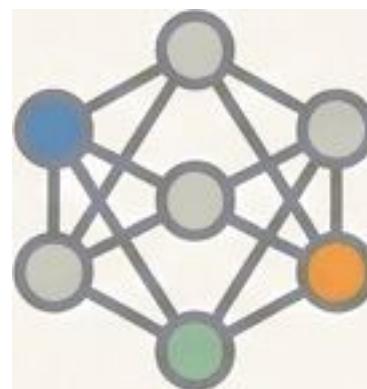


Recommender Systems: Un estudi comparatiu de Sistemes de Recomanació

Pere Maeso, David Miquel,
Sergi Escudero i Marc Arroyo

La nostra missió: Estudiar diferents recomanadors a través de l'experimentació

El nostre objectiu és una anàlisi empírica i comparativa dels mètodes de recomanació,avaluant com el seu rendiment es veu afectat per quatre factors clau:



Mètodes
d'aprenentatge
computacional
diversos



Dos datasets
amb diferents
volums de dades



Impacte del
filtratge de
dades



L'ajust fi
d'hiperparàmetres

Un ventall de tècniques de recomanació

Hem implementat tres famílies principals d'algorismes:



Filtratge Col·laboratiu (User-User & Ítem-Ítem)

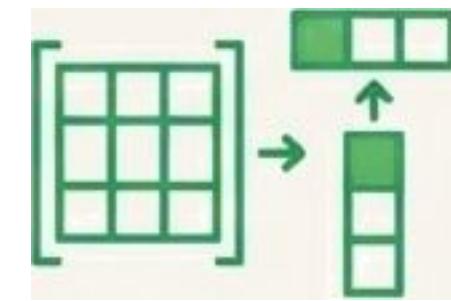
Implementacions: Llibreria
vs. Manual.

Mètriques de similitud:
Pearson, Cosinus Ajustat.



Content-Based

Recomanacions basades
en les característiques
dels ítems.

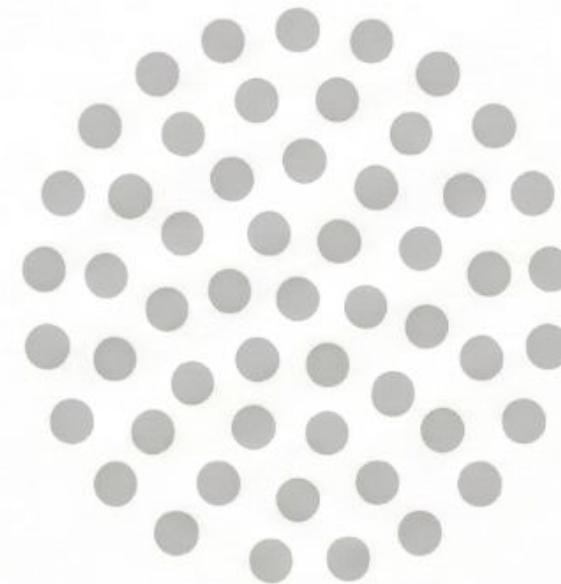


Factorització Matricial (Funk-SVD)

Descobriment de
factors latents en les
dades.

Els datasets: Dos escenaris per posar a prova els models

MovieLens Small

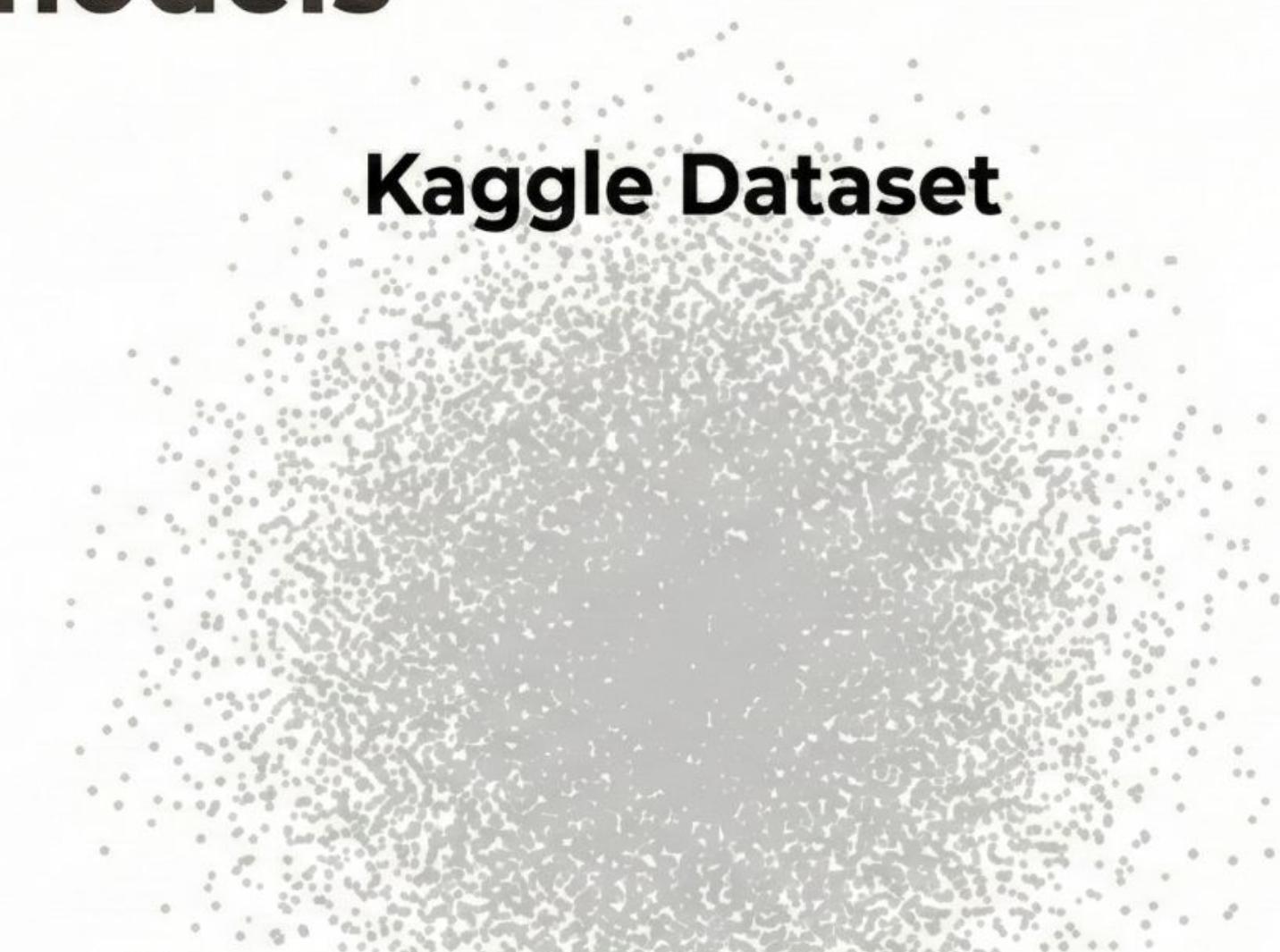


Entorn controlat, ideal per a proves ràpides.

Content-Based basat en **gèneres**.

~100k valoracions

Kaggle Dataset

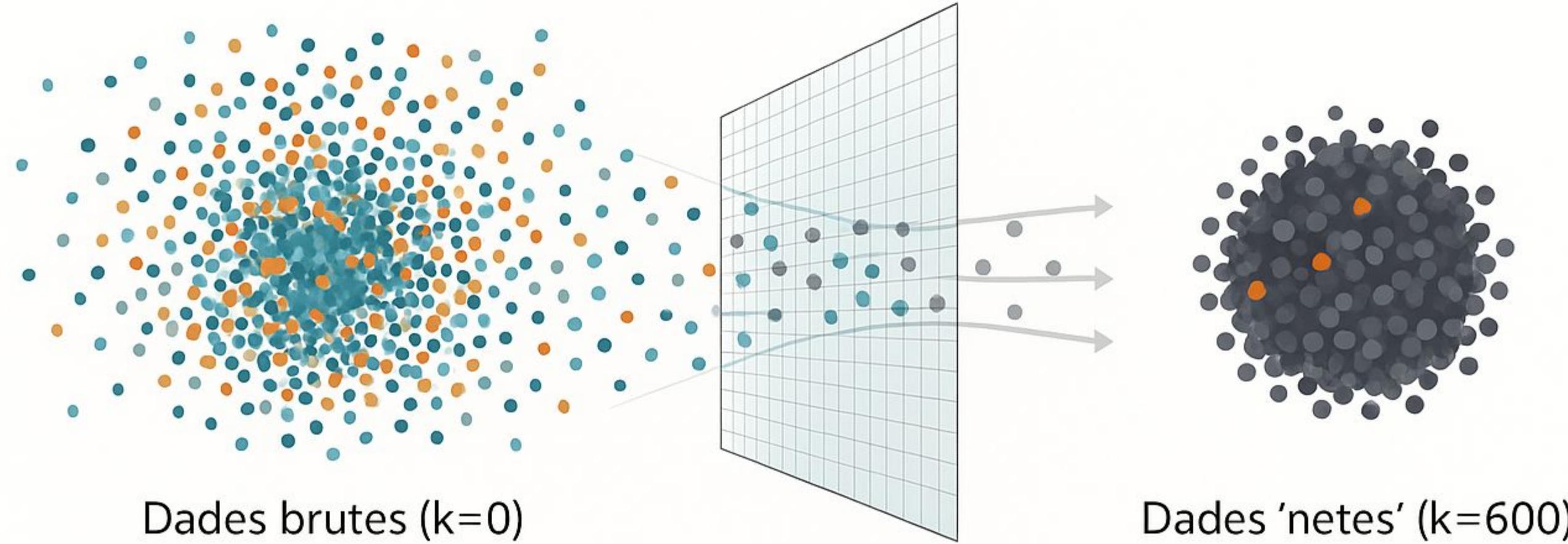


Entorn realista, complex i a gran escala.

Content-Based basat en **keywords amb TF-IDF**.

~26M valoracions

Una variable oculta: L'impacte del filtratge de dades



Què fem?

Eliminem usuaris i pel·licules amb poques valoracions (filtre ' k ').

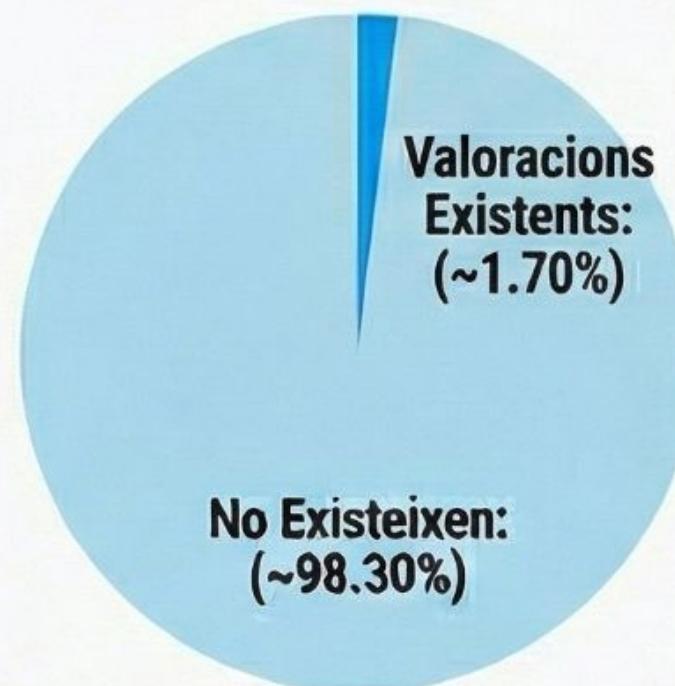
Per què?

Per reduir el soroll.

