Analisi dell’efficacia del data clustering sull’addestramento di Recommender System

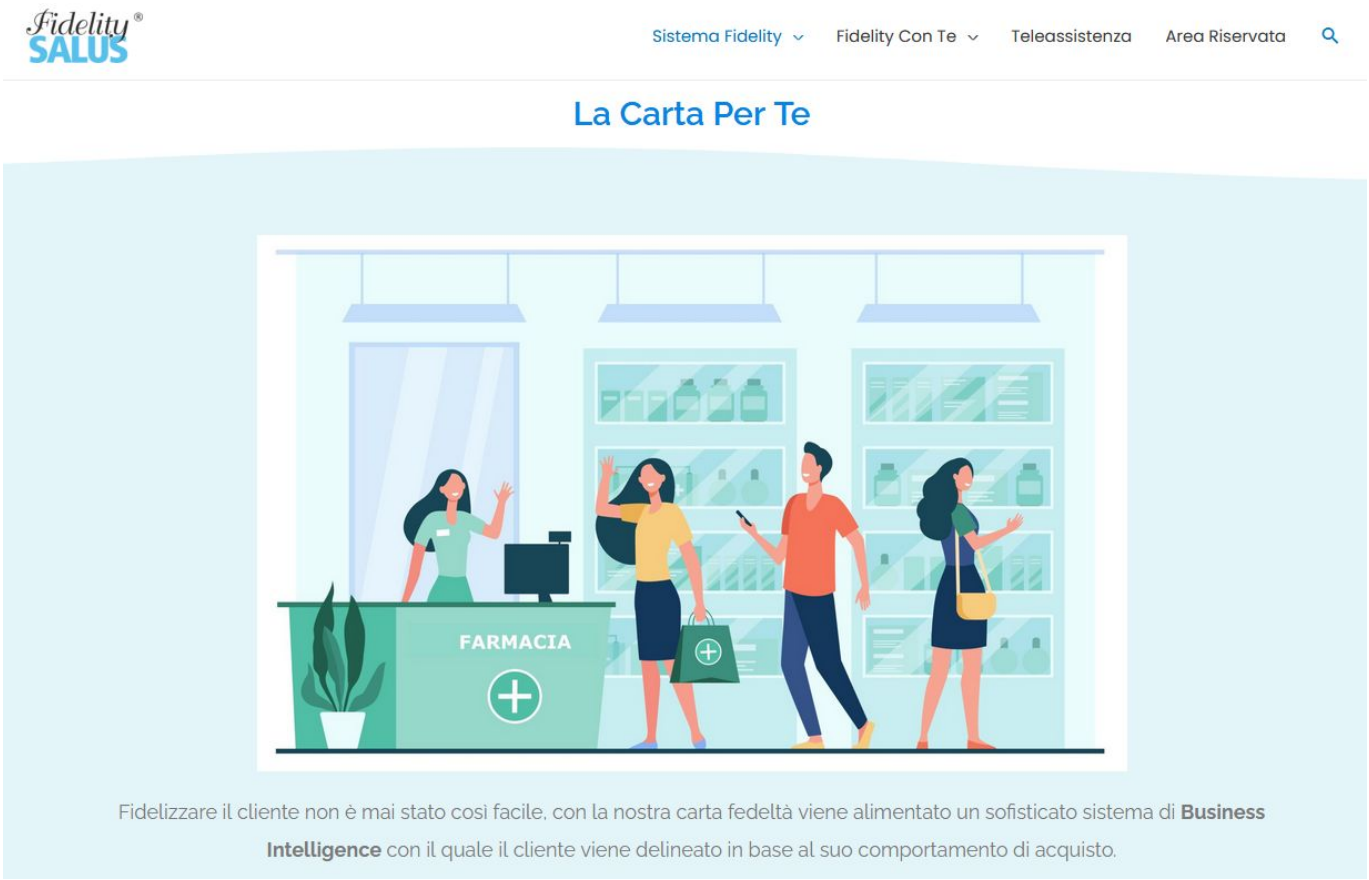
Laureando:
Andrea Ricci

Relatori:
Valentina Poggioni
Alina Elena Baia

29 aprile 2021
A.A. 2019/2020

UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Fidelity Salus



- Erogazione e gestione delle Fidelity card
- Sconti riservati ai possessori
- Campagne pubblicitarie personalizzate

Sommario

Obiettivi del lavoro di tesi:

- Implementazione di un Recommender System che consigli alle farmacie nuovi prodotti da inserire in magazzino
- Applicazione di tecniche di clustering atte a verificare se le performance del Recommender System possano essere migliorate

Sommario

- Recommender System
  - Collaborative Filtering
  - Allenamento del modello: Matrix Factorization e Bayesian Personalized Ranking
  - Prestazioni del sistema

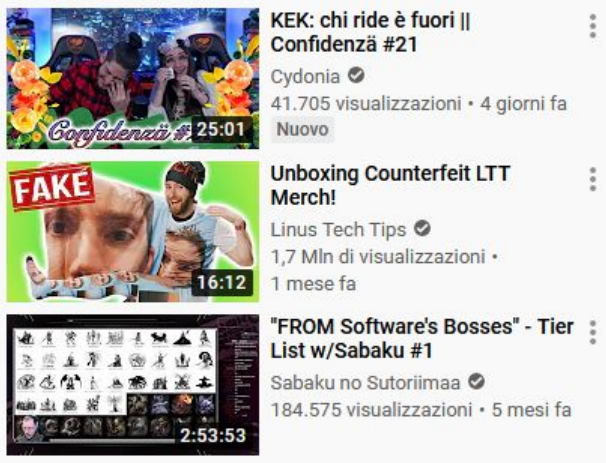
- Clustering
  - Selezione degli attributi
  - Applicazione degli algoritmi di clustering
  - Variazioni delle prestazioni del Recommender System

- Conclusioni e sviluppi futuri

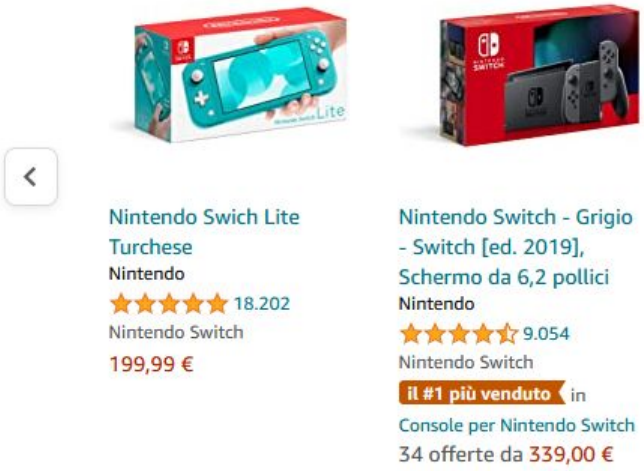
Recommender System

Recommender System

Un recommender system è un software di filtraggio dei contenuti che crea delle raccomandazioni personalizzate specifiche per l'utente così da aiutarlo nelle sue scelte.



