Clustering via Non-Negative Matrix Factorization (NNMF)

Projet réalisé par :
Fezoui Yacine - Paul-Arthur NGUYEN - Stephane WU
Master 1 Informatique 2024/2025

Introduction

La segmentation des données est une approche essentielle pour identifier des groupes au comportement similaire dans des jeux de données complexes. La méthode NNMF (Non-Negative Matrix Factorization) permet de réduire la dimensionnalité tout en conservant la structure sous-jacente des données. Ce projet vise à explorer l'application de la NNMF combinée au clustering pour analyser les comportements des clients du Mall Customers Dataset, en mettant en évidence des tendances utiles pour la prise de décision.

Dataset :

Le jeu de données utilisé, Mall Customers Dataset, comporte 200 enregistrements décrivant des clients d'un centre commercial à l'aide de plusieurs caractéristiques. Ces variables incluent le genre, l'âge (en années), le revenu annuel (en milliers de dollars) et le spending score, une évaluation sur une échelle de 1 à 100 mesurant les habitudes de dépense. Pour notre étude, seules les variables âge, revenu annuel et spending score ont été sélectionnées, car elles permettent d'identifier des groupes distincts en fonction des comportements démographiques et économiques. Le prétraitement des données a permis de standardiser ces valeurs et de faciliter l'analyse.

	Age	Annual Income (k$)	Spending Score (1-100)
0	19	15	39
1	21	15	81
2	20	16	6
3	23	16	77
4	31	17	40

Figure 1 : Visualisation des données brutes du dataset

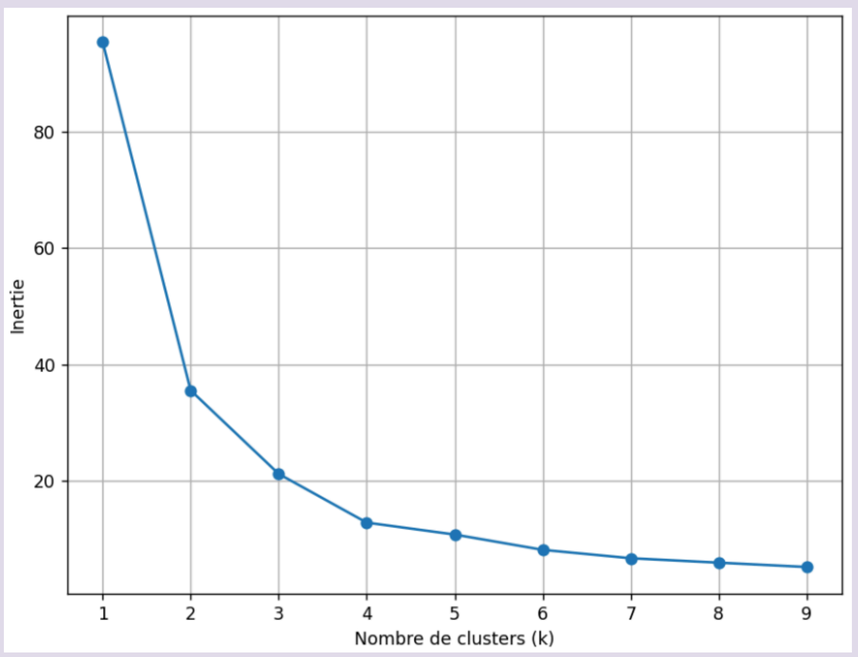

Figure 3 : Courbe de la méthode du coude

Méthodologie :

Pour analyser les données, un prétraitement a été effectué en remplaçant les valeurs manquantes et les outliers par la moyenne, suivi d'une standardisation des valeurs :

Age	Annual Income (k$)	Spending Score (1-100)
0.500000	0.567568	0.193878
0.423077	0.648649	0.122449
0.961538	0.297297	0.479592
0.307692	0.792793	0.224490
0.326923	0.531532	0.724490

Figure 2 : Visualisation des données après prétraitement

Deux méthodes de NMF (Factorisation en Matrice Non-Négative) ont été testées, produisant des résultats similaires. Ensuite, pour déterminer le nombre optimal de clusters, deux approches ont été utilisées :

- La première, **la méthode du coude**, a suggéré 3 clusters, comme le montre le graphe affiché à gauche
- La seconde **méthode du score de silhouette** a confirmé ce choix, garantissant ainsi une segmentation cohérente.

Les librairies principales utilisées incluent pandas pour manipuler les données, matplotlib pour les visualisations, et sklearn pour la NMF et le clustering KMeans.

