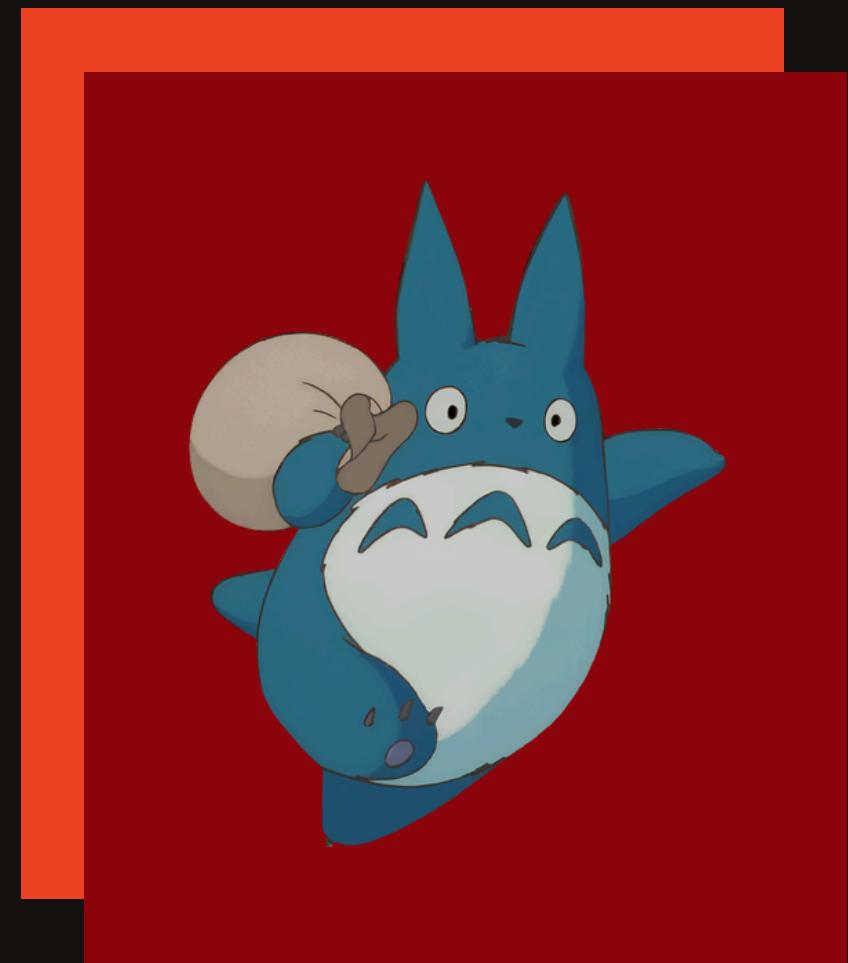




fidelio

Chi sono



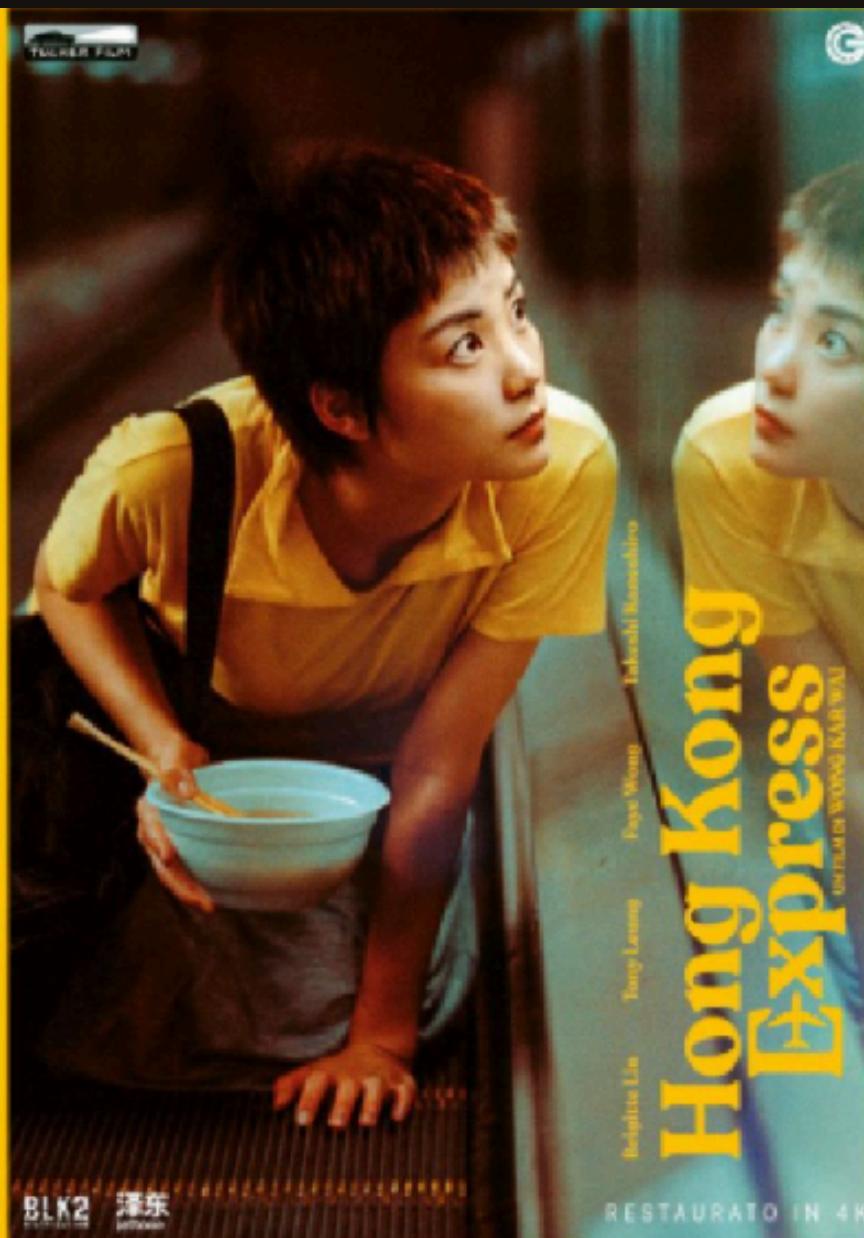
Tullo Nikolas
Matr: 0512119040

Il paradosso della scelta

L'aumento esponenziale dei film disponibili sulle piattaforme digitali ha reso il processo di scelta sempre più complesso per lo spettatore.



Obiettivi

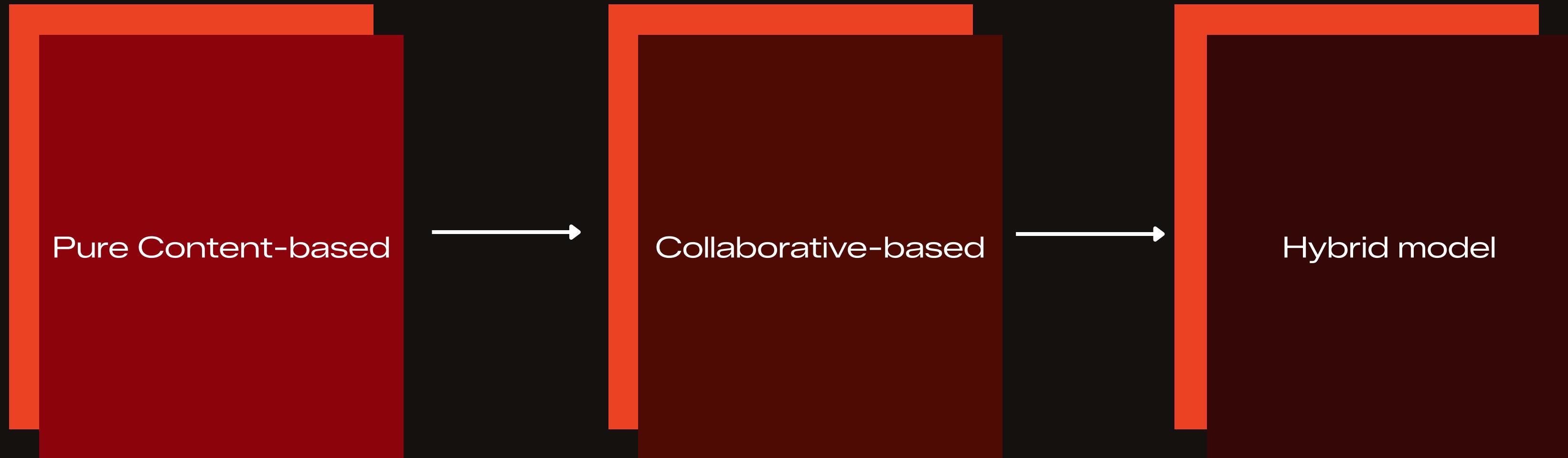


Creare un modello di Intelligenza Artificiale che modelli le preferenze dell'utente e utilizzi queste informazioni per ordinare e suggerire i film più rilevanti, come:

- Integrando collaborative filtering, content-based e user/critic features
- Ottimizzando MAP@K, Precision@K e Recall@K
- Valutando il contributo di ogni componente tramite ablation study
- Garantendo scalabilità su dataset reali

Approccio

Modelli



O1. MovieLens + RottenTomatoes



Il dataset di interazioni di MovieLens fornisce informazioni su circa 3Mln di informazioni raccolte dai rating degli utenti sui film, sul quale è stato fatto un merge con il dataset da 1.4Mln di interazioni di RottenTomatoes



O2. TMDB API



Attraverso TMDB API sono stati recuperati originariamente 85,235 record di metadati su film.

DATASET





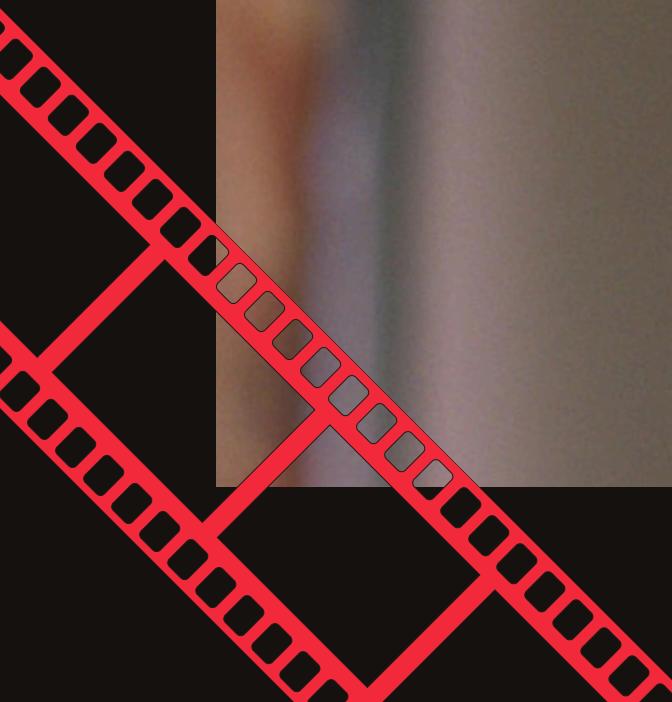
O1. Intersezione tra dataset

Ho effettuato una inner join tra il dataset di film e quello dei credits, senza questa operazione il modello content based non funzionerebbe.

O2. Filtraggio cast

Sono stati mantenuti solo i top 5 attori per film. Poiché in media sono quelli più influenti nella scelta, (Protagonisti, Antagonisti, ruoli di supporto primari etc...)

PREPROCESSING

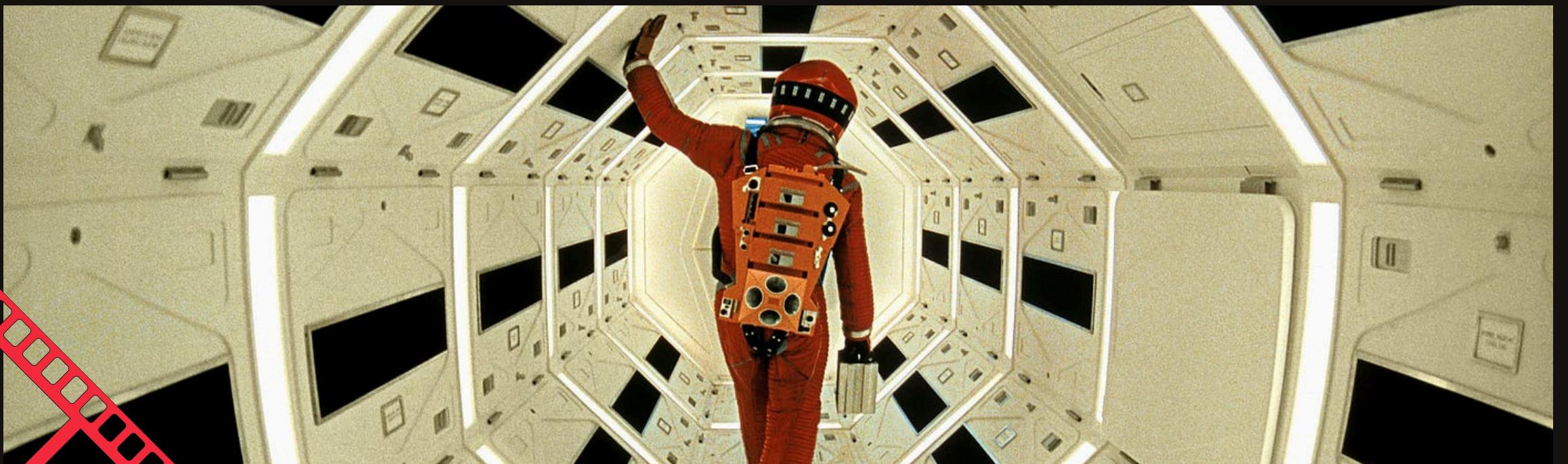


Osservando il dataset delle critic reviews mi sono accorto che molti record, seppur contenenti la recensione in forma testuale non contenevano il voto finale assegnato al film.

Ho imputato questi dati mancanti usando **VADER** (Valence Aware Dictionary and Sentiment Reasoner) un lexicon-rule based sentiment analyzer che produce un Compound Sentiment Score C in [-1 , 1], che va da estremamente negativo ad estremamente positivo in base al contenuto della recensione. E' stata poi applicata una normalizzazione allo score nella nostra scala.