Youtube

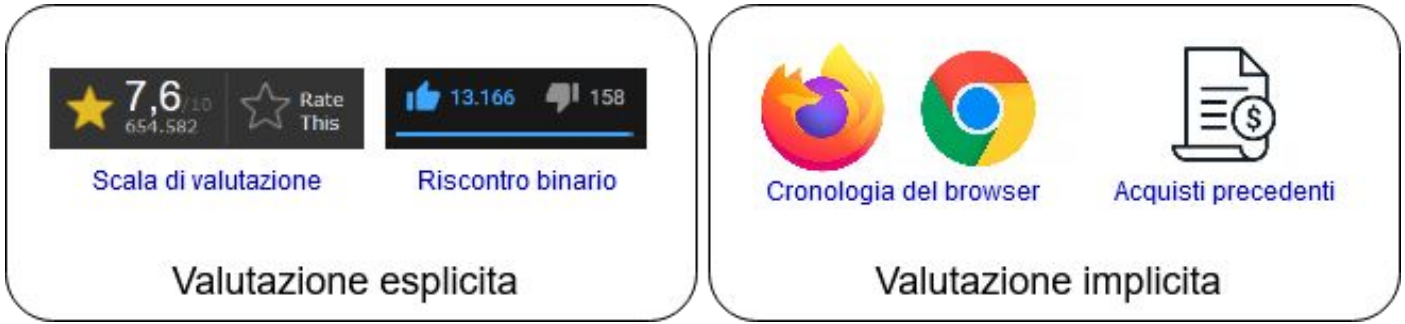


Amazon

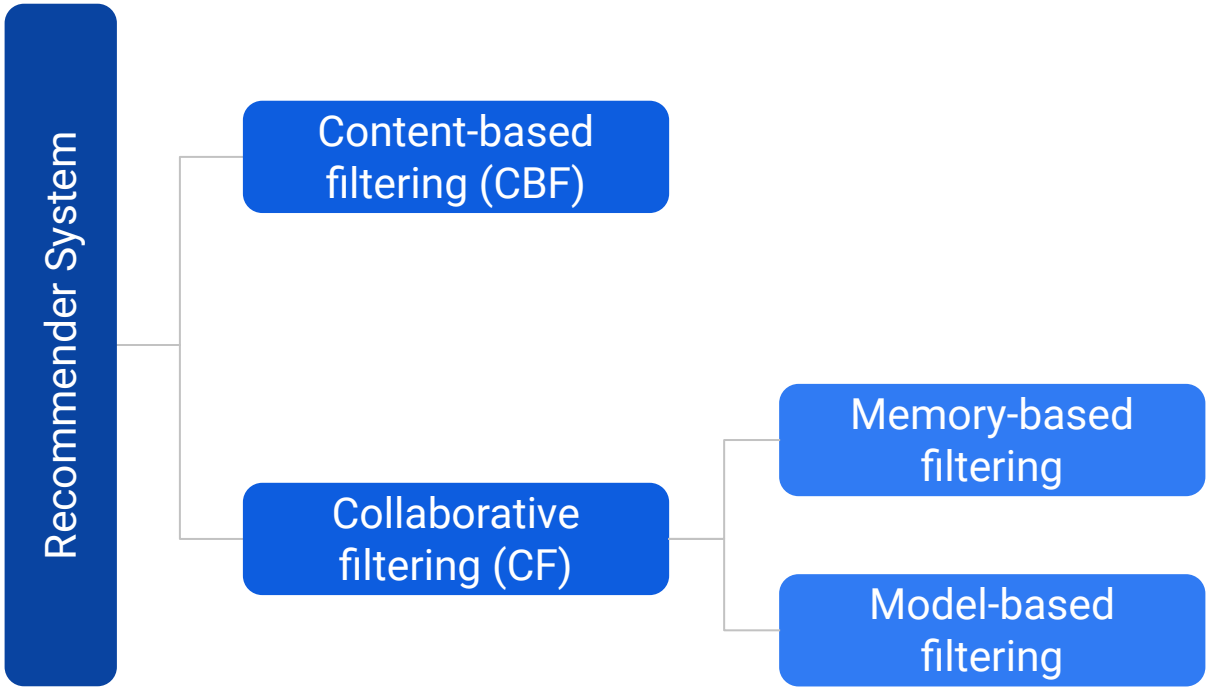
Feedback implicito ed esplicito

Tipologie di raccolta delle informazioni:

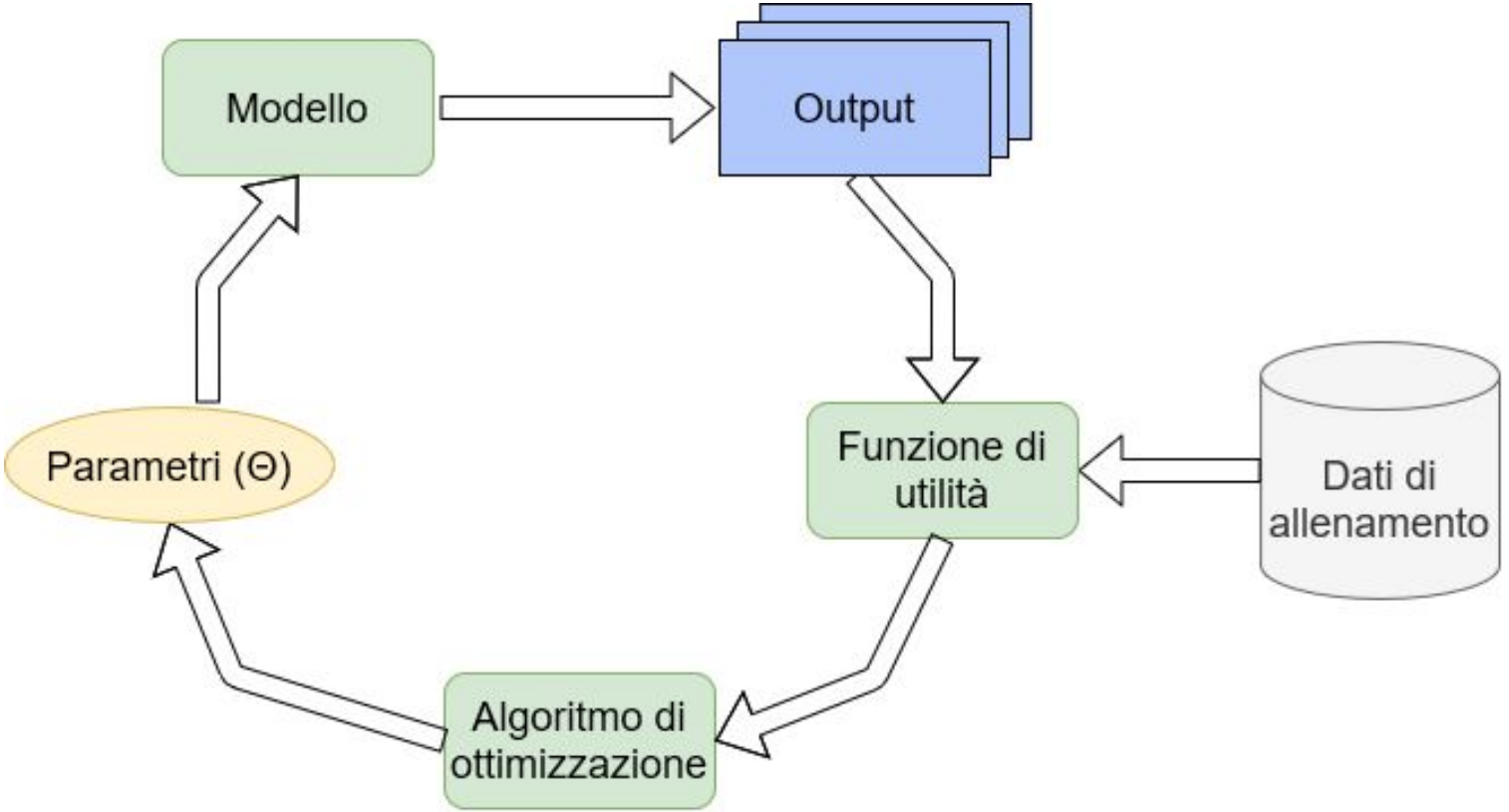
- Feedback esplicito: l'utente assegna esplicitamente le valutazioni agli oggetti
- Feedback implicito: viene studiato il comportamento dell'utente per derivarne le preferenze



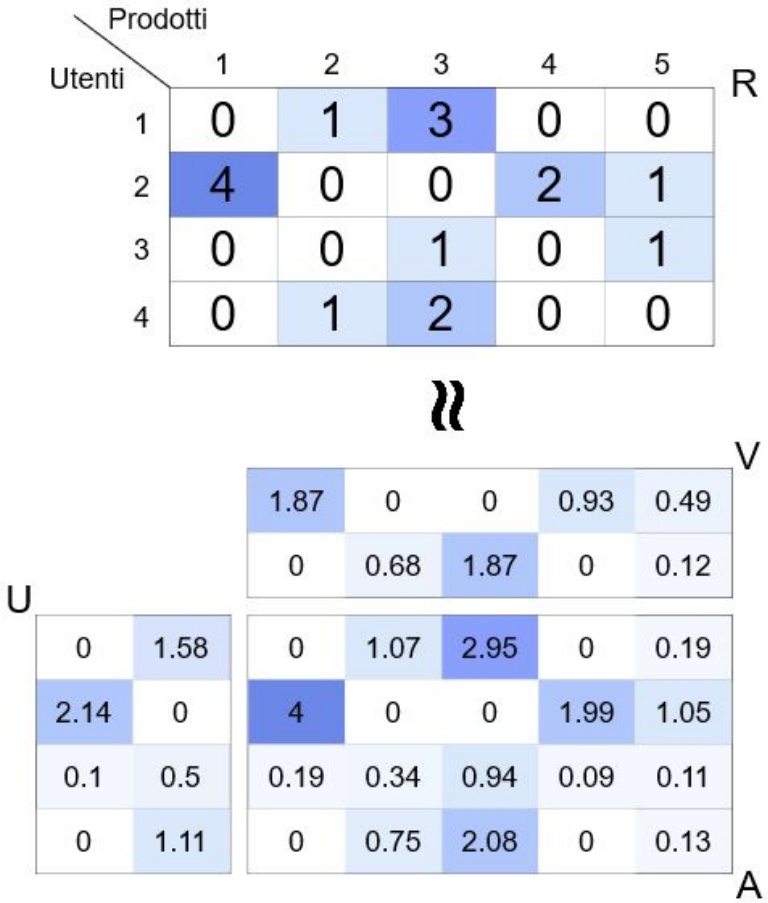
Tipologie di recommender system



Allenamento del modello



Matrix Factorization



- f = numero fattori latenti
- $U \in \mathbb{R}^{m \times f}$ $V \in \mathbb{R}^{n \times f}$
- $A \approx UV^T$ contiene la preferenza predetta dal sistema per ogni coppia utente/oggetto
- Obiettivo: garantire che *A* rappresenti una buona approssimazione di *R*

