Master Data Science: Projet Data Mining

1 M Movies Dataset

Clustering et Recommandation

Etudiantes:

Hayet Danoun

Serine Issaad

Enseignant:

Rémy Cazabet

2024-2025

SOMMAIRE

Introduction

Dans le cadre de ce projet, nous serons amenés à faire une étude sur un dataset afin de pouvoir en extraire des informations ou découvrir des patterns qui nous aideront à mieux comprendre la nature de nos utilisateurs et leurs préférences et à s'adapter à celles-ci en concevant des modèles de prédiction.

Répartition des tâches :

- Préparation de données : tâche commune discutée et faite ensemble en présentiel.
- Clustering: Serine Issaad.
- Système de recommandation: Hayet Danoun.

I. Présentation des données

1. Choix de la base de données

Notre étude sera tournée autour des films. Nous avons essayé de trouver une base de données riche qui contient à la fois des informations démographiques sur les utilisateurs et le genre et le rating des films d'une variété de films.

Remarque: le Dataset utilisé a été publié en 2003. Il est fort probable qu'il ne reflète pas les utilisateurs d'aujourd'hui, vu les normes qui ont changé en termes de société et de qualité cinématographique. Cependant, sa richesse nous est la raison pour laquelle nous l'avions choisie, nous permettant ainsi de montrer comment mener une étude.

Lien vers le DataSet

2. Structure des données

Le jeu de données utilisé se compose de trois ensembles distincts :

Movies: 3952 films, chacun identifié par un ID unique, un titre et un ou plusieurs genres. La multiplicité des genres pour certains films ajoute une dimension supplémentaire à notre analyse, mais offre une meilleure précision.

Format: MovieID::Title::Genres

Users: 6040 utilisateurs avec leurs informations démographiques telles que le sexe, l'âge, la profession et le code postal. On note une prédominance masculine (4331 hommes contre 1709 femmes), ce qui pourrait influencer les recommandations en fonction des préférences liées au genre.

Format: UserID::Gender::Age::Occupation::Zipcode

Ratings: 1000210 notes de films, sur une échelle de 1 à 5 (entiers). Chaque évaluation est associée à un utilisateur et à un film, avec un horodatage correspondant. Le nombre minimum d'évaluations pour un film est de 20.

Format: UserID::MovieID::Rating::Timestamp

Aucune valeur manquante n'a été détectée dans les trois ensembles de données, ce qui simplifie le traitement ultérieur en évitant le recours à des techniques d'imputation.

3. Exploration des données

Distribution des notes : La médiane est de 4, indiquant une tendance générale des utilisateurs à attribuer des évaluations positives. Cette concentration peut compliquer la distinction entre les films véritablement appréciés et ceux jugés simplement satisfaisants.

Répartition des âges des utilisateurs : les utilisateurs sont répartis sur une large gamme d'âges, avec une concentration notable dans la tranche des 25-34 ans, qui est également la plus active en termes de nombre de notations.

Genres de films : Les genres les plus représentés sont le drame, la comédie et l'action. Cette diversité, bien que certaines catégories soient plus dominantes, permet de capturer une variété de préférences pour affiner les recommandations.

II. Préparation des données

1. Filtrage des utilisateurs et des films

Pour optimiser la performance du modèle et réduire sa complexité, nous avons appliqué des critères de filtrage :

Utilisateurs : Dans la partie EDA, nous avons remarqué que la médiane des évaluations pour chaque portion d'âge était de 120. Nous avons décidé alors d’exclure ceux ayant noté plus de 120 films afin de se concentrer sur l'utilisateur moyen et d'éviter que les super-utilisateurs n'influencent disproportionnellement notre clustering ou recommandation.

Films : dans la même partie (EDA), nous remarquons que le nombre minimal et maximal d'évaluations pour un film est de 1 et 3428, respectivement. Vu que plus le film est populaire, plus sa notation est plus fiable et donc pesant plus, nous avons choisi d'éliminer les films ayant reçu moins de 600 évaluations. Ceci devrait augmenter la fiabilité de nos résultats.

Remarque: le filtrage des films peut être vu d’une manière complètement différente: il est plus probable que les utilisateurs donnent une bonne note à un film très connu, voire même s’ils ne l’ont pas regardé. D'où l'élimination des films ayant un nombre d'évaluations supérieur à un seuil précis. Toutefois, notre but est surtout d'étudier les genres préférés par différents types d’utilisateurs. Ceci dit, un utilisateur qui note un film d’action 2⁄5 est probablement un utilisateur qui aime les films d’action (étant donné qu’il a choisi de le regarder) mais qui ne l’a pas aimé. Cette filtration sera surtout utile pour le clustering.

Résultats du filtrage :

- Nombre d'utilisateurs réduit de 6040 à 3346.

- Nombre de films réduit de 3952 à 499.
- Nombre d'évaluations réduit de 1000210 à 295943.

2. Comparaison des distributions avant et après filtrage

La distribution des notes est restée similaire après filtrage, toujours centrée sur les notes de 3 et 4. Le filtrage n'a donc pas altéré la tendance générale des évaluations, mais a permis d'affiner le jeu de données en le rendant plus représentatif.

3. Ingénierie des caractéristiques (Feature Engineering)

Nous nous interessons seulement aux variables catégorielles sexe, âge et profession. Zip-code et Timestamp sont à supprimer.