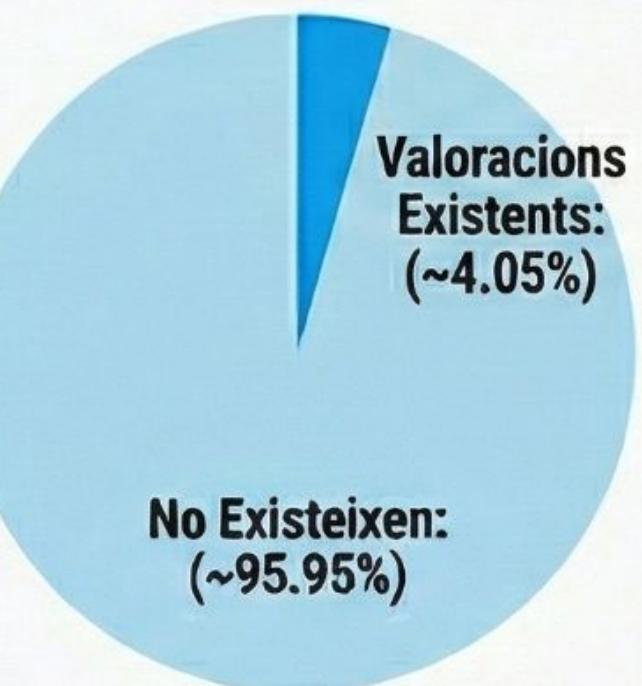
Percentatge de Valoracions Existents Pre- i Post-Filtratge Comparació

Impacte del Filtratge al Dataset MovieLens (Simple)

Abans del Filtratge



Després del Filtratge

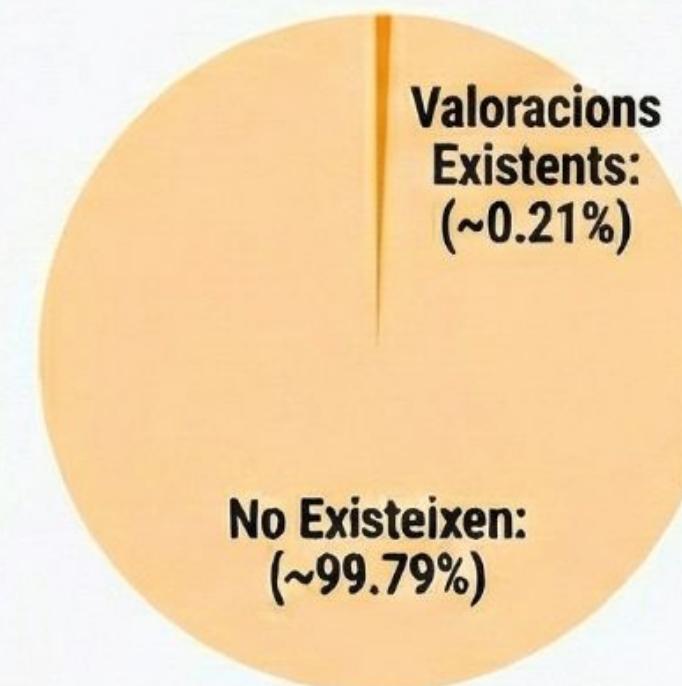


Pel·lícules: 9742, Usuaris: 610
Valoracions: 100.8k

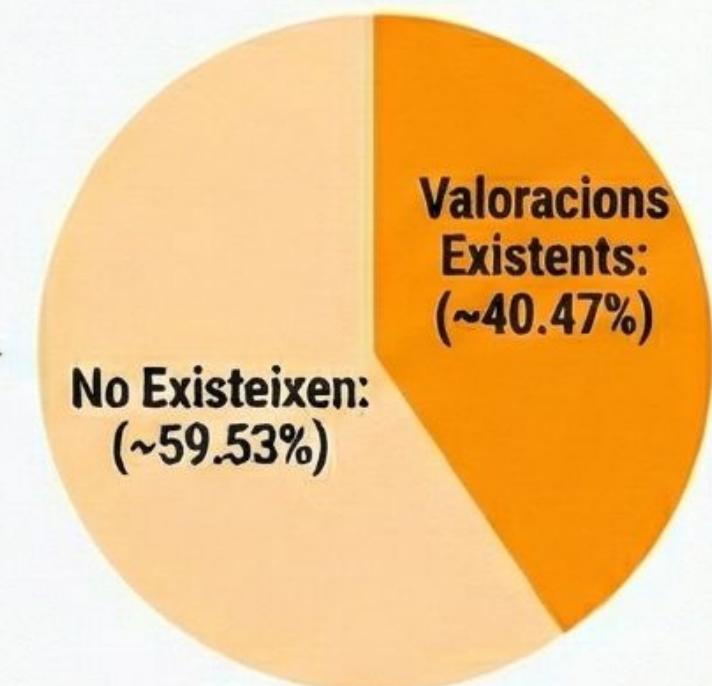
Pel·lícules: 3650, Usuaris: 610
Valoracions: 90.3k

Impacte del Filtratge al Dataset Kaggle

Abans del Filtratge



Després del Filtratge



Pel·lícules: 46,483, Usuaris: 270,896
Valoracions: 26.02M

Pel·lícules: 2,084, Usuaris: 3,312
Valoracions: 2.79M

LEAVE-ONE-OUT CROSS-VALIDATION

ORIGINAL DATASET

| User ID | Movie ID | Rating | Timestamp |
|---------|----------|--------|------------|
| 1 | 101 | 5 | 1234567890 |
| 1 | 102 | 4 | 1234567900 |
| 1 | 103 | 3 | 1234568000 |
| 2 | 104 | 5 | 1234568100 |
| 2 | 105 | 4 | 1234568200 |
| 2 | 106 | 2 | 1234568300 |

TRAINING SET (Two ratings per user)

| User ID | Movie ID | Rating | Timestamp |
|---------|----------|--------|------------|
| User 1 | 101 | 5 | 1234567890 |
| User 1 | 102 | 4 | 1234567900 |
| User 2 | 104 | 5 | 1234568100 |
| User 2 | 105 | 4 | 1234568200 |

TEST SET (One rating left out per user)

| User ID | Movie ID | Rating | Timestamp |
|---------|----------|--------|------------|
| User 1 | 103 | 3 | 1234568000 |
| User 2 | 106 | 2 | 1234568300 |

Un sistema interactiu per a l'exploració i la predicción

Tota l'anàlisi s'ha integrat en un menú interactiu que permet:

- Seleccionar el dataset (`Small` / `Kaggle`).
- Definir el nivell de filtratge (`k`).
- Predir valoracions individuals.
- Generar recomanacions Top-N personalitzades.
- Calcular totes les mètriques i gràfics presentats.

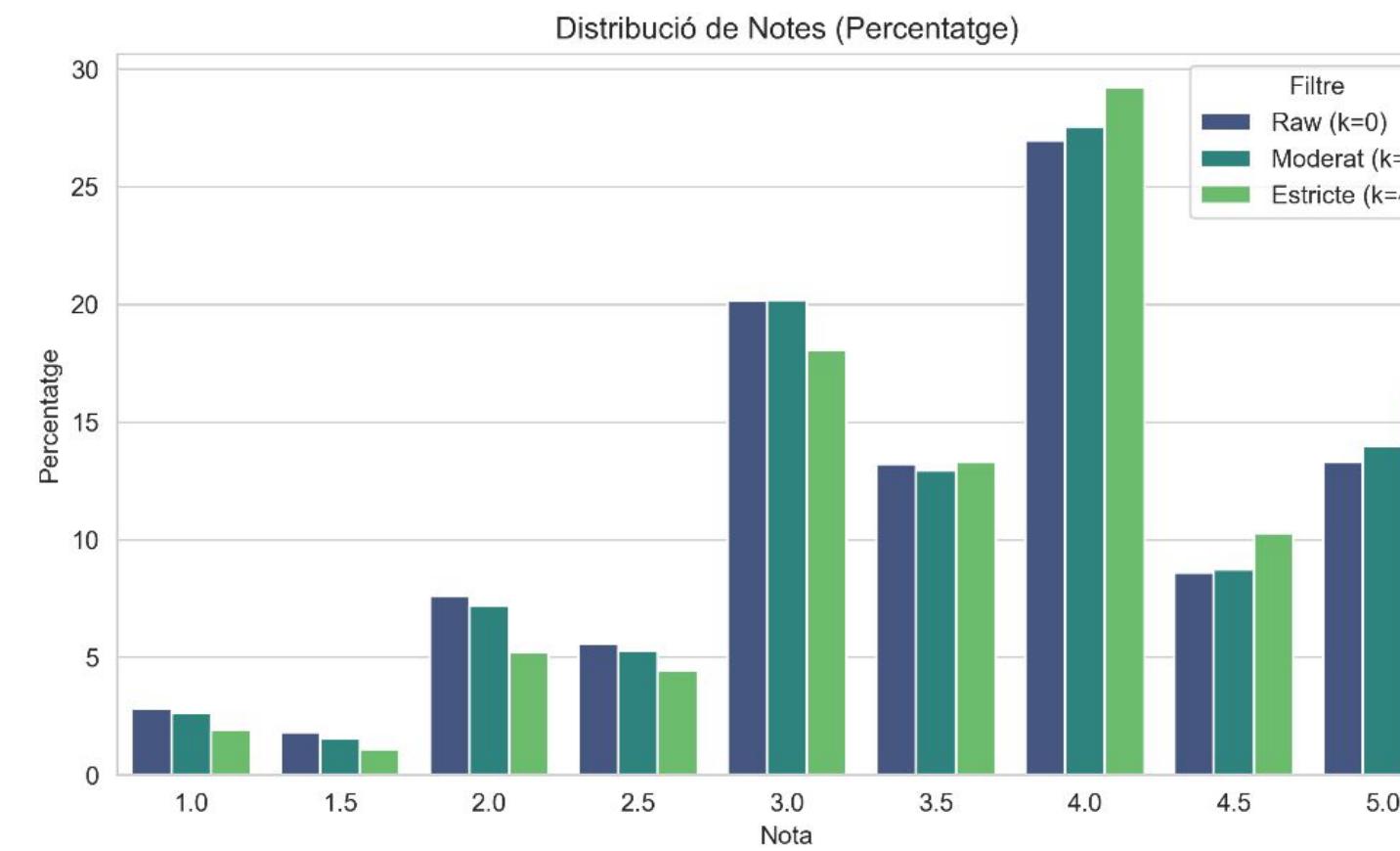
==== SISTEMA MULTI-DATASET (Actual: Kaggle (k=300)) ===

1. Seleccionar Dataset i Carregar (Small / Kaggle)
2. Predir Valoració
3. Recomanar Top-N
4. Mètriques i Gràfics
5. Calcular Matrius Manuals (Lent)
6. Sortir

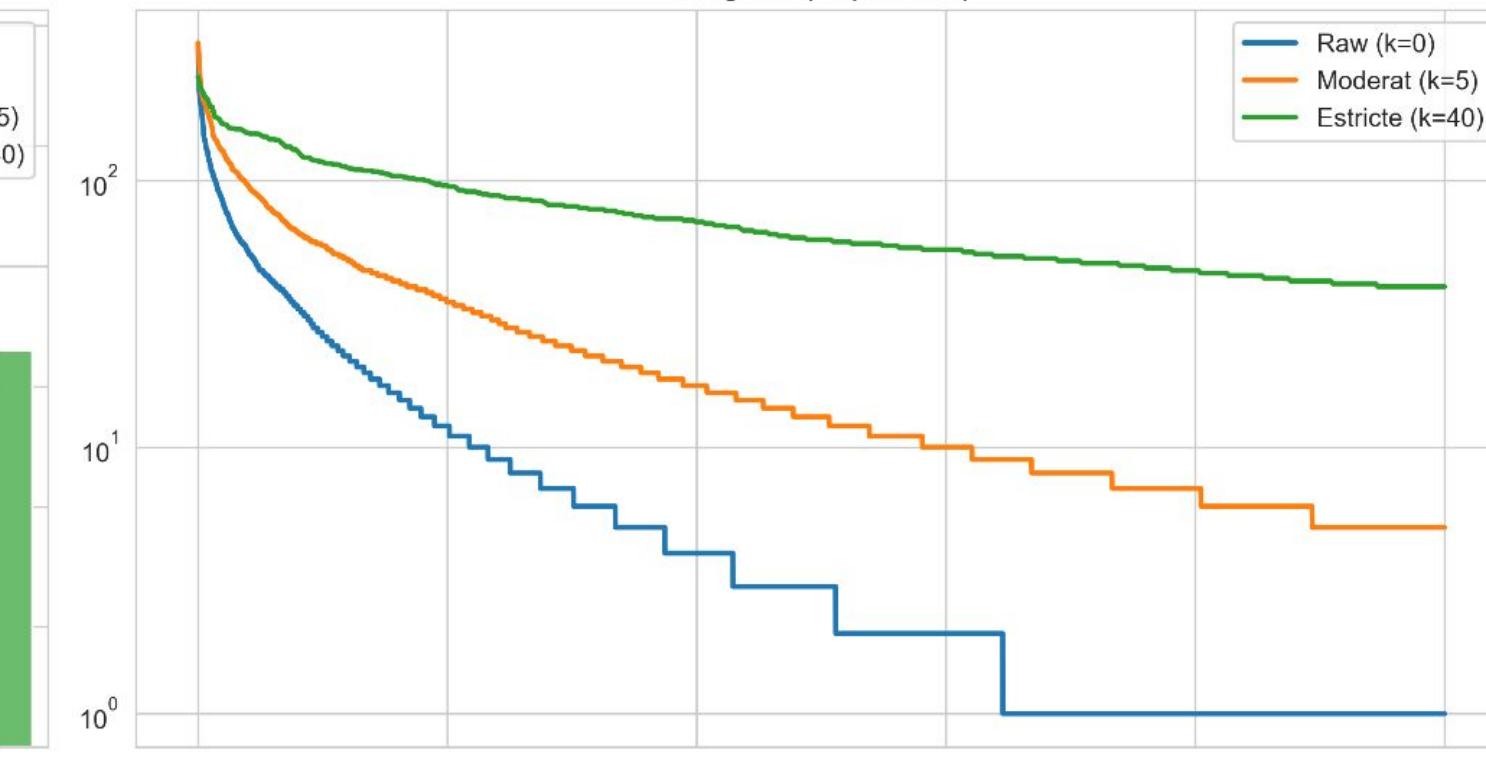
Opció: ■

Estudi biax de les dades (Small)

