



Progetto Metodi Informatici per la Gestione Aziendale

SVILUPPO DI DIVERSE TIPOLOGIE DI DI RECOMMENDATION SYSTEM

AMAZON REVIEWS - CD AND VINYL




Andrea Vasciminno 904899

Matias Maciej Bonoli 912941



INTRODUZIONE

- 
- Processo di **pulizia del dataset**
 - **Analisi esplorativa**
 - **Preparazione dei dati per utilizzo con modelli collaborativi**
 - **Generazione delle raccomandazioni**
 - **Cluster & PCA Visualization**
 - **Preprocessing dati testuali**
 - **Confronto tecniche di raccomandazione**
 - **Progetto Avanzato**



PROCESSO DI PULIZIA DEL DATASET

Abbiamo lavorato sul dataset per ridurlo di dimensione e renderlo utilizzabile con costi computazionali sostenibili

01

RIDUZIONE DEGLI UTENTI MENO ATTIVI

Si considerano solamente gli utenti che abbiano lasciato almeno 12 recensioni

02

ELIMINAZIONE PRODOTTI POCO RECENSITI

Si considerano solamente i prodotti con più di 22 recensioni

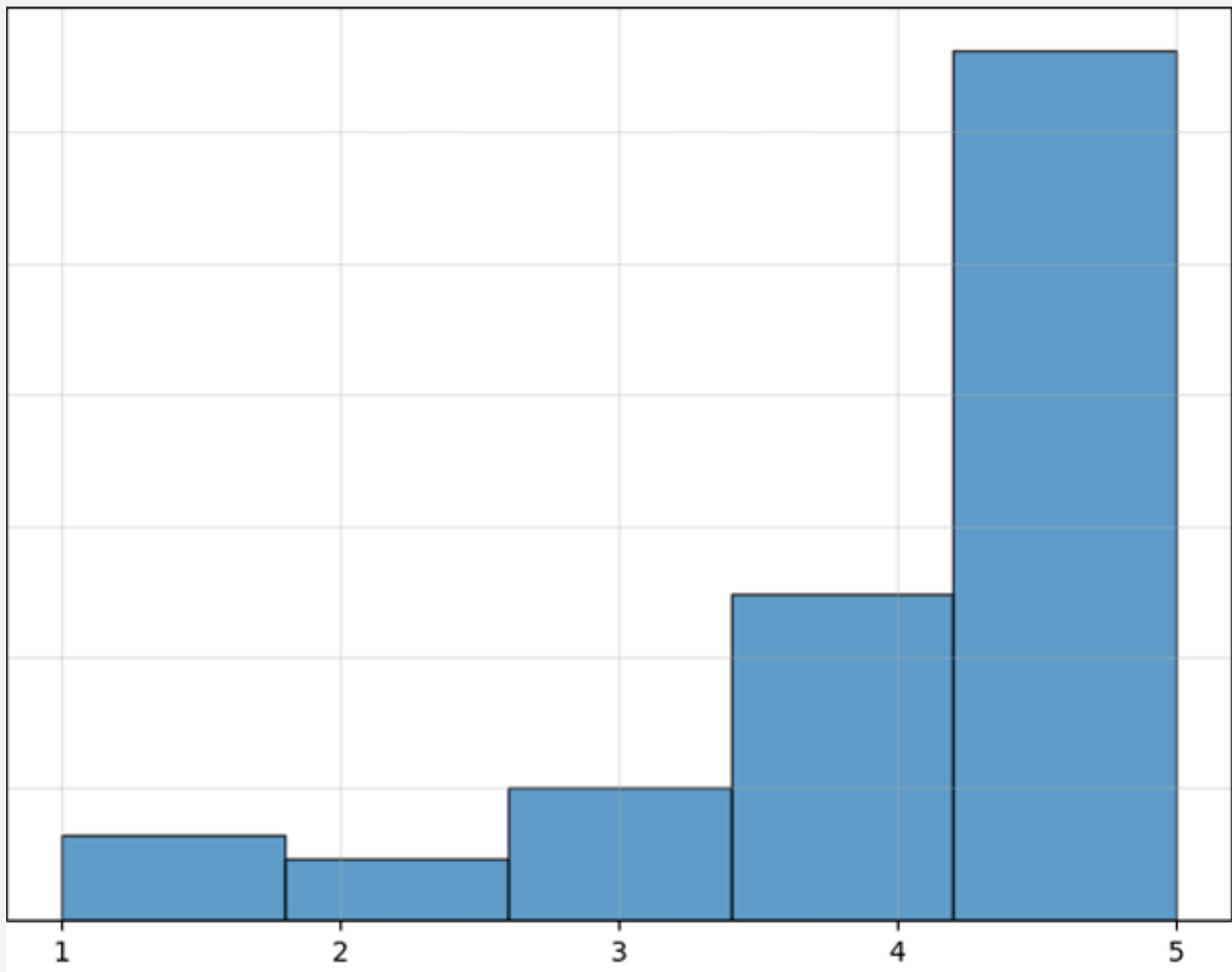
03

RISULTATO:

