

Système de Recommandation de films

Projet réalisé dans le cadre du bootcamp
organisé par AKADEMI, Option data science + IA

SOMMAIRE

Introduction	1
Exploration et Prétraitement des Données	2
Méthodologie	3
Évaluation des Performances	4
Génération des Recommandations	5
Visualisations	6
Recommandations Commerciales	7
Conclusion	8

1. Introduction

Aperçu

Le projet consiste à concevoir un système de recommandation de films capable de proposer à chaque utilisateur les films qu'il est le plus susceptible d'apprécier. Pour ce faire, nous avons utilisé l'ensemble de données MovieLens 100k, fourni par le laboratoire GroupLens de l'Université du Minnesota.

- L'ensemble de données contient 100 000 évaluations provenant de plusieurs milliers d'utilisateurs et de milliers de films.
- Chaque évaluation est notée sur une échelle de 0.5 à 5 étoiles.
- Le projet utilise une approche de filtrage collaboratif, complétée par un filtrage basé sur le contenu pour créer un modèle hybride capable de gérer le problème du démarrage à froid.

Objectif

Objectif principal :

- Fournir à chaque utilisateur une liste personnalisée de 5 films recommandés basés sur ses préférences passées et les comportements d'autres utilisateurs.

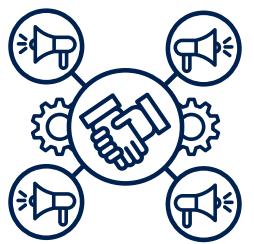
Objectifs secondaires :

- Évaluer la performance du modèle à l'aide de métriques standards telles que RMSE et MAE.
- Explorer la combinaison de méthodes collaboratives et basées sur le contenu pour améliorer la précision et la ro