IMPUTAZIONE: VADER

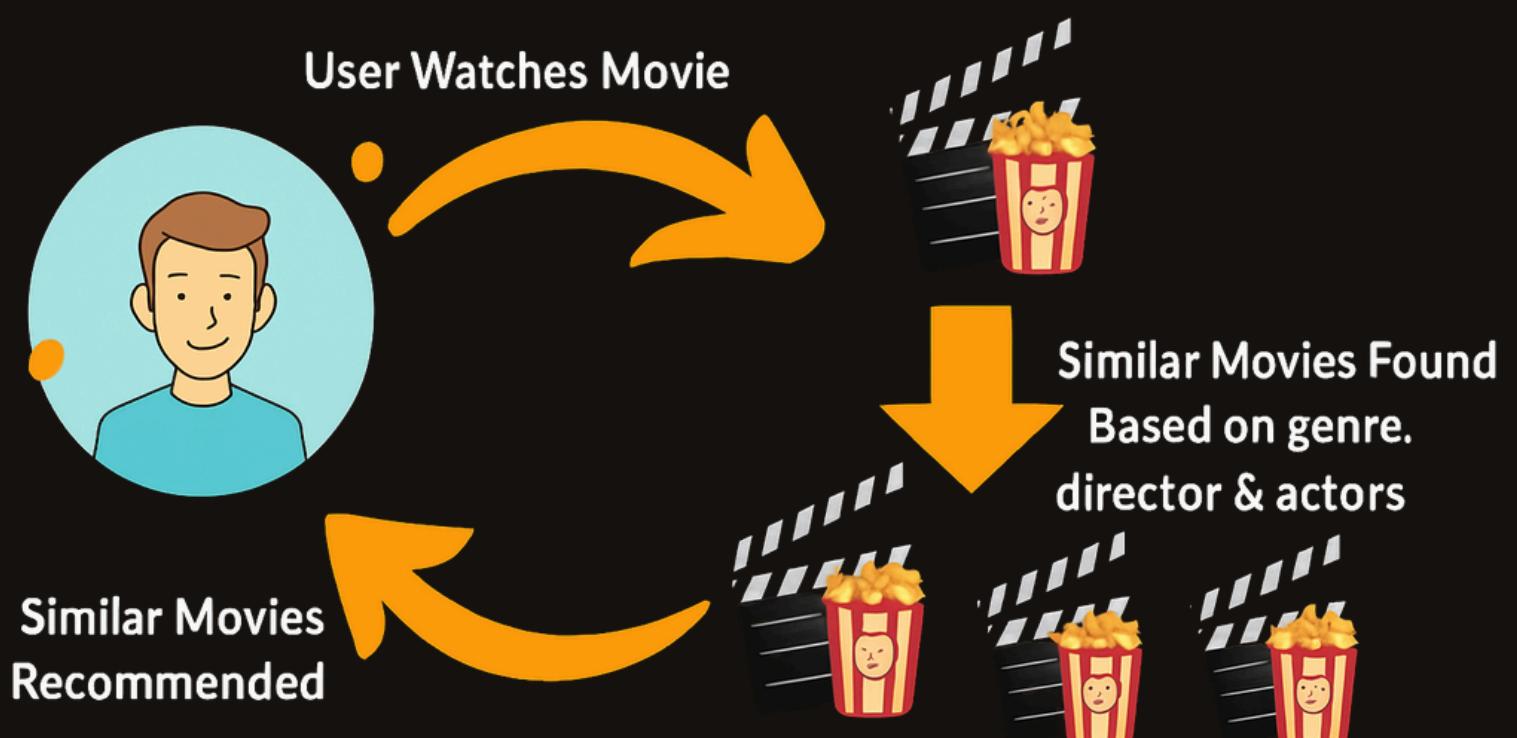
$$\hat{y} = \frac{C + 1}{2} \times 5$$



1. Content-based filtering

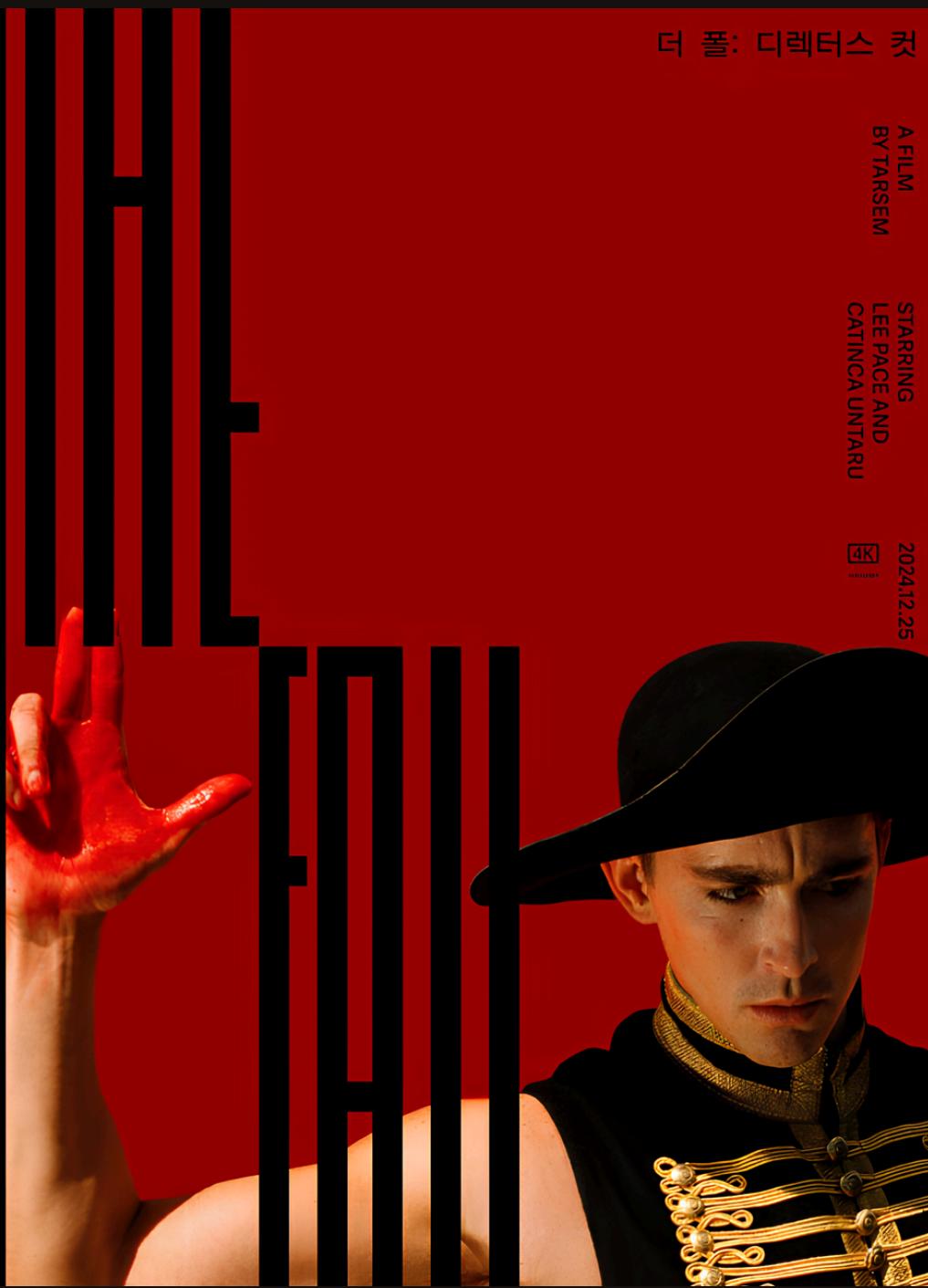
L'approccio content-based raccomanda film simili a quelli apprezzati dall'utente, basandosi su attributi come generi, cast e regista.

Crea una "zuppa" testuale unificata per ogni film, la vettorializza, costruisce un profilo utente come media pesata dei film rated, e seleziona i top K non visti.



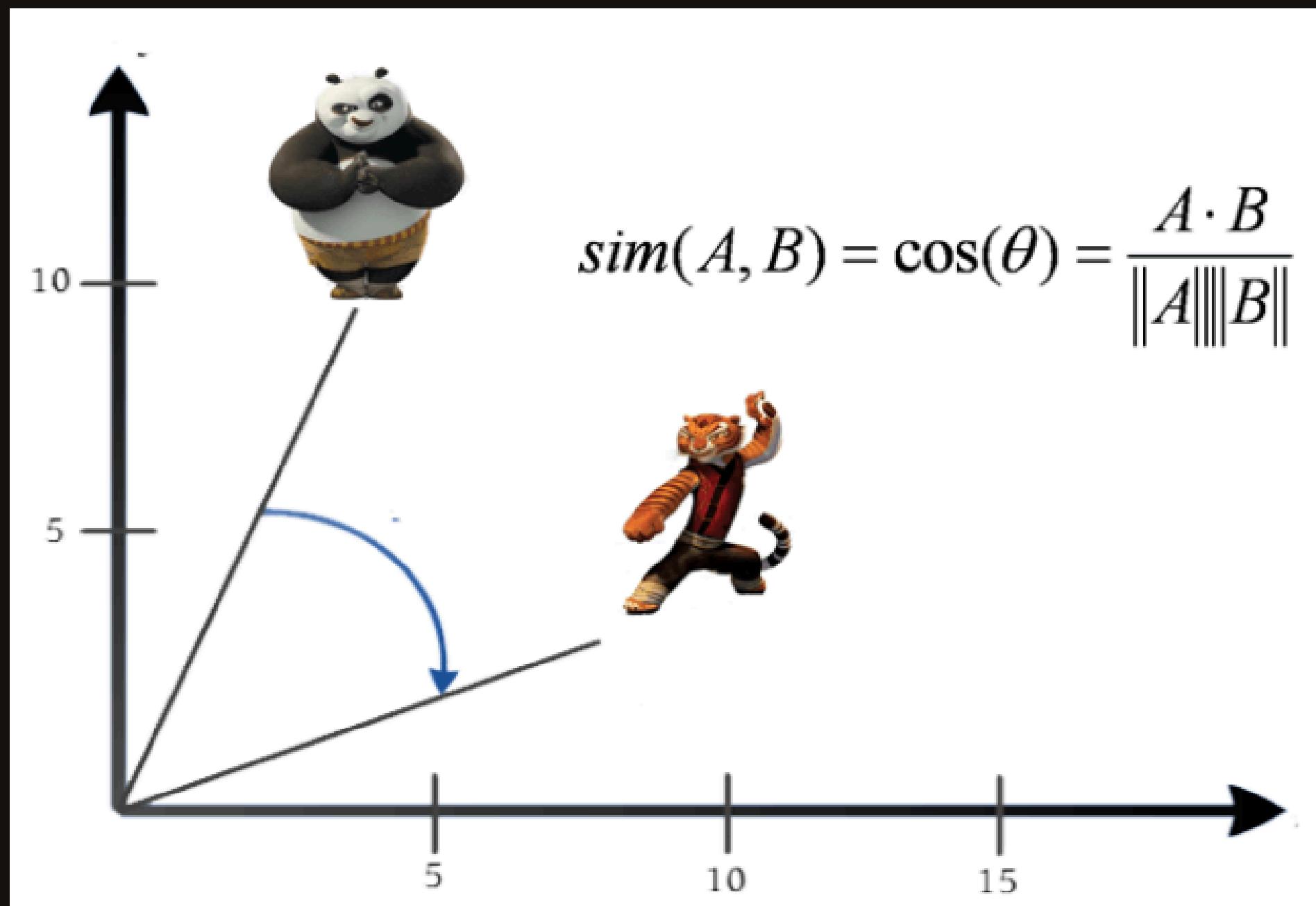
Come funziona?