Bayesian Personalized Ranking

Siano:

- U ed I l'insieme di tutti gli utenti ed oggetti
- $I_u^+ = \{i \in I : (u,i) \in S\}$ l'insieme degli oggetti con i quali l'utente *u* ha interagito

L'obiettivo dell'algoritmo è quello di trovare una classifica personalizzata $>_u \subset I^2$ per ogni utente $u \in U$ ed ogni coppia di oggetti $(i,j) \in I^2$ che soddisfi le proprietà di un ordine totale.

Bayesian Personalized Ranking

Dati di allenamento:

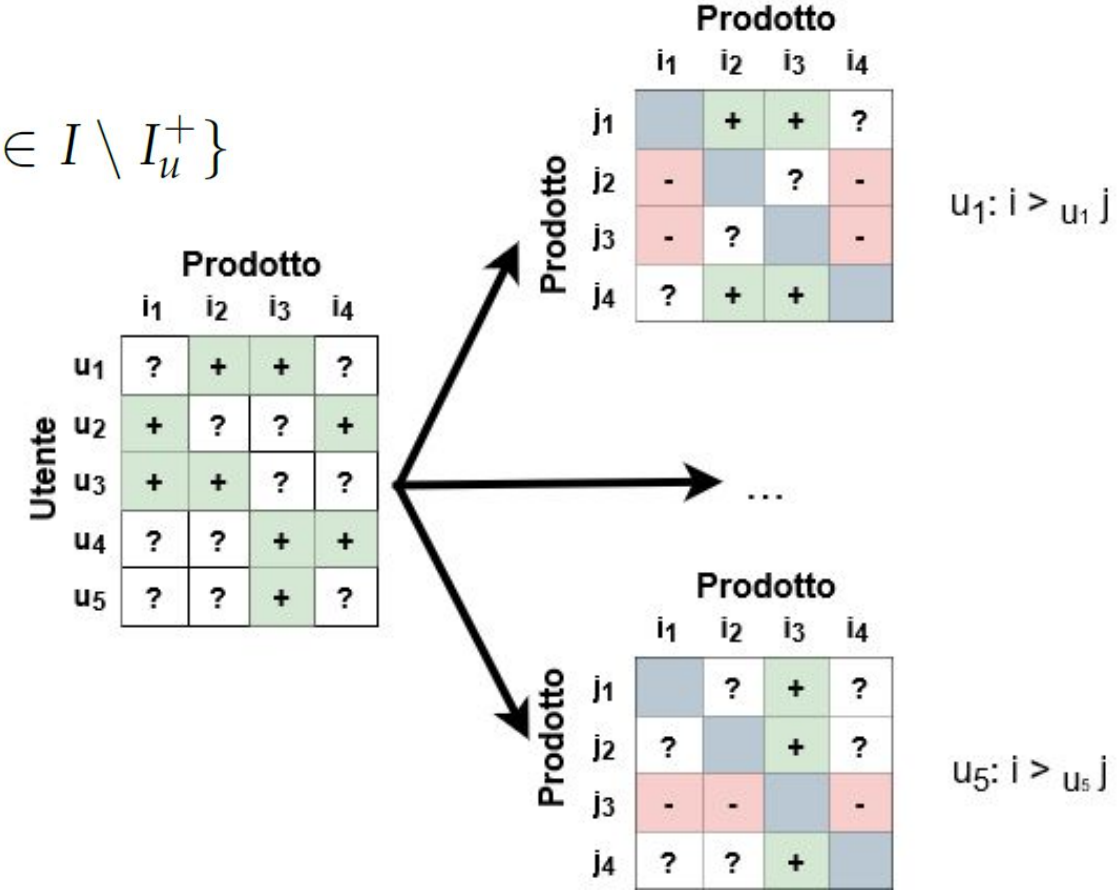
$$D_S := \{(u,i,j) | i \in I_u^+ \wedge j \in I \setminus I_u^+\}$$

dove

$$(u,i,j) \in D_S$$

$$\Updownarrow$$

$$i >_u j$$



Criterio di ottimizzazione BPR-OPT

BPR-OPT massimizza la probabilità a posteriori $p(\Theta| >_u)$.

$$
\begin{aligned}
BPR-OPT &:= \ln p(\Theta |>_u) \\
&= \dots \\
&= \sum_{(u,i,j)\in D_S} \ln \sigma(\hat{x}_{uij}) - \lambda_\Theta \|\Theta\|^2
\end{aligned}
$$

Se il modello scelto è la Matrix Factorization, allora $\hat{x}_{uij} = \hat{x}_{ui} - \hat{x}_{uj}$

Struttura del dataset degli acquisti

UserID	1397958
ItemID	971989823
ItemName	LFP FERMENTIFLUID 10X10ML
CatCode	4AA2F35
CatName	FERMENTI LATTICI
Quantity	1
PharmacyID	512

Record di esempio del dataset degli acquisti riferito all'anno 2019

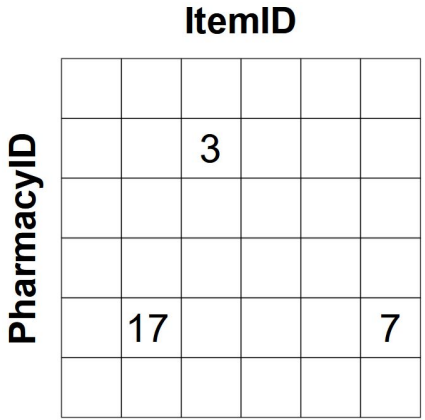
Matrice degli acquisti

UserID	ItemID	Quantity	PharmacyID
18	1017	7	15
745	1017	3	137
18	96	1	137
111	1017	10	15
1048	96	6	137