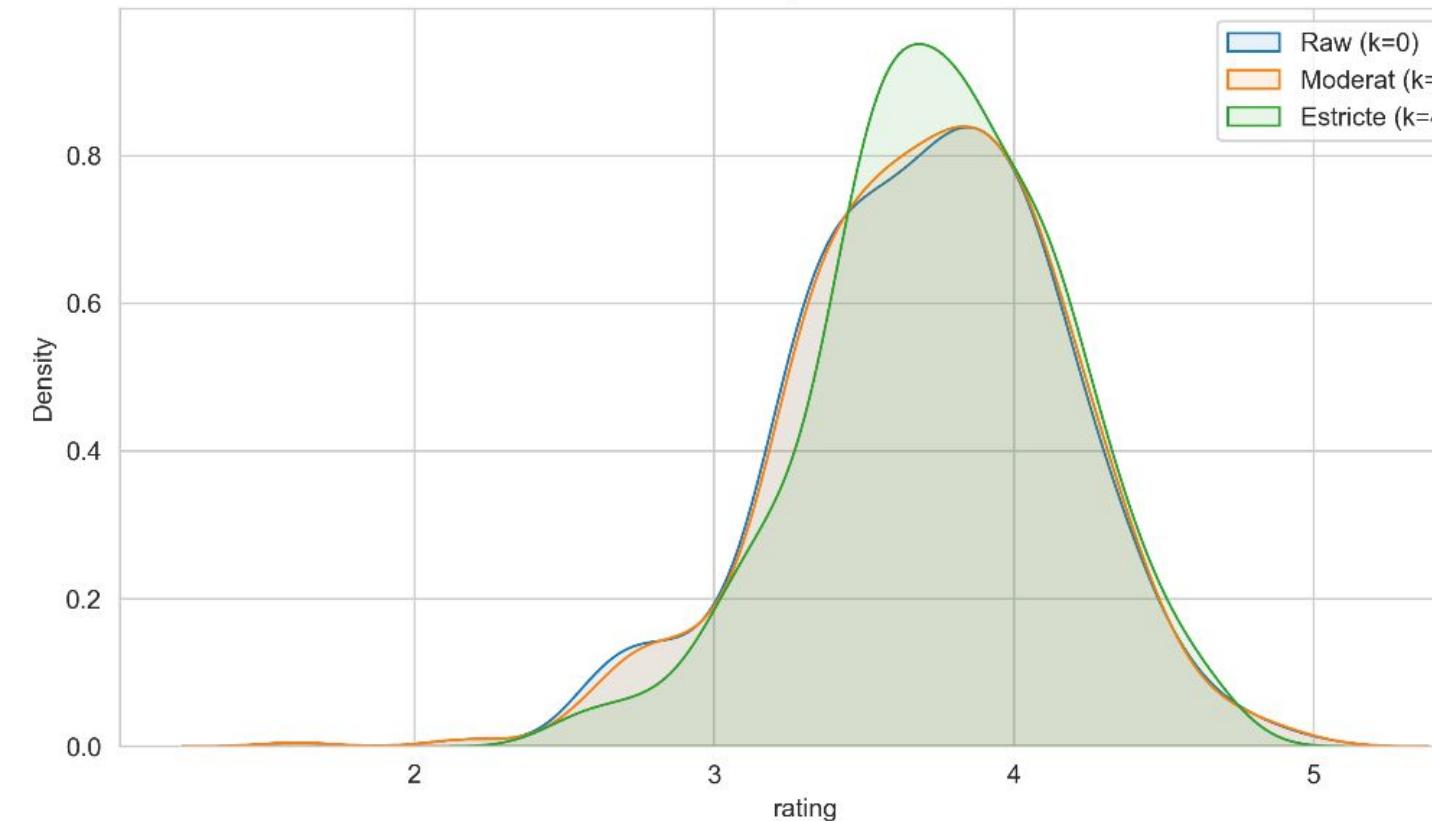
Impacte del Filtratge en el Biaix del Dataset



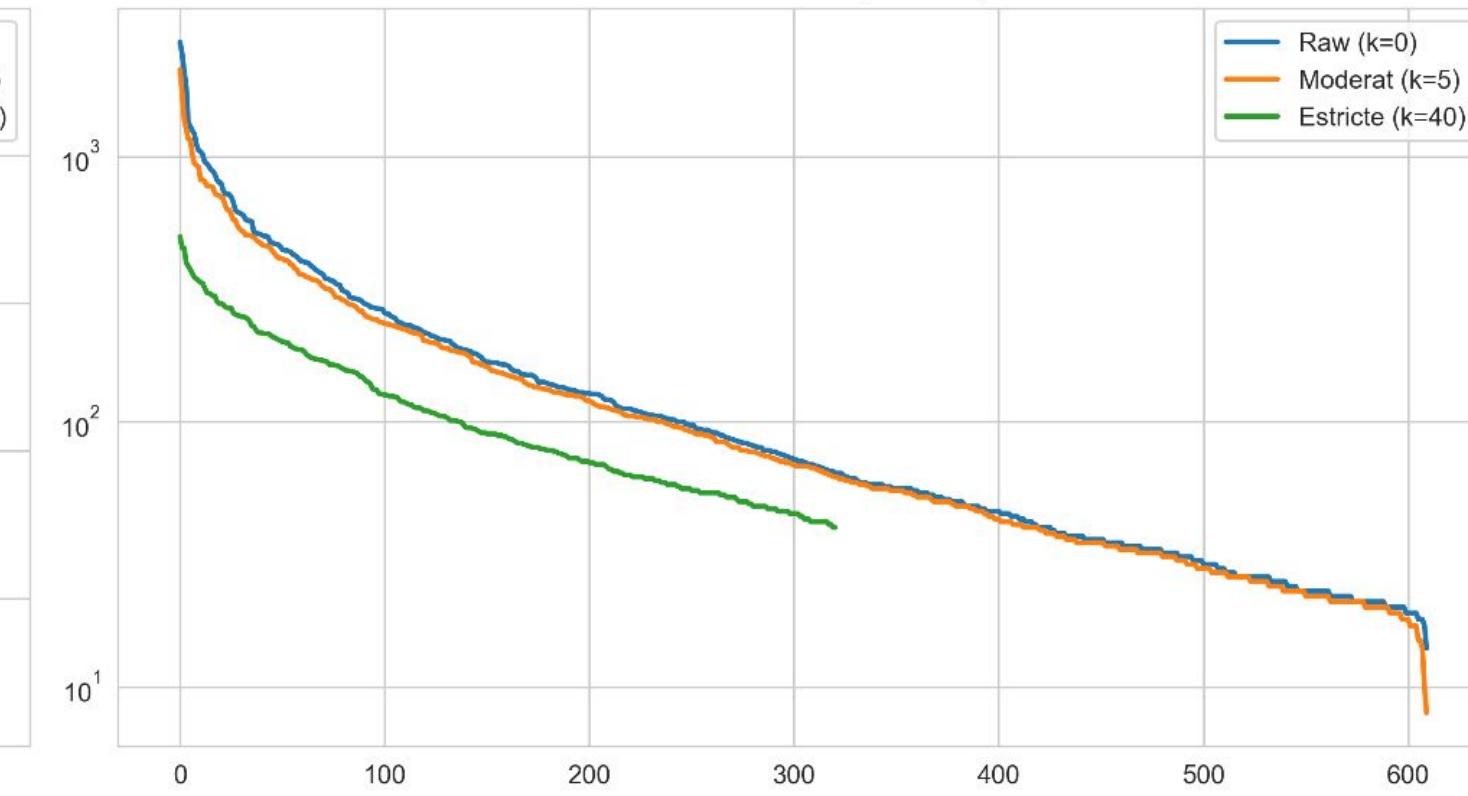
Long Tail (Popularitat)



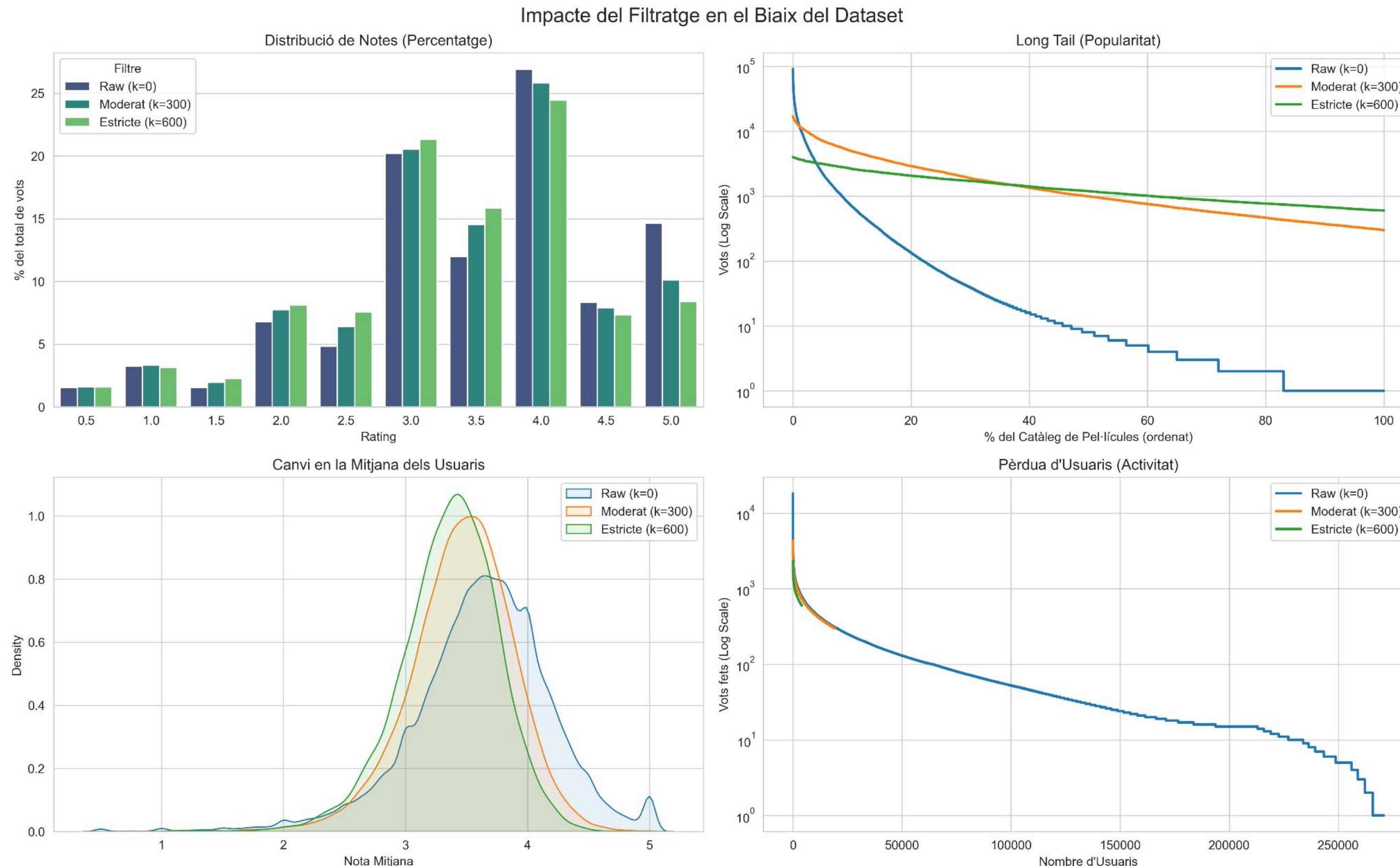
Canvi en la Mitjana dels Usuaris



Pèrdua d'Usuaris (Activitat)

