Machine Learning - Rapport du projet 2

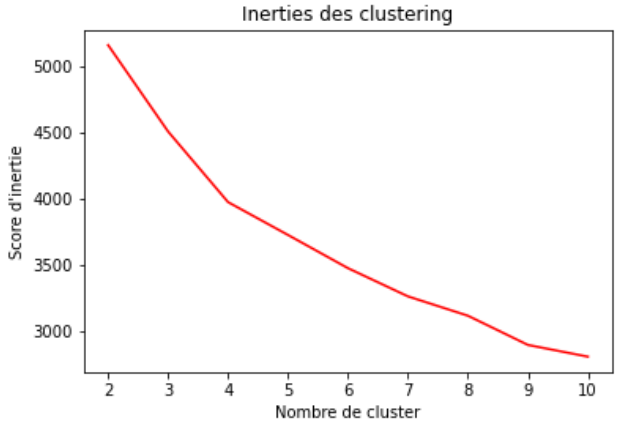
Benjamin Cochez

Décembre 2020

1 Clustering/Visualisation

1.1 Choix du nombre de cluster

Pour me donner une idée, j'ai décidé d'utiliser la méthode du coude qui pourrait m'aider à choisir plutôt tel ou tel cluster. Voici le tableau qui reprend les différents clustering et les scores d'inertie correspondants à chaque clustering donné par le modèle KMeans :



On peut observer qu'en dessous de 4 clusters il y a une trop grosse erreur. On pourrait donc estimer qu'il faille créer au moins 4 clusters pour intérpréter au mieux les données. Cependant, cette méthode est plus une idée qu'une obligation puisque l'interprération des données est très subjective. C'est-à-dire qu'une personne pourrait donc très bien faire une interprétation de ces données avec un clustering à 3 clusters.

1.2 Justification du choix des différents clustering

Note : j'ai mis les photos dans le dossier du rapport pour pouvoir zoomer si besoin, les noms des pays sont peut être trop petits.

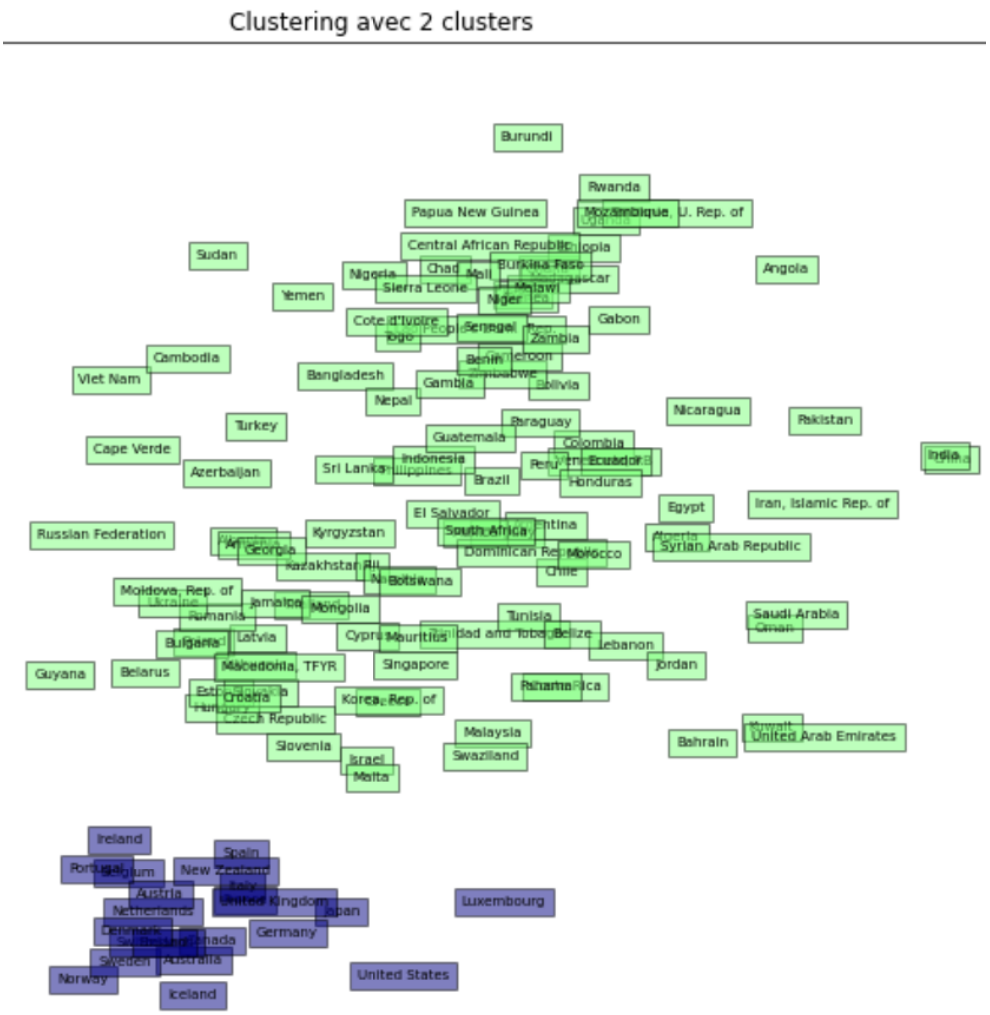
1.2.1 Clustering qui ne contient pas assez de cluster