Pour le clustering, nous avons fusionné Users avec Movies (sans titres) et Ratings. Les genres ont été transformés en colonnes qui ont pour valeur le rating donné par l'utilisateur pour le film évalué. Ensuite, pour chaque utilisateur, nous avons calculé la moyenne de son évaluation pour chacun des genres. Nous obtenons le DataFrame en-dessous.

	userid	Action	Adventure	Animation	Children's	Comedy	Crime	Documentary	Drama	Fantasy	...	Musical	Mystery	Romance	Sci-Fi
0	3	4.12500	4.105263	4.0	4.000000	3.954545	0.00	0.0	4.000000	4.5	...	4.0	3.0	3.800000	3.833333
1	6	4.00000	3.800000	4.0	4.000000	3.875000	2.00	0.0	3.642857	3.0	...	4.5	0.0	4.090909	3.500000
2	12	3.80000	5.000000	0.0	5.000000	3.600000	3.50	0.0	4.333333	0.0	...	5.0	3.0	3.000000	0.000000
3	13	3.45614	3.436364	3.0	3.428571	3.000000	3.25	0.0	3.894737	3.4	...	4.0	3.0	3.375000	3.575758
4	20	4.31250	3.750000	0.0	0.000000	4.000000	4.75	0.0	4.600000	0.0	...	0.0	4.5	5.000000	4.000000

5 rows × 22 columns

III. Clustering

Notre dataset contient désormais plusieurs variables. Nous devrions donc réduire sa dimensionnalité.

1. PCA

PCA ne semble pas adéquat vu le nombre de variables catégorielles que nous possédons. La variance ne pourra pas être capturée. Nous avons quand même essayé de l'appliquer sur les variables représentant les genres seulement.

Résultat

Les valeurs des Eigenvectors et les ratios des variances capturées montrent que PCA a échoué à capturer environ 50% de la variance.

2. T-SNE

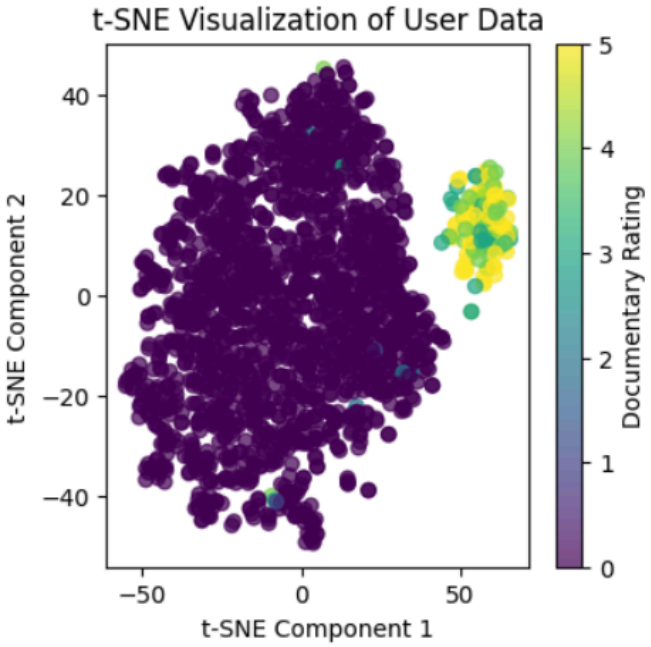
T-SNE est adapté aux variables mixtes (continues et catégorielles). Ce qui est notre cas. Nous avons joué sur deux paramètres afin d'obtenir des clusters significatifs: la perplexité et les variables utilisées.

Lors de l'inclusion des composantes PCA, nous rajoutons une seule variable categoricale (Sexe puis Age). Autrement, les composantes t-SNE auraient capturé les similarités locales entre profession, âge et sexe. Intéressant, mais ce n'est pas le but.

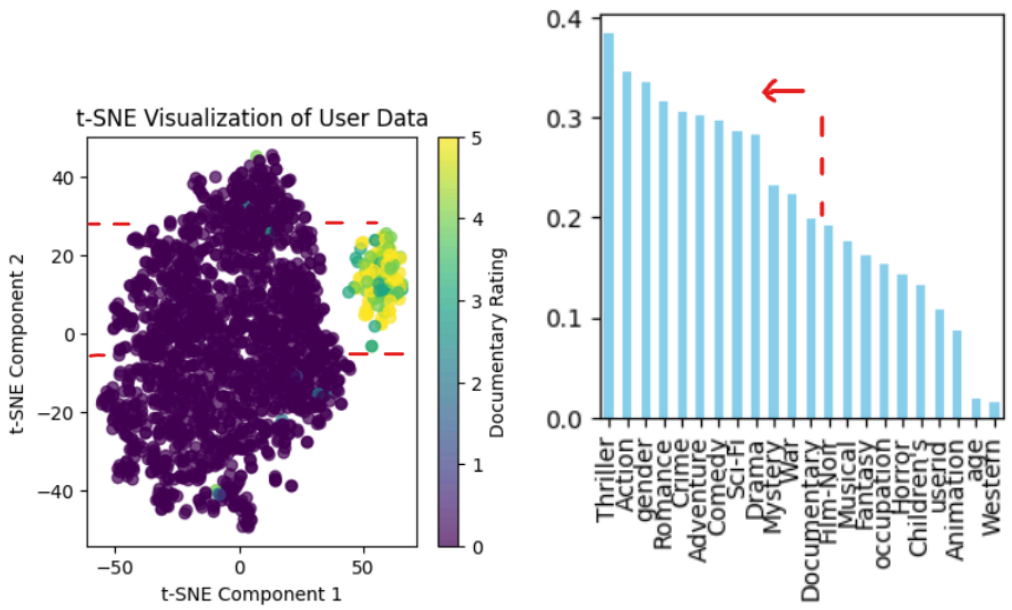
Voici les résultats pour chaque combinaison de paramètres:

1. Perplexity = 25, t-SNE appliqué sur toutes les variables originales

- Trustworthiness: 0.9630063642307438
- Graphe 1 : Coloré selon la variable Documentary

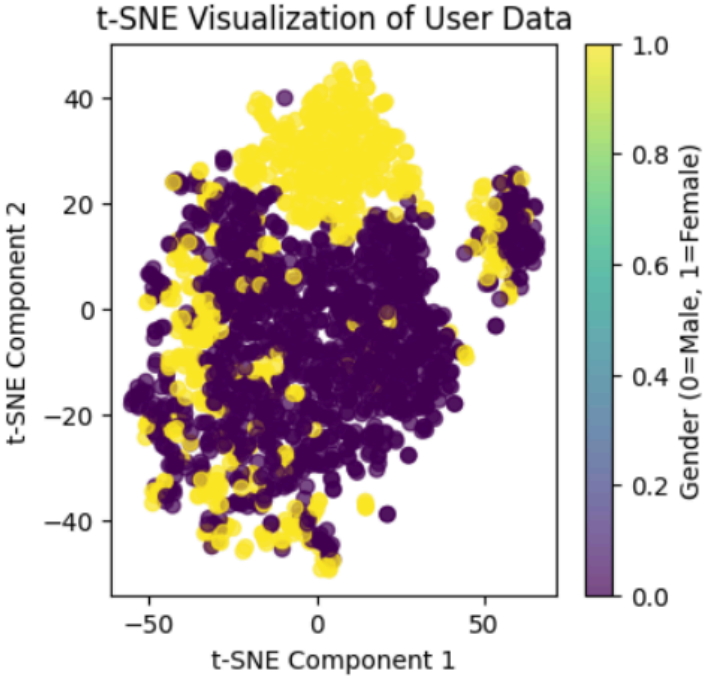


- Interprétation 1



1. Le cluster violet indique un rating de 0, il représente donc les utilisateurs qui n'ont jamais regardé de documentaires. Toutefois, la majorité de ceux qui en ont regardé un l'ont évalué entre 3 et 5: le cluster à droite.
2. Tous les genres montrés en bars ont été évalués positivement entre 3 et 5 (tous les bars avant le genre Documentary de Thriller à War) par les fans des documentaires. Ceci n'empêche pas que d'autres utilisateurs les ont aussi évalués du même score.
   Nous pouvons le vérifier en changeant la coloration du graphe de selon le documentaire à selon le genre en question. Nous constatons que la couleur majoritaire des deux clusters reste verte-jaune.

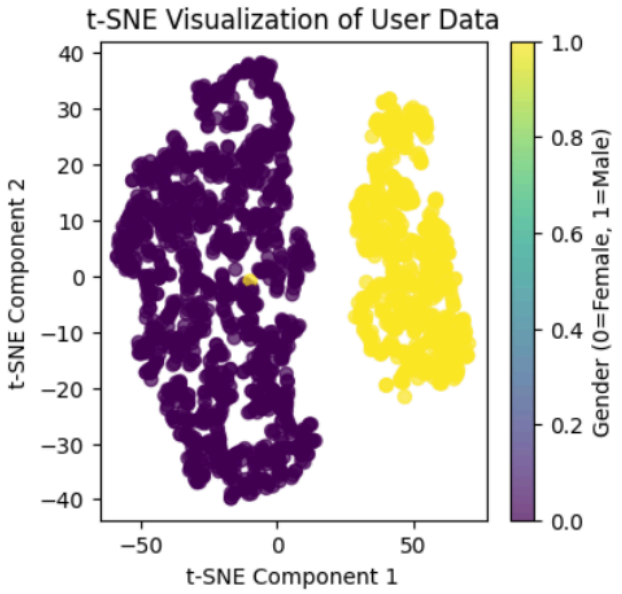
- Graphe 2: coloré selon la variable Sexe



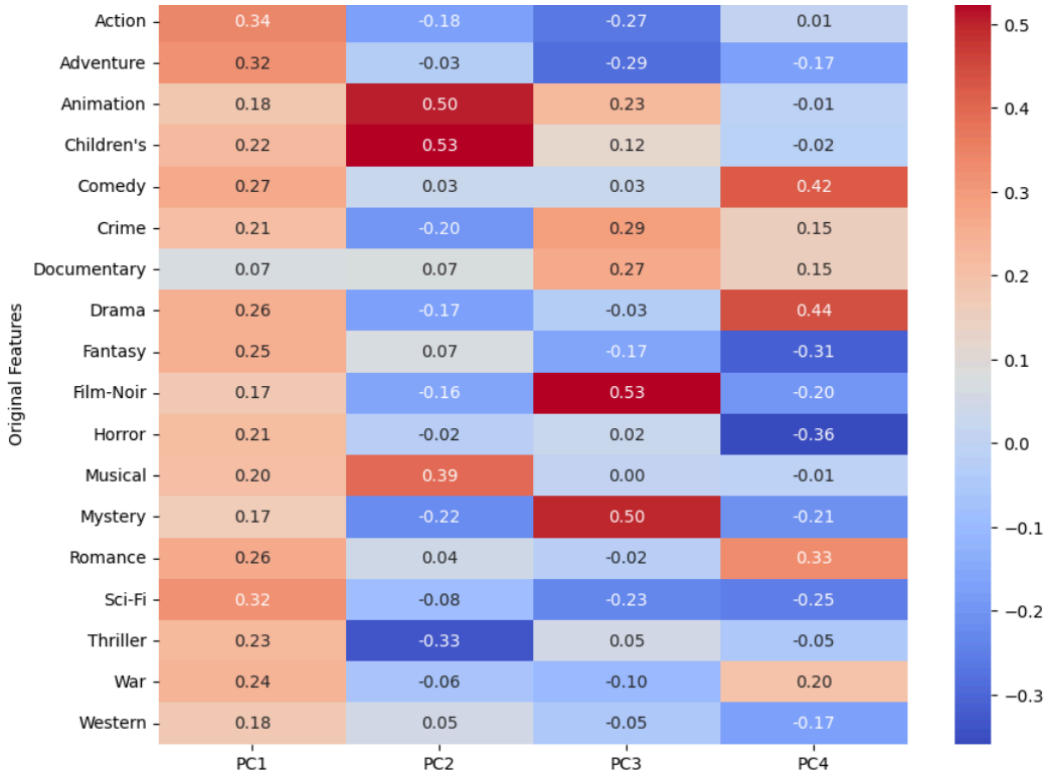
- Interprétation 2
  Les femmes regardent et aiment (rating élevé) plus de genres (genres représentés par la composante t-SNE 2) comparées aux hommes.