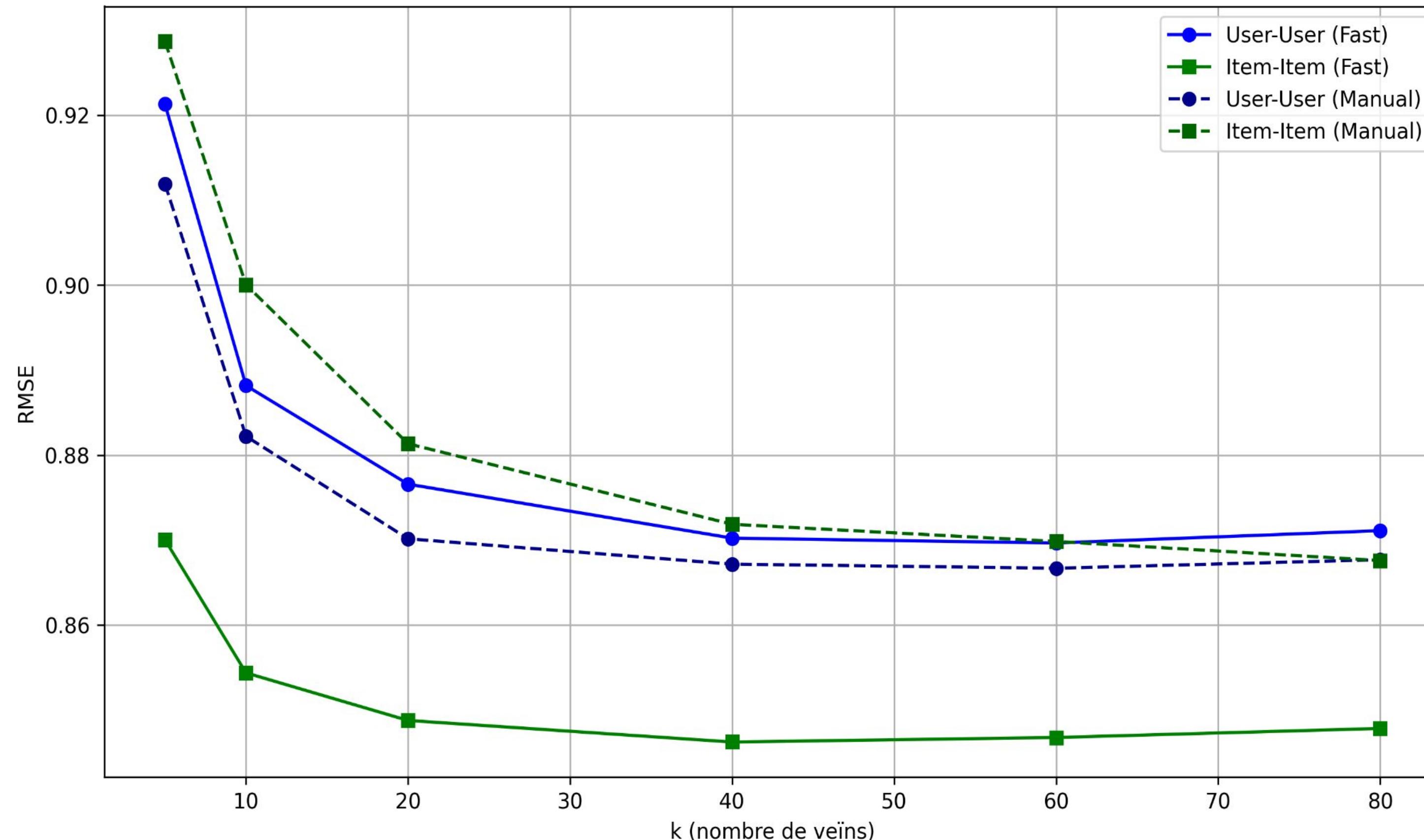


Estudi biax de les dades (Kaggle)



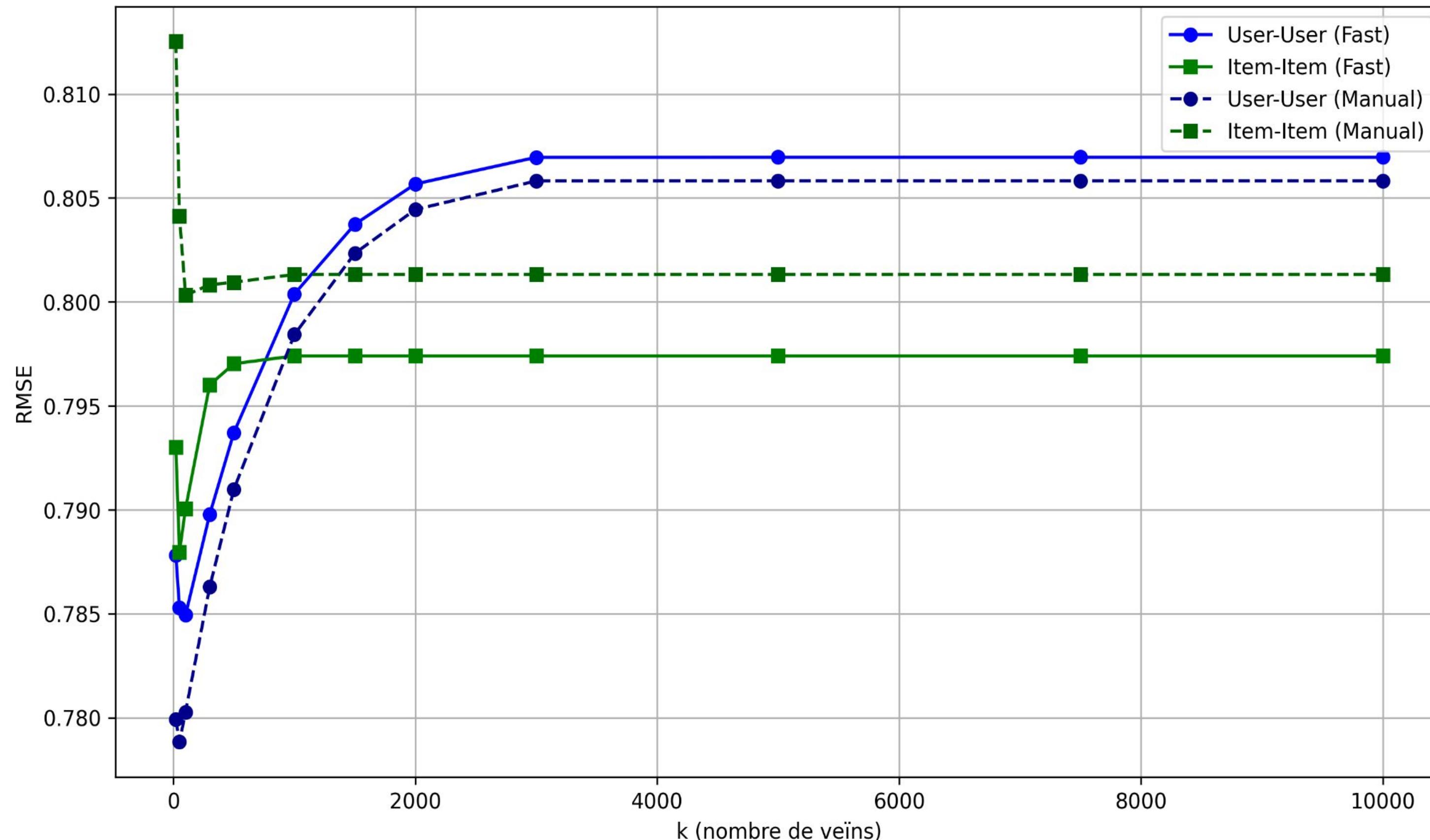
Estudi Hiperparàmetre K (Recomanador Col·laboratiu (Small))

RMSE vs k (veïns) - Comparació Fast vs Manual



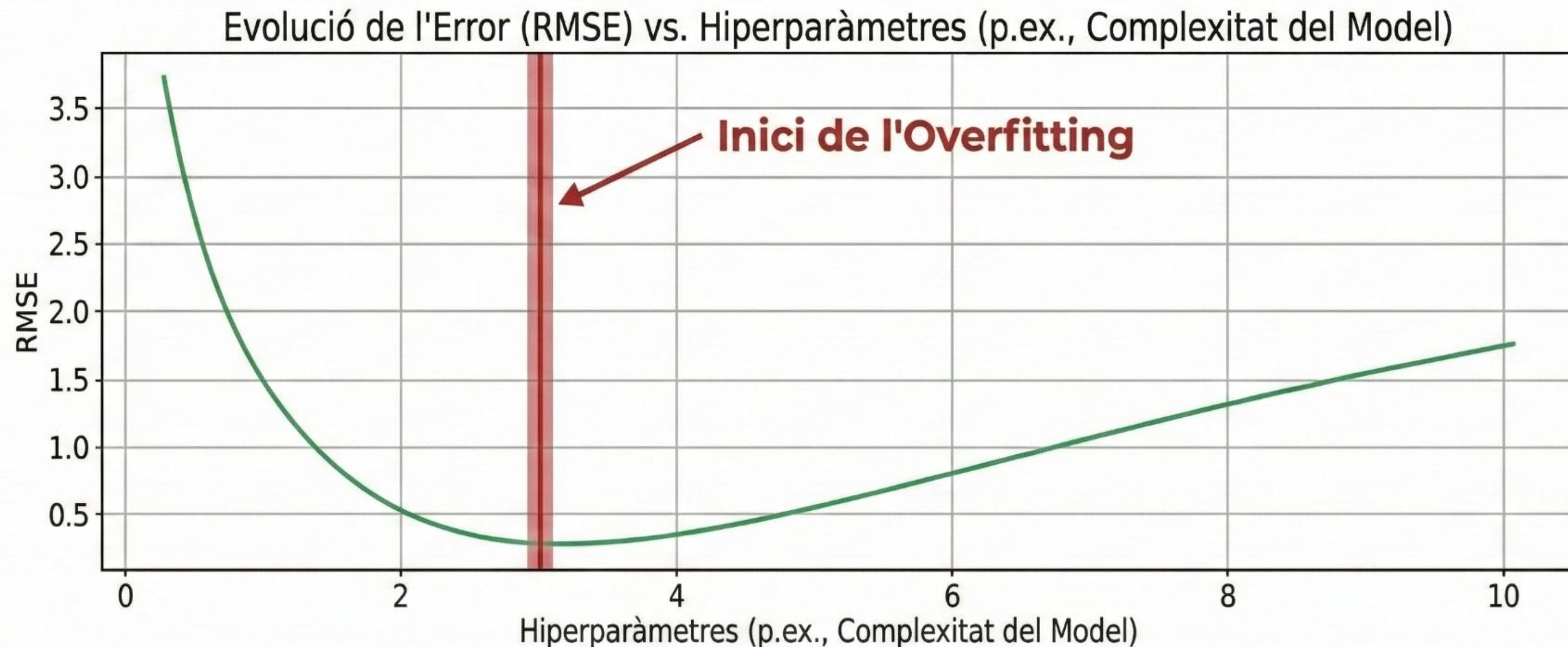
Estudi Hiperparàmetre K (Recomanador Col·laboratiu (Kaggle))