Content Based filtering



- 1) Costruisce una rappresentazione testuale unificata, "zuppa", degli attributi dei film (genere, regista, durata, attori...). Ex: actionadventure christophernolan christianbale michaelcaine.
- 2) Trasformo la zuppa in un vettore impiegando TF-IDF (Term Frequency - Inverse Document Frequency). Il peso di una parola all'interno di un documento è inversamente proporzionale al suo numero di presenze. I vettori vengono poi normalizzati attraverso L2.
- 3) Confrontiamo i vettori risultanti stabilendone la Cosine Similarity
 - 0 per vettori scorrelati (ortogonali).
 - 1 per vettori positivamente correlati (parallel).
 - -1 per vettori negativamente correlati (opposti).

Ad esempio...



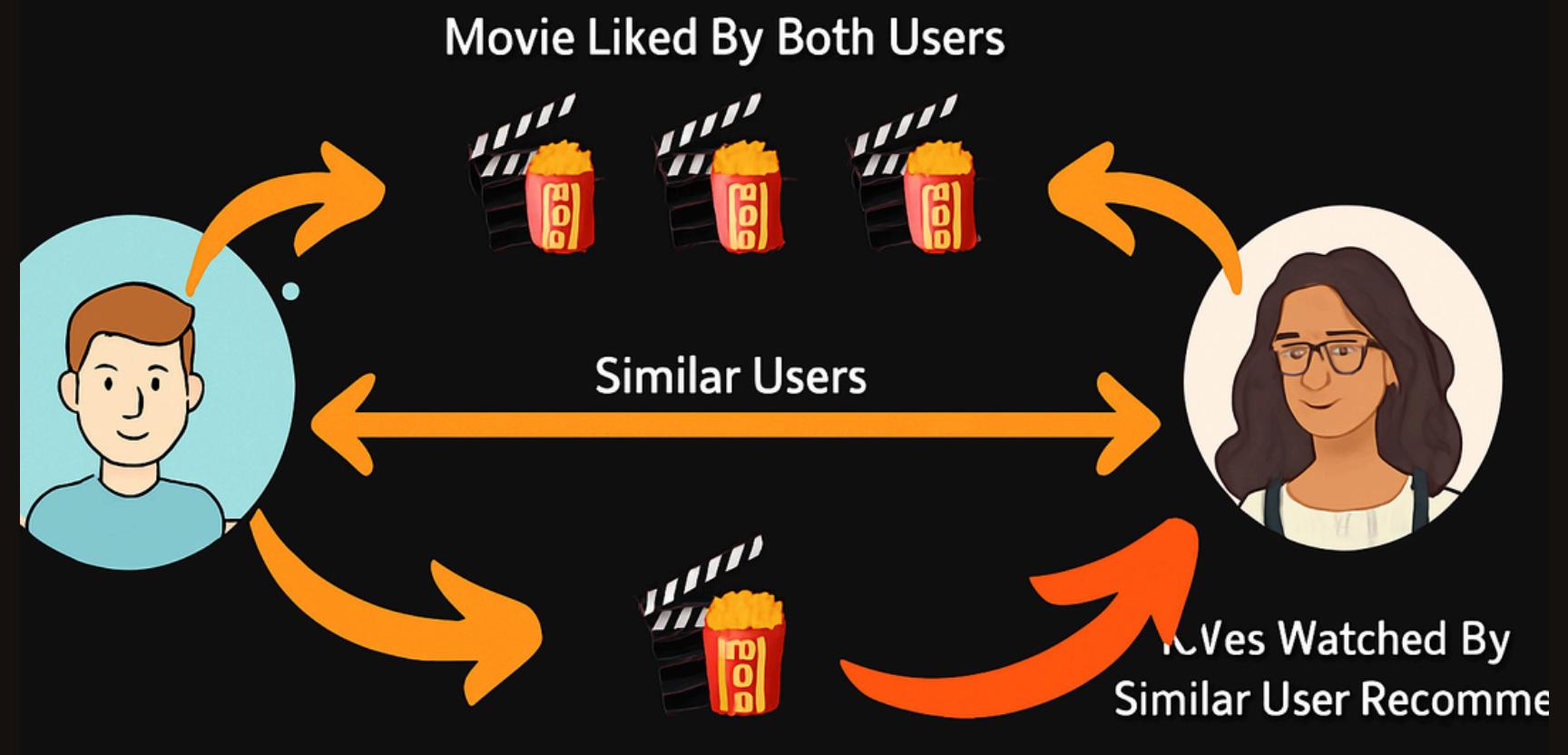
RISULTATI

Metric	Value	Interpretation
Catalog Coverage	34.04%	Moderate coverage
Unique Items Recommended	4,058 / 12,468	
Avg. Recommendation Confidence	0.4238 ± 0.1560	
Baseline (Random Pairs)	0.0284	
Lift over Random	1,393.7%	Strong signal
Intra-List Diversity (ILD)	0.4991 ± 0.2788	Moderate diversity
Hubness (Top 1% of Catalog)	41.63%	High bias
Hubness (Top 1% of Recommended)	24.08%	
Gini Coefficient	0.5445	Moderate inequality
Normalized Entropy	0.8796	Good distribution

2. Collaborative-filtering

L'approccio collaborative raccomanda film che piacciono a persone con gusti simili a quelli dell'utente in questione, guardando solo come gli utenti hanno votato i film.

Trova pattern nascosti e prevede quanto piacerebbe un film che l'utente non ha ancora visto, consigliando quelli con il voto previsto più alto.



Come funziona?

Collaborative filtering



Viene utilizzata la tecnica SVD++ (SIngular Value Decomposition plus plus), una tecnica di fattorizzazione di matrice che incorpora segnali di feedback impliciti per catturare informazioni riguardo alle preferenze dell'utente.

