Projet du cours Intelligence Artificielle et apprentissage

Lefebvre Florian

May 29, 2023

Contents

1 Introduction

1.1 Résumé de l'objectif du projet

Il faut au départ choisir une bases de données dan le site de datasets de kaggle[1]. Ensuite, il faut appliquer (au moins) trois algorithmes de clustering distincts et étudier le nombre de cluster idéal et la qualité des clusters. Il est

[1] https://www.kaggle.com/datasets?search=clustering

possible dorénavant d'utiliser les librairies disponibles[2], voire les comparer avec nos implémentations.

1.2 Le choix du projet

J'ai choisi la base de données suivante accessible à cette url[3]. Le but de celle-ci est de catégoriser les pays à l'aide de facteurs socio-économiques et sanitaires qui déterminent le développement global du pays. HELP International est une ONG humanitaire internationale qui s'engage à lutter contre la pauvreté et à fournir aux habitants des pays arriérés des commodités et des secours de base en cas de catastrophes et de calamités naturelles. Le PDG doit prendre la décision de choisir les pays qui ont le plus besoin d'aide.

2 Le choix du projet

2.1 Les raisons de ce choix

J'ai trouvé que le fait de partir sur un cluster non supervisé offrait un challenge plus important étant donné que les données ne sont pas étiquetées. Il est donc impossible à l'algorithme de calculer de façon certaine un score de réussite. Mais c'est à l'algorithme de découvrir les structures sous-jacentes à ces données. Malgré tout, à l'aide de connaissances personnelles sur l'économie de ces pays, il est possible de vérifier si les résultats obtenus sont fiables ou non. Le fait de pouvoir suggérer à un PDG des pays à aider économiquement est en soi très motivant. Nous posons l'hypothèse que les méthodes d'apprentissage rassemblerons en 1 cluster les pays à aider en priorité.

2.2 Les outils nécessaires pour le projet

Pour ce projet j'ai utilisé l'IDE Spyder[4] car il est considéré comme l'un des meilleurs outils pour faire du Machine Learning avec Python. Spyder est une interface de développement spécialement conçue pour les projets orientés data. En ce qui concerne les bibliothèques, j'ai utilisé scikit-learn[5] pour les algorithmes principaux de Machine Learning, scipy[6] pour des calculs de mathématiques complexes, numpy[7] pour gérer des tableaux, pandas pour

[2]https://scikit-learn.org/stable/modules/clustering.html
[3]https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data
[4]https://www.spyder-ide.org/
[5]https://scikit-learn.org/stable/
[6]https://scipy.org/
[7]https://numpy.org/

gérer des DataFrames[8], et matplotlib[9] pour créer des visualisations de données.

3 Les algorithmes

3.1 Les méthodes d'évaluation

Dans les algorithmes que nous allons appliquer à notre cluster, deux méthodes d'évaluation seront utilisées pour permettre d'étudier le nombre de cluster idéal et leur qualité. La première est la méthode Elbow aussi appelée du coude. La méthode consiste à tracer la variation expliquée en fonction du nombre de clusters, et à choisir le coude de la courbe comme le nombre de clusters à utiliser.

La seconde est la méthode silhouette qui est également une méthode pour trouver le nombre optimal de clusters. Le coefficient de silhouette est une mesure de la similitude d'un point de données à l'intérieur d'un groupe par rapport à d'autres groupes. Ce coefficient peut varier entre -1 et +1. Un coefficient proche de +1 signifie que l'observation est située bien à l'intérieur de son propre cluster, tandis qu'un coefficient proche de 0 signifie qu'elle se situe près d'une frontière ; enfin, un coefficient proche de -1 signifie que l'observation est associée au mauvais cluster.

3.2 L'algorithme kmeans

La méthode des kmeans permet de regrouper les objets en K clusters distincts. La méthode des kmeans repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque objet et le centroïde de son cluster. La figure 1 est une évaluation de k means avec la méthode elbow, pour chercher le nombre optimal de clusters .La méthode elbow semble indiquer un nombre minimal de clusters entre 4 et 5, tandis que au dessus de 8 clusters, le gain est faible. La figure 2 est une évaluation de kmeans avec la méthode silhouette. Nous pouvons constater que pour 4, 5 ou 6 clusters, les taux d'erreurs sont élevés. Pour 7 ou 8 clusters, les résultats sont plus qualitatifs. Au dessus de 8 clusters, les résultats baissent en qualité. Les 2 méthodes d'évaluation sont donc cohérentes entre elle.