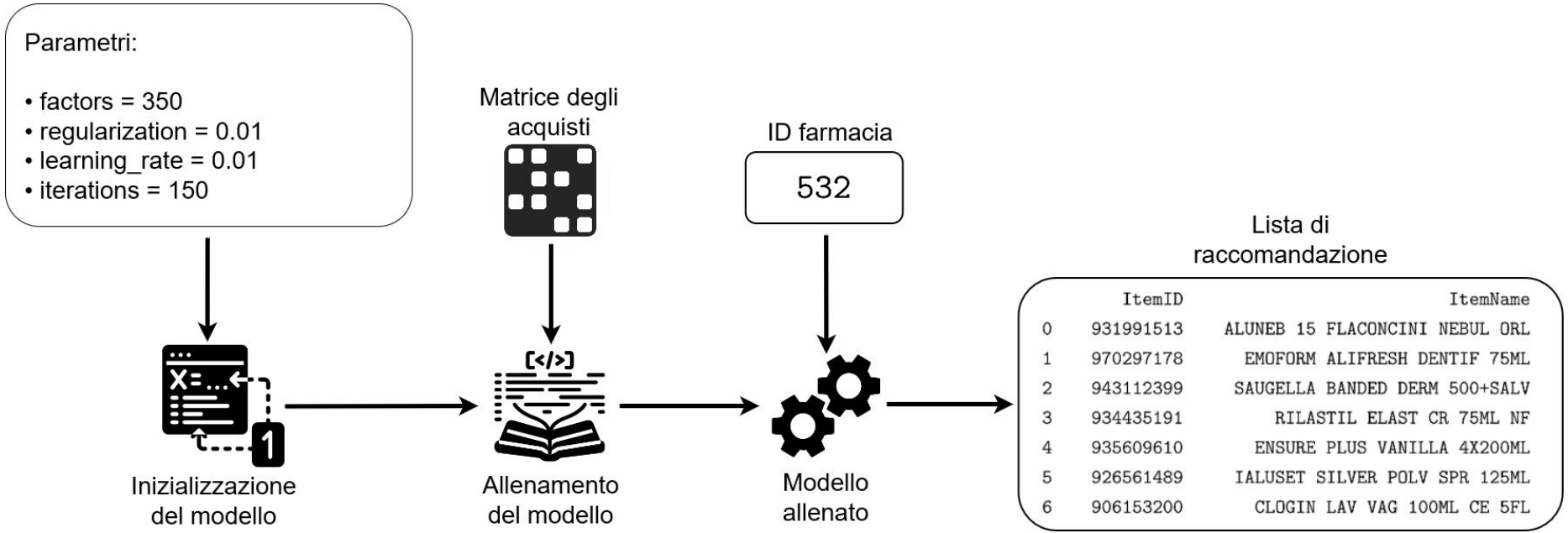
≈9.5 milioni di record

PharmacyID	ItemID	Quantity
15	1017	17
137	1017	3
137	96	7

≈ 3 milioni di record



Implementazione del Recommender System



	ItemID	ItemName
0	931991513	ALUNEB 15 FLACONCINI NEBUL ORL
1	970297178	EMOFORM ALIFRESH DENTIF 75ML
2	943112399	SAUGELLA BANDED DERM 500+SALV
3	934435191	RILASTIL ELAST CR 75ML NF
4	935609610	ENSURE PLUS VANILLA 4X200ML
5	926561489	IALUSET SILVER POLV SPR 125ML
6	906153200	CLOGIN LAV VAG 100ML CE 5FL

https://github.com/benfred/implicit

Testing del Recommender System

	ALS	BPR
Sparsity	97.68%	
Precision@5	23.01%	76.45%
Precision@10	23.42%	73.18%
MAP@5	14.45%	71.16%
MAP@10	12.04%	65.57%
nDCG@5	23.08%	77.39%
nDCG@10	23.34%	74.78%

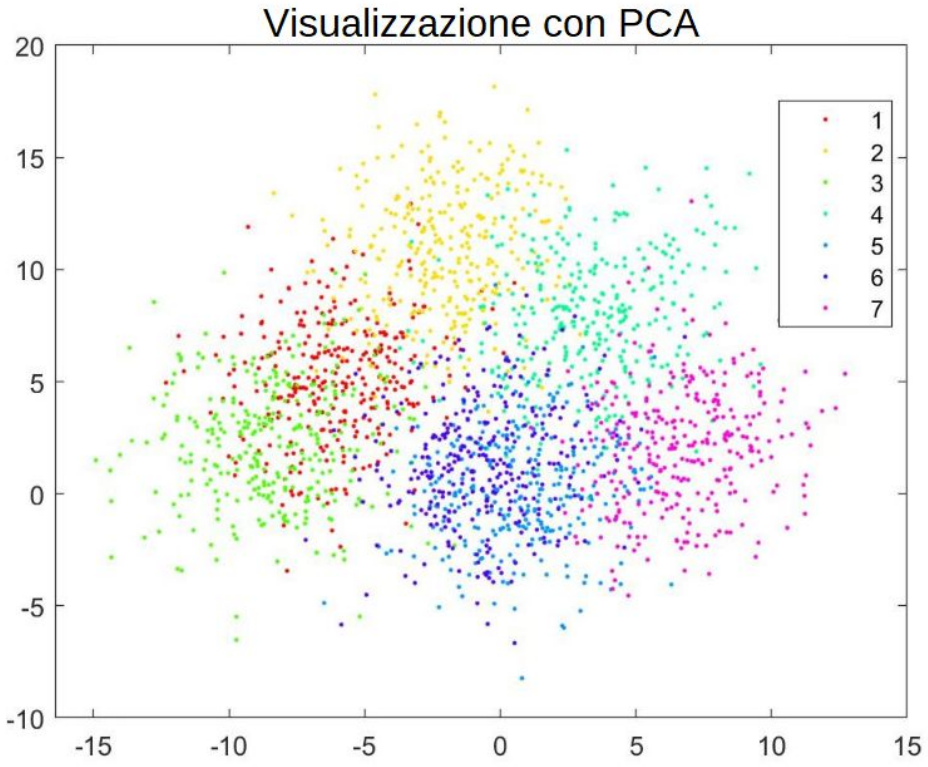
Suddivisione dei record nella fase di testing:

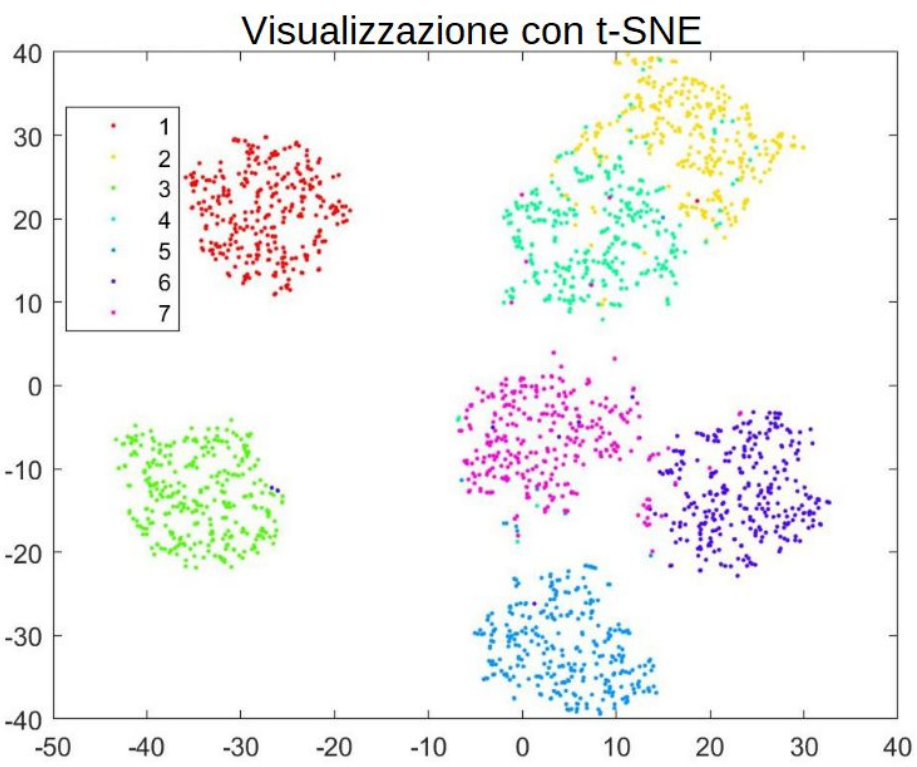
- 80% assegnati al training set
- 20% assegnati al test set

$$\text{Precision} = \frac{\#\text{ previsioni corrette}}{\#\text{ totale previsioni effettuate}}$$

Clustering

Riduzione della dimensionalità





Struttura del dataset delle farmacie

PharmacyID	1172
PharmacyName	FARMACIA XY
GroupID	61
GroupDES	FARMACIA XY
ConsortiumCode	5
ConsortiumDES	UNICONS
City	ROMA
CAP	00132
Province	RM
StartupDate	2014-01-01
Tutor	mariorossi
Revenue	734427.50
UsersNumber	938

Selezione degli attributi

Numero totale di farmacie: 893

	Numero di etichette
CAP	610
City	611
Province	94
GroupID	615
ConsortiumCode	9
Tutor	7
StartupDate	575

Numero di etichette uniche per attributo categorico

Selezione degli attributi

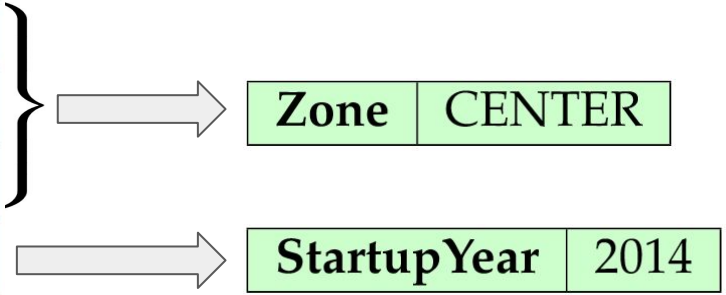
Numero totale di farmacie: 893

	Numero di etichette
CAP	610
City	611
Province	94
GroupID	615
ConsortiumCode	9
Tutor	7
StartupDate	575

