## TP 2 - Panorama des méthodes de fouille de donnée

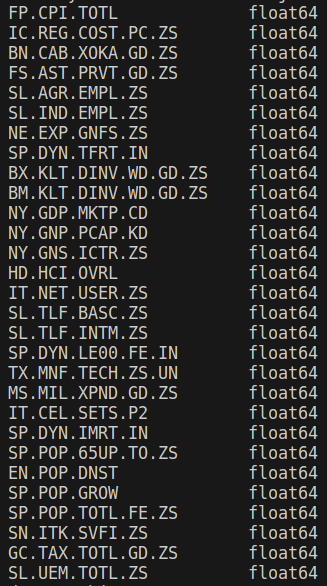
### 1 Analyse exploratoire

* 1. **Classification / Clustering**

1. Ouvrez le fichier Data\_World\_Development\_Indicators2.csv à l'aide de la bibliothèque pan-

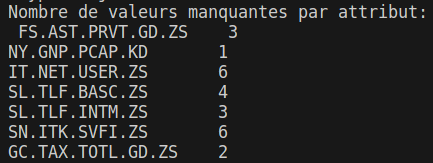
das. Combien y a-t-il d'attributs ? De quels types sont-ils ?

Il y a 31 attributs :



2. Y a-t-il beaucoup de valeurs manquantes dans les données ? Comment ces valeurs sont-elles

représentées dans le ficher ?



Les valeurs manquantes sont représentées par NaN.

Remplacez chaque valeur manquante par la médiane obtenue par chaque attribut en utilisant SimpleImputer. Quel est l'intérêt d'utiliser ici la médiane plutôt que la moyenne ?

- Sensibilité aux valeurs aberrantes : La médiane est moins sensible aux valeurs aberrantes que la moyenne. Si vos données contiennent des valeurs extrêmes, la médiane peut offrir une meilleure représentation de la tendance centrale, car elle ne dépend pas autant de ces valeurs extrêmes.

- Distribution asymétrique : En présence d'une distribution asymétrique, la médiane peut être un meilleur indicateur de la tendance centrale. Par exemple, dans le cas d'une distribution fortement asymétrique, la moyenne peut être tirée vers la queue de la distribution, tandis que la médiane reste au centre.

- Caractéristiques discrètes : Pour les variables discrètes ou catégorielles, la médiane peut être plus appropriée que la moyenne. Si une variable a des valeurs entières et que la moyenne n'a pas de sens dans ce contexte, la médiane peut être plus significative.

3. Il est souvent utile d'appliquer un filtre de normalisation sur tous les attributs avant d'utiliser

des méthodes de clustering. Quel est l'effet du filtre StandardScaler sur les données ? A-t-on

besoin ici d'employer ce filtre sur les données étudiées ?

Le filtre StandardScaler permet de transformer les données en ajustant la moyenne à zéro et l'écart type à un. L'effet du StandardScaler sur les données est de les ramener à une échelle commune, éliminant ainsi les différences d'échelle entre les attributs. Cela est particulièrement important lors de l'utilisation de méthodes de clustering, car beaucoup d'entre elles, comme les algorithmes basés sur la distance (par exemple, k-means), sont sensibles aux différences d'échelle entre les attributs. Si un attribut a une échelle beaucoup plus grande que les autres, il peut dominer l'influence du clustering.