RMSE vs k (veïns) - Comparació Fast vs Manual

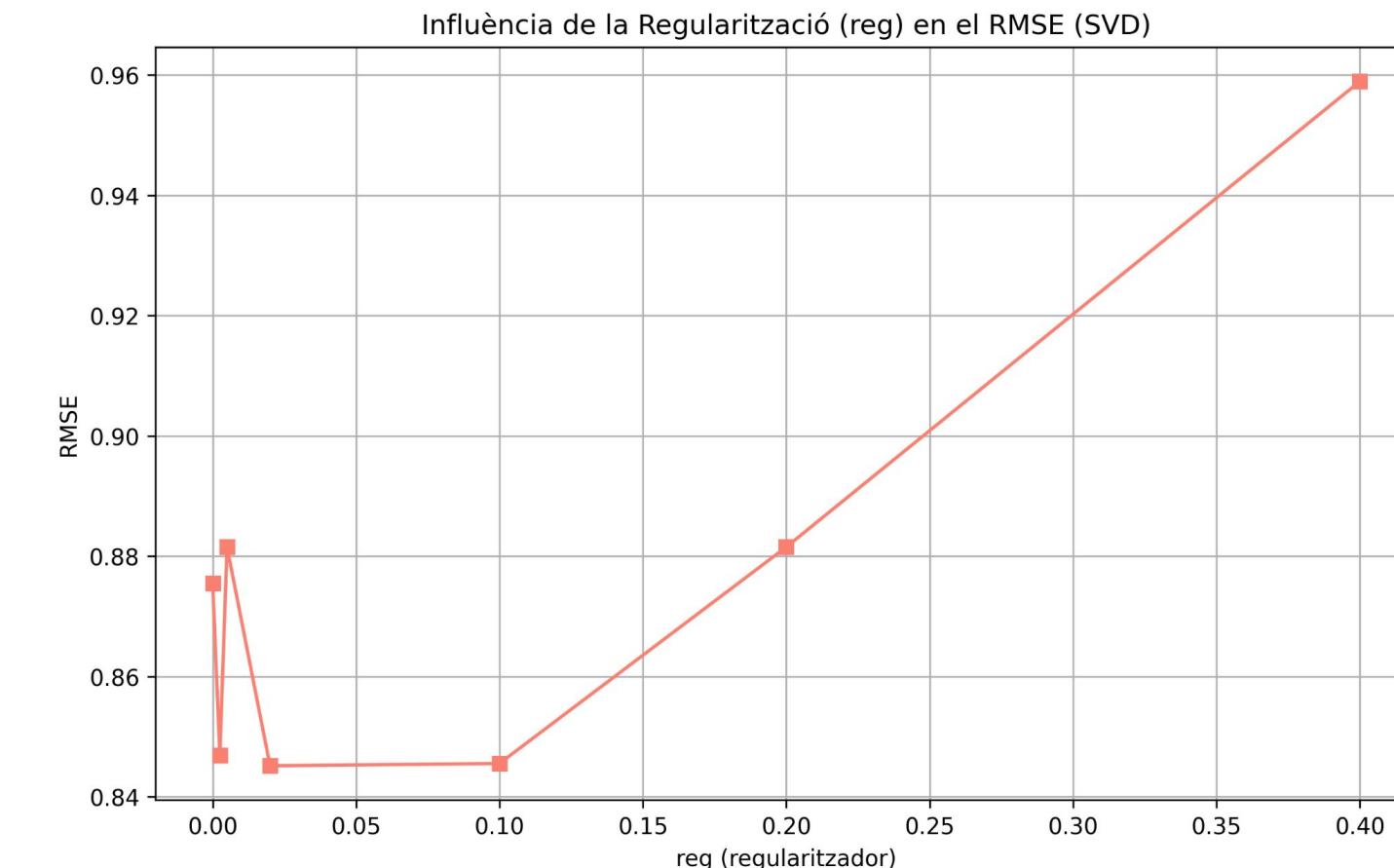
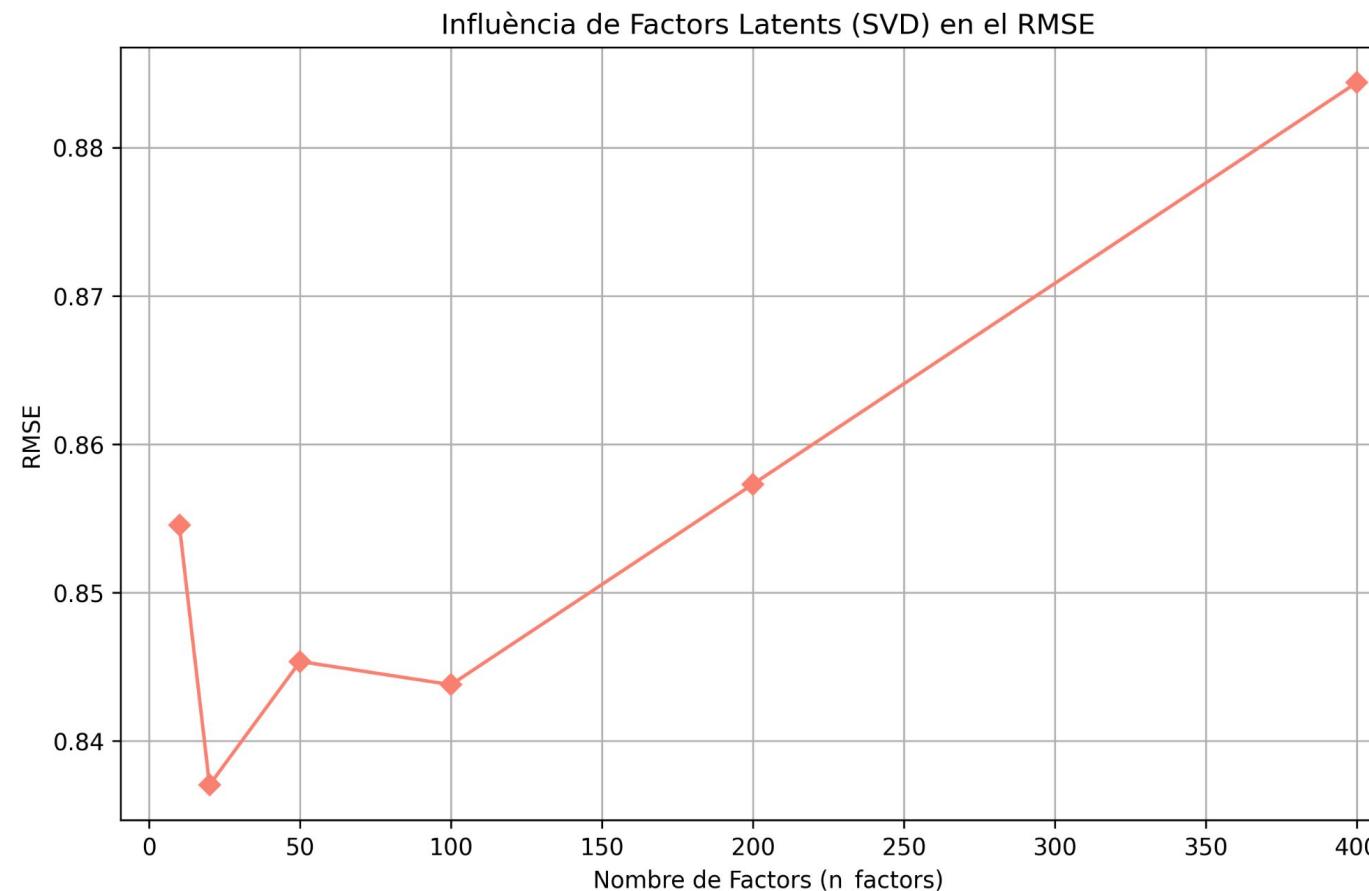
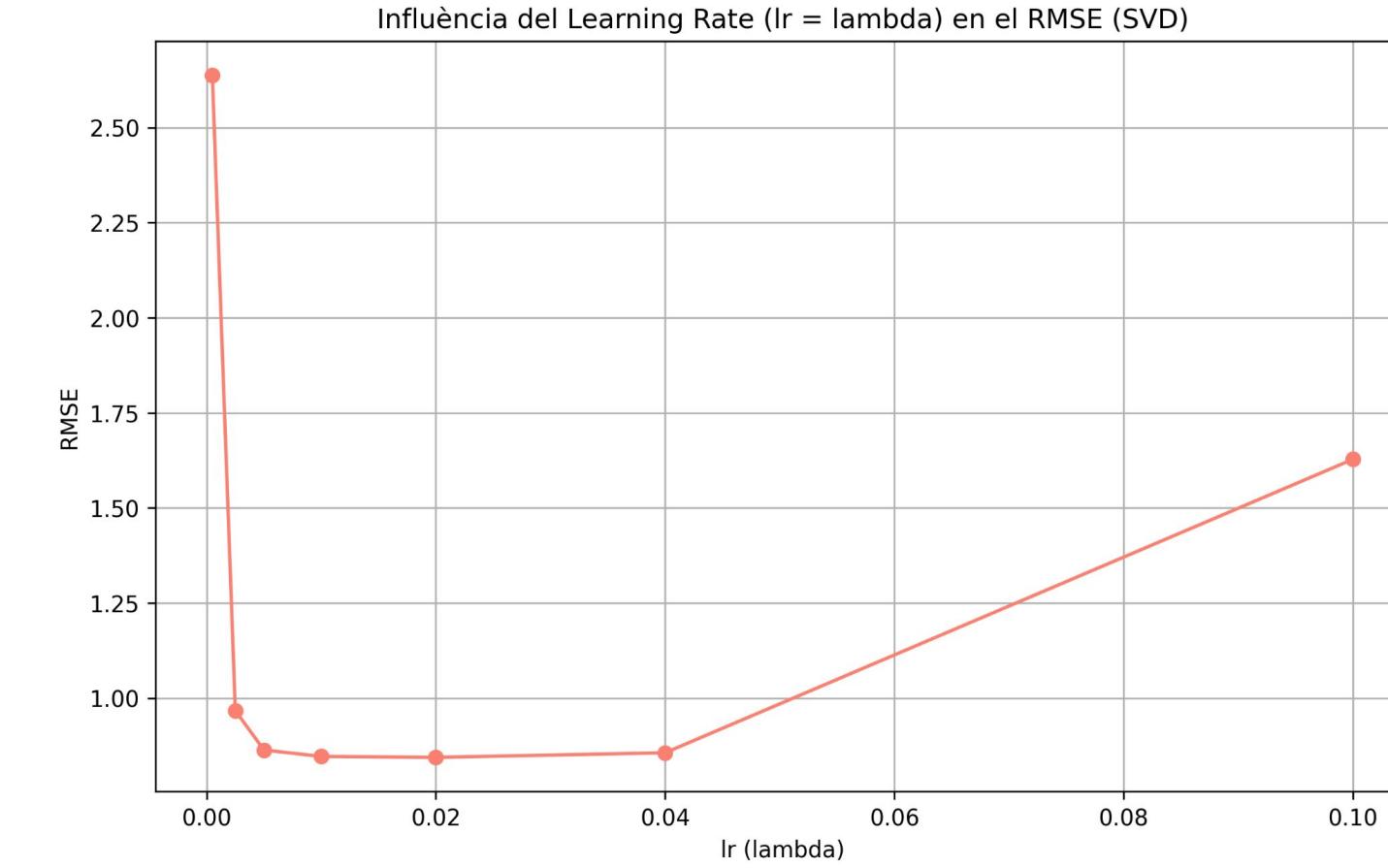
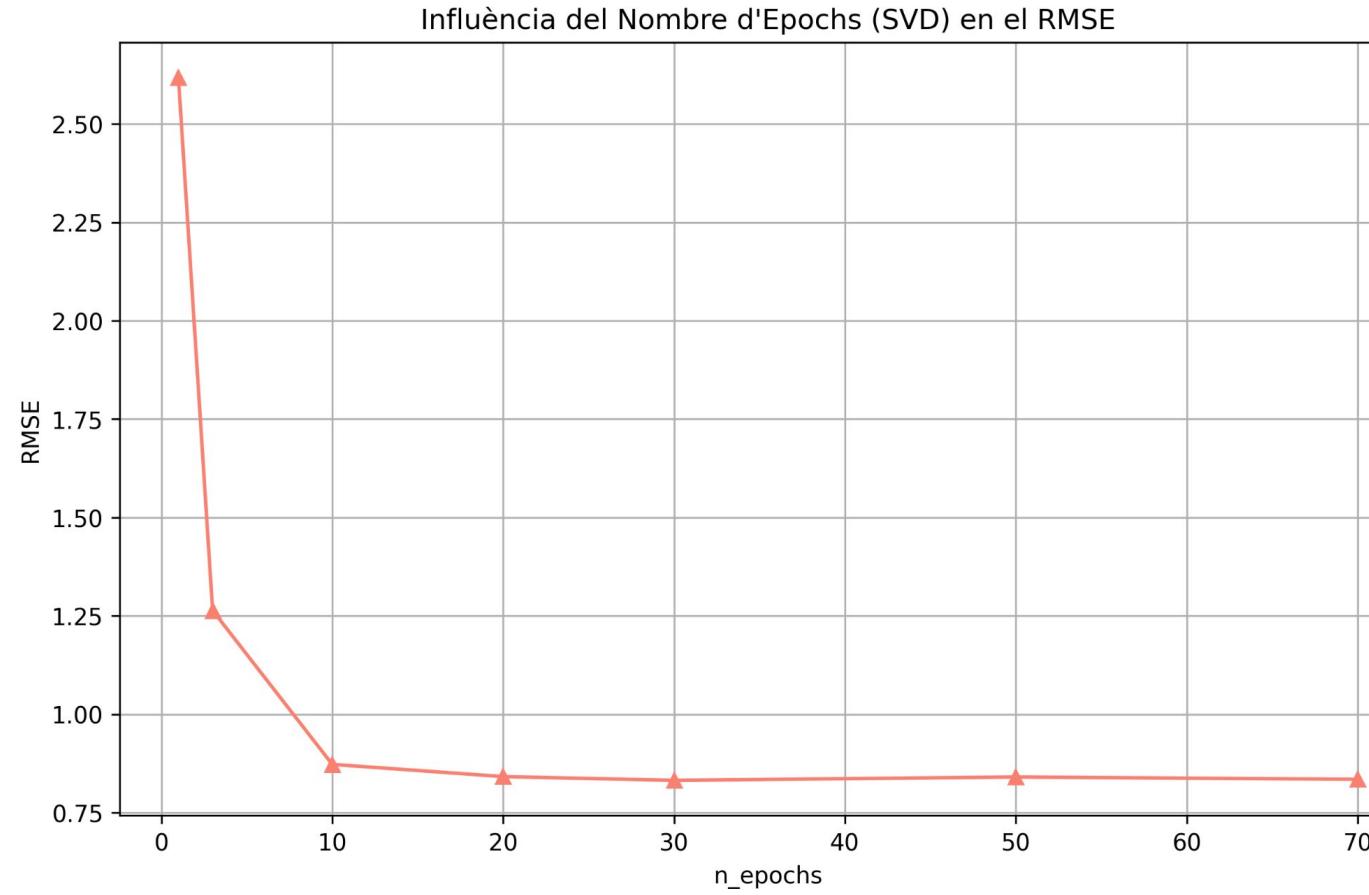