Numero di etichette uniche per attributo categorico

Selezione degli attributi

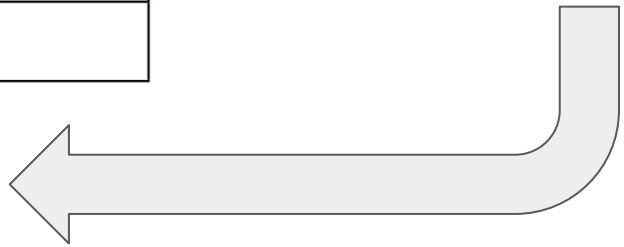
PharmacyID	1172
GroupID	61
ConsortiumCode	5
City	ROMA
CAP	00132
Province	RM
StartupDate	2014-01-01
Tutor	mariorossi
Revenue	734427.50
UsersNumber	938



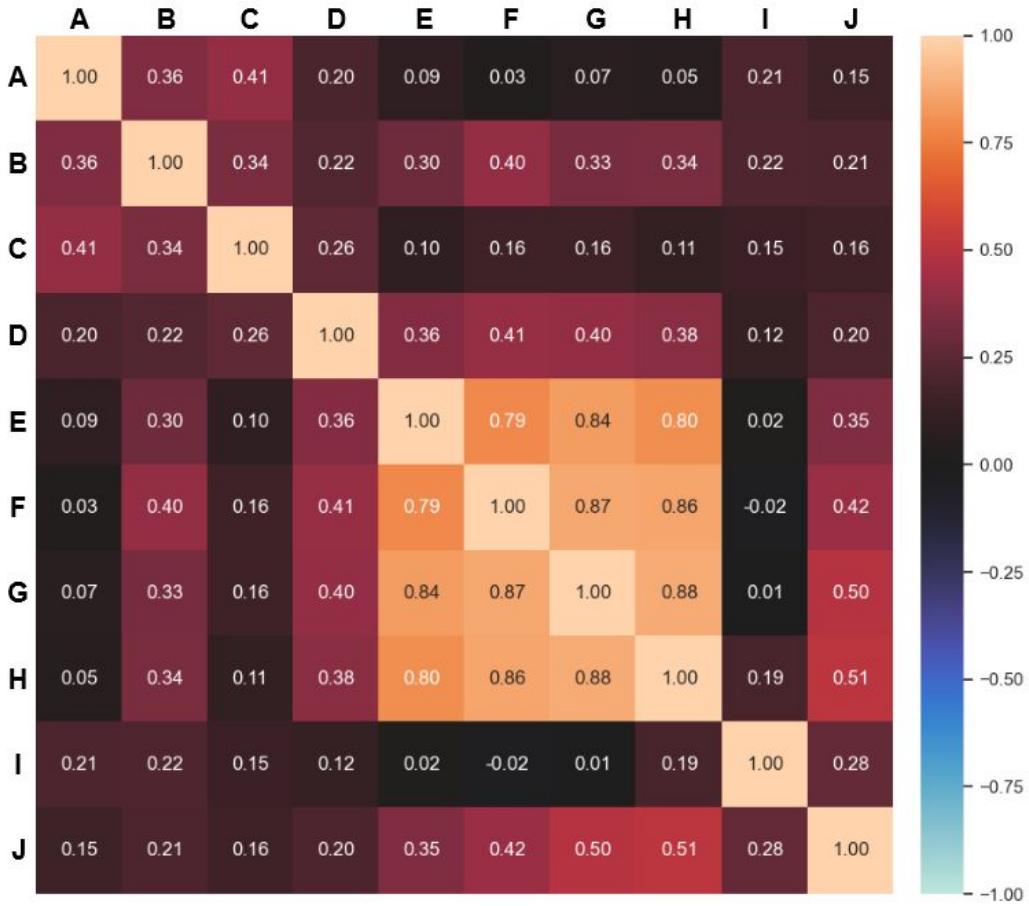
Selezione degli attributi

PharmacyID	1172
ConsortiumCode	5
Zone	CENTER
StartupYear	2014
Tutor	mariorossi
Revenue	734427.50
UsersNumber	938

ItemsNumber	2636
TotalQuantity	13810
SalesMean	1.7
SalesMax	38



Selezione degli attributi



A: *Zone*
B: *ConsortiumCode*
C: *Tutor*
D: *StartupYear*
E: *Revenue*
F: *UsersNumber*
G: *ItemsNumber*
H: *TotalQuantity*
I: *SalesMean*
J: *SalesMax*

Attributi rimossi:
- *UsersNumber*
- *ItemsNumber*
- *TotalQuantity*

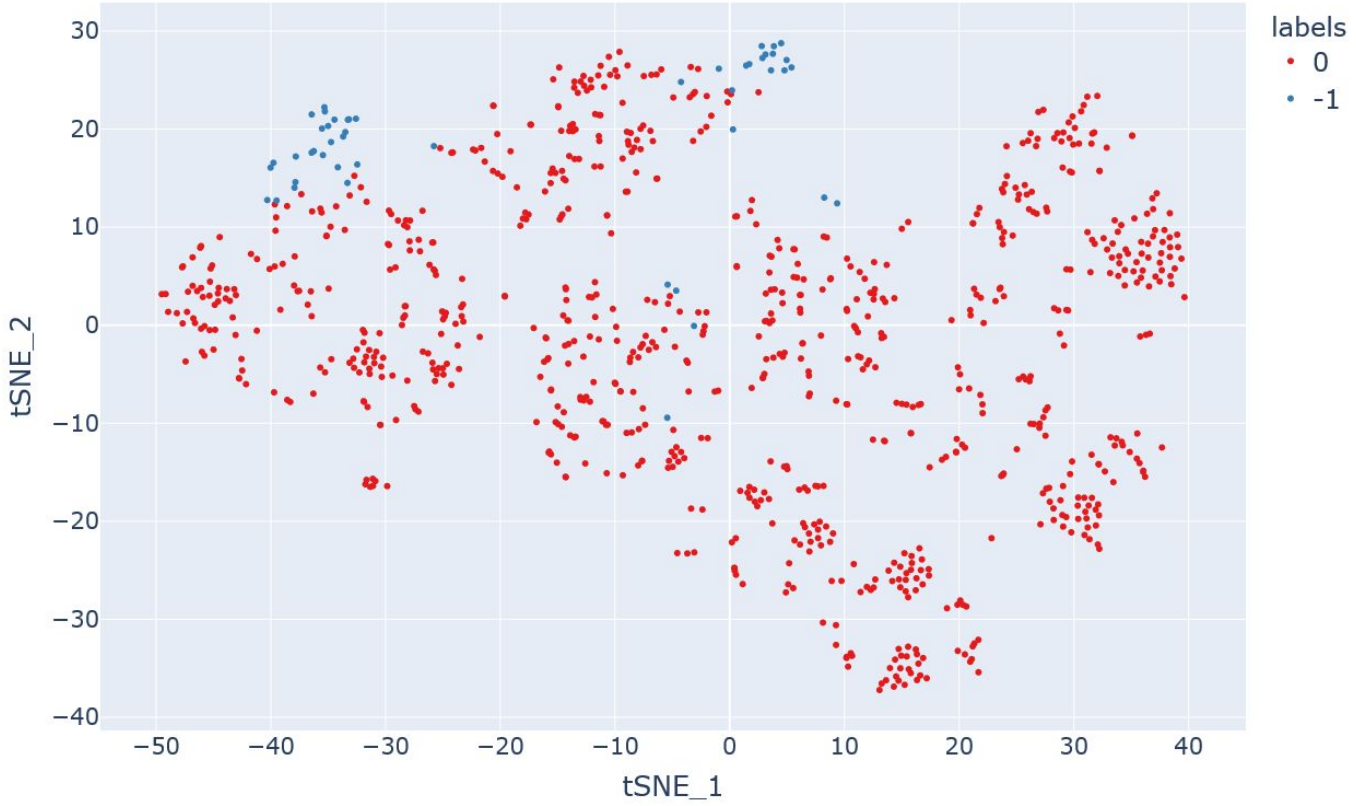
Clustering

Algoritmi testati:

- DBSCAN (density-based)
- K-means (partizionale)
- Agglomerative Clustering (gerarchico)