	rating
count	112.140
mean	4,25
std	1,14

ANALISI ESPLORATIVA - RATING

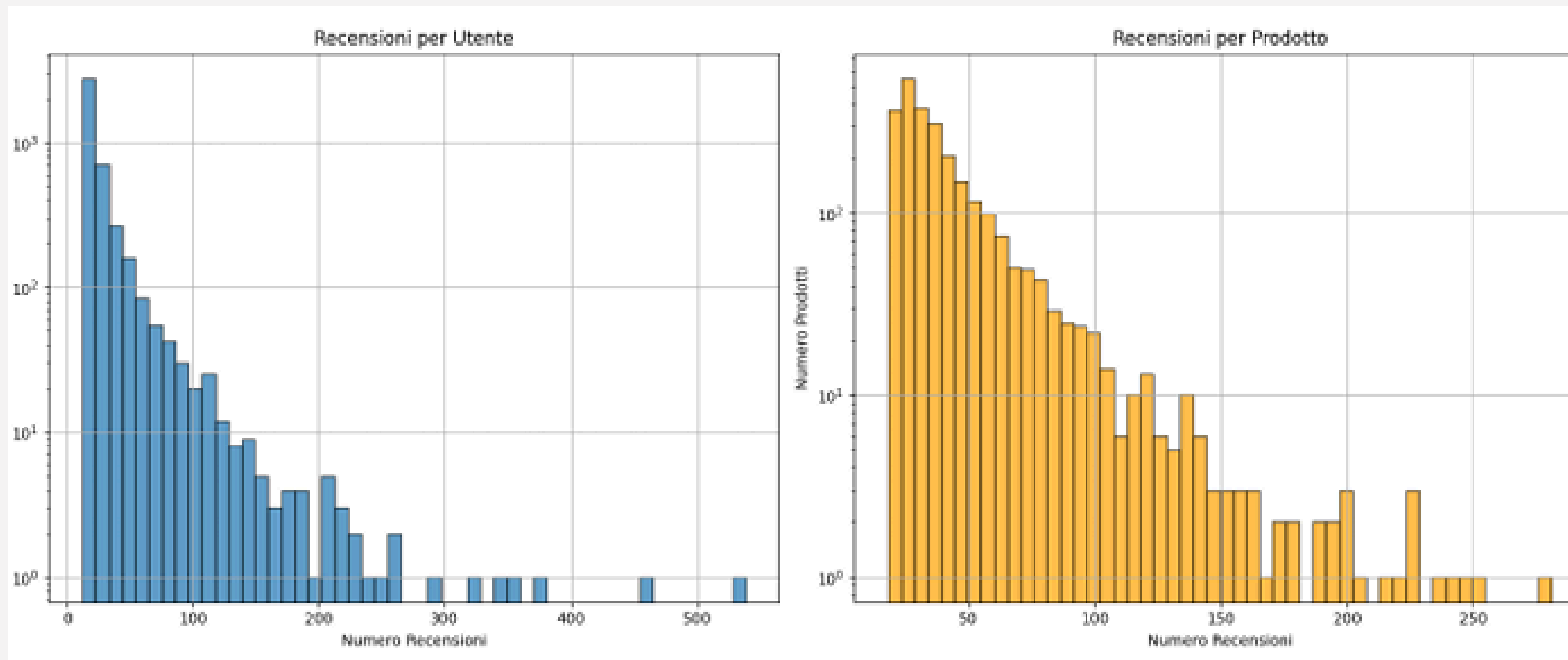
Abbiamo effettuato un analisi esplorativa dei dati, per vedere varie statistiche sui vari campi del dataset



Rating	Count	Percentuale
1	6.404	6,86%
2	4.595	4,65%
3	10.064	9,91%
4	24.818	22,18%
5	66.259	56,40%