[8]https:https://pandas.pydata.org/
[9]https://matplotlib.org/

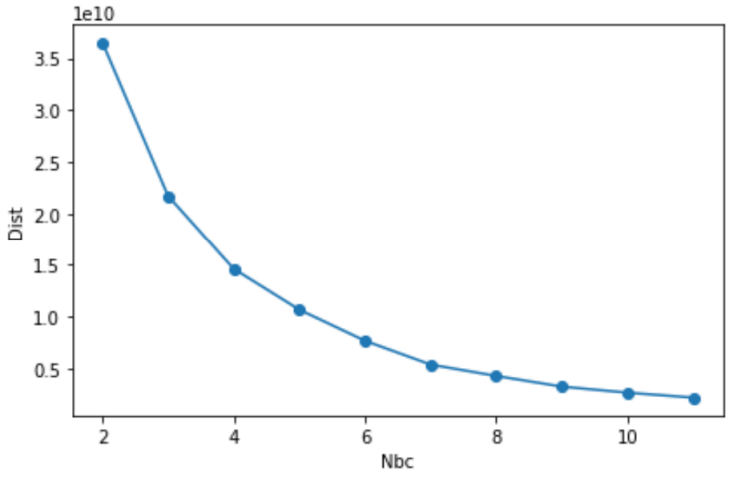

Figure 1: Evaluation de l'algorithme kmeans avec la méthode Elbow

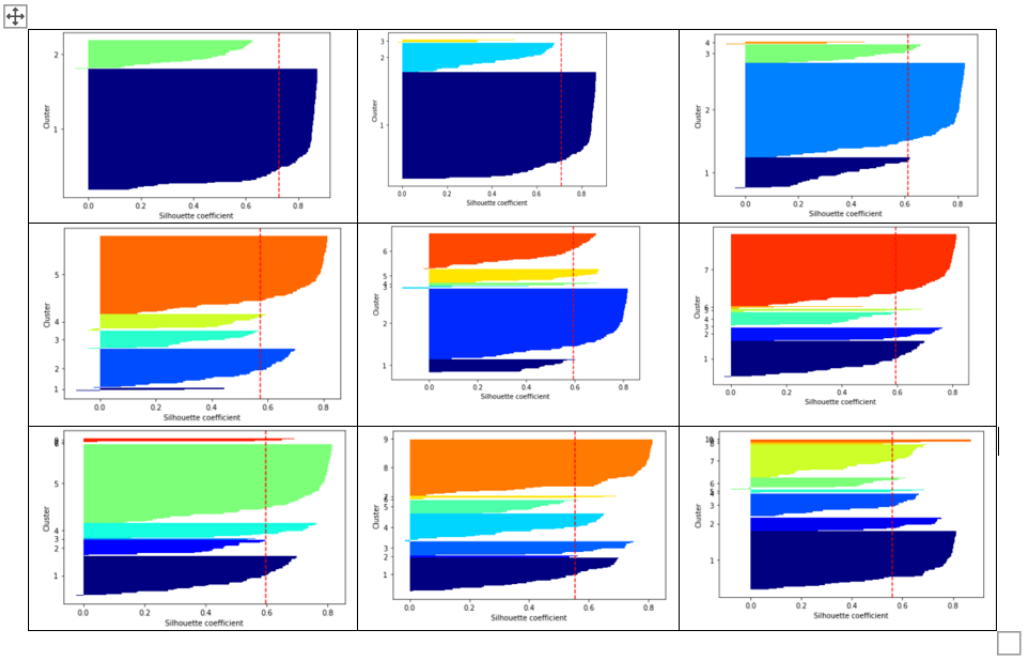

Figure 2: Kmeans avec la méthode silouhette de 2 à 10 clusters

3.3 L'algorithme cluster hiérarchique

Cet algorithme vise à trouver un regroupement naturel en fonction des caractéristiques des données. L'algorithme de clustering hiérarchique permet

de trouver des groupes imbriqués de données en construisant la hiérarchie. Nous avons évalué cette méthode avec différents critères de distance. Parmi les différents critères de distance, de nombreux résultats étaient redondants. Nous avons donc gardé les distances 5000, 10000 et 15000. Ces résultats sont visible dans la figure 3, où les clusters ont été évalué avec la méthode Silhouette. Les meilleures résultats sont obtenus avec une distance de 15000. Avec cette distance, le cluster hiérarchique génère 14 clusters.