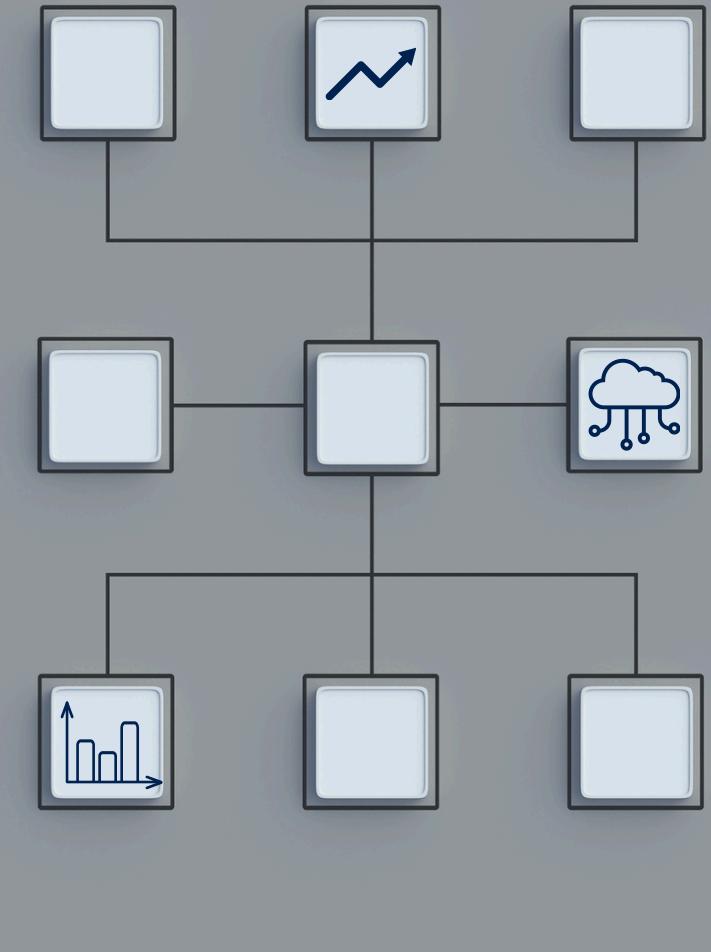
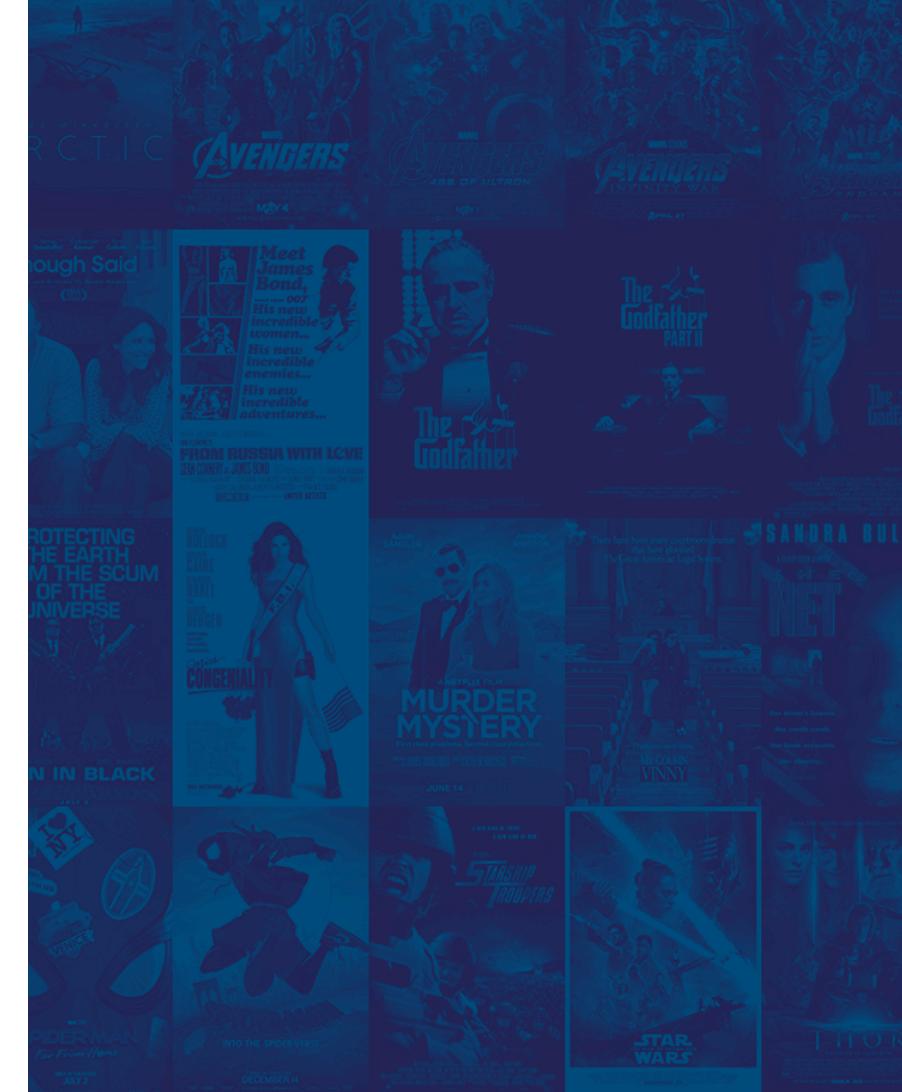
Problématique métier :

- « Comment maximiser la satisfaction des utilisateurs en anticipant les films qu'ils aimeront, tout en gérant les nouveaux films et nouveaux utilisateurs sans données historiques ? ».

2. Exploration et Prétraitement des Données



Les données sont complètes et cohérentes, permettant de construire un système de recommandation hybride, combinant les préférences explicites des utilisateurs (notes) et les caractéristiques des films (genres et tags).



1

Fichiers chargés

movies.csv : contient les films avec movield, title et genres (plusieurs genres possibles, séparés par |).

ratings.csv : contient les notes des utilisateurs (userId, movield, rating, timestamp).

links.csv : relie chaque film à ses identifiants IMDb et TMDb, utile pour enrichir les données.

tags.csv : mots-clés attribués par les utilisateurs pour décrire les films, pouvant améliorer le filtrage basé sur le contenu.