ANALISI ESPLORATIVA - UTENTI E PRODOTTI

Abbiamo effettuato un'analisi esplorativa dei dati, per vedere varie statistiche sui vari campi del dataset



UTENTI

- Media recensioni per utente: **26,58**
- Mediana: **18,0**
- Massimo: **539** recensioni

PRODOTTI

- Media recensioni per prodotto: **43,11**
- Mediana: **34,0**
- Massimo: **281** recensioni

PREPARAZIONE DEI DATI PER I MODELLI COLLABORATIVI

SPLIT DI TRAINING E TEST SET

Il dataset è stato diviso in un set di training(80%) e un set di test(20%)

OTTIMIZZAZIONE DEI PARAMETRI

Ottimizzazione dei parametri per KNN ed SVD tramite Grid Search

TEST INIZIALE CON KNN E SVD

SVD		KNN	
Metrica	Valore	Metrica	Valore
MSE	0,8232	MSE	0,9647
RMSE	0,9073	RMSE	0,9822

TEST CON PARAMETRI OTTIMIZZATI

Algoritmo	MSE	RMSE
K-NN (ottimizzato)	0.9705	0.985
SVD (ottimizzato)	0.7793	0,882
Miglioramento	19,7%	10,4%

GENERAZIONE DELLE RACCOMANDAZIONI



Predizione dei rating per tutti gli item non recensiti (filling **Raccomandation Matrix**)

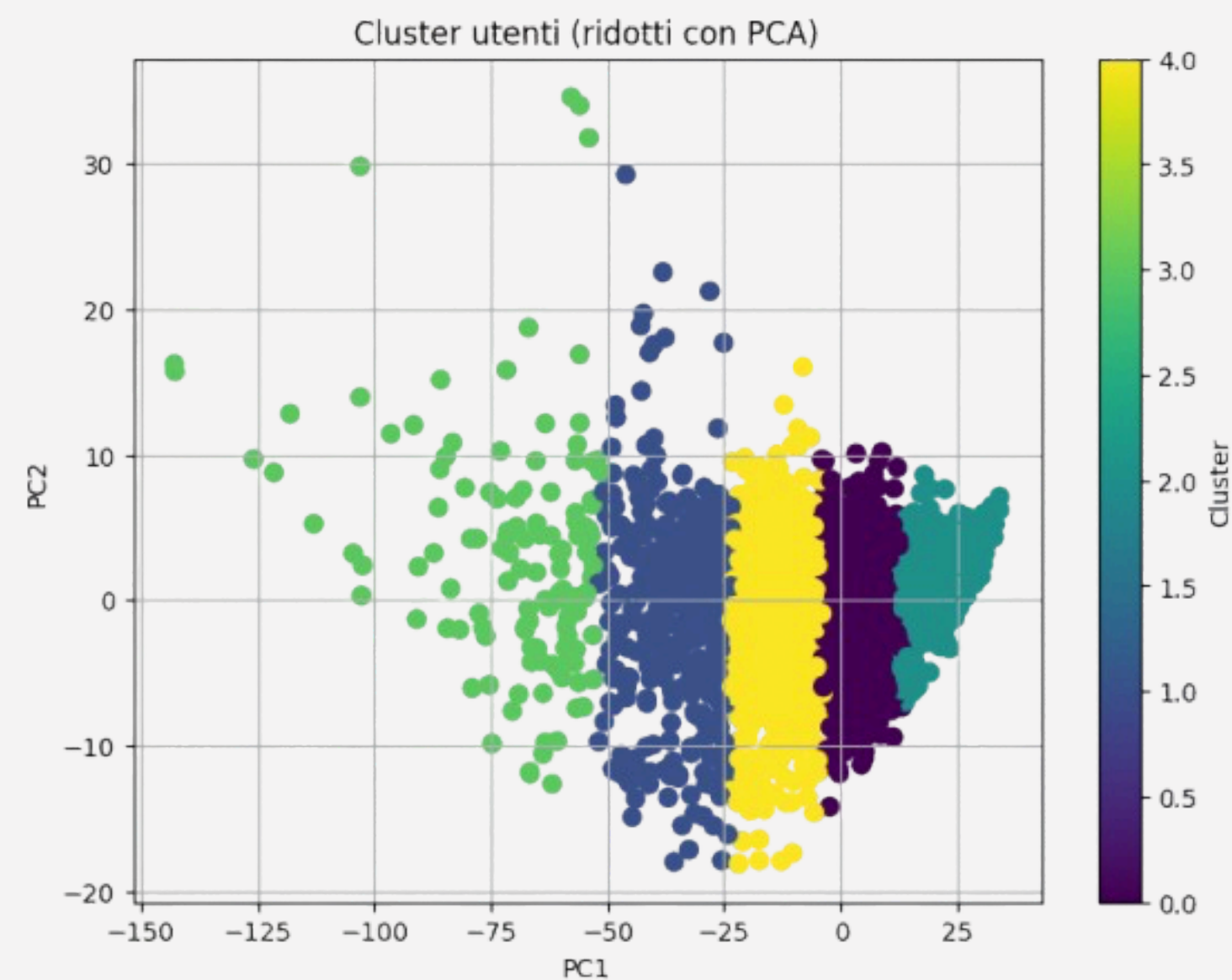
Esclusione degli item già valutati dall'utente