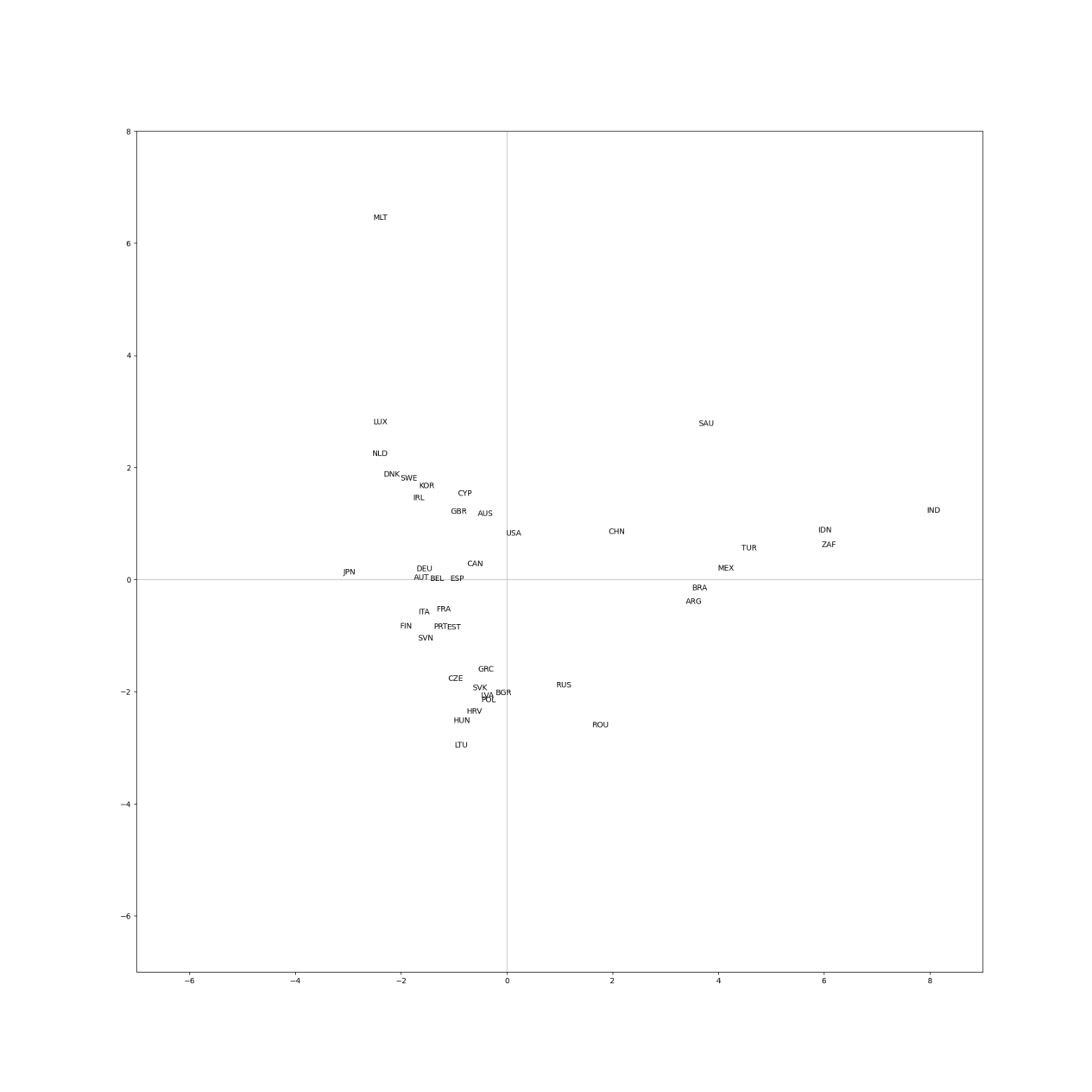
Nos valeurs ont différent ordre de grandeur selon l’attribut donc **oui** il est nécessaire d’appliqué le filtre.