2. Perplexity = 25, t-SNE appliqué sur la variable Sexe + les composantes PCA

- Trustworthiness: 0.9973405193356247
- Graphe



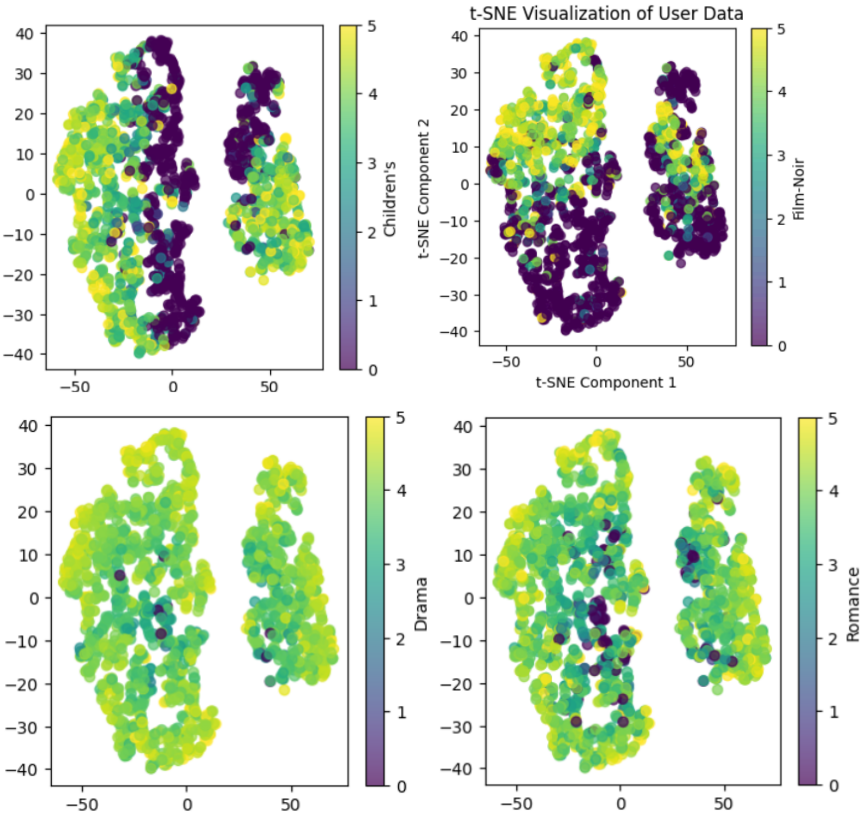
- Interprétation



La première composante t-SNE capture les relations entre le sexe et PC_4, tandis que la deuxième composante capture celles entre PC_1, PC_2 et PC_3. Ces dernières capturent la variance entre multiples genres. Ce qui nous pousse à affirmer ces relations, c'est bien la trustworthiness élevée de nos composantes qui s'élève à 0,997.

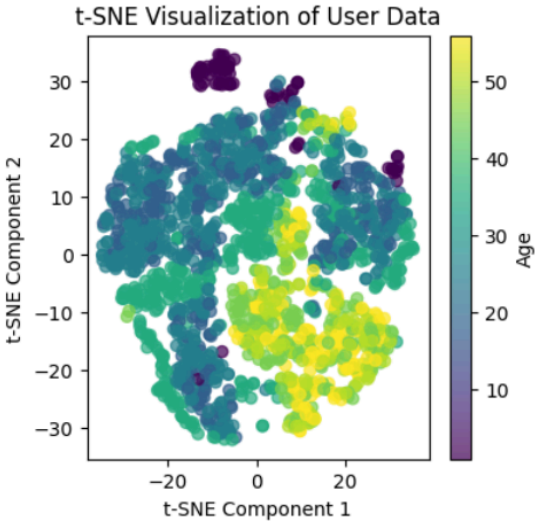
En considérant le fait que le nombre de mâles soit environ deux fois le nombre de femmes, nous concluons que

1. La proportion féminine qui aime ou qui n'aime pas ce genre de films est égale à la proportion masculine.
2. Les deux sexes se comportent similairement en termes de rating des genres. Exemple:

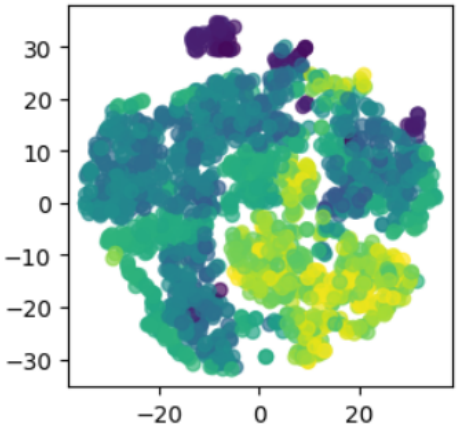


3. Perplexity = 50, t-SNE appliqué sur la variable Age + les composantes PCA

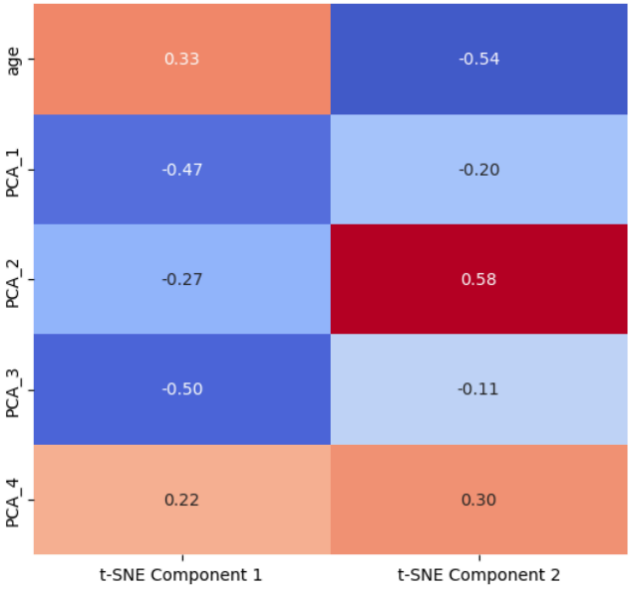
- Trustworthiness: 0.9927715295147949
- Graphe



- Interprétation
  En représentant les deux composantes selon: Romance, War, Comedy, Drama et Age, nous obtenons le graphe suivant:



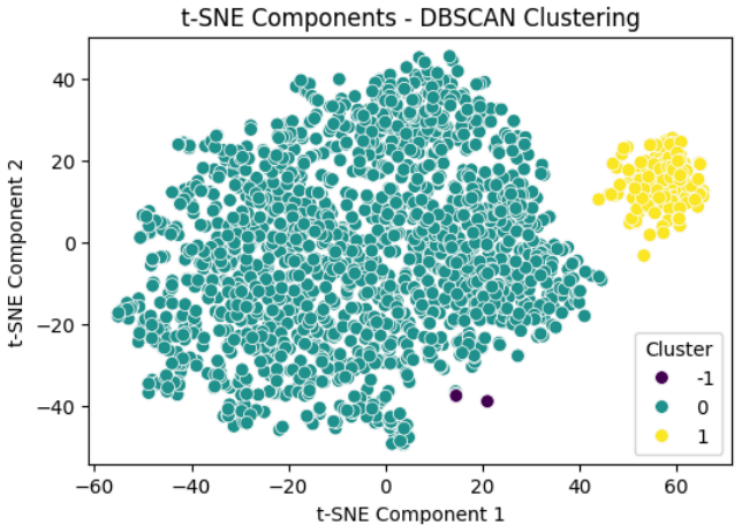
Ceci est justifié par la corrélation entre les composantes PCA. On donne à titre d'exemple: pour avoir le cluster vert-jaune, il faut maximiser PCA_4 (sur X) et minimiser PCA_1 et PCA_3 (sur Y).



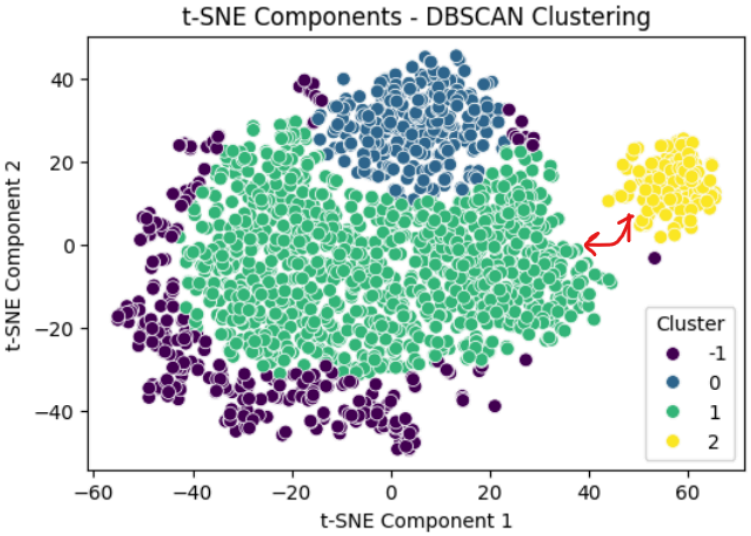
Nous en déduisons que les gens plus âgés (à partir de 45 ans) sont satisfaits par les films de genre Romance, War, Drama et Comedy. (1er cluster)
Les adolescents (moins de 18 ans) regardent les genres Animation, Children's et Musical (PCA_2) plus que d'autres. (2e cluster)
Les gens entre 18 et 45 ans préfèrent et aiment regarder le reste des genres.

3. DBScan

En observant nos données, il est évident que DBScan serait le bon modèle à utiliser pour regrouper nos données. Nous avons réussi à trouver les bons paramètres pour cela.