Detectant l'overfitting: Quan el model comença a ‘memoritzar’ en lloc d’aprendre

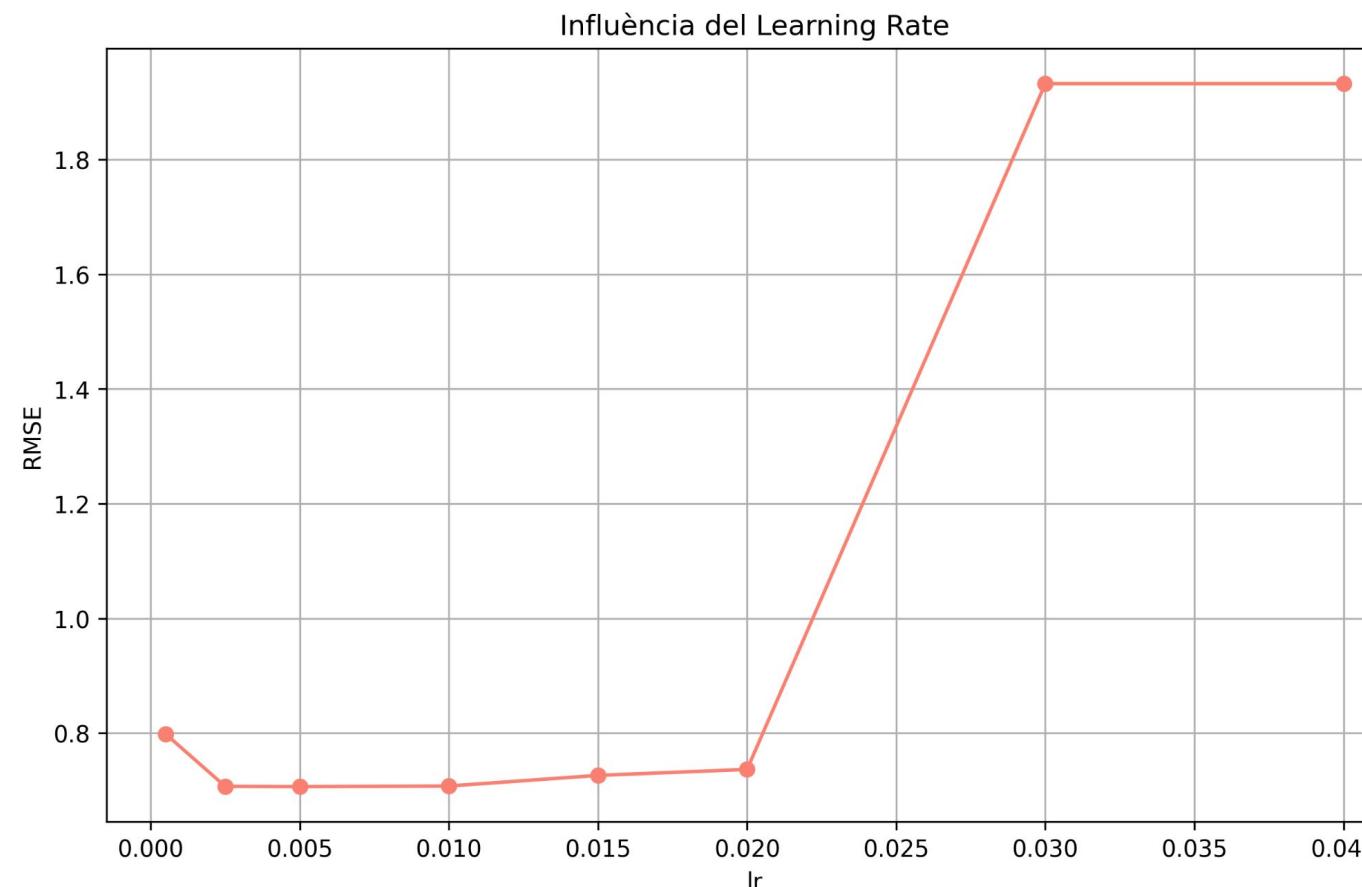
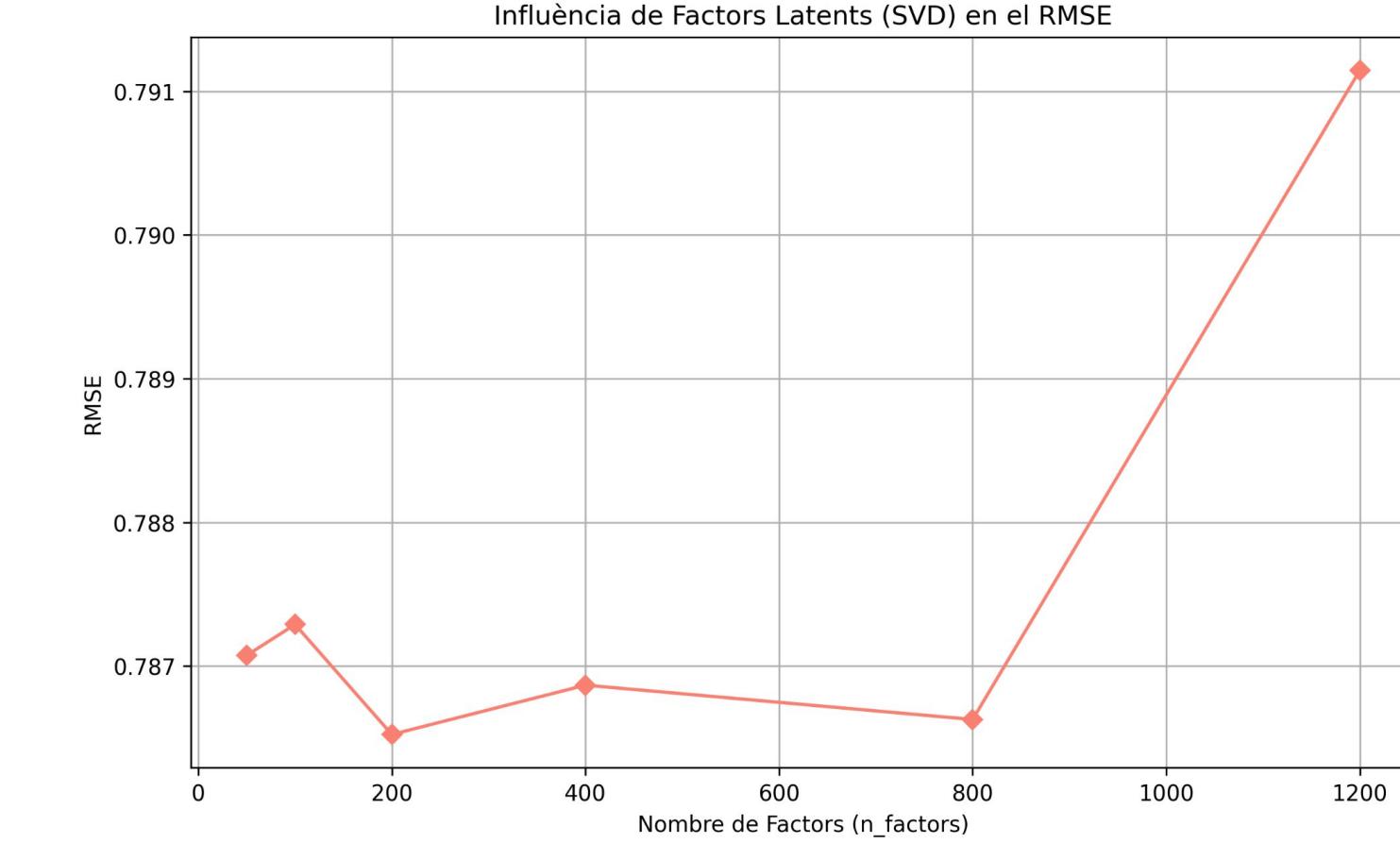
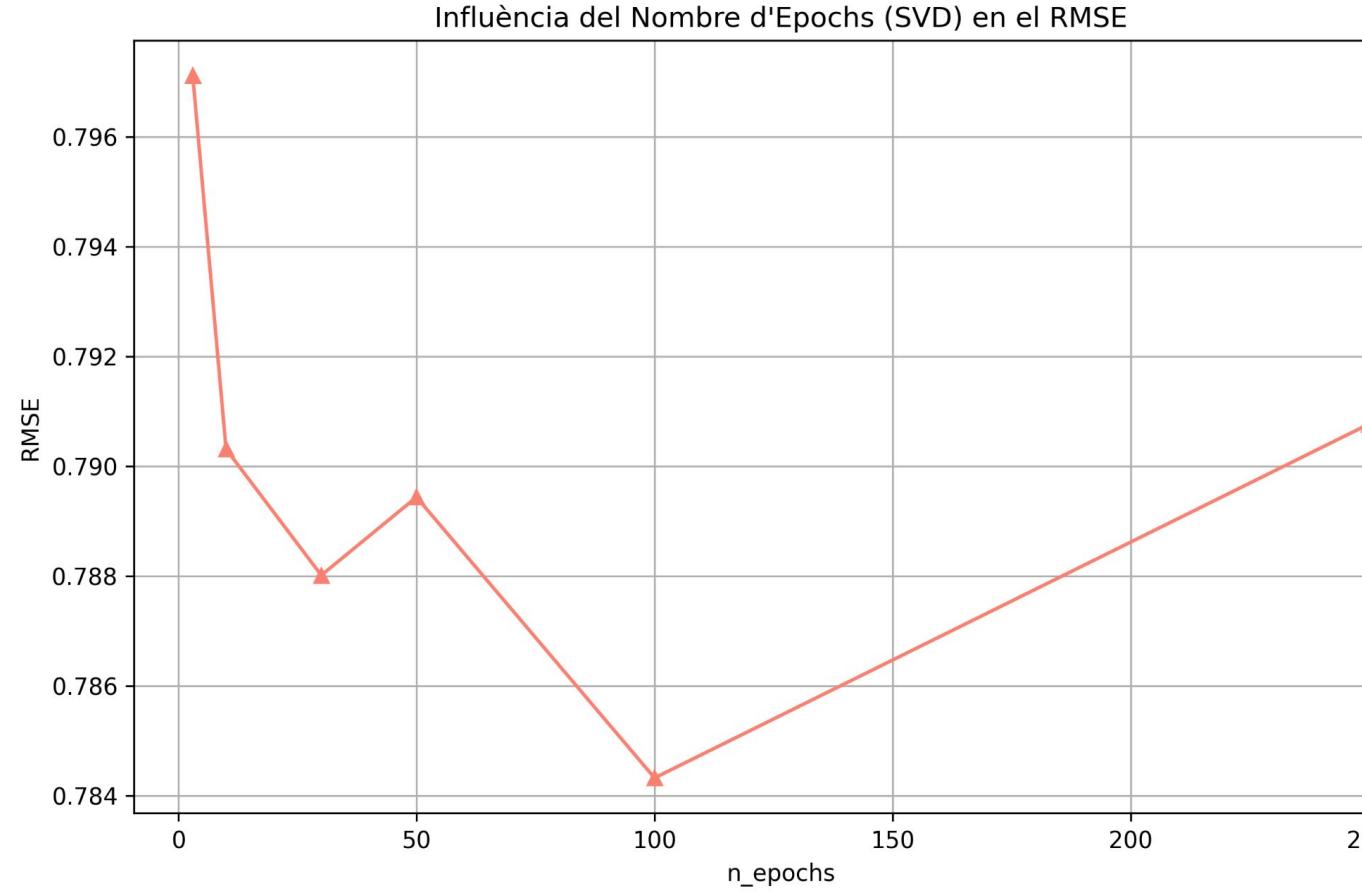
A mesura que ajustem els hiperparàmetres per augmentar la complexitat del model, l’error en el conjunt de test comença a empitjorar. Aquest és el senyal inequívoc que el model ha perdut la seva capacitat de generalitzar.