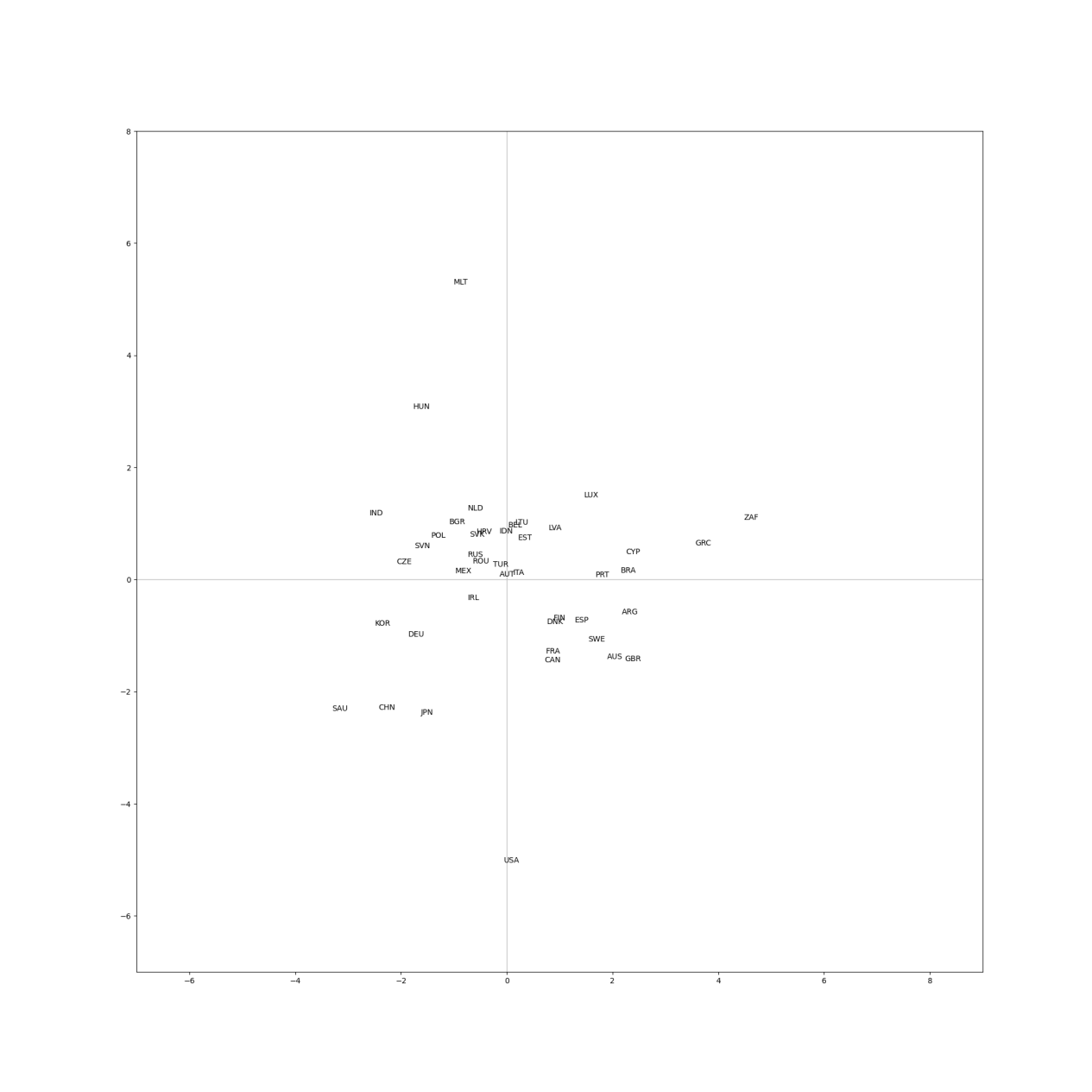
4. Effectuez une analyse en composantes principales (ACP) sur les données à l'aide de la biblio-

thèque scikit-learn.

Joignez au rapport l'affichage des instances étiquetées par le code du pays suivant les 2 facteurs

principaux de l'ACP, puis suivant les facteurs 3 et 4 de l'ACP.