2

Structure et observations

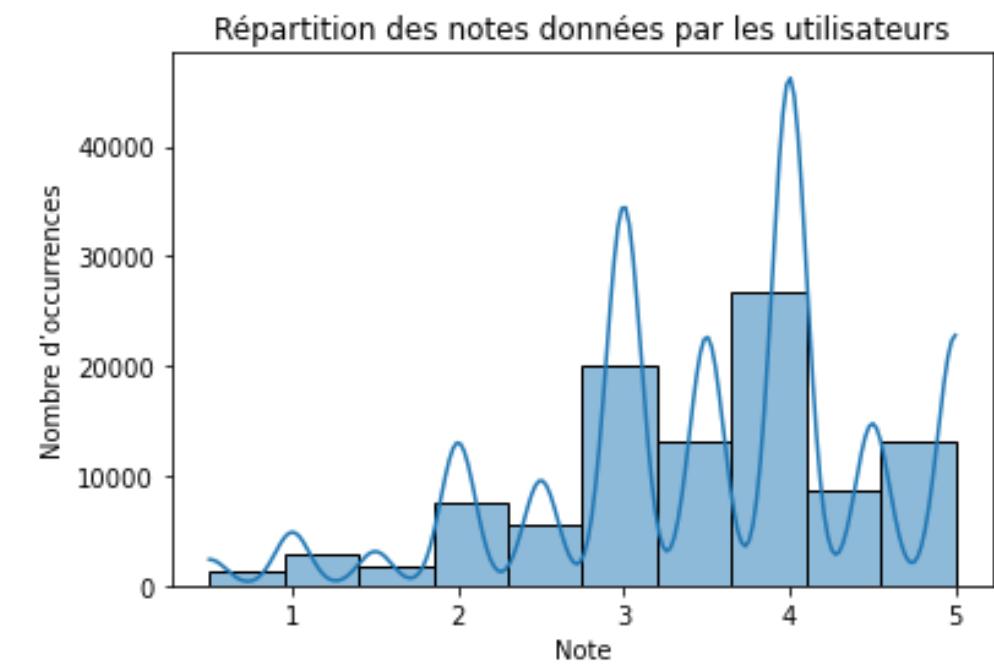
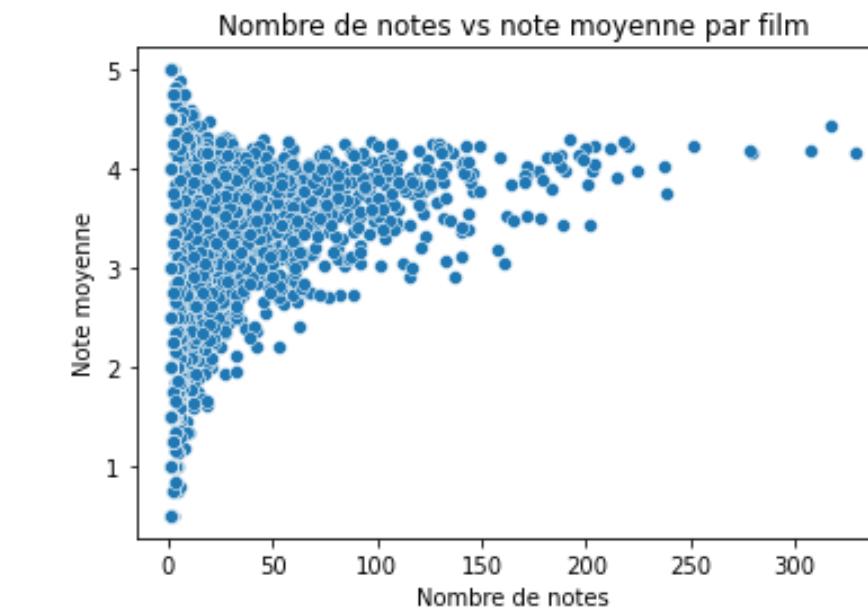
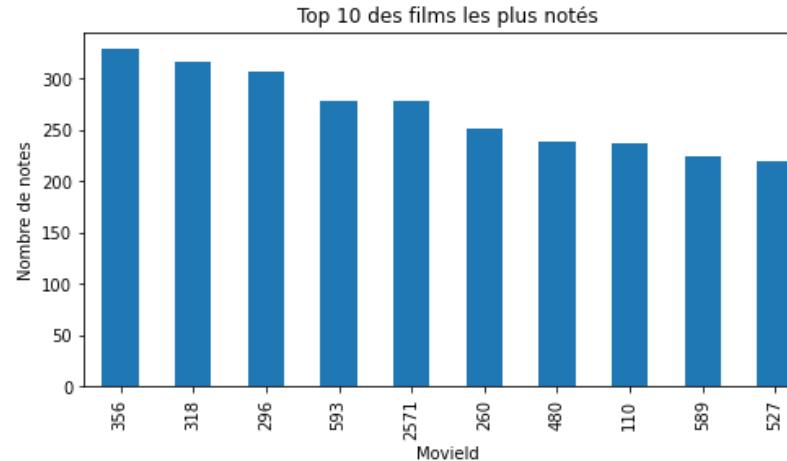