Ordinamento per rating predetto decrescente

Selezione dei top 20 item per utente

User ID	Top-3 Raccomandazioni	Rating
AFW2PDT3AMT4X3PYQG7FJZH5FXFA	1. Ultimate Rascals, The 2. Carole King Tapestry 3. Kind Of Blue	5.0 5.0 5.0
AE7BV6IMNPZ3F266H7PXMH3BZQNG	1. Innervisions Remastered 2. Dirt (Explicit Lyrics) 3. At Folsom Prison	5.0 5.0 5.0
AGTMZCWIWBH43TCW7UKG2YV2EKKA	1. Innervisions Remastered 2. Dirt (Explicit Lyrics) 3. In Step	5.0 5.0 5.0
AFCU2ZFZ2ZLMM5YX2MXUOV52WMKQ	1. At Folsom Prison 2. Message In A Box: Complete Recordings 3. Fear Of A Black Planet	5.0 5.0 5.0
AGWDYYVVWM3DC3CASUZKXK67G6IA	1. The Immaculate Collection 2. Harem 3. Scarlet's Walk	5.0 5.0 5.0

CLUSTER & PCA VISUALIZATION



Cluster	Utenti	Rating Medio	Std	Profilo
0 (Viola)	1307 (31.0%)	4.370	0.373	Utenti positivi
1 (Blu)	417 (9.9%)	3.608	0.448	Utenti moderati
2 (Teal)	1403 (33.3%)	4.679	0.328	Utenti entusiasti
3 (Verde)	125 (3.0%)	2.903	0.554	Utenti critici
4 (Giallo)	967 (22.9%)	4.027	0.408	Utenti equilibrati

PREPROCESSING DATI TESTUALI

Per trasformare le informazioni testuali non strutturate in rappresentazioni numeriche utilizzabili dai modelli di machine learning, è stata implementata una pipeline completa di Natural Language Processing (NLP)

01

COMBINAZIONE CAMPI TESTUALI

Unione dei campi title e description del dataset

02

TOKENIZZAZIONE

Suddivisione del testo in token

03

RIMOZIONE DI STOPWORDS

Rimozione di elementi non informativi

04

LEMMATIZZAZIONE

Riduzione dei token alla loro forma base

CONFRONTO TECNICHE DI RACCOMANDAZIONE

Approccio	Algoritmo	MSE	RMSE	Δ vs Best CF
Collaborative Filtering	K-NN	0.971	0.985	+24.6%
	SVD	0.779	0.883	baseline
Content-Based	TF-IDF + K-NN	1.030	1.020	+32.2%
	Transformer + K-NN	1.010	1.000	+29.7%

SENTIMENT ANALYSIS

Abbiamo implementato tecniche di NLP per classificare automaticamente le recensioni in sentiment positivo, neutro o negativo utilizzando Random Forest con embedding TF-IDF e Transformer.

TF-IDF

Metrica	Valore Globale	Negative	Neutral	Positive
Accuracy	0.857	-	-	-
RMSE	0.563	-	-	-
MAE	0.201	-	-	-
Precision	-	0.77	0.71	0.86
Recall	-	0.48	0.08	0.99
F1-Score	-	0.60	0.15	0.92

TRANSFORMER

Metrica	Valore Globale	Negative	Neutral	Positive
Accuracy	0.830	-	-	-
RMSE	0.646	-	-	-
MAE	0.252	-	-	-
Precision	-	0.87	0.72	0.83
Recall	-	0.18	0.04	1.00
F1-Score	-	0.29	0.08	0.91

**GRAZIE
DELL'ATTENZIONE**