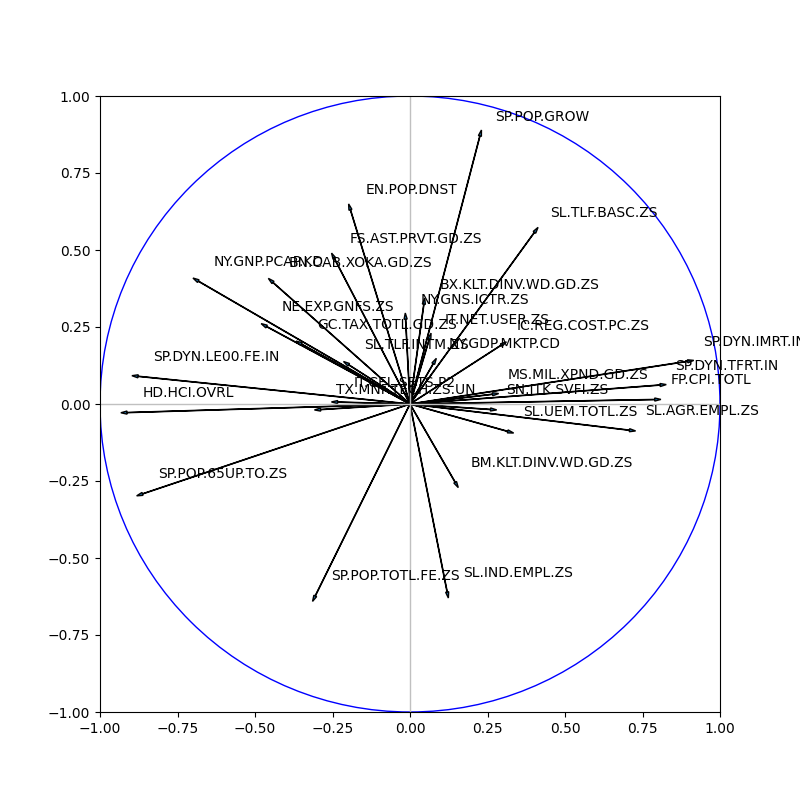




5. Que représentent les 4 premiers facteurs de l'ACP ? Vous pourrez répondre à cette question

en calculant la variation des valeurs propres en fonction du rang et en traçant le cercle de

corrélation entre les facteurs et les variables originales.



Commentez les deux graphiques obtenus à la question précédente (affichage dans les deux

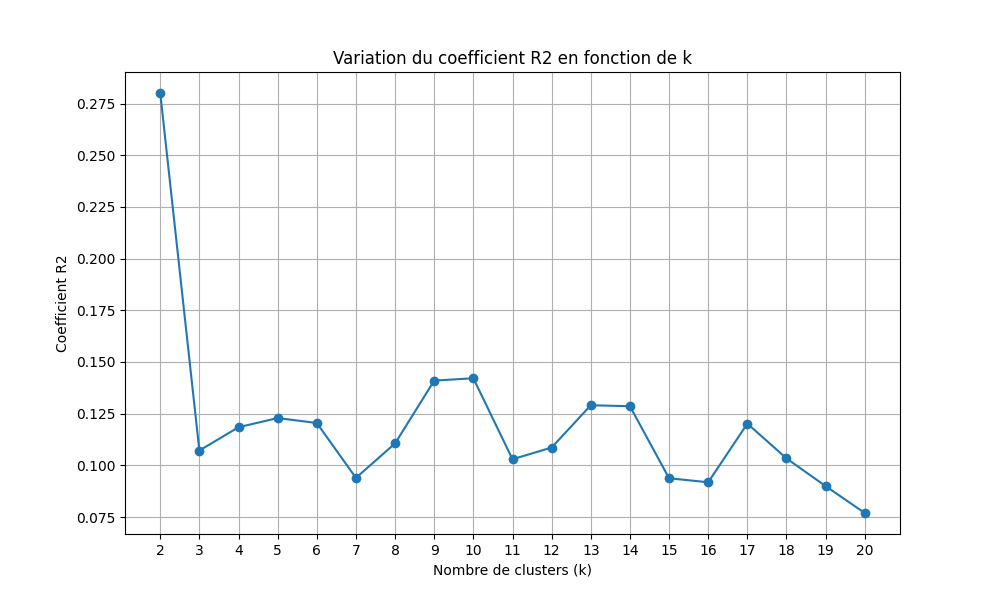
premiers plans de l'ACP) en discutant la proximité entre les pays.

6. Reprenez les données telles qu'elles étaient avant leur transformation par l'ACP. En utilisant

KMeans de la bibiothèqe scikit-learn, réalisez un clustering avec la méthode des k -moyennes,

où k représente le nombre souhaité de clusters. Tracez la variation du R2 pour k variant de 2

à 20. D'après cette courbe, quelle valeur de k reteneriez-vous ?