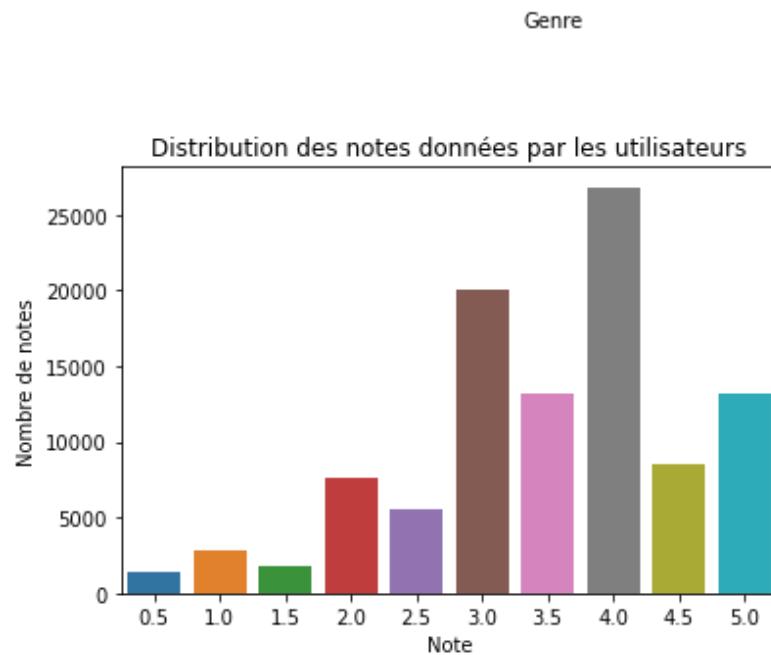
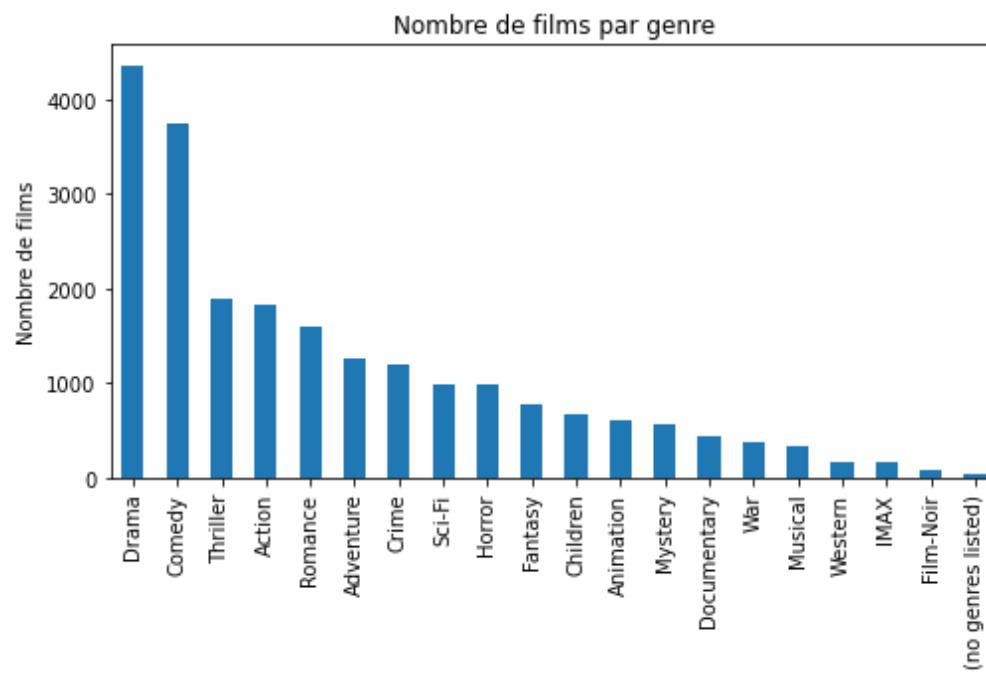
movies : 3 683 films, genres multiples → utile pour analyser les préférences par genre.

