TP

Classification de pays selon le niveau d'aide nécéssaire

Table des matières

Introduction

Dans ce TP nous nous intéressons à un problème de clustering de pays en besoin d'aide humanitaire. Afin de pouvoir attribuer des fonds aux pays en ayant le plus besoin, il est nécéssaire de classifier les pays selon leur niveau de développement en 3 classes:

- Pas d'aide nécéssaire
- Peut avoir besoin d'aide
- À besoin d'aide

Pour ce faire on se basera sur les données suivantes qui donnent des indicvateurs sur chaque potentiel pays en besoin d'aide:

Country-data.csv 9.0KB

data-dictionary.csv 0.8KB

Matériel

Installation python avec les packages suivants:

- gestion des données : numpy, scipy et pandas
- visualisation : matplotlib, seaborn, geopandas
- machine learning : Scikit-learn

Rendu

À l'issu de ce TP, vous devez rendre les deux éléments suivants:

- Un rapport détaillant votre travail selon les éléments suivants:
  - Présentation du dataset
  - Visualisation des features les plus importantes, et relations entre features
  - choix de features importantes
  - présentation de la (des) pipeline(s) Scikit learn utilisées avec hyperparamètres utilisés
  - visualisation du résultat de clustering
  - Analyse post-clustering des classes
  - Ce que vous avez appris
- Un programme python (ok pour un notebook jupyter) permettant de reproduire les visualisations de la présentation et qui effectue le clustering

> ℹ️ On s'attachera à justifier les éléments suivants:
> * Comment on a choisi les features importantes pour notre problème
> * Comment on a choisi l'algorithme d'apprentissage et tuné les hyperparamètres
> * Si le clsutering obtenu fais sens

Ce n'est pas grave si le résultat final n'est pas très bon ou pas cohérent. Du moment que l'on est capable d'analyser pour dire ce qui ne vas pas et qu'on a compris ce qu'on a fait (et qu'on peut l'expliquer), c'est l'essentiel. La note ne prends pas en compte un clustering attendu.

Délai : 21 Février au soir

Conditions : Par groupe de 2

Lecture et visualisation

Dans un premier temps nous allons lire les données et les mettre en mémoire.

1. Télécharger les données et les mettre dans un dossier de travail
2. Utiliser la librairie pandas pour mettre en mémoire les données
3. Avec la documentation de pandas, essayer de calculer les statistiques suivantes:
    a. Moyenne et variance par catégorie d'information (feature)
    b. Matrice de corrélation entre features
4. Avec la libraire seaborn, visualiser la distribution des valeurs pour chaque variable selon tous les pays.
5. Faire une fonction qui pour un choix de feature, visualise sous la forme d'histogramme les X pays ayant les valeurs les plus élevées (ou plus faibles). Utiliser cela pour intuiter quel jeux de variables permettent de mieux décrire la situation d'un pays.
6. Essayer de trouver avec ces visualisations quelques exemples de pays dans chaque classe (pas besoin d'aide, peut avoir besoin d'aide, a besoin d'aide).

Choix d'un algorithme et pipeline

1. Familiariser vous avec Scikit-learn (cf TD)
2. Construire une approche de type "Pipeline" qui à partir des données en mémoire fera les choses suivantes:
    a. Prétraitements (minmax, standardscaler, etc)
    b. réduction de dimension (combinaison de feature, pca, ou autre)
    c. clustering
3. Essayer plusieurs pipelines avec des hyperparamètres différents et choisir un résultat selon un critère que vous définirez (google: algo + hyperparametre tuning)

Analyse des résultats

1. Afficher les résultats de clustering selon une forme compréhensible
2. (Bonus) Afficher les résultats sous la forme d'une carte de couleurs avec geopandas
3. Vérifier la cohérence des classes obtenues en regardant les valeurs des variables selon les classes obtenues. (On pourra s'aide de visualisation de distribution, ou en regardant des statistiques de type moyenne, variance par classe)