Lo scopo è predire il rating che l'utente darà ai film che non ha visto.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + |R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} y_j \right)$$

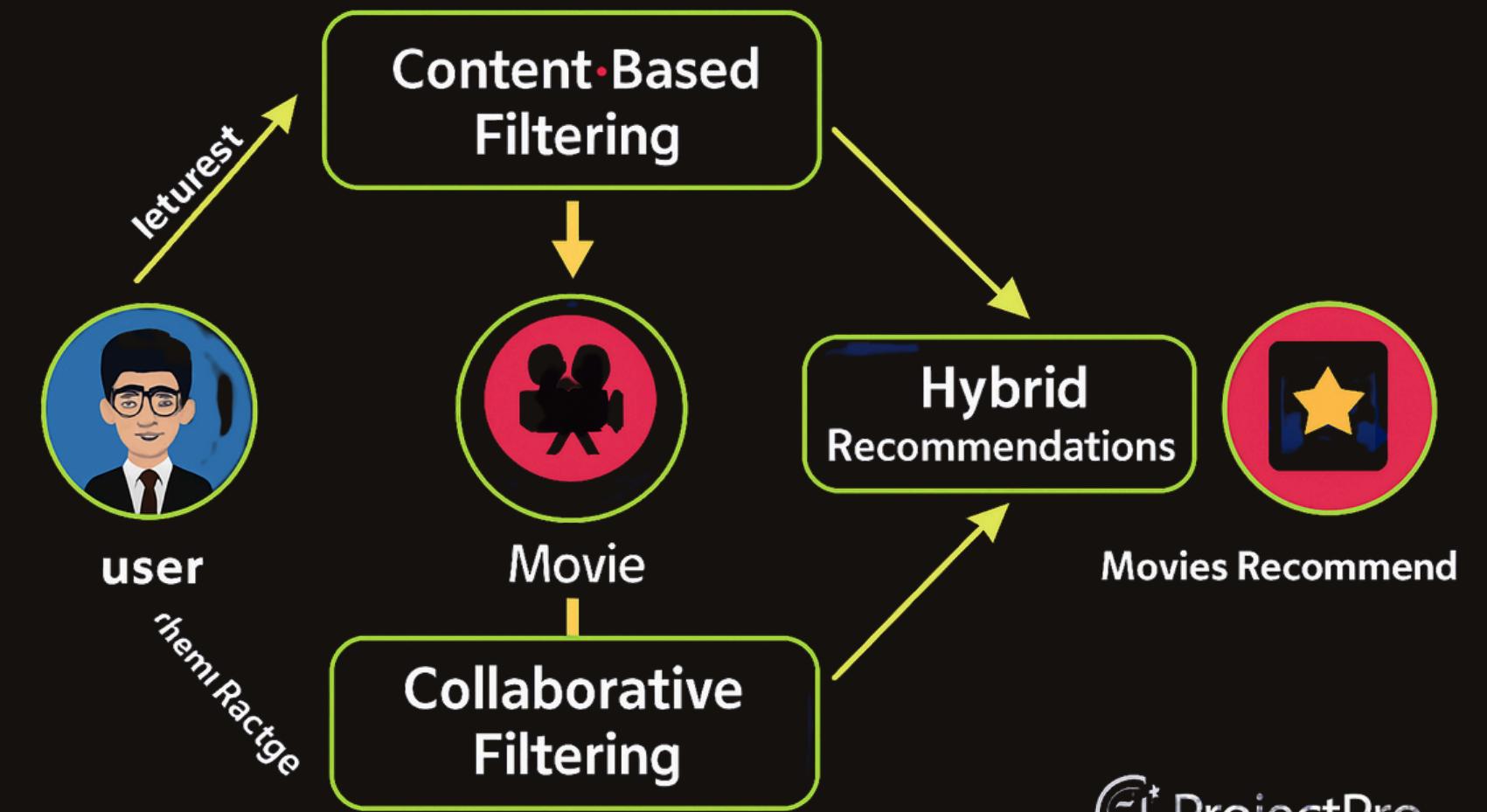
RISULTATI

K	Precision@K	Recall@K	F1@K	Hit Rate@K	AUC-ROC	Coverage@K
3	0.8215	0.1943	0.2882	0.9762	0.7439	0.1299
5	0.7970	0.3065	0.3994	0.9901	0.7439	0.1679
7	0.7758	0.4101	0.4798	0.9945	0.7439	0.2005
9	0.7530	0.5015	0.5359	0.9963	0.7439	0.2283
11	0.7309	0.5808	0.5756	0.9975	0.7439	0.2538
13	0.7122	0.6405	0.6033	0.9976	0.7439	0.2718
15	0.6973	0.6883	0.6241	0.9978	0.7439	0.2873
17	0.6850	0.7274	0.6404	0.9979	0.7439	0.3004
19	0.6737	0.7588	0.6522	0.9981	0.7439	0.3104
21	0.6650	0.7860	0.6625	0.9981	0.7439	0.3178

Users evaluated: 10,000 (0 skipped)

3. Hybrid

In un Weighted Hybrid model si prendono i punteggi generati separatamente dal modello content-based e da quello collaborative, si normalizzano e poi si combinano definendo un output finale.



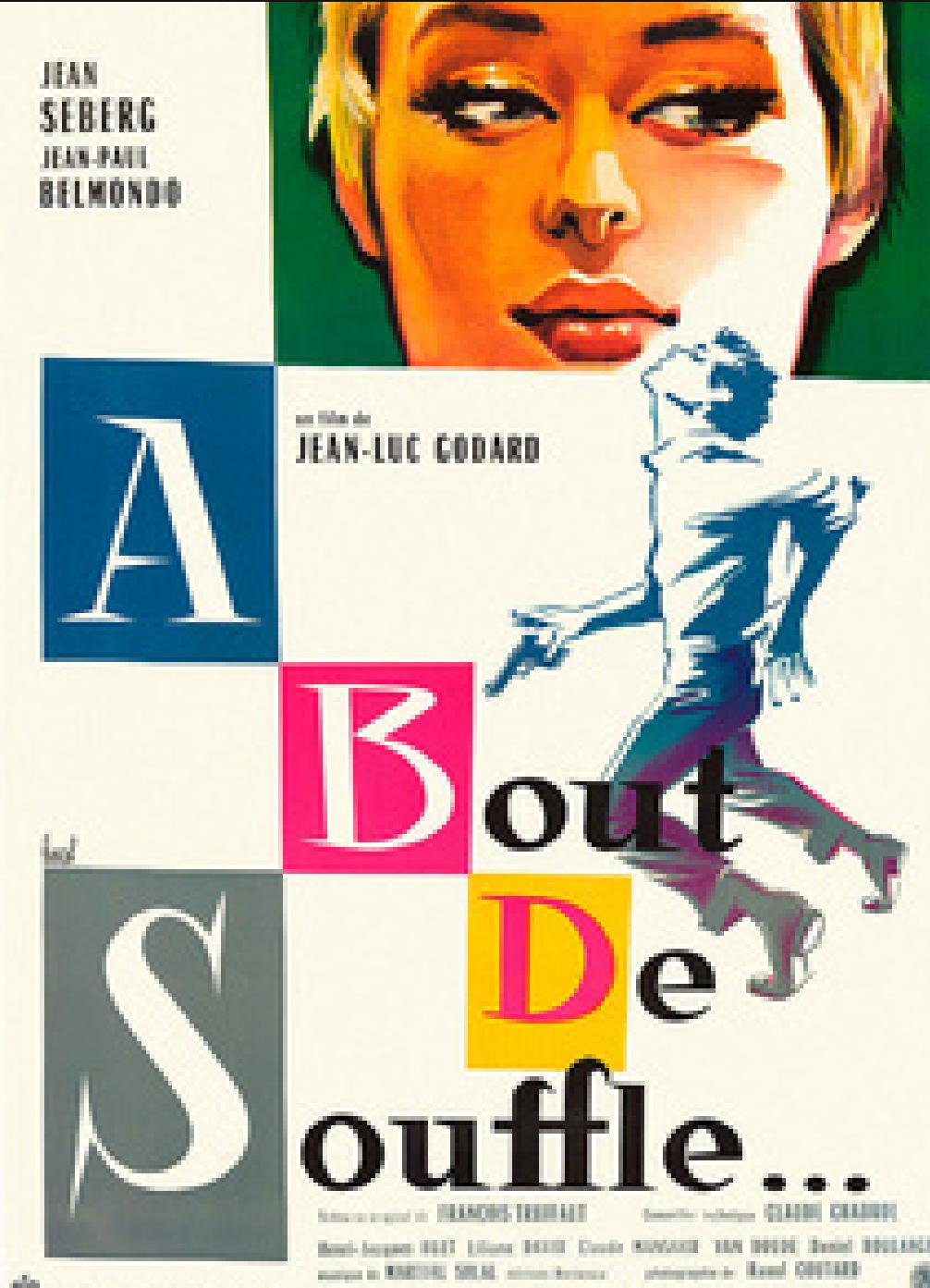
Histogram Gradient Boosting Regressor

Ho deciso di utilizzare un regressore lineare
a discesa del gradiente.