ratings : beaucoup de lignes → capture les interactions utilisateur-film, indispensable pour le filtrage collaboratif.

tags : informations qualitatives → possibilité d'ajouter des recommandations basées sur le contenu.

3. Méthodologie

- Statistiques générales :** Les notes sont globalement équilibrées avec une moyenne de 3,5 (écart-type 1,04), une médiane identique (3,5) et des valeurs allant de 0,5 à 5 .
- Utilisateurs & Films :** Le dataset couvre 610 utilisateurs uniques et 9 724 films uniques , ce qui garantit une bonne diversité.
- Répartition des notes :** La majorité des évaluations se situent entre 3 et 4 , confirmée par les histogrammes et countplots.
- Comportement des utilisateurs :** La plupart des utilisateurs notent moins de 50 films , mais certains sont très actifs.



- Genres :** Les plus représentés dans le catalogue sont Action, Comédie et Drame .
- Films populaires :** Les plus notés incluent des classiques comme Braveheart , Star Wars , Pulp Fiction , Shawshank Redemption et Forrest Gump .
- Films peu vus mais bien notés :** Plusieurs films moins connus affichent une note $\geq 4,5$ malgré <20 évaluations , révélant des « perles cachées ».
- Corrélations :** Les films très vus ont une note moyenne stable autour de 3,5 , tandis que certains films peu vus concentrent des évaluations exceptionnellement positives.

Évaluation

4. des Performances

Nous avons construit un modèle de filtrage collaboratif basé sur les utilisateurs avec la librairie surprise. Les données ('userId', 'movieId', 'rating') ont été scindées en trainset (80%) et testset (20%). Le modèle KNNBasic a appris les similarités entre utilisateurs pour prédire les évaluations manquantes.

L'évaluation sur le testset donne des performances solides pour ce type de modèle : RMSE ≈ 0.94 , MAE ≈ 0.72 , montrant que les prédictions sont proches des notes réelles. Les premières prédictions illustrent la capacité du modèle à estimer des notes même pour des utilisateurs ou films peu connus, tout en signalant les cas impossibles.

5. Génération des Recommandations



Top 5 recommandations pour l'utilisateur 1 :

Cry, the Beloved Country (1995)

Lamerica (1994)

Heidi Fleiss: Hollywood Madam (1995)