J'ai décidé de prendre un premier clustering avec 2 clusters comme clustering qui ne contiendrait pas assez de cluster à mes yeux. Plusieurs points ont motivé mon choix.

Premièrement, l'erreur d'inertie qui est donné avec un clustering de 2 clusters est vraiment très grande, ce qui indique que certains pays sont très loin du centre de leur cluster et donc qui ont probablement de grandes différences par rapport à un pays qui serait proche du centre de son cluster.

Deuxièmement, quand on regarde la visualisation en 2D, on a du mal à faire ressortir des informations qui nous apprendrais des choses sur le Dataset. La seule chose

qu'on pourrait dire c'est que les pays les plus développés se retrouvent dans un cluster et dans l'autre ce sont tous les pays développés et en voie de développement, mais même là c'est assez vague comme interprétation.



Les pays les plus proches du centre de leur cluster sont : United Kingdom, El Salvador.

1.2.2 Clustering qui contient assez de cluster

Ensuite, j'ai décidé de prendre un deuxième clustering à 5 clusters. Dans ce clustering, nous pouvons dégager beaucoup plus de choses intéressantes que dans le clustering à 2 clusters.

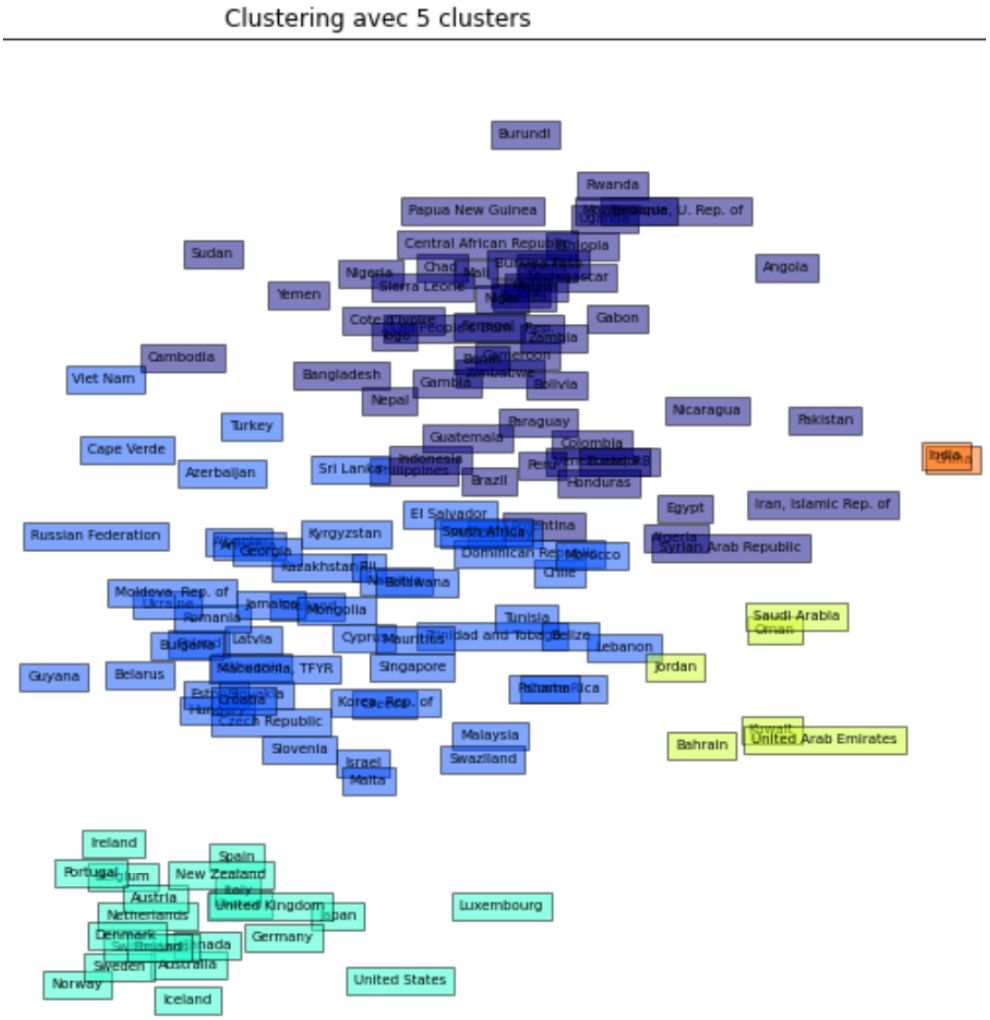
Premièrement, nous pouvons distinguer des ensembles de pays avec un mode de vie différent, que ça soit religieux, ou au niveau des habitudes alimentaires (accès à l'eau et à la nourriture) ou le fait qu'il y ait plus de femme qui travaille, ect. Par exemple, les pays d'Arabie qui sont regroupés dans le cluster jaune ont un mode de vie très différent des pays Européens, Africains ou de l'Est de l'Europe. On peut faire le même constat pour certains pays asiatiques comme la Chine ou l'Inde.