C’est la valeur k=2 qui maximise le coefficient R2 c’est donc la valeur que je retiendrais.

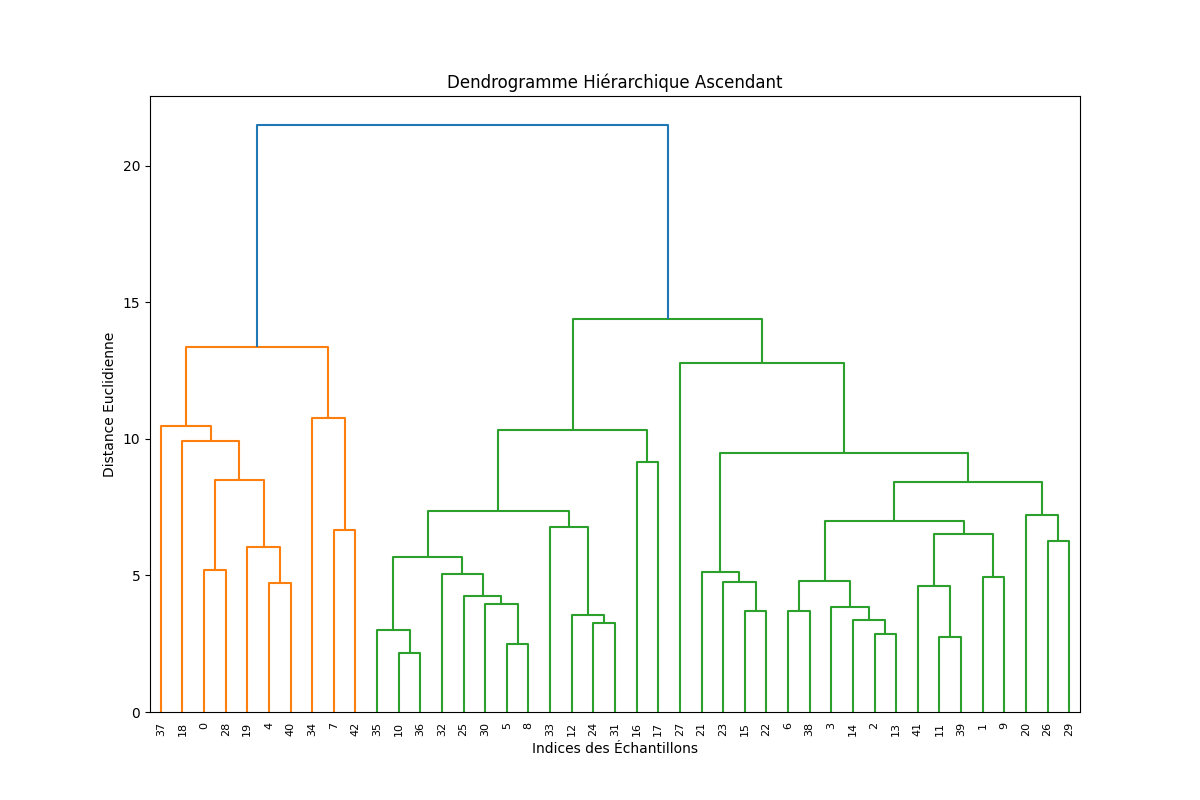
7. En fixant k à 8, commentez les profils des groupes contenant la France, le Mexique et la

Bulgarie.

8. Réalisez un clustering au moyen d'une méthode hiérarchique ascendante avec la classe dendo-

gram de la bibliothèque scipy.

Joignez au rapport l'affichage du dendogramme obtenu.



9. Commentez l'arbre obtenu précédemment en discutant des similitudes entre pays. À quel

niveau de l'arbre feriez-vous une coupure ? Combien de groupes de pays seraient alors créés ?

Je couperai vers 20 pour avoir deux clusters bien séparés. On aura 2 groupe de pays créés

**1.2 Recherche de règles d'association**

1. Ouvrez les données à l'aide de la bibliothèque csv. Quels sont les types des attributs ? Que

représentent les valeurs présentes dans le fichier ? Comment agit les commandes du script

fourni, permettant la discrétisation ?