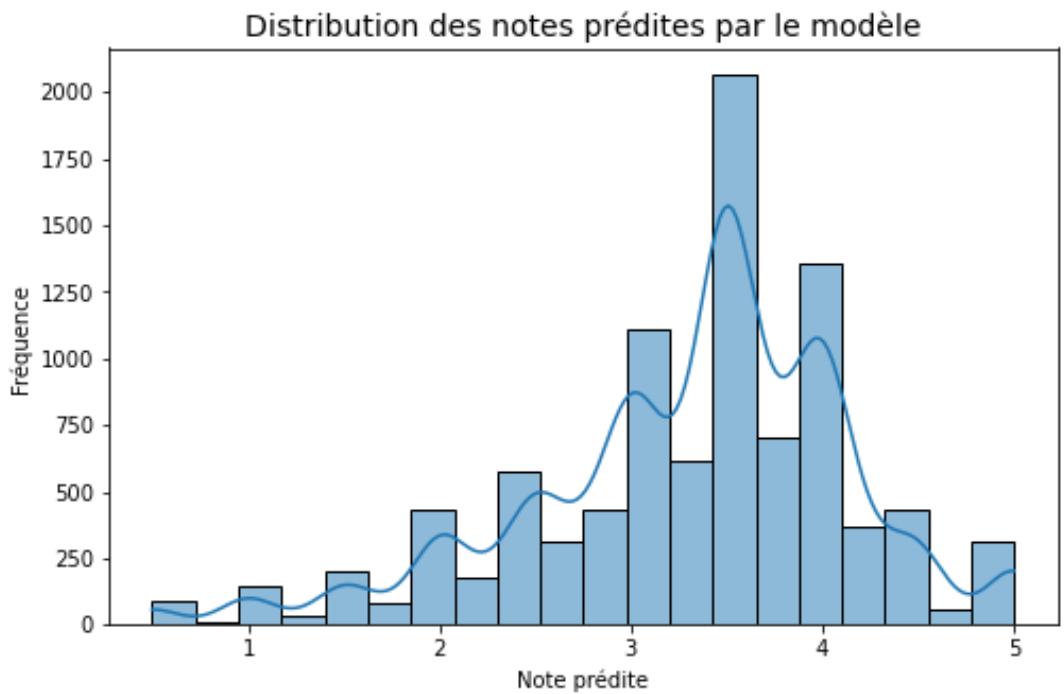
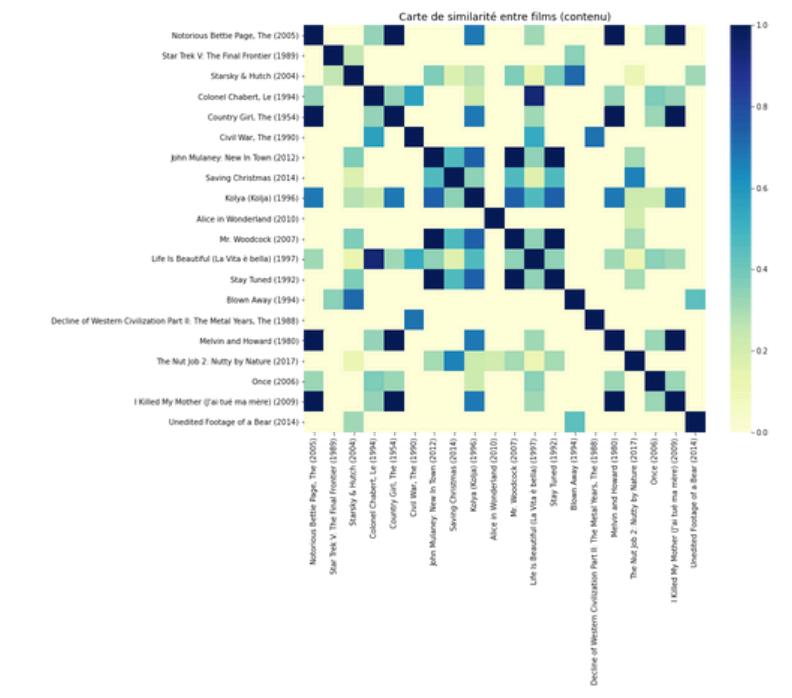
