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

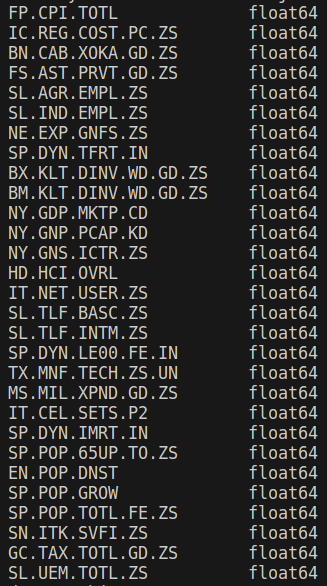
### 1 Analyse exploratoire

* 1. **Classification / Clustering**

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

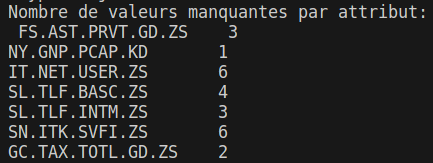
das. Combien y a-t-il d'attributs ? De quels types sont-ils ?

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.

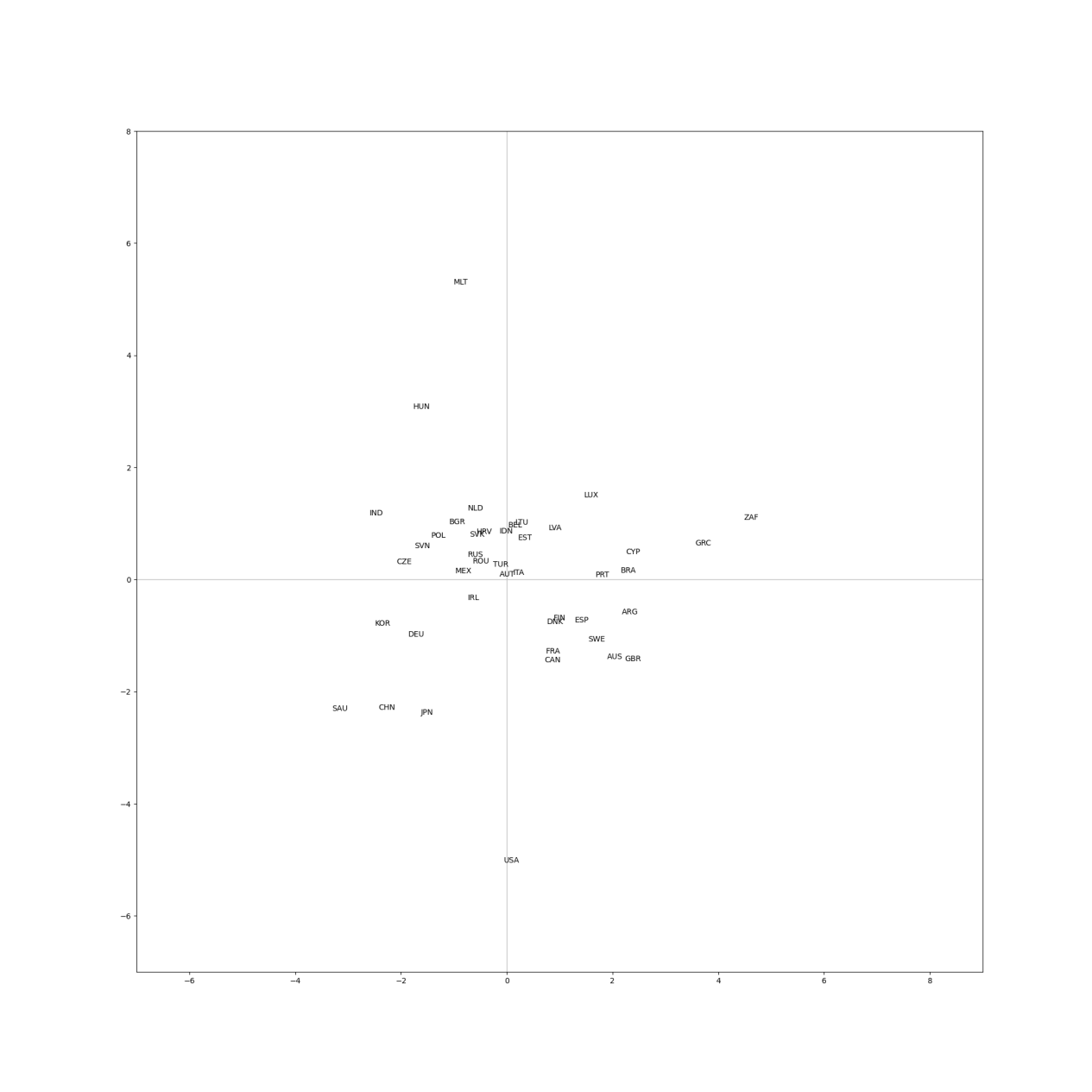
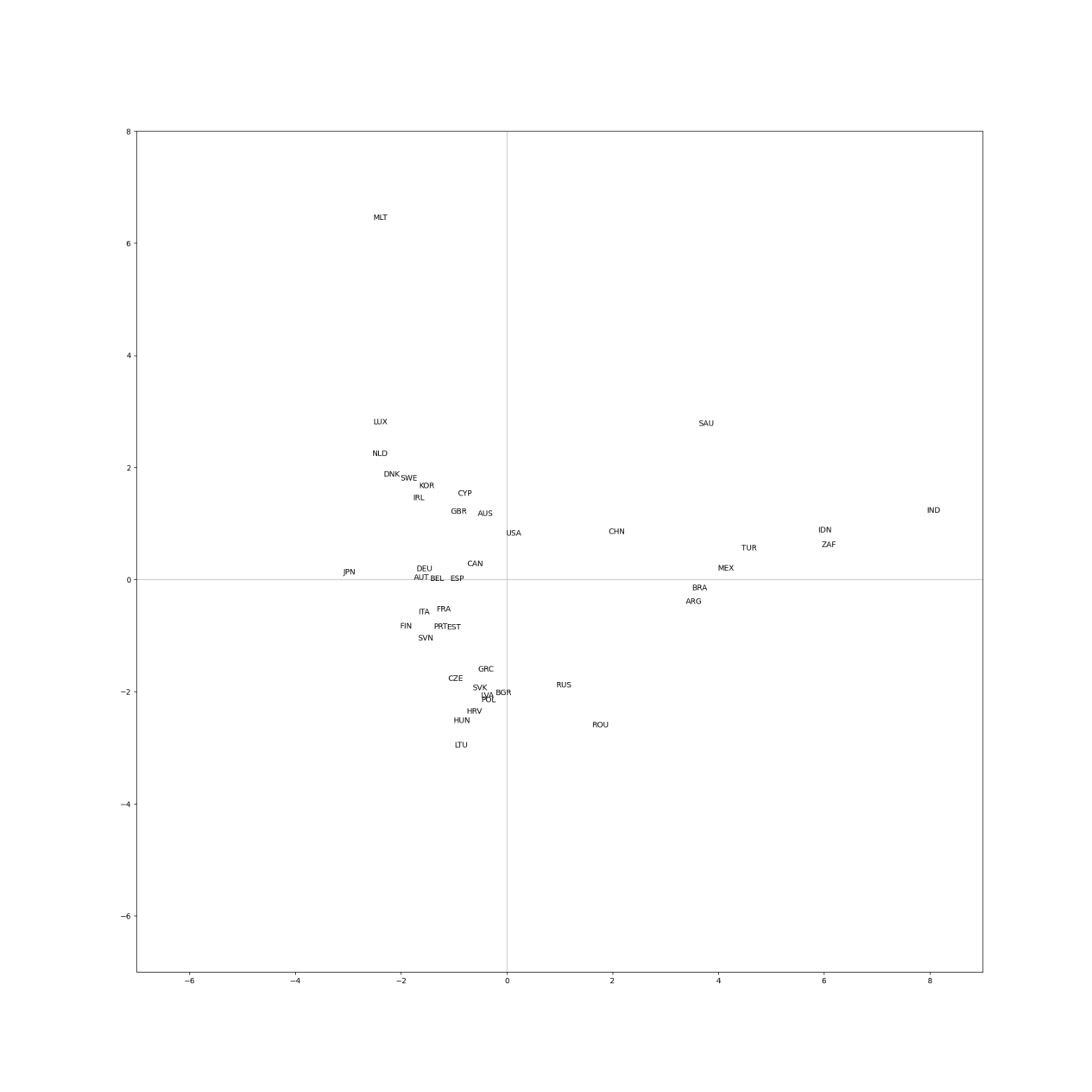
Nos valeurs ont différent ordre de grandeur selon l’attribut donc **oui** il est nécessaire d’appliqué le filtre.

