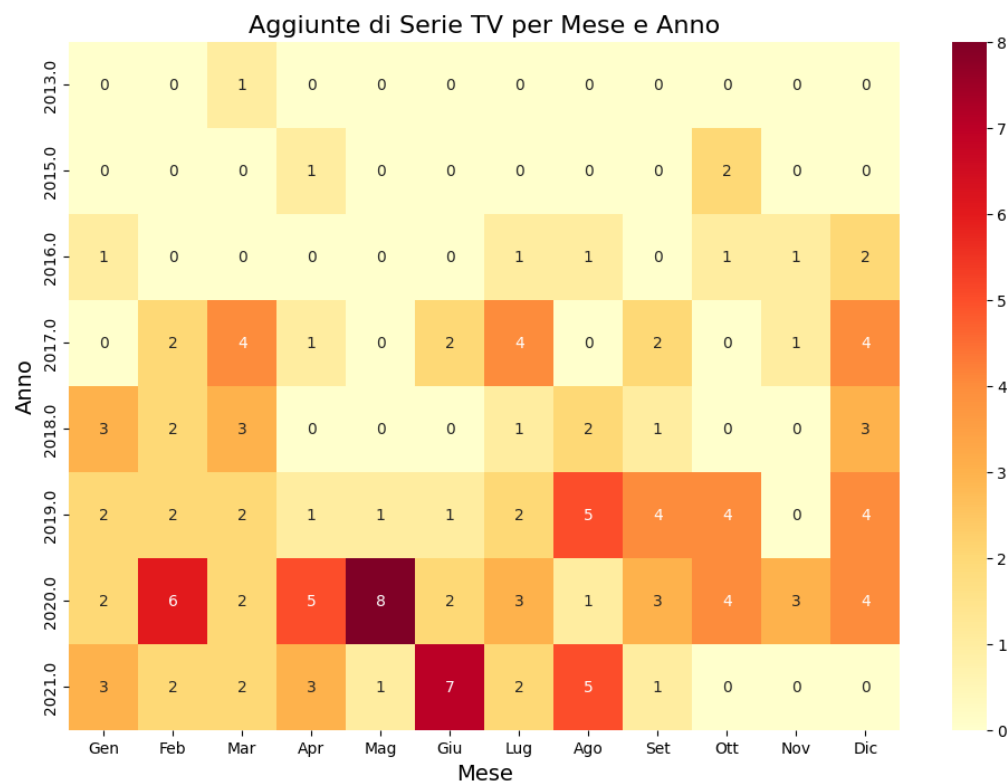
3. Andamento Mensile delle Aggiunte di Serie TV (grafico in basso):

Mostra la distribuzione delle nuove Serie TV aggiunte mese per mese.

Le barre arancioni indicano che le aggiunte sono distribuite durante tutto l'anno, ma con alcuni mesi leggermente più “ricchi”, suggerendo una certa stagionalità nelle pubblicazioni.



Plot : Matrice di confusione



Heatmap uscite Serie TV (2013–2021)