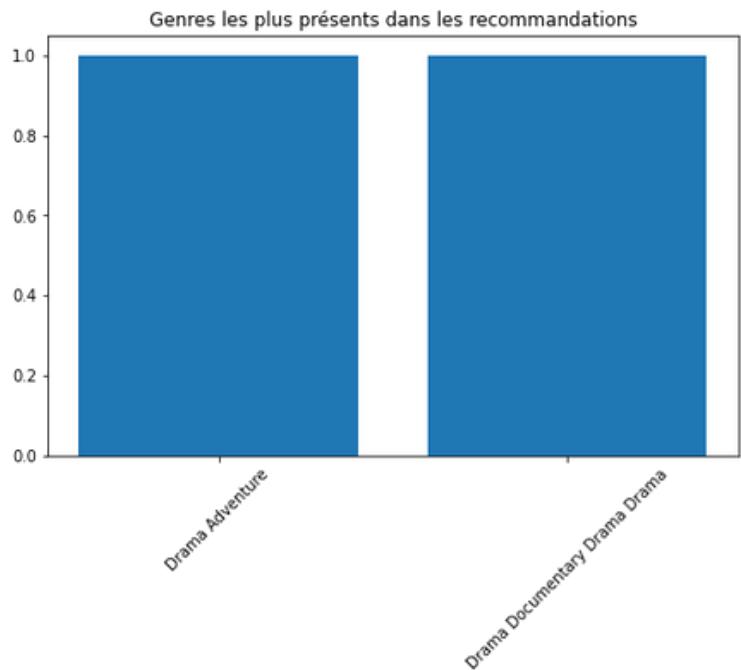
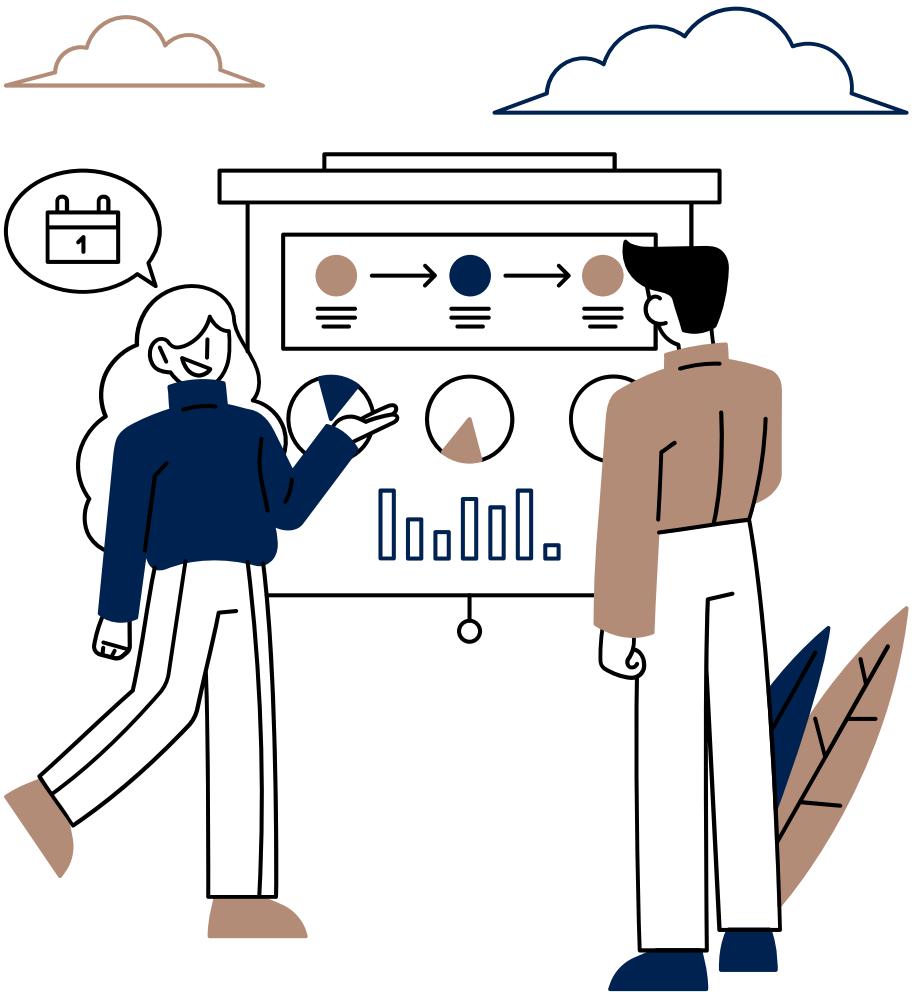
4. Awfully Big Adventure, An (1995)