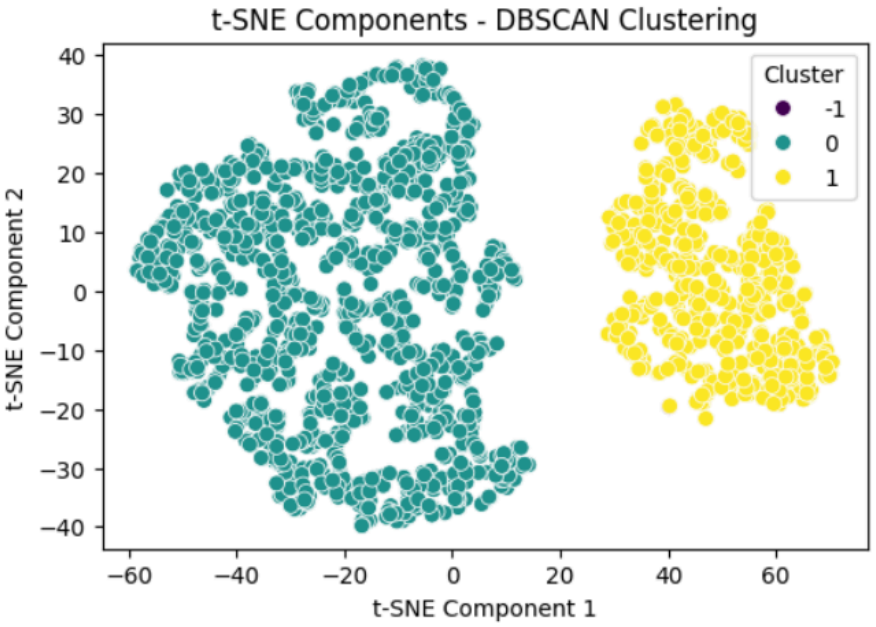


Cluster Documentaire

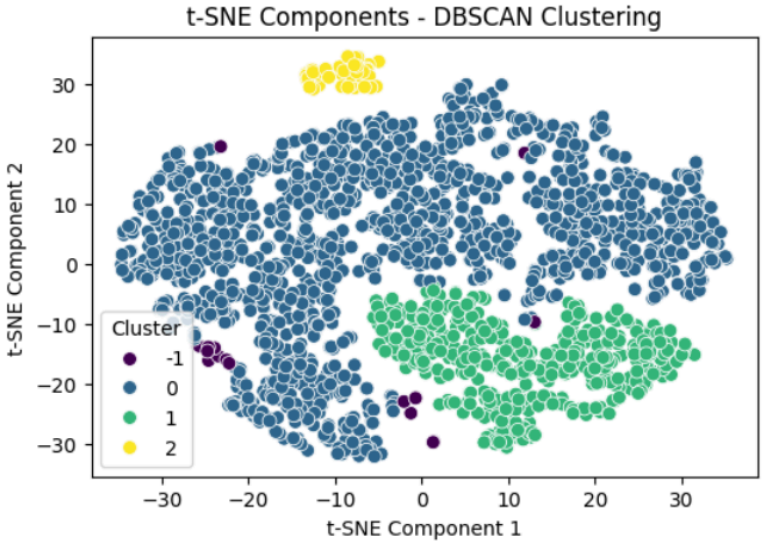


Cluster preferences selon sexe

Pour ce cluster, il faudrait considérer dans les futures utilisations que le cluster jaune et vert appartiennent au même cluster. De même pour les clusters violet et bleu.



Cluster Comportement Sexe



Cluster Préférences selon âge

Conclusion

Cette étude nous a permis d'explorer nos données d'une manière à pouvoir découvrir des pattern et des relations entre les différentes variables. Des informations pertinentes ont été déduites et d'autres pourront être découvertes grâce à cette approche. Cela est très important pour avoir une meilleure compréhension des utilisateurs qui nous permettra de s'adapter à leur nature et à leurs exigences.

IV. Système de Recommandation

Après une exploration approfondie des données et un clustering des utilisateurs via t-SNE et DBSCAN, nous avons développé plusieurs systèmes de recommandation basés sur différents principes. Ces méthodes sont ensuite comparées pour évaluer leur performance et pertinence.

1. Recommandation Basée sur les Clusters

Pour générer des recommandations personnalisées, nous avons employé une approche de clustering, qui consiste à regrouper les utilisateurs selon leurs préférences cinématographiques, puis à recommander des films populaires dans chaque groupe. Voici un résumé de la méthodologie utilisée :

Méthode de Clustering :

1. Nous avons commencé par préparer les données en calculant pour chaque utilisateur la note moyenne globale et la note moyenne par genre. Les genres ont été transformés en colonnes grâce à un encodage one-hot, puis normalisés pour obtenir des caractéristiques cohérentes.
2. Pour simplifier la structure des données, nous avons appliqué une Analyse en Composantes Principales (PCA), ce qui nous a permis de capturer les préférences globales des utilisateurs dans un espace réduit tout en conservant l'essentiel de l'information.
3. Un algorithme de clustering K-Means a été appliqué aux composantes principales issues de la PCA. Nous avons fixé le nombre de clusters à 5 pour regrouper les utilisateurs en fonction de leurs préférences similaires.
4. Les utilisateurs ont ensuite été visualisés dans l'espace des composantes principales (PC1 et PC2) avec des couleurs distinctes pour chaque cluster, permettant de mieux comprendre la segmentation des utilisateurs en groupes distincts.

Génération de Recommandations :

1. Pour chaque utilisateur, nous identifions son cluster en fonction des caractéristiques calculées.
2. Nous récupérons les films bien notés par les autres membres du même cluster, en calculant la note moyenne de chaque film au sein du cluster pour identifier les plus populaires.
3. Afin de proposer uniquement des recommandations nouvelles, les films déjà vus par l'utilisateur cible sont exclus.