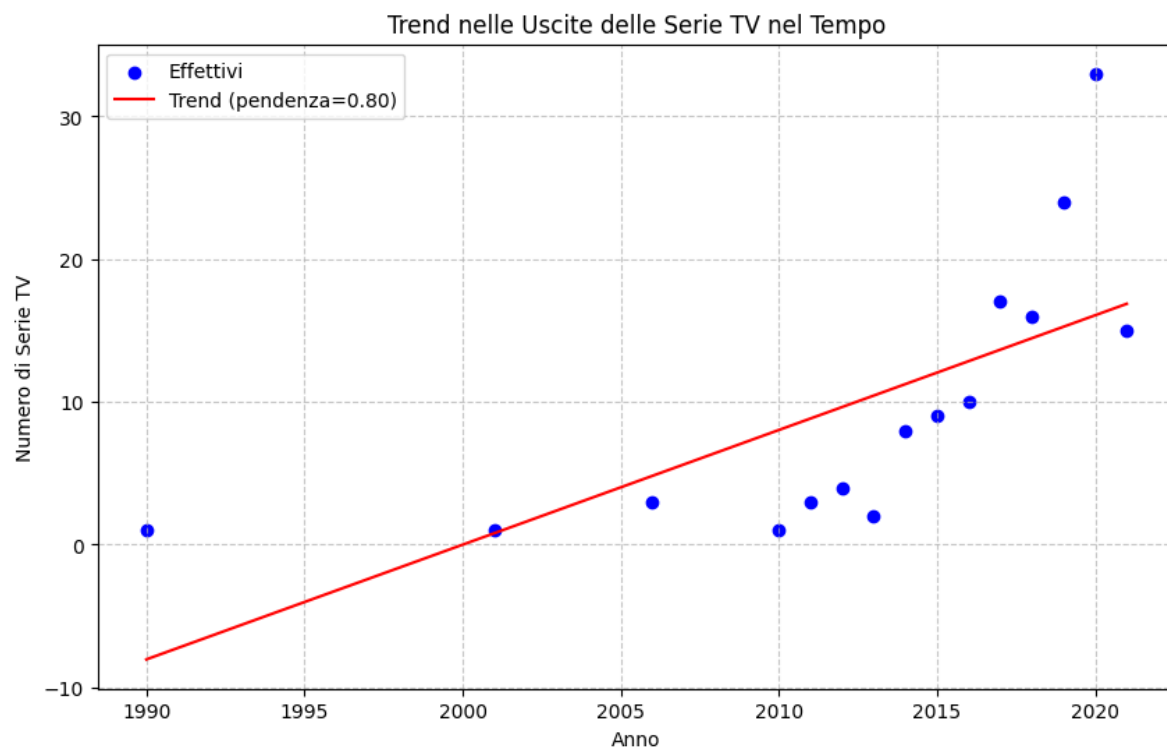
Come leggerla:

- Righe = anni
- Colonne = mesi
- Colore = numero di nuove Serie TV

Cosa mostra:

- **Picchi evidenti** in mesi specifici (es. maggio–giugno 2020)
- **Crescita nel tempo**: più uscite negli anni recenti
- **Stagionalità**: alcuni mesi più “ricchi” di aggiunte
- **Conclusione**: la produzione di Serie TV è aumentata negli ultimi anni, con pattern stagionali ben visibili.

Plot: Trend uscite delle serie TV nel tempo



•Grafico

- Punti blu = numero di Serie TV per anno
- Linea rossa = trend (regressione lineare, pendenza $\approx +0.8$ Serie TV/anno)

•Cosa evidenzia

- Crescita costante delle uscite negli ultimi anni
- Trend positivo = Netflix investe sempre di più nelle Serie TV
- Serie TV diventano parte **centrale** del catalogo
- Si può dedurre la produzione di Serie TV su Netflix è aumentata in modo continuo e significativo.



Conclusione 5° quesito

- Picchi produttivi individuati con $z\text{-score} > 2$
- Percentuale Serie TV in crescita costante sul totale
- Regressione lineare : coefficiente positivo = trend crescente
- Analisi mensile + heatmap : evidenziano stagionalità nelle pubblicazioni
- **Conclusione:** Netflix ha investito sempre di più nelle Serie TV, rendendole una parte centrale del catalogo.



Tirando le somme

Evoluzione produzione

- Crescita costante dei contenuti
- Serie TV sempre più centrali (soprattutto dopo il 2010)

Tipo ↔ Rating

- Relazione significativa (Chi-quadro)
- Rating diversi per Film e Serie TV → strategie di targetizzazione

Predizione rating

- Random Forest = modello più accurato
- Anno, durata e tipo → già predittivi del rating

Clustering contenuti

- 5 cluster naturali (testo + variabili categoriali)
- Utile per raccomandazioni e analisi di mercato

Picchi Serie TV

- Anni con z-score > 2 → strategie aziendali/mercato
- Presenza di stagionalità nelle aggiunte

Sintesi finale: Netflix ha ampliato e diversificato l'offerta, puntando sempre più sulle Serie TV e adattando i contenuti a rating e target.