Deuxièmement, on pourrait également distinguer les pays par leur économie et démographie. Par exemple, en Chine et en Inde il y a énormément de personnes qui y vivent et leur croissance économique en 2004 étaient relativement proches et assez importante. De plus, on pourrait dire que les pays dans le cluster Violet ont une croissance démographie très forte mais une économie plutôt faible, dans

le cluster bleu clair, les pays ont une faible croissance démographique mais une économie plutôt bonne et dans le cluster bleu foncé ont peu voir une démographie importante et une économie partagée. Pour les pays du cluster jaune, l'économie est assez particulière puisqu'elle n'est pas basée sur la culture (très aride) mais plus sur l'exploitation de ressource comme le pétrole.

Par ailleurs, on pourrait déduire que tout en bas (cluster bleu clair) on va retrouver des pays très développés, au milieu (cluster bleu foncé/jaune/orange) des pays plus ou moins développés et tout en haut des pays en voie de développement (cluster violet).

Troisièment, d'un point de vu de la santé on pourrait émettre l'hypothèse qu'il est préférable de vivre dans un pays du cluster bleu clair (accès aux soins médicaux, espérence de vie plus élevé, plus de règles de sécurité, moins de maladie), que dans les pays du cluster bleus foncés/jaune qui est un compromis entre bonne et mauvaise santé et que dans les pays du cluster violet/orange où la santé est plutôt problématique.