2. Le niveau de couverture est très dépendant des données. Établissez l'ensemble des itemsets

fréquents à l'aide de la fonction apriori() de mlxtend en fixant un support minimal à 0,002.

Expliquez ce que signifie cette valeur de 0,1. Combien d'itemsets obtenez-vous ? Donnez leur

fréquence en fonction du nombre d'items qu'ils contiennent.

3. Construisez les règles à partir de l'ensemble des itemsets précédemment obtenus en recourant à la fonction association\_rules() de mlxtend. Pour ne pas avoir un nombre trop important de règles générées, fixez un seuil minimal de 0,5 pour la valeur de confiance.

Combien de règles obtenez-vous ? Commentez la première règle obtenue.

4. Apriori possède plusieurs critères pour évaluer la pertinence d'une règle, parmi lesquelles

confidence et lift 2 . Quelle est la meilleure règle pour l'ensemble grocery pour chacune des métriques confidence et lift ? Commentez les scores et règles obtenus.

5. Un dirigeant du supermarché souhaite améliorer ses ventes en réorganisant le positionnement

des produits dans ses rayons. D'après votre analyse du panier de consommation fourni, quels

produits sont souvent achetés conjointement avec du yaourt et et du café ?

### 2 Analyse prédictive

**2.1 Classement**

1. Ouvrez les données à l'aide de la bibliothèque pandas. Combien y a-t-il d'attributs dans ces

fichers ? Quels sont leurs types ? Quelle est la variable à prédire ? Quels attributs, indiqués

comme numériques, prennent deux valeurs et sont donc en réalité binaires ? Transformez ces

attributs pour qu'ils deviennent catégoriels. Vous pourrez également supprimer la colonne ID

qui identifie de manière unique chaque client, mais n'aide pas à la prédiction sur l'efficacité

de la campagne marketing.

2. Les classes sont-elles équilibrées dans les fichiers ? Y a-t-il des valeurs manquantes dans les

données ? Comment ces valeurs sont-elles représentées dans le ficher ? Quels sont les attributs

concernés par des valeurs manquantes ?

1. Expliquez en une phrase en quoi consiste la méthode élémentaire de classement DummyClassifier utilisant la stratégie most\_frequent. Dans quel cas ce classifieur pourrait donner les meilleurs résultats parmi ceux testés ?

4. Dans un premier temps, seules les données numériques sont considérées pour prédire la classe. Remplacez chaque valeur manquante par la moyenne obtenue par chaque attribut en utilisant SimpleImputer. Normalisez ensuite les valeurs numériques à l'aide de StandardScaler.

Calculez les performances des 6 méthodes de classement étudiées sur les données marketing en réalisant une validation croisée en 5 blocs sur les données. Parmi les méthodes testées, quelle

est la plus performante par rapport au taux de classification ?

5. Dorénavant, on se propose de n'employer que les attributs catégoriels. Réalisez une disjonction des variables catégorielles à l'aide de la classe OneHotEncoder de la bibliothèque scikit-learn.

Calculez les performances de classement des 6 méthodes étudiées, en réalisant à nouveau une

validation croisée en 5 blocs. Observez-vous des performances similaires à ce que vous aviez à

la question précédente ?

6. Utilisez enfin l'ensemble des colonnes, c-à-d en prenant en compte les attributs numériques et catégoriels. Calculez à nouveau les performances de classement en validation croisée à 5 blocs

et commentez les résultats obtenus.

7. Jusqu'à présent, la fonction accuracy\_score fournie dans le script ne calcule que le taux de

classification (accuracy ). Pourquoi cette métrique ne permet-elle pas d'évaluer correctement

les méthodes pour des classes déséquilibréesire. Construisez un nouveau jeu de données équi-

libré sur les 2 classes par cette méthode (vous pourrez utiliser la fonction resample du package sklearn.utils ou bien RandomOverSampler de la bibliothèque imbalanced-learn). Relancez l'évaluation des methodes de classement sur ce nouveau jeu de données.