Clustering con DBSCAN

Visualizzazione t-SNE

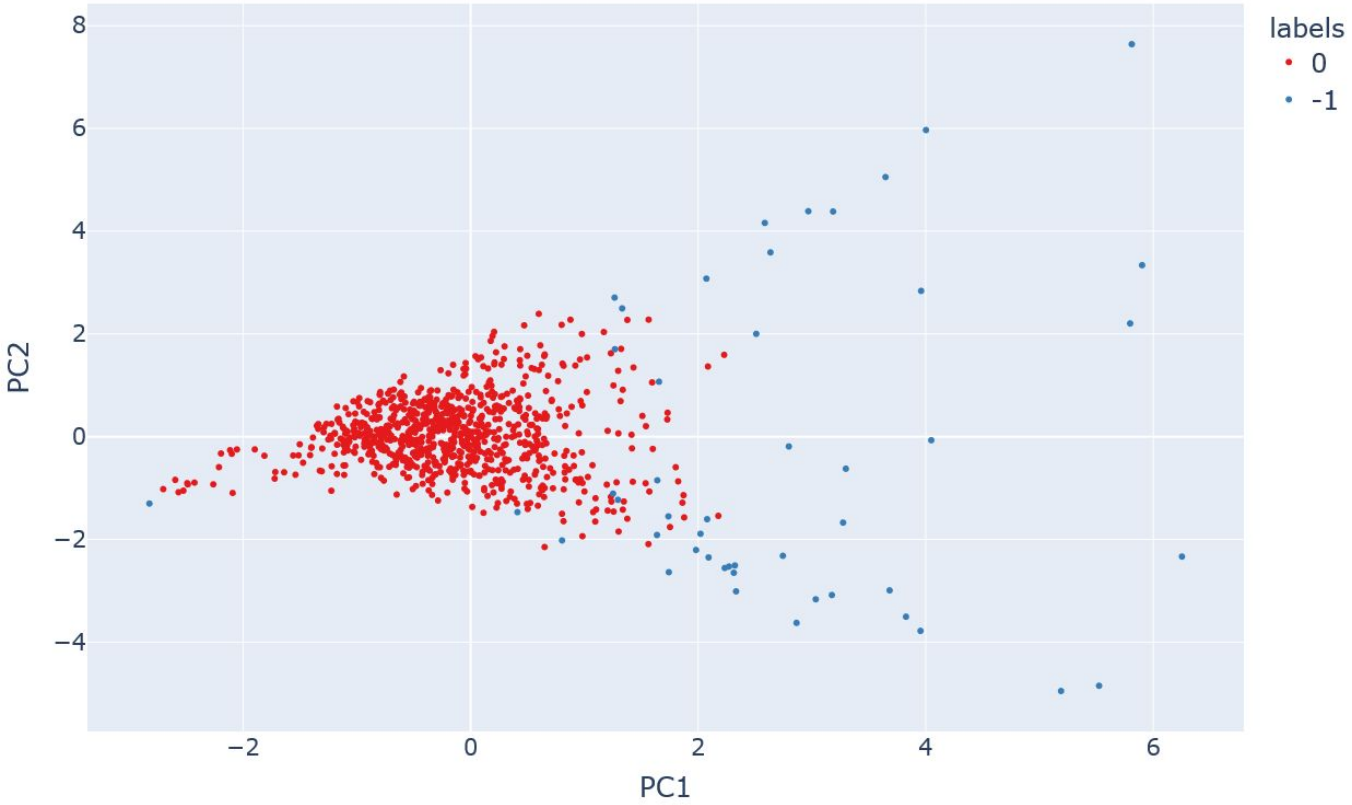


Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

Parametri:
- *minPts=10*
- *eps=1.85*

Clustering con DBSCAN

Visualizzazione PCA

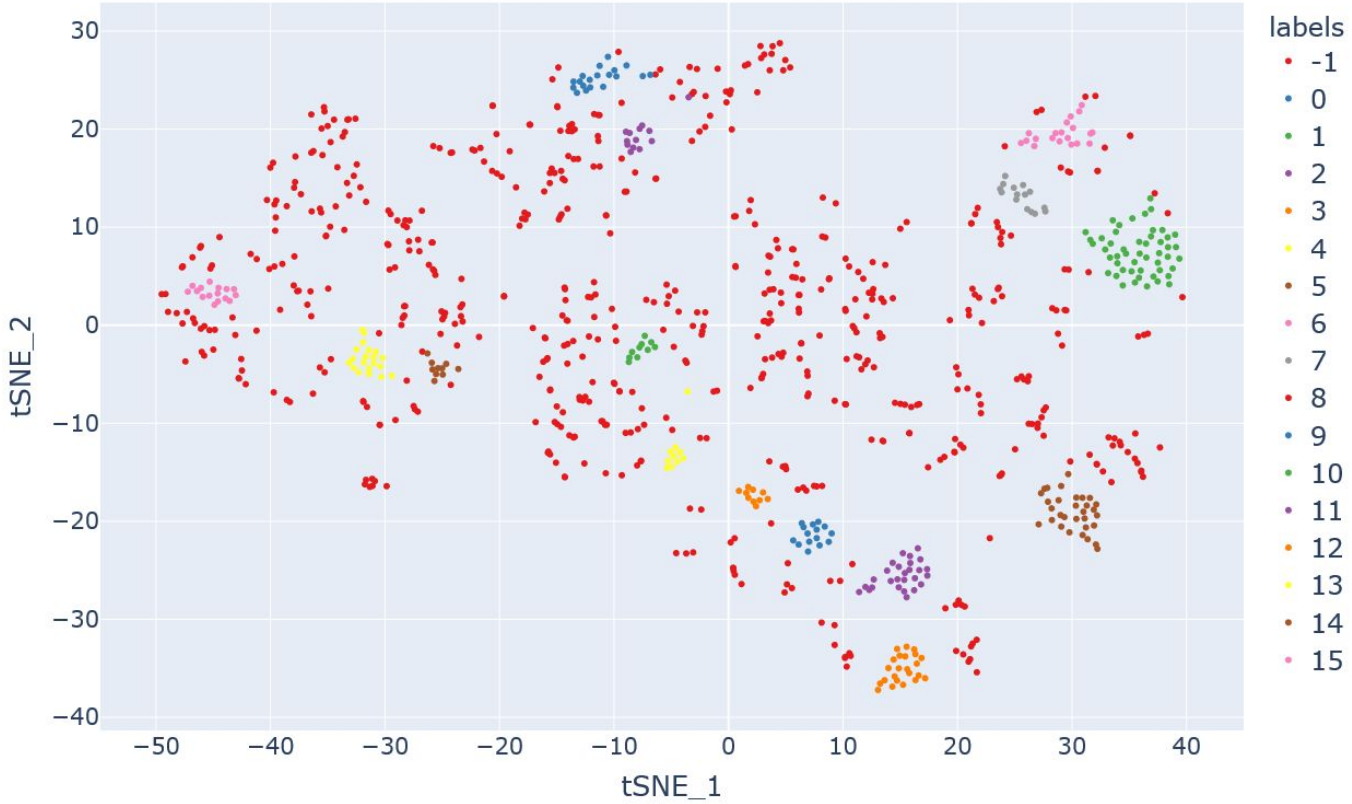


Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

Parametri:
- *minPts=10*
- *eps=1.85*

Clustering con DBSCAN

Visualizzazione t-SNE



Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

Parametri:
- *minPts=10*
- *eps=1.25*

Clustering con K-means

Visualizzazione PCA

Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