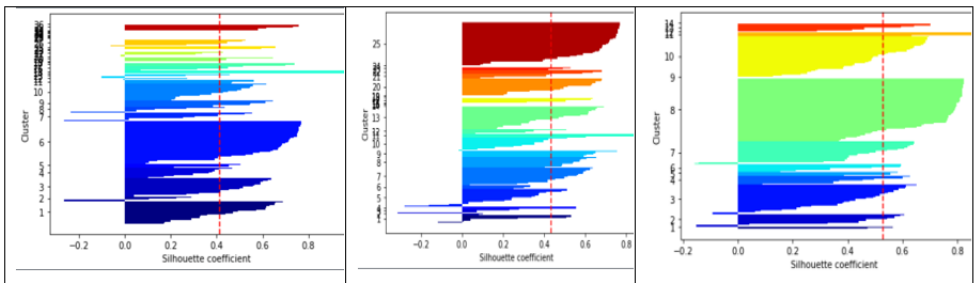

Figure 3: cluster hiérarchique avec la méthode silouhette

3.4 L'algorithme Dbscan

DBSCAN (density-based spatial clustering of applications with noise) est un algorithme de partitionnement de données. Il s'agit d'un algorithme fondé sur la densité dans la mesure qui s'appuie sur la densité estimée des clusters pour effectuer le partitionnement. Cependant, cet algorithme isole les outliers. Cependant avec nos donnés, la plupart des pays peu développés sont considérés comme des outliers. De plus, malgrès des tentatives avec de nombreux paramètres différentes; nous avons obtenu des résultats peu qualitatifs, comme visible dans la figure 4.

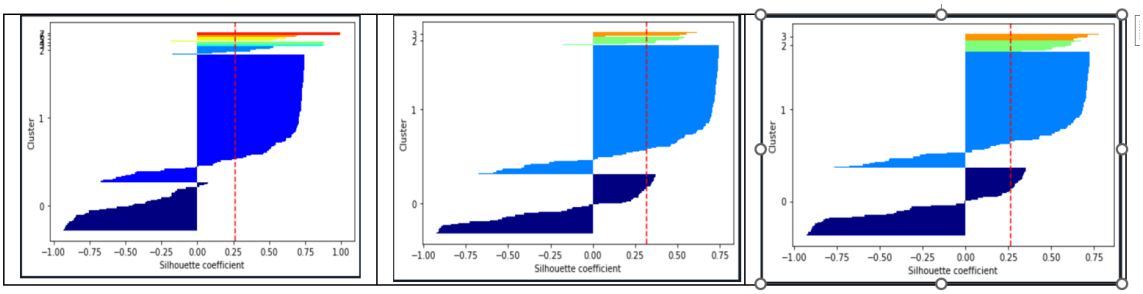

Figure 4: Résultats avec l'algorithme Dbscan

4 Conclusions

Parmi les différentes méthodes qu'on a testées, il y en a certaine qui ont donné des résultats et d'autres non. Toutes les méthodes ne sont pas adaptées à un même exercice. Les choix de la méthode semble être un aspect fondamental de l'analyse de donnés.

Dans cet exercice spécifique, l'objectif était d'identifier des pays à aider en priorité. Nous avons observé dans le clustering que les pays peu développés sont plus proches des pays moyennement développés que les pays moyennement développés sont proches des pays très développés. Les méthodes de clustering ont donc tendance à isoler les pays les plus riches. Augmenter le nombre de clusters ne fait que subdiviser les pays riches. Les pays en difficulté sont donc dilués au milieu des pays moyennement développés. Cela réfute donc notre hypothèse de départ: il n' y a pas de clusters rassemblant uniquement les pays en difficultés.