Résultats et Discussion :

Les résultats de la segmentation des clients sont présentés dans un tableau, qui présente les moyennes des variables pour chaque cluster, ainsi que le nombre d'éléments et la répartition par sexe. Ce tableau permet de mieux comprendre les caractéristiques distinctives de chaque groupe de clients, avec des informations sur l'âge, le revenu annuel, le score de dépenses, et la répartition hommes-femmes dans chaque cluster. Ces données offrent une base pour l'analyse approfondie des comportements et des tendances au sein des groupes segmentés.

Cluster	Age	Annual Income (k$)	Spending Score (1-100)	Nombre d'éléments	% male	% female
0	41.112245	58.704082	47.489796	98	37.755102	62.244898
1	30.196429	68.017857	82.357143	56	42.857143	57.142857
2	44.565217	55.434783	16.826087	46	58.695652	41.304348

Figure 4 : Tableau des moyennes, des pourcentages par cluster

- **Cluster 0 (48.5%) :** Ce groupe, principalement composé de personnes de classe moyenne, préfère les produits au bon rapport qualité/prix et effectue des achats réfléchis. Ils sont prêts à dépenser plus pour un besoin essentiel. Il serait efficace de promouvoir des produits de qualité et des offres limitées pour les inciter à des achats plus importants.

- **Cluster 1 (28.5%) :** Ce groupe représente des jeunes adultes, souvent débutants dans la vie active, qui dépensent impulsivement, attirés par les tendances et les marques populaires. Les réseaux sociaux, les produits en vogue et les promotions limitées seraient des leviers efficaces pour capter leur attention.

- **Cluster 2 (23%) :** Ces clients dépensent peu et privilégient les achats ciblés ou complémentaires. Ils recherchent des promotions, des produits de niche ou des articles difficiles à trouver. Proposer des offres et des programmes de fidélité pourrait les encourager à acheter davantage et à revenir plus souvent.

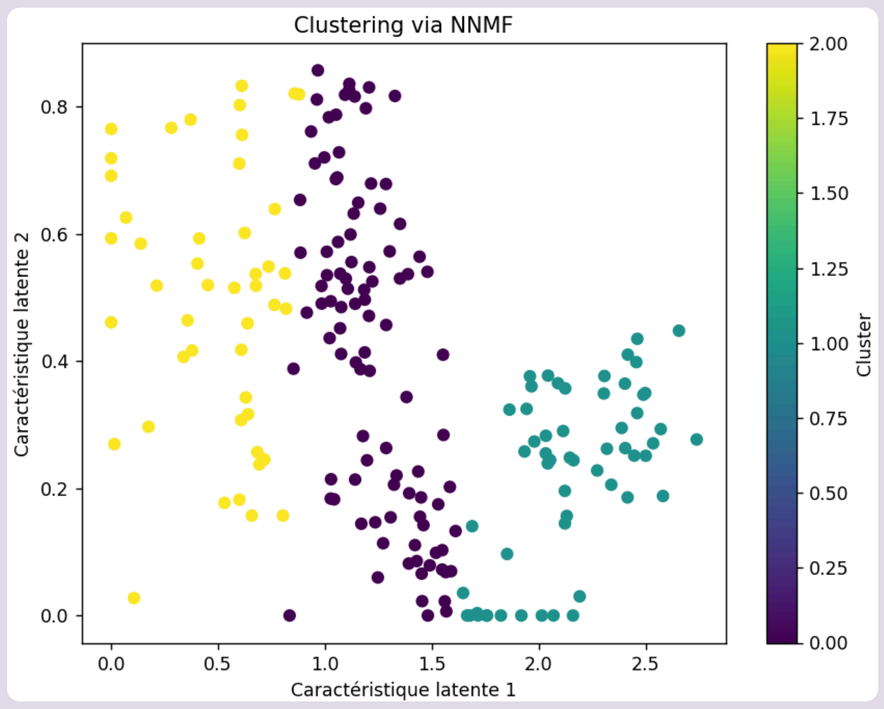

Figure 5 : Visualisation des clusters sur les deux premières dimensions latentes

Conclusion

L'intégration de la NNMF et du clustering a permis d'identifier trois groupes distincts de clients, mettant en lumière des tendances comportementales importantes. Ces résultats peuvent être utilisés pour cibler efficacement les campagnes marketing et adapter les stratégies commerciales. La robustesse de cette approche peut être explorée davantage en comparant avec d'autres méthodes comme les GMM ou DBSCAN.