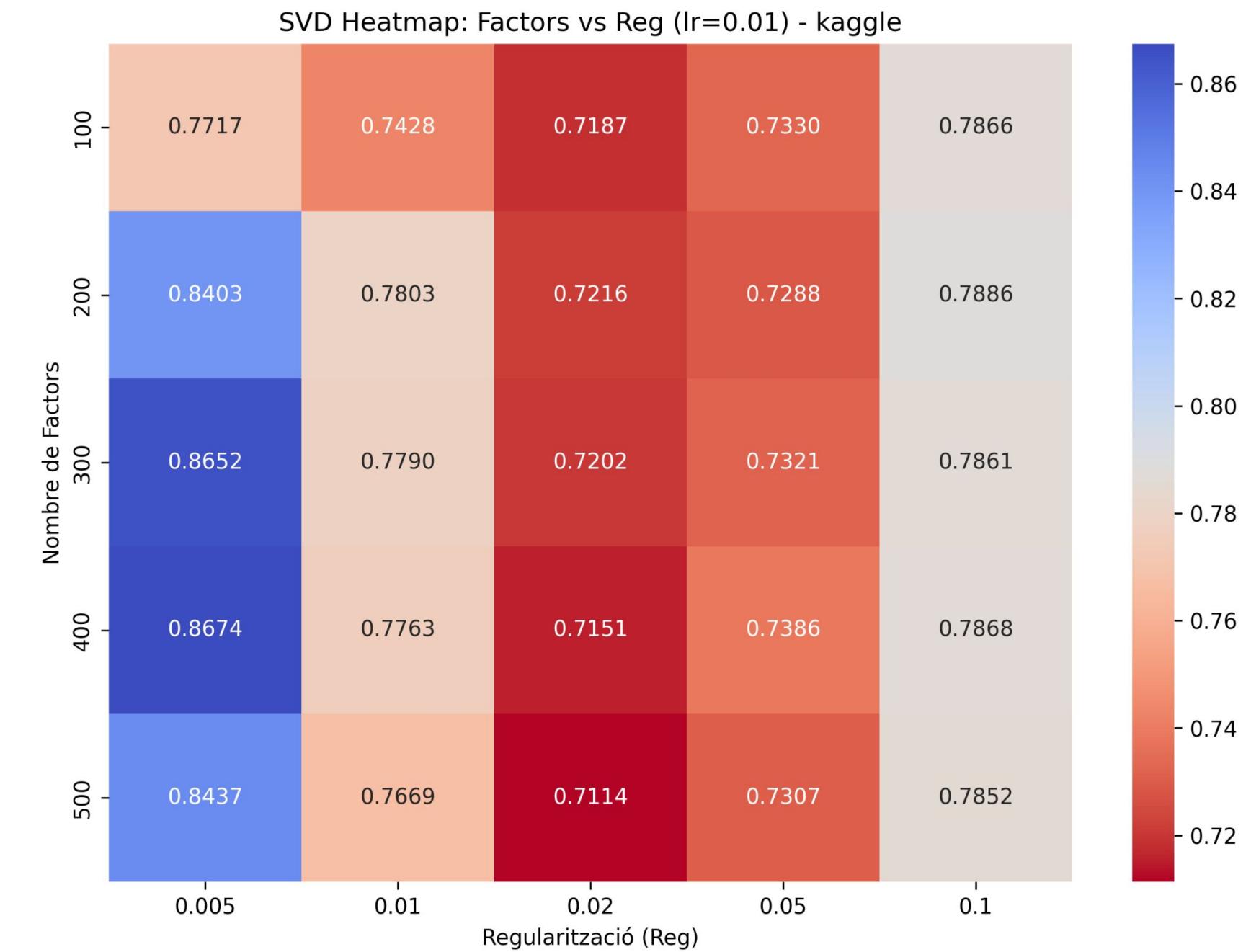
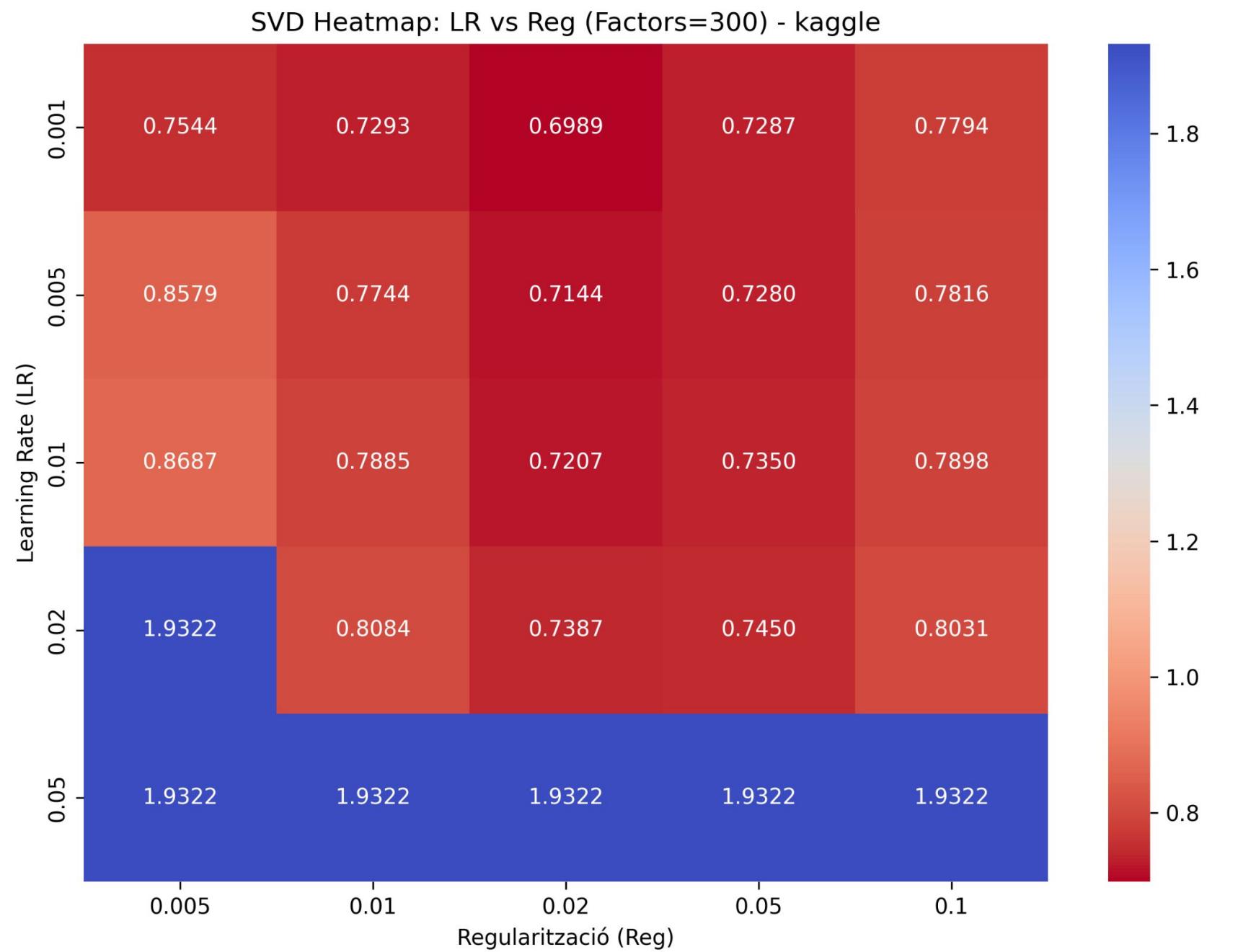
Estudi hiperparàmetres Funk-SVD (Small)



Estudi hiperparàmetres Funk-SVD (Kaggle)

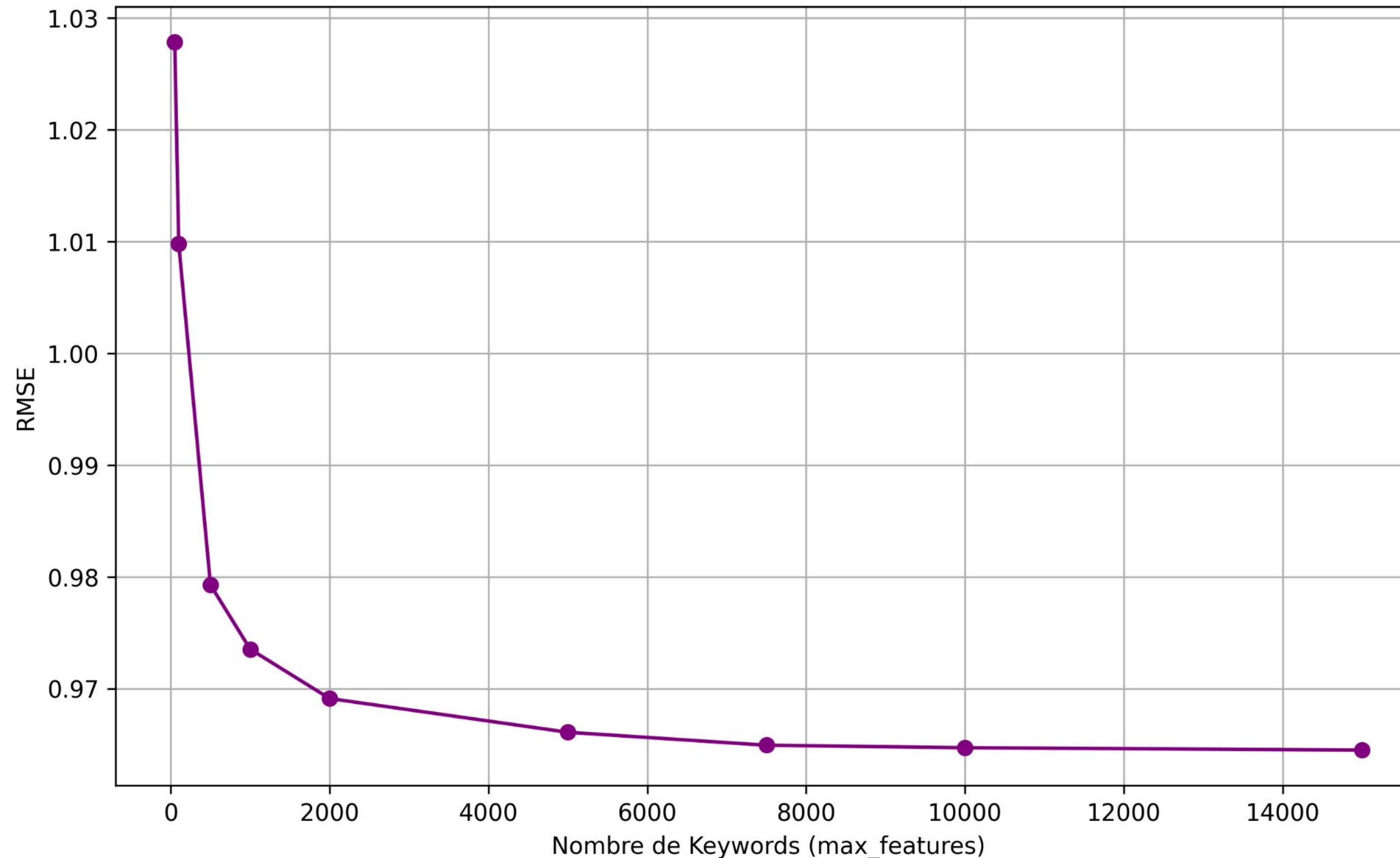


Comparació Heatmap (Kaggle)