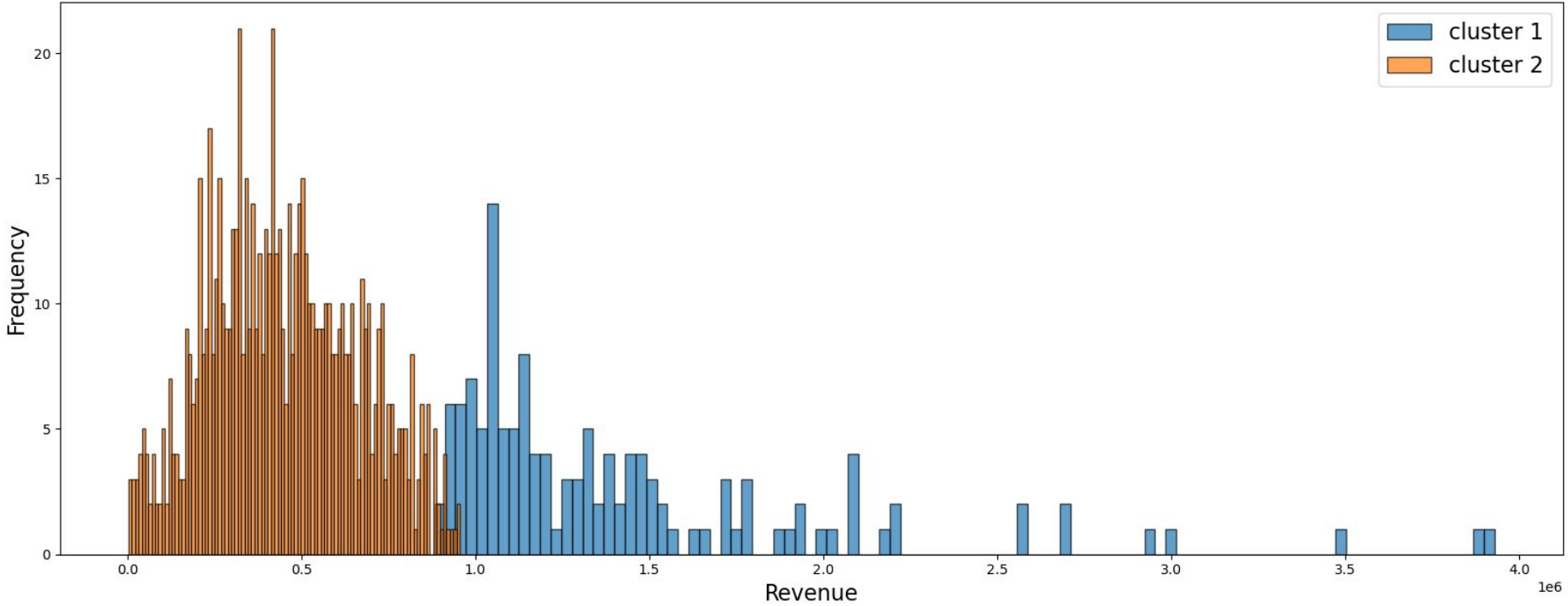
Parametri:
- *K = 2*

Clustering con K-means

	Cluster 1	Cluster 2
N° di elementi	131	762
Precision@5	89.16%	75.58%

Risultati per suddivisione in 2 cluster

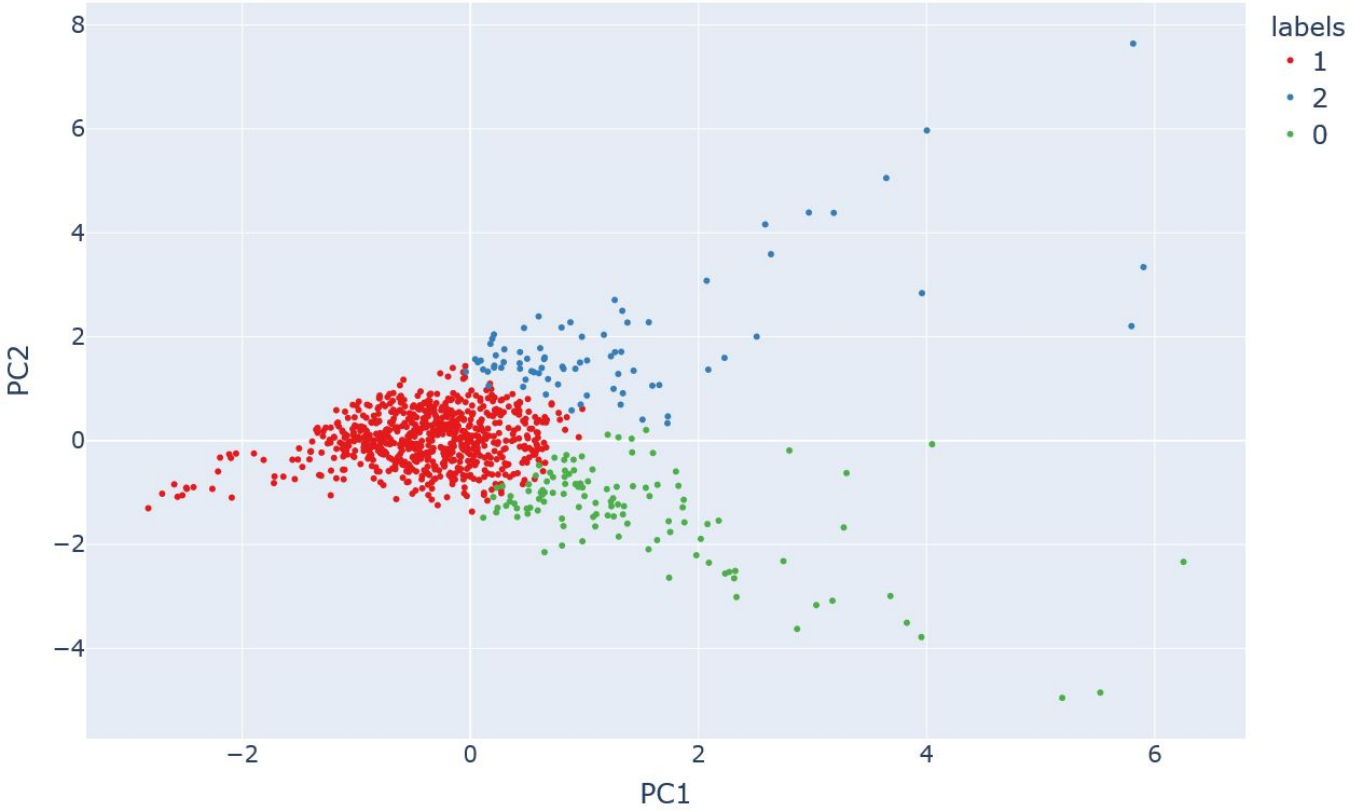
Clustering con K-means



Distribuzione della variabile Revenue

Clustering con K-means

Visualizzazione PCA



Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

Parametri:
- *K = 3*

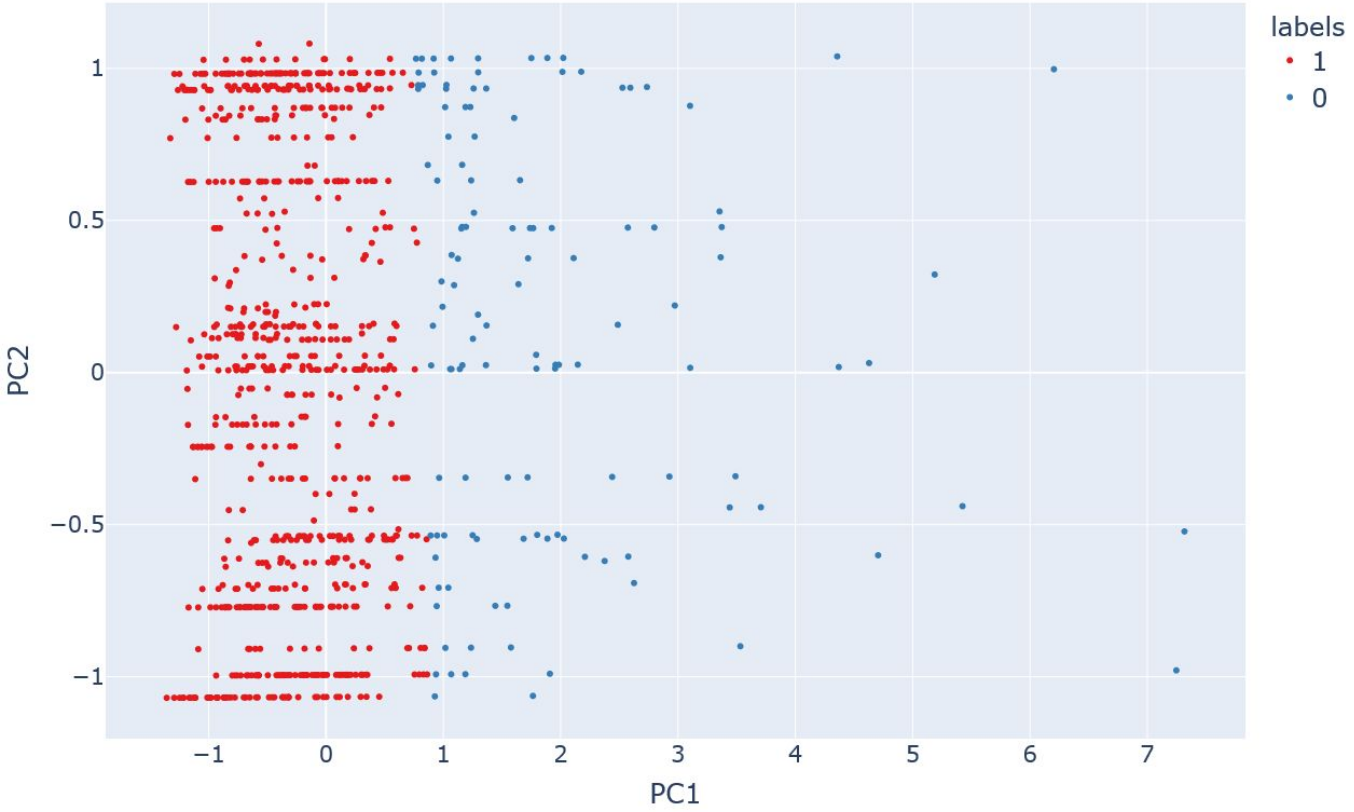
Clustering con K-means

	Cluster 1	Cluster 2	Cluster 3
N° di elementi	116	700	77
Precision@5	88.96%	75.42%	56.81%