- Scalabilità
- Integrazione Naturale di Feature Eterogenee
- Apprendimento Automatico dei Pesi
- Ottimizzazione Globale



Come funziona?



- 1) Applico TF-IDF ai film ottenendo una matrice sparsa.
- 2.a) Opero poi con SVD++ per effettuare la riduzione di dimensionalità della matrice, ottenendo per ogni film 20 componenti latenti.
- 2.b) Simultaneamente utilizzo il modello collaborativo per generare uno score predittivo per ogni coppia (u,i) user-item che userò come prima input feature per il prossimo step.



Come funziona?

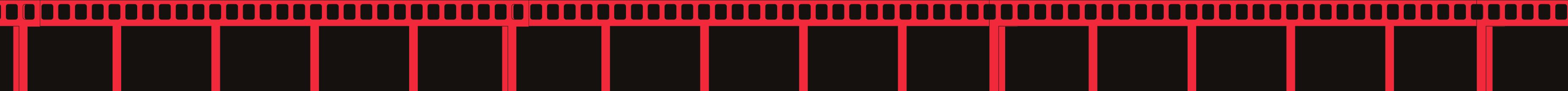
seconda pipeline

- 3) Il risultato di queste operazioni è un vettore X

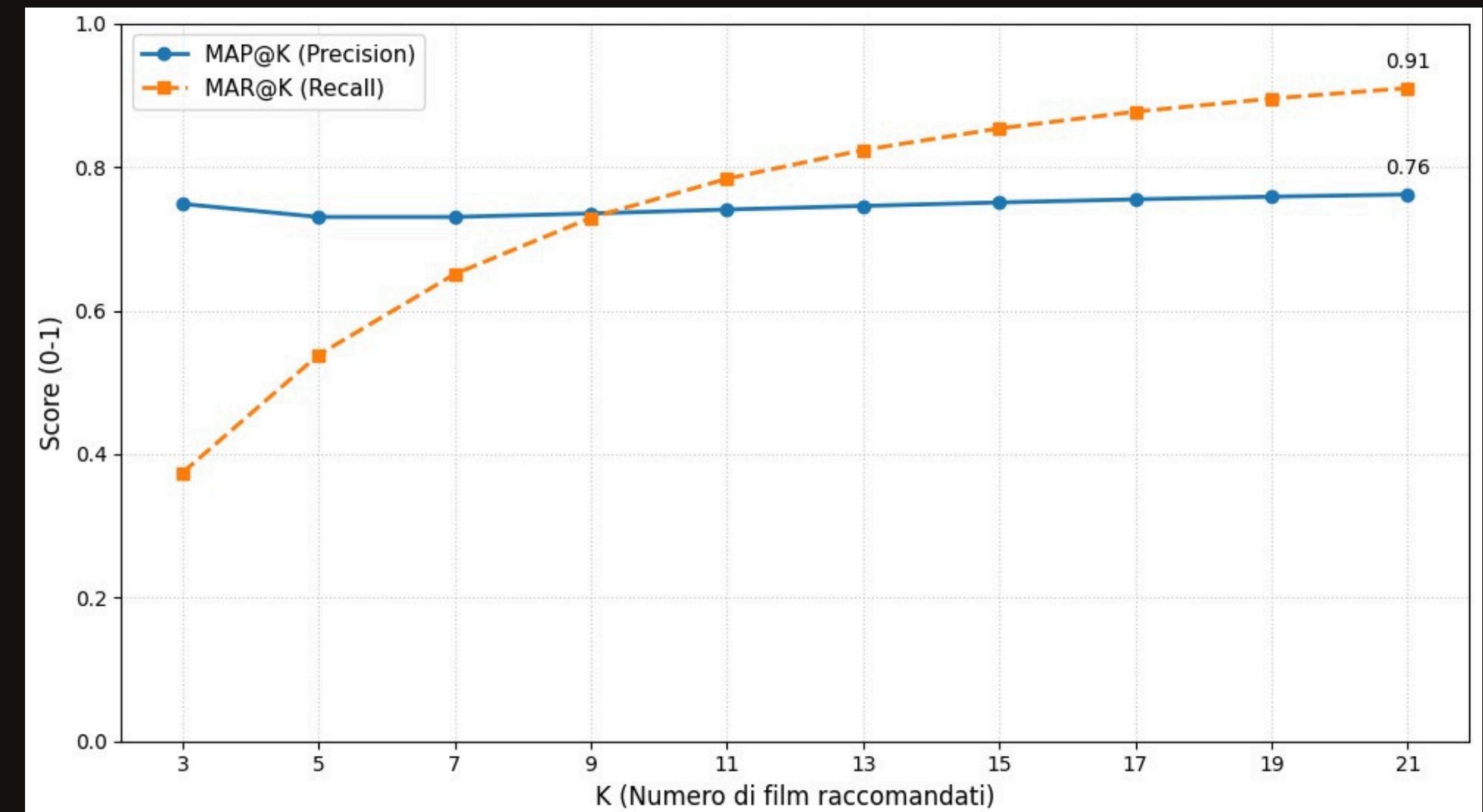
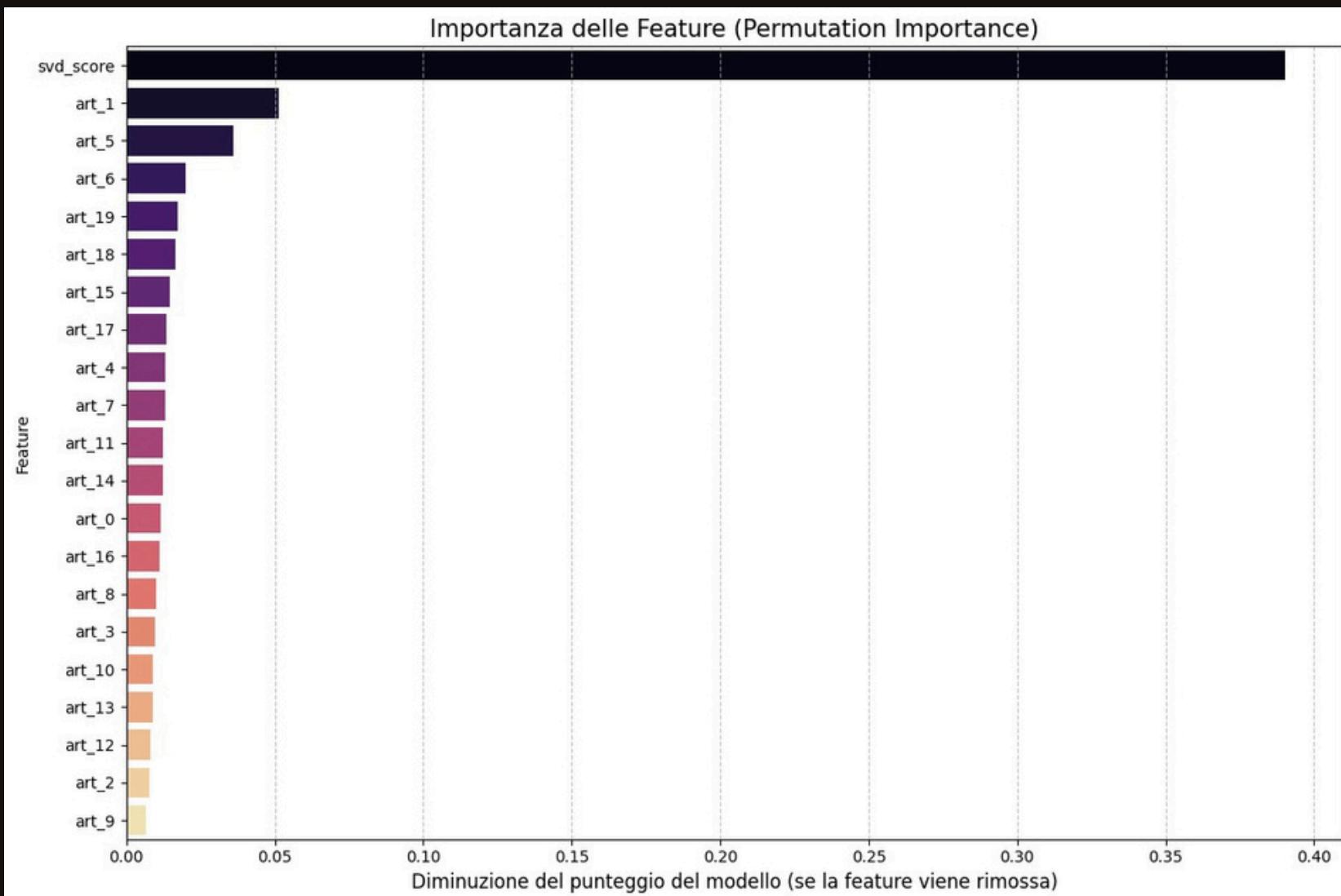
$$X_{ui} = [\underbrace{\hat{r}_{cf}}_{\text{User Preference}} , \underbrace{e_0, e_1, \dots, e_{19}}_{\text{Artistic Embeddings}}]$$

- 4) Il regressore predirra il valore di rating di un film dato X

Il modello unisce i valori in histogrammi iniziando da una stima iniziale. In seguito, aggiunge un piccolo albero decisionale per sistemare l'errore di quello prima. Ogni albero nuovo prova a seguire il percorso (gradiente) che abbassa l'errore rimasto.