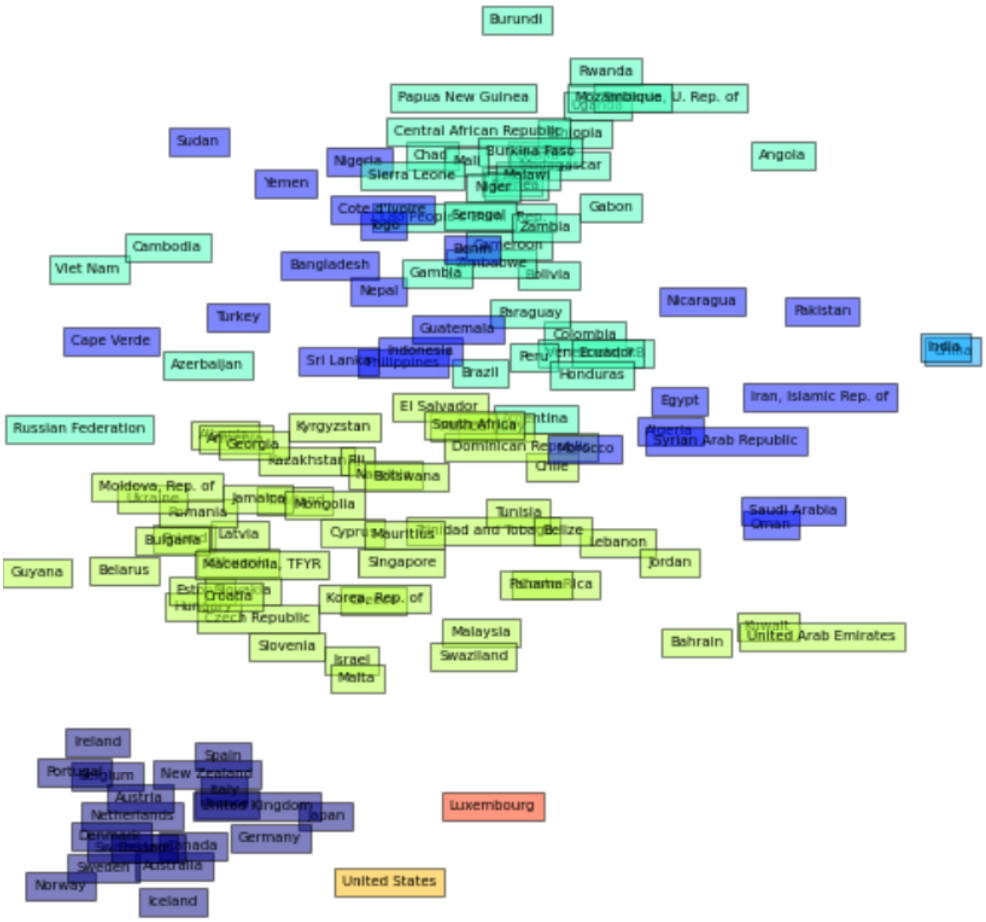
Les pays les plus proches du centre de leur cluster sont : Niger, Cyprus, United Kingdom, Saudi Arabia, China.

1.2.3 Clustering qui contient trop de cluster

Finalement, j'ai décidé de prendre un clustering à 7 clusters. Je trouve que ce clustering devient plus approximatif. Dans ce clustering, on a déjà plus de mal

à sortir de nouvelles généralités. Bien que le Luxembourg ait une économie bien particulière et certaines caractéritisiques qui le sépare des autres pays, peut-on se dire qu'il est vraiment très différent des pays du cluster bleu clair ? Il est vrai que nous pourrions encore tirer des informations de cette visualisation mais cela ne nous apporterais pas vraiment plus d'information que le clustering à 5 clusters, puisqu'ici cela devient de plus du plus en cas par cas et donc, faire plusieurs clusters d'un seul pays ne nous aidera pas plus à comprendre les regroupements. On voit d'ailleurs qu'avec la méthode du coude, l'erreur diminue mais cela est vraiment infime par rapport à un clustering à 5 clusters.

Clustering avec 7 clusters



Les pays les plus proches du centre de leur cluster sont : Finland, Guatemala, China, Guinea, Cyprus, United States, Luxembourg.

1.3 Explication sur la réduction de dimension

Pour la réduction de dimension, j'ai décidé de régler le méta paramètre de « perplexity » à 40 (ni trop peu, ni trop grand) qui me permet d'avoir une visualisation en 2D de mes données en haute dimension relativement fiable et stable. J'ai dû relancer plusieurs fois l'algorithme t-SNE pour essayer d'atteindre une erreur la plus basse possible et j'ai décidé de m'arrêter à une erreur de 0.217. Il y a donc certes de la perte d'information mais très légère (assurée par l'algorithme t-SNE qui utilise la descente de gradient pour atteindre un minima local).