4. Les films les mieux notés, mais non visionnés par l'utilisateur, sont ensuite suggérés en tant que recommandations personnalisées.

Résultat :

Nous avons testé cette méthode sur l'utilisateur 1. Voici les résultats :

```
Recommandations basées sur les clusters pour l'utilisateur 1 :
                                                  title
1107                          Perfect Candidate, A (1996)
1309   Blood For Dracula (Andy Warhol's Dracula) (1974)
2493                                       Bandits (1997)
2774  Black Cat, White Cat (Crna macka, beli macor) ...
3212                    Brandon Teena Story, The (1998)
```

Les films recommandés à l'utilisateur 1 sont ceux qui sont très bien notés par les autres membres du même cluster, mais que l'utilisateur 1 n'a pas encore vus.

Les recommandations incluent une variété de films, comme *Smoke (1995)* et *How Green Was My Valley (1941)*, ce qui peut indiquer un goût diversifié au sein du cluster.

2. Recommandation Basée sur la Similarité Utilisateur

Dans cette approche, nous avons mis en place un système de recommandation fondé sur la similarité entre utilisateurs. L'objectif est de recommander des films appréciés par des utilisateurs ayant des goûts similaires à ceux de l'utilisateur cible. Pour se faire, nous avons calculé similarité cosinus entre les utilisateurs pour repérer les plus proches "voisins".

Calcul de la Similarité :

1. Nous avons utilisé la similarité cosinus pour mesurer la proximité entre les profils des utilisateurs. Cette mesure permet d'identifier les utilisateurs ayant des préférences similaires en comparant leurs évaluations de films.
2. Pour chaque utilisateur, nous avons calculé la similarité cosinus avec tous les autres utilisateurs du dataset, en se basant sur leurs notes attribuées aux films.

Génération de Recommandations :

1. Après avoir identifié les "k" utilisateurs les plus similaires pour chaque utilisateur cible, nous avons collecté les films bien notés par ces voisins.
2. Les films que l'utilisateur cible a déjà vus sont exclus des recommandations.
3. Enfin, les films les plus fréquemment appréciés par les utilisateurs similaires sont recommandés à l'utilisateur cible, en pondérant les scores de recommandation selon le niveau de similarité.

Résultat :

Voici le résultat de la simulation effectuée sur l'utilisateur 1 :

```
Recommandations pour l'utilisateur 1 :
                                                title     score
0                    Silence of the Lambs, The (1991)  0.238859
1  Star Wars: Episode V - The Empire Strikes Back...  0.277165
2                       Raiders of the Lost Ark (1981)  0.244006
3  Star Wars: Episode VI - Return of the Jedi (1983)  0.240380
4                               American Beauty (1999)  0.291350
```

Les films recommandés à l'utilisateur 1 sont ceux que ses "voisins" ou utilisateurs similaires apprécient le plus. Chaque film est associé à un score de recommandation, reflétant la pertinence du film par rapport aux goûts de l'utilisateur cible.

Cette méthode personnalise les recommandations selon les goûts spécifiques de chaque utilisateur. Cependant, elle peut restreindre la diversité des suggestions (ex : ici on recommande Star Wars 4 et 5) en se concentrant sur un petit groupe d'utilisateurs similaires.

3. Recommandation par Décomposition en Valeurs Singulières (SVD)

Pour proposer des recommandations adaptées aux préférences individuelles des utilisateurs, nous avons employé la décomposition en valeurs singulières (SVD), une méthode efficace pour identifier les goûts latents.

Méthodologie et Choix des Hyperparamètres

Pour offrir des recommandations plus précises, nous avons utilisé la décomposition en valeurs singulières (SVD), une technique de factorisation de matrice qui identifie les facteurs latents influençant les évaluations des films par les utilisateurs. Afin de maximiser la performance du modèle, nous avons appliqué une recherche par grille (Grid Search) pour optimiser les hyperparamètres clés de SVD :

- **n_factors** : Nombre de facteurs latents pour la décomposition.
- **n_epochs** : Nombre d'itérations d'entraînement.
- **lr_all** : Taux d'apprentissage pour la descente de gradient.
- **reg_all** : Paramètre de régularisation pour éviter le surapprentissage.

Après plusieurs essais, les meilleurs paramètres obtenus étaient :

```
Meilleurs paramètres pour SVD :
{'n_factors': 250, 'n_epochs': 150, 'lr_all': 0.007, 'reg_all': 0.15}
Meilleur RMSE : 0.9407
Meilleur MAE : 0.7445
RMSE: 0.9249
MAE:  0.7324
0.7324470583987116
```

Ces valeurs ont été choisies pour minimiser les erreurs de prédiction, mesurées par le RMSE (Root Mean Squared Error) et le MAE (Mean Absolute Error). L'optimisation du modèle a permis

d'atteindre un **RMSE** de 0.9249 et un **MAE** de 0.7324, confirmant la précision de cette approche pour une personnalisation fine et des recommandations pertinentes, même pour des films moins évidents.

Processus de Recommandation

Pour évaluer l'efficacité de la recommandation, un utilisateur fictif a été ajouté avec des évaluations fictives pour divers films. Ensuite, le modèle SVD a été entraîné sur l'ensemble de données pour prédire les notes des films non encore évaluées par cet utilisateur. La fonction de recommandation sélectionne ensuite les films ayant les scores prédits les plus élevés.