RISULTATI



RISULTATI

K	MAP@K	MAR@K
3	0.7487	0.3741
5	0.7305	0.5384
7	0.7306	0.6510
9	0.7355	0.7294
11	0.7409	0.7839
13	0.7459	0.8238
15	0.7508	0.8539
17	0.7551	0.8773
19	0.7588	0.8955
21	0.7619	0.9102

...studio di ablazione

Uno studio di ablazione nel machine learning è un esperimento in cui si rimuovono o modificano componenti specifici di un modello (come layer o feature) per valutare il loro impatto sulle prestazioni complessive, aiutando a identificare cosa è essenziale.

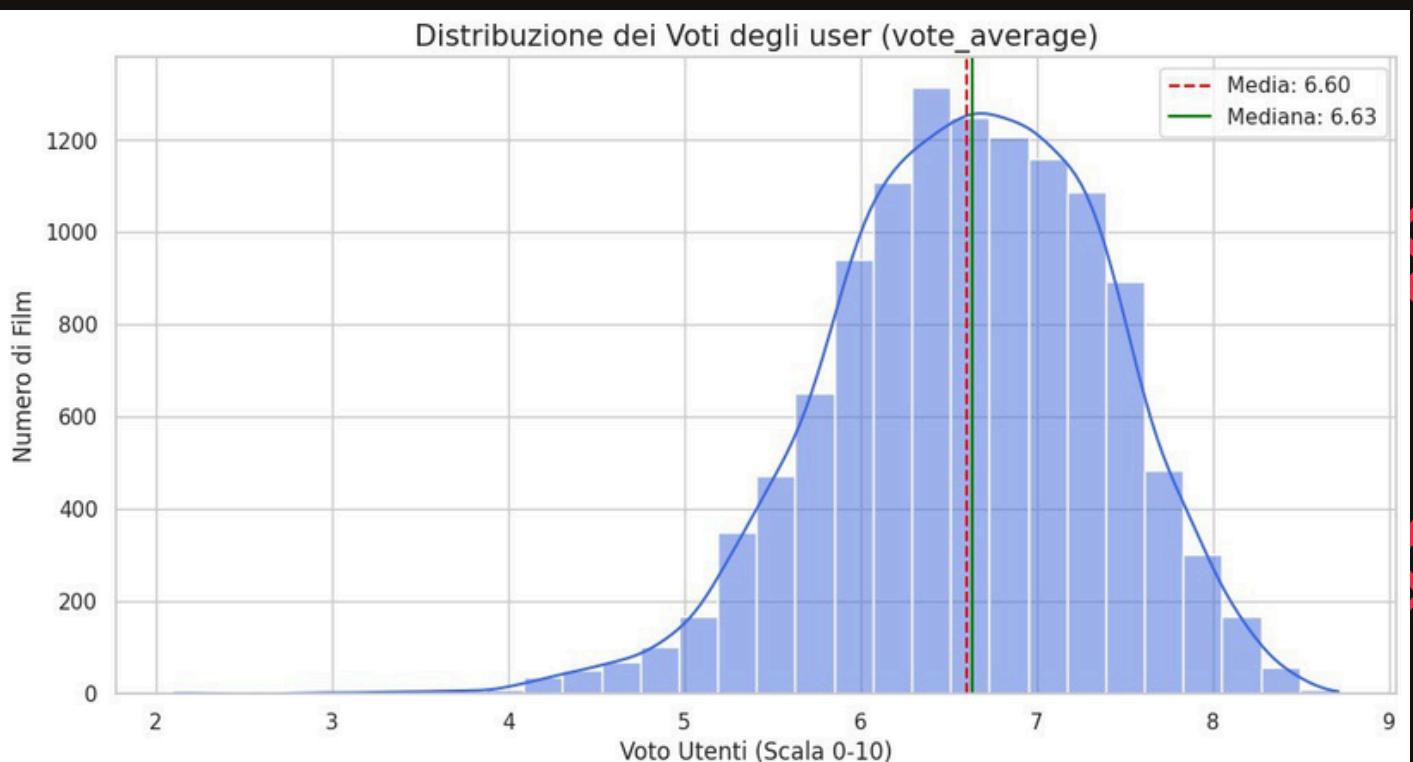
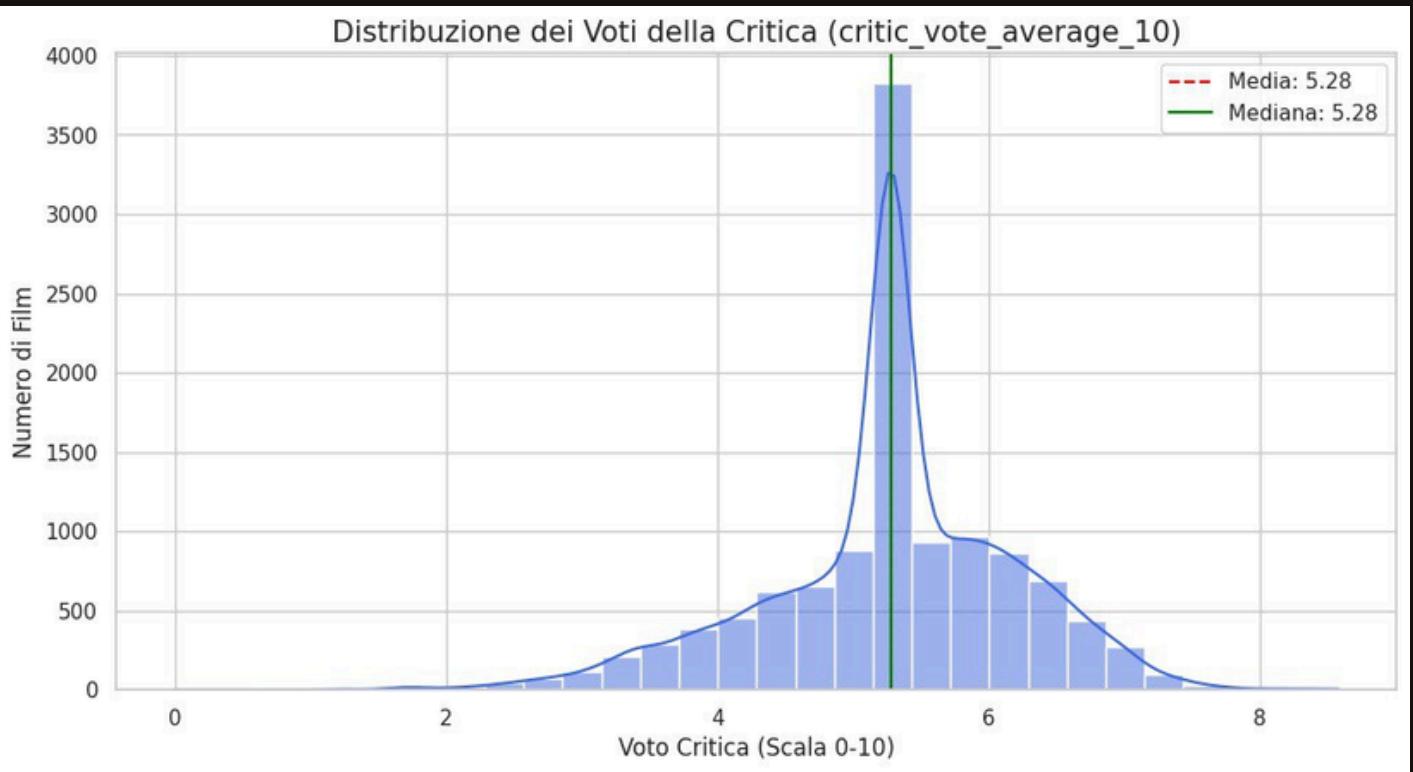
In questo caso ho studiato il comportamento del modello hybrid aggiungendo ai dati tenuti in considerazione:

- 1) vote average
- 2) vote count
- 3) popularity score

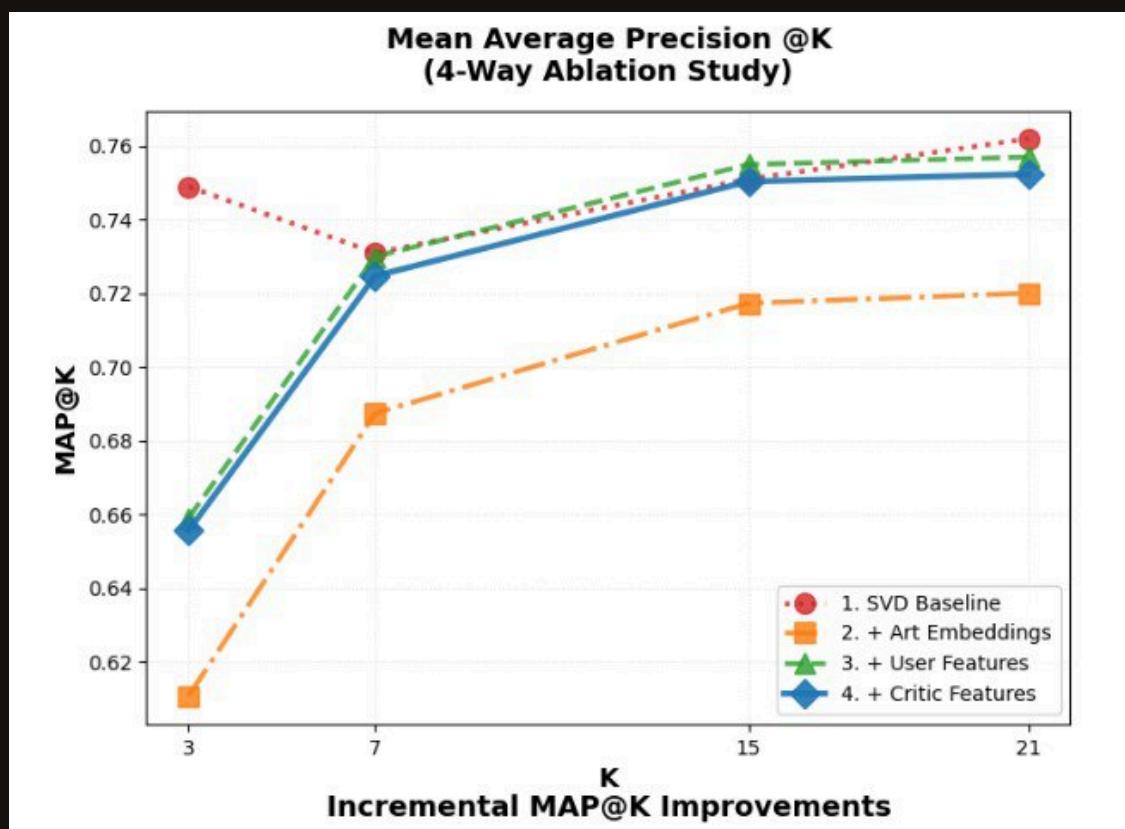
e in secondo luogo:

- 4) Critic vote average

ottenuti da TMDB.



RISULTATI

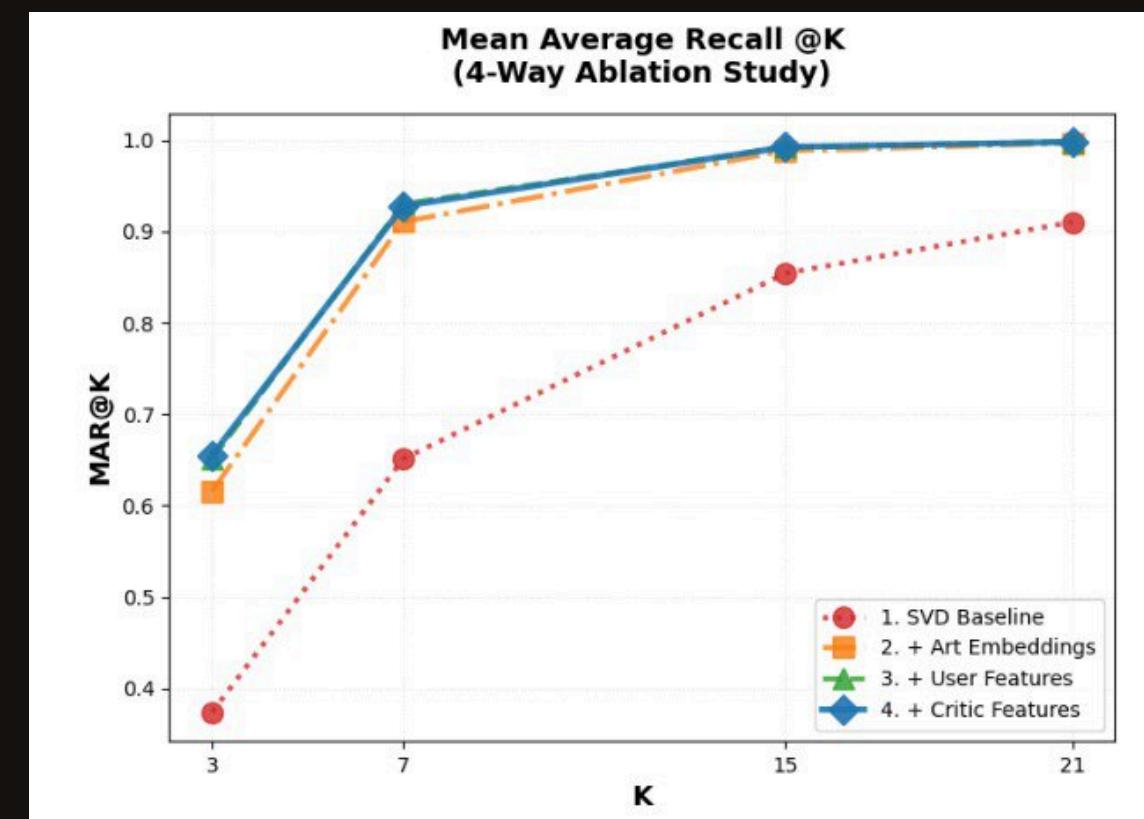


MAP@21 (Precision):

#	Modello	Valore	Δ
1	SVD Baseline	0.7620	-
2	+ Art Embeddings	0.7201	-0.0419
3	+ User Features	0.7570	+0.0369
4	+ Critic Features	0.7524	-0.0046
Miglioramento totale (1→4): -0.0096 (-1.3%)			

MAR@21 (Recall):

#	Modello	Valore	Δ
1	SVD Baseline	0.9100	-
2	+ Art Embeddings	0.9963	+0.0863
3	+ User Features	0.9980	+0.0017
4	+ Critic Features	0.9976	-0.0004
Miglioramento totale (1→4): +0.0876 (+9.6%)			



RMSE (Errore Predittivo):

Modello	Valore
SVD Baseline	1.0261
+ Art Embeddings	0.922
+ User Features	0.8794
+ Critic Features	0.8765

Conclusioni

E' stata esperienza altamente formativa, non sono mancate difficoltà e momenti di frustrazione, soprattutto legati alle limitate risorse computazionali a disposizione, che hanno imposto compromessi sull' operato, tuttavia mi ha permesso di comprendere meglio i sistemi di raccomandazione che ci circondano.





fidelio

Tullo Nikolas