Risultati per suddivisione in 3 cluster

Clustering con K-means

Visualizzazione PCA

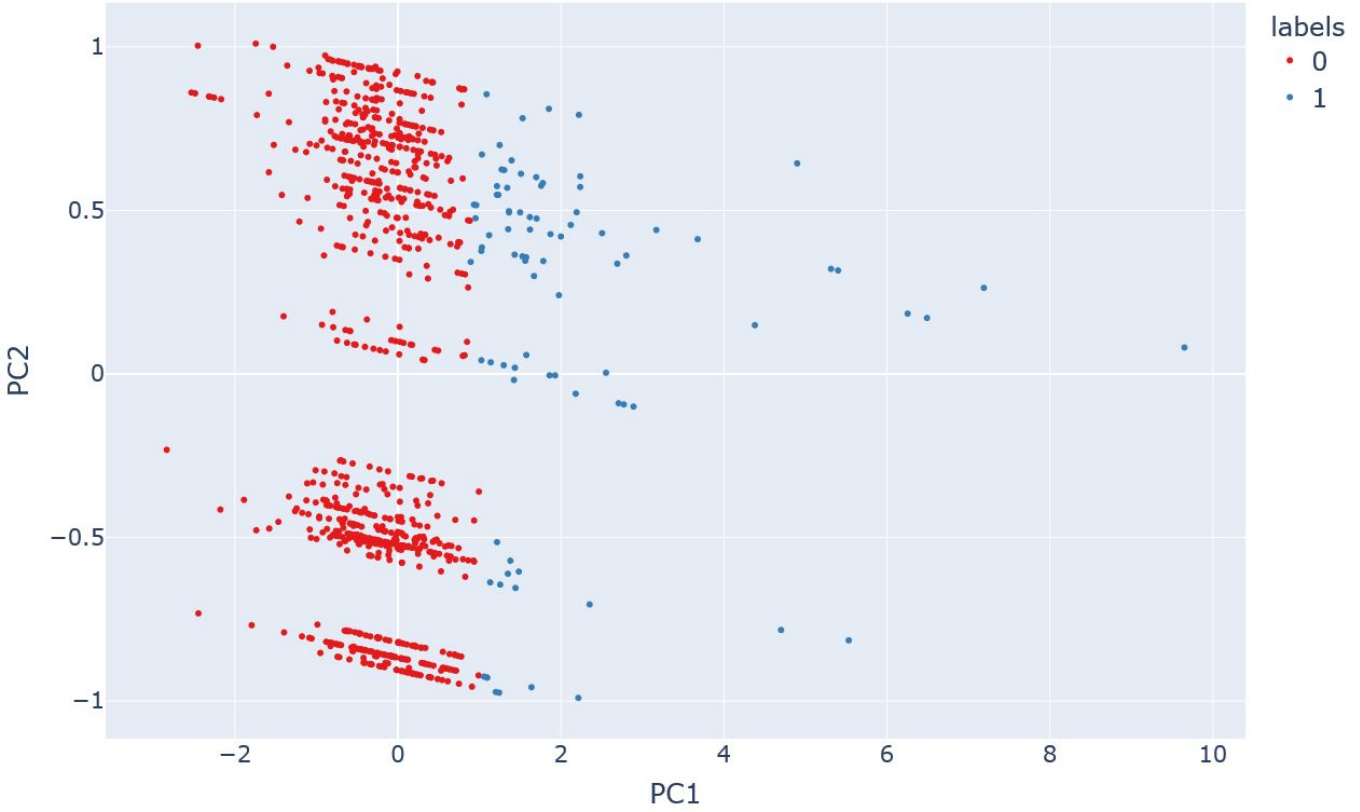


Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *Revenue*

Parametri:
- *K = 2*

Clustering con K-means

Visualizzazione PCA



Attributi:
- *Tutor*
- *Zone*
- *StartupYear*
- *SalesMean*

Parametri:
- *K = 2*

Clustering con K-means

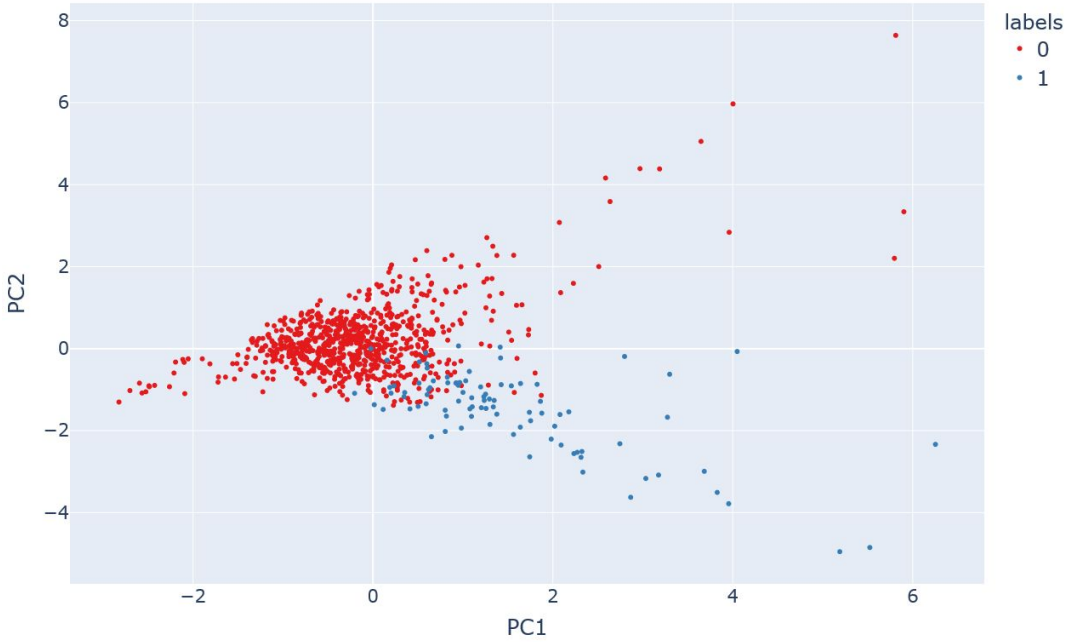
	Cluster 1	Cluster 2
N° di elementi	123	770
Precision@5	85.85%	75.53%

Risultati per suddivisione in 2 cluster con attributi *Tutor, ConsortiumCode, Zone, Revenue*

	Cluster 1	Cluster 2
N° di elementi	87	806
Precision@5	61.45%	75.69%

Risultati per suddivisione in 2 cluster con attributi *Tutor, Zone, StartupYear, SalesMean*

Agglomerative Clustering



Attributi:
- *Tutor*
- *ConsortiumCode*
- *Zone*
- *StartupYear*
- *Revenue*
- *SalesMean*

Parametri:
- *K = 2*

	Cluster 1	Cluster 2
N° di elementi	90	803
Precision@5	88.99%	75.87%

Conclusioni

- Risultati prima parte:
  - L’algoritmo BPR ha offerto ottimi risultati con un valore di Precision@5 del sistema superiore al 75%.

- Risultati seconda parte:
  - L’algoritmo DBSCAN si è rivelato inefficace.
  - L’applicazione degli algoritmi K-means ed Agglomerative Clustering hanno portato ad un miglioramento delle performance del Recommender System, seppure parziale. Possibilità di utilizzare RS specializzato per punti vendita ad alta Revenue.

Sviluppi futuri

- Arricchimento del dataset con attributi più significativi per le operazioni di clustering.

- Studio del funzionamento di algoritmi di clustering con una funzione obiettivo personalizzata, che utilizzi come valore da massimizzare la *Precision@5* restituita in output dal test delle performance del Recommender System.

Grazie per l’attenzione