Résultats et Interprétation

```
Préférences initiales de l'utilisateur fictif (9999) :
Toy Story (1995) : 5 étoiles
Twelve Monkeys (1995) : 4 étoiles
Star Wars: Episode IV - A New Hope (1977) : 3 étoiles
Terminator 2: Judgment Day (1991) : 5 étoiles
Godfather, The (1972) : 4 étoiles

Recommandations pour l'utilisateur fictif (9999) :
                              title  estimated_rating
0  Shawshank Redemption, The (1994)          4.820868
1           Schindler's List (1993)          4.679536
2                     Sanjuro (1962)          4.668301
3         When We Were Kings (1996)          4.619757
4         American History X (1998)          4.607653
```

Les meilleures recommandations pour l'utilisateur fictif incluent des titres populaires tels que *Shawshank Redemption* (4.82 étoiles) et *Schindler's List* (4.67 étoiles). Ces scores montrent que le modèle SVD parvient à identifier des films alignés avec les préférences implicites de l'utilisateur, capturant des facteurs latents tels que le genre ou le style apprécié.

4. Recommandation Basée sur la Similarité dans l'Espace PCA

Dans cette approche, nous avons utilisé l'espace PCA pour calculer la similarité entre utilisateurs, permettant de capturer des préférences globales et variées. La similarité cosinus est appliquée aux vecteurs de chaque utilisateur dans cet espace réduit pour identifier les profils similaires.

Méthodologie :

Après avoir identifié les utilisateurs similaires, nous recommandons des films populaires parmi leurs évaluations, en excluant ceux déjà vus par l'utilisateur cible. Les films sont ensuite pondérés par le degré de similarité entre utilisateurs, assurant des recommandations pertinentes.

Résultat :

```
Recommandations basées sur la similarité PCA pour l'utilisateur 1 :
                                        title     score
0                            It Takes Two (1995)  1.569934
1           Amazing Panda Adventure, The (1995)  1.361051
2                     So Dear to My Heart (1949)  2.273032
3                      Dog of Flanders, A (1999)  1.056131
4  Mating Habits of the Earthbound Human, The (1998)  1.423582
```

Les films recommandés à l'utilisateur 1 semblent variés, allant de films familiaux (*It Takes Two*) au films d'aventure (*Amazing Panda Adventure*). Cela montre que cette méthode capture bien des préférences globales et propose des choix qui peuvent enrichir l'expérience utilisateur. En pondérant les notes par similarité, cette méthode aligne les recommandations avec les goûts de l'utilisateur tout en lui permettant de découvrir de nouveaux films. Ces suggestions personnalisées reflètent des préférences globales partagées, offrant ainsi une expérience de découverte enrichissante pour l'utilisateur 1.

5. Analyse de Réseau avec NetworkX

Pour approfondir l'analyse des relations entre utilisateurs et films, nous avons utilisé NetworkX pour modéliser le système de recommandation sous forme de graphe bipartite. Cette approche permet d'explorer la centralité des utilisateurs et des films dans le réseau, offrant des informations complémentaires pour la recommandation.

Méthode de l'analyse

Nous avons construit un graphe où chaque nœud représente un utilisateur ou un film, et chaque lien correspond à une évaluation. Les mesures de centralité calculées incluent la centralité de degré, qui identifie les utilisateurs les plus actifs et les films les plus populaires, et la centralité d'intermédiarité, qui révèle les nœuds jouant un rôle de pont dans le réseau. Les utilisateurs avec une forte centralité de degré sont de bons candidats pour des recommandations diversifiées, tandis que les films ayant une forte centralité sont populaires et adaptés à des recommandations grand public.

Résultats de l'analyse

Voici les résultats pour les utilisateurs et les films ayant les valeurs de centralité de degré les plus élevées.

```
Top utilisateurs par centralité de degré :
[(2858, 0.23750000000000002), (2762, 0.17047872340425532), (3175, 0.12207446808510639), (3114, 0.11329787234042553), (2716, 0.10851063829787234)]

Top films par centralité de degré :
[(2858, 0.23750000000000002), (260, 0.17739361702127666), (1210, 0.17154255319148937), (2762, 0.17047872340425532), (1196, 0.16276595744680852)]
```

Les utilisateurs 2858, 2762, 3175, 3114, 2716 ont évalué un grand nombre de films, ce qui les rend influents dans le réseau. Ils peuvent être considérés comme des "super-utilisateurs" ou des "power users" en raison de leur activité élevée.

Les films 2858, 260, 1210, 2762, 1196 ont été évalués par un grand nombre d'utilisateurs, ce qui signifie qu'ils sont populaires et susceptibles de plaire à un large public. Ils pourraient être des candidats sûrs pour les recommandations, car ils ont déjà attiré l'intérêt d'un grand nombre d'utilisateurs.

Conclusion

Chaque méthode de recommandation présente des avantages et des limites en termes de personnalisation et de diversité des suggestions. Le clustering offre une bonne adéquation pour des préférences générales mais manque de finesse dans la personnalisation. La similarité entre utilisateurs permet des recommandations hautement personnalisées, bien que cela puisse restreindre la diversité en se concentrant uniquement sur des utilisateurs aux goûts similaires. La SVD combine personnalisation et exploration, en facilitant la découverte de nouveaux films grâce aux facteurs latents. Enfin, la similarité dans l’espace PCA propose des recommandations variées en capturant des préférences globales, enrichissant ainsi l’expérience utilisateur. L’intégration de ces approches pourrait créer un système de recommandation plus complet et polyvalent, capable de répondre aux attentes variées des utilisateurs et d'optimiser leur expérience. Les prochaines étapes consisteraient à évaluer ces méthodes dans un contexte réel et à ajuster le modèle selon les retours des utilisateurs afin de maximiser l’efficacité et la satisfaction.