Estudi hiperparàmetre (Recomanador Content-Based)

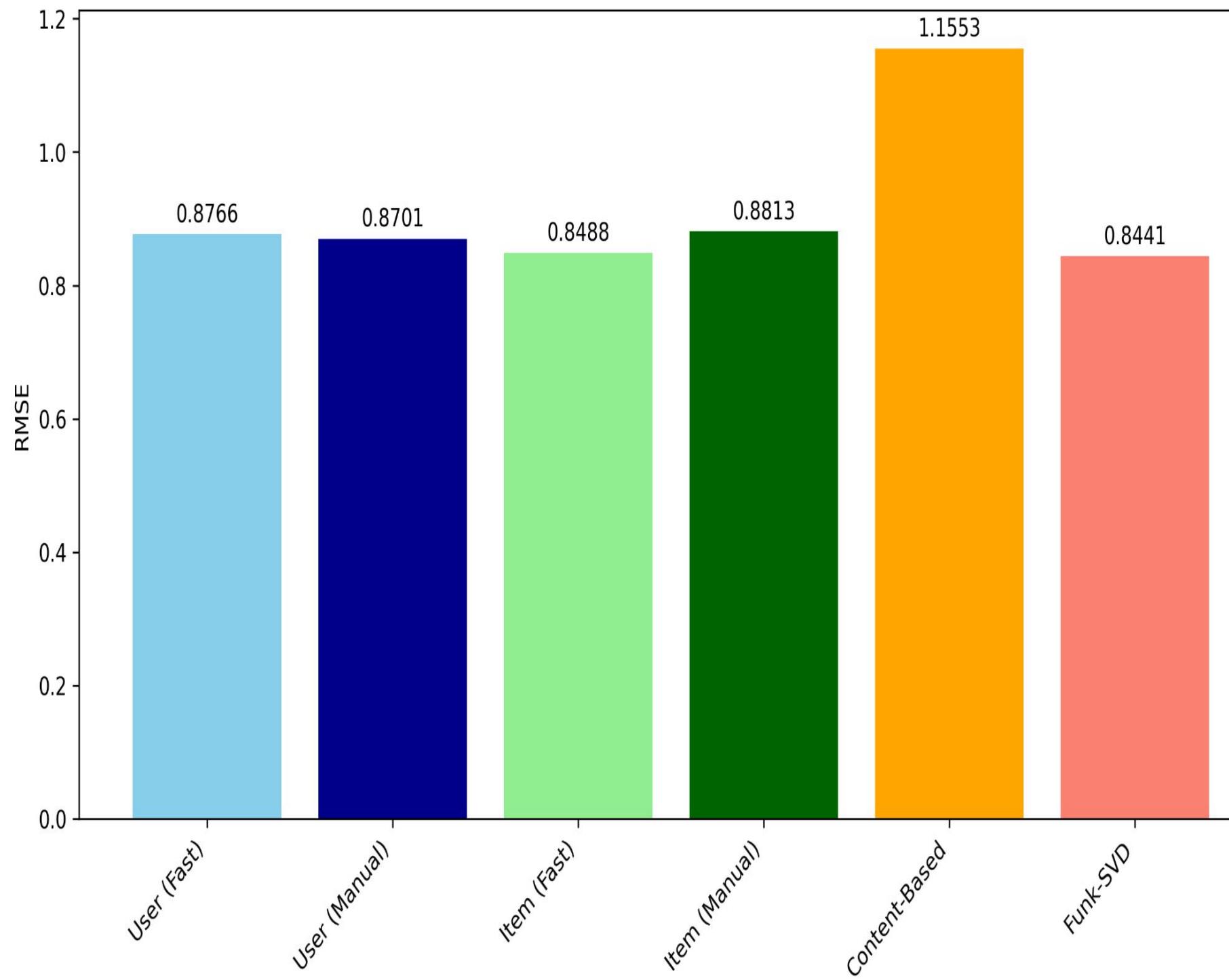
Influència del Nombre de Keywords (Content-Based) en el RMSE



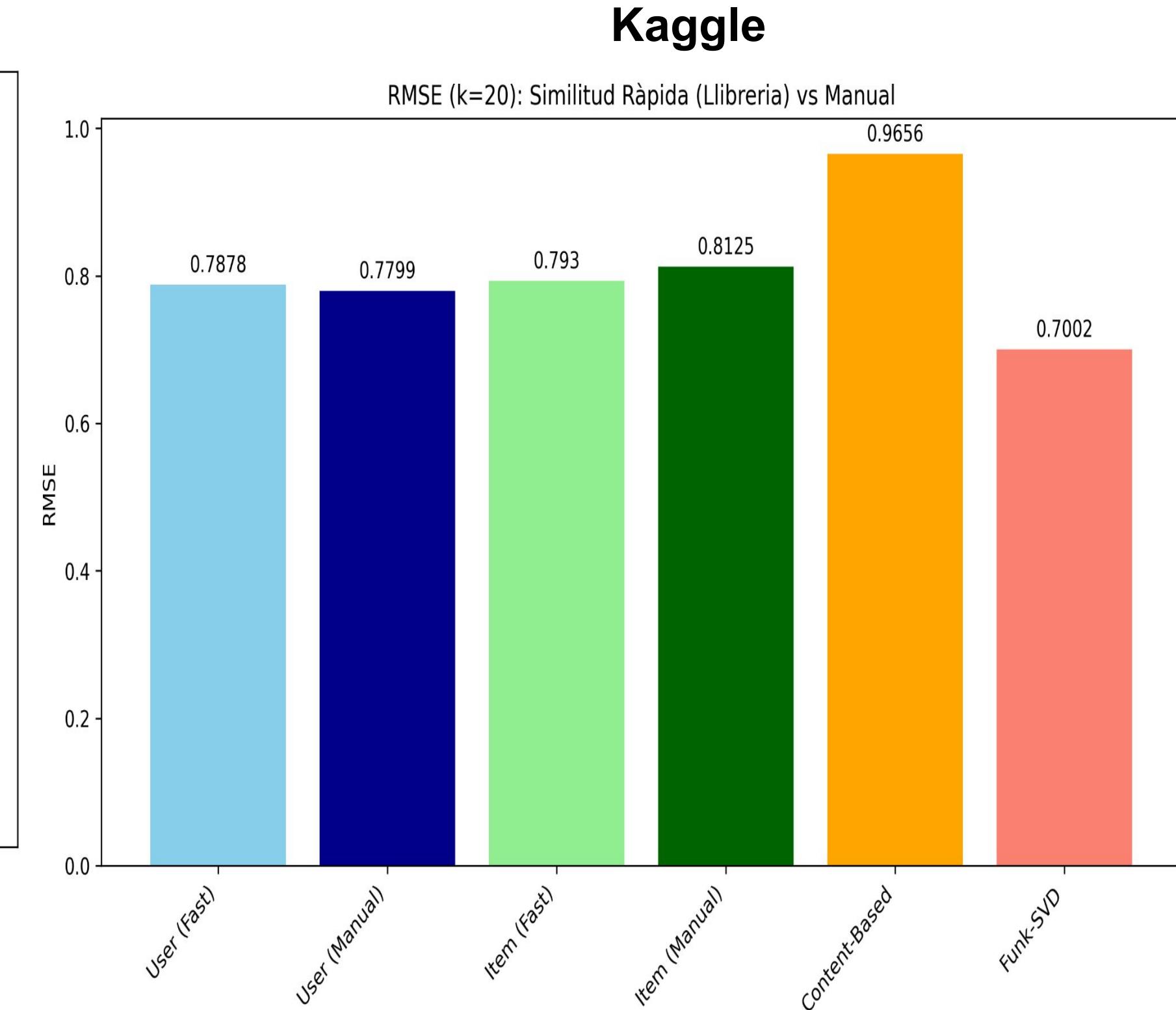
Comparativa global de rendiment (RMSE)

L'error (RMSE) varia significativament entre mètodes i datasets.

MovieLens Small



Kaggle

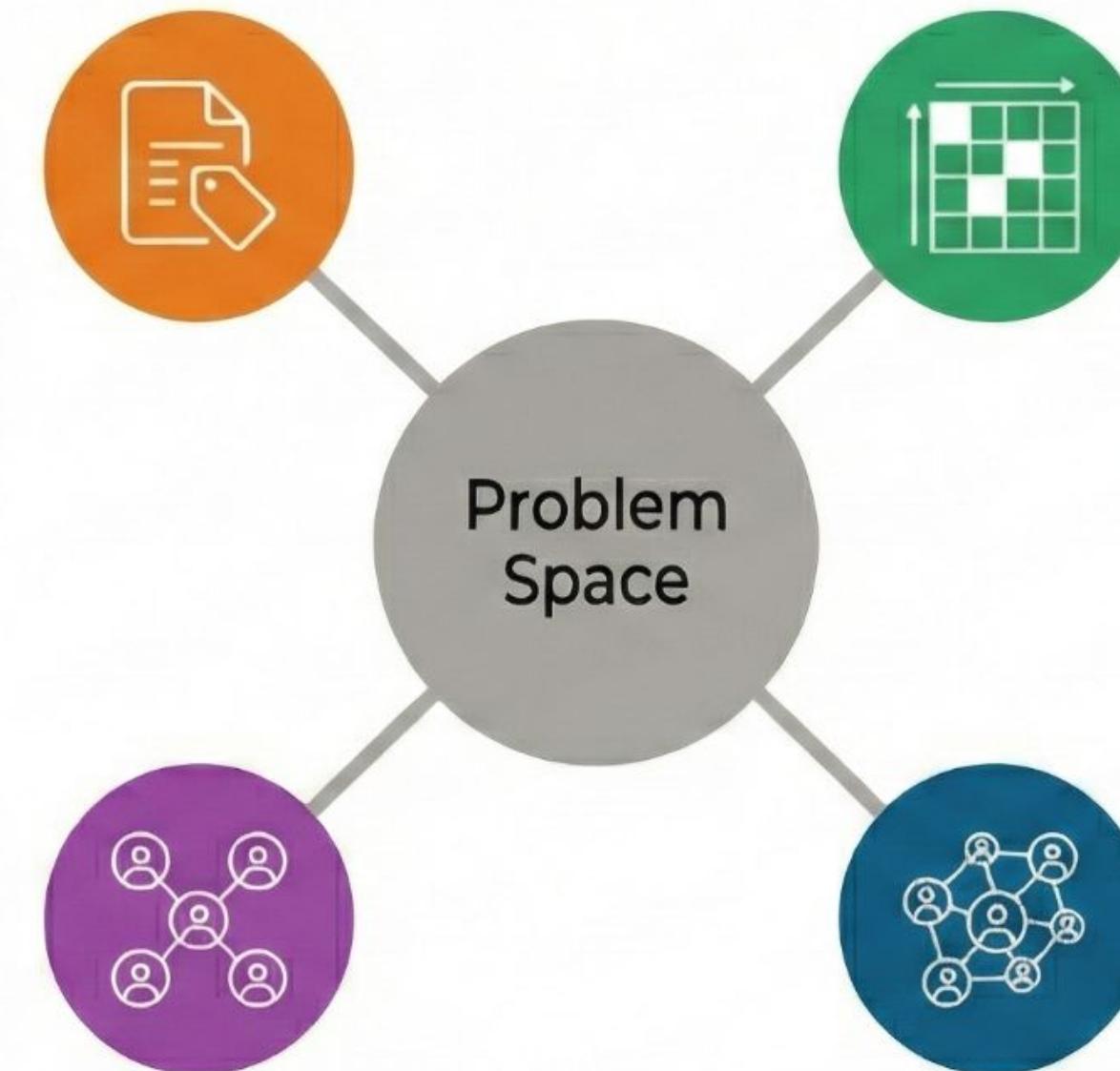


Conclusió: No existeix un únic mètode òptim

El rendiment superior d'un algorisme depèn fortemet de l'escenari:

Content-Based (TF-IDF)

Valuós en datasets amb metades riques (com el de Kaggle).



Funk-SVD

Excel·lent en datasets densos i ben filtrats, però vulnerable a l'overfitting si no es calbra bé.

User-User (Pearson)

Eficaç en datasets més petits, però pateix problemes d'escalabilitat i dispersió de dades.

Item-Item (Cosine Ajustat)

Un competitor robust i sovint més estable, especialment quan el nombre d'items és menor que el d'usuaris.

Preguntes?

Moltes gràcies per la vostra atenció.