Priest (1994)

6 Visualisations

Je vais conclure mon projet avec une étape finale visuelle. Cela va permet de :

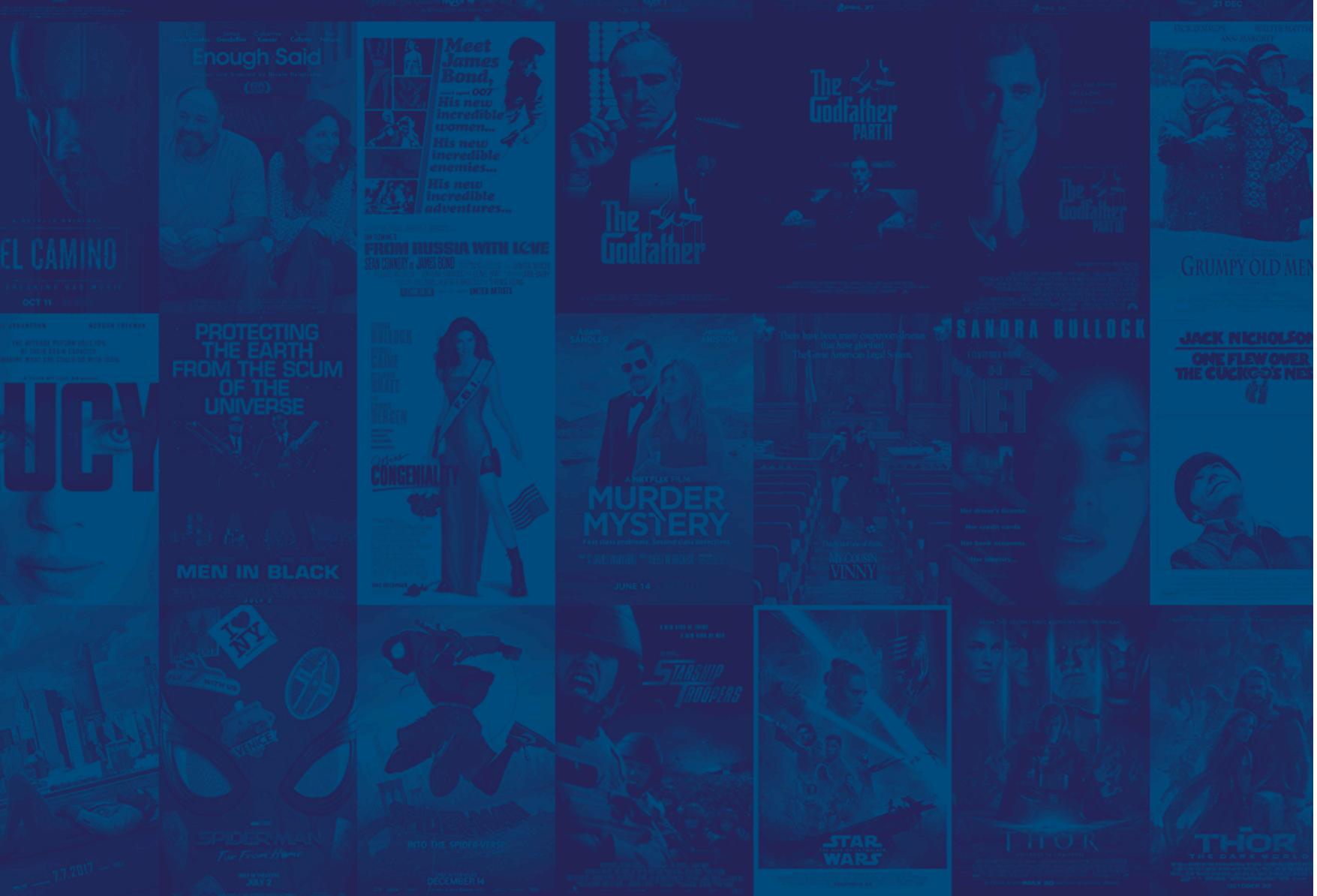
- Rendre les résultats concrets (on ne reste pas que dans les chiffres).
 - Mieux communiquer mes trouvailles à un public non-technique.
 - Donner une valeur ajoutée business à mon notebook.



7 Recommandations Commerciales

Optimisation de l'expérience utilisateur : Les recommandations générées (top-5 par utilisateur) montrent une personnalisation efficace. Intégrer ce système dans une plateforme réelle permettrait aux utilisateurs de découvrir plus facilement des films correspondant à leurs goûts.

Segmentation et ciblage marketing : L'analyse des genres prédominants dans les films recommandés offre un aperçu des préférences des utilisateurs. Les équipes marketing peuvent utiliser ces informations pour proposer des promotions ou contenus adaptés à chaque segment.



Fidélisation et engagement : En combinant filtrage collaboratif et contenu, le système propose à la fois des films similaires à ceux appréciés et des suggestions nouvelles. Cela encourage la rétention des utilisateurs et augmente le temps passé sur la plateforme.

4. Conclusion

Le projet démontre que même un modèle simple de filtrage collaboratif, enrichi par un filtrage basé sur le contenu, peut produire des recommandations pertinentes et exploitables. Les visualisations finales (histogramme des notes, carte de similarité, distribution des genres) rendent les résultats tangibles et communicables à un public non technique. Ce travail constitue une base solide pour un déploiement commercial et offre des perspectives d'amélioration avec des données plus riches ou des modèles hybrides plus sophistiqués.

Merci

Projet réalisé dans le cadre du boot camp organisé
par AKADEMI, option data science + IA

Contact : germodee@gmail.com