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

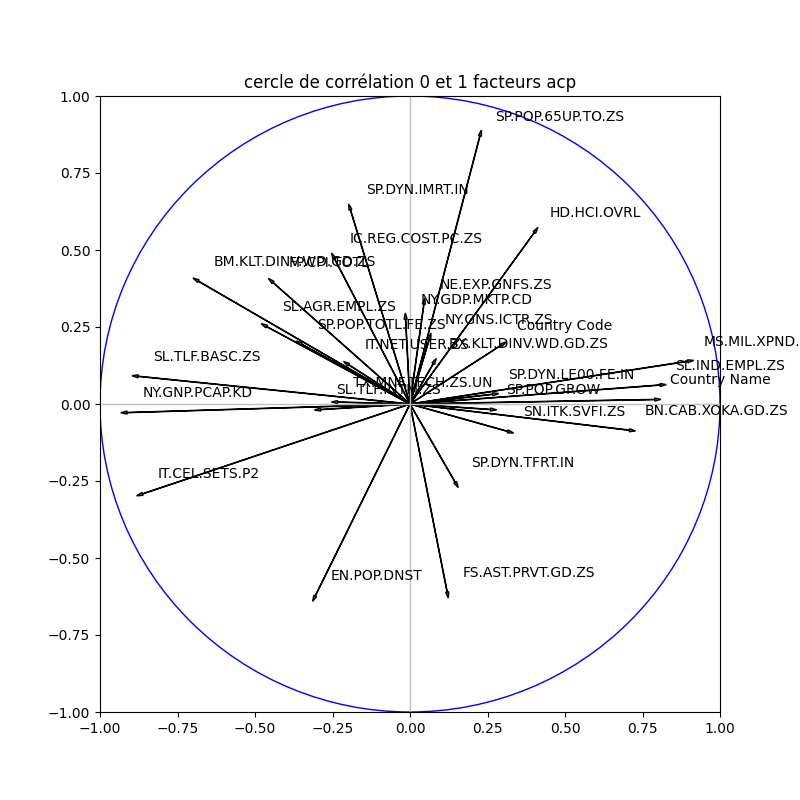
principaux de l'ACP, puis suivant les facteurs 3 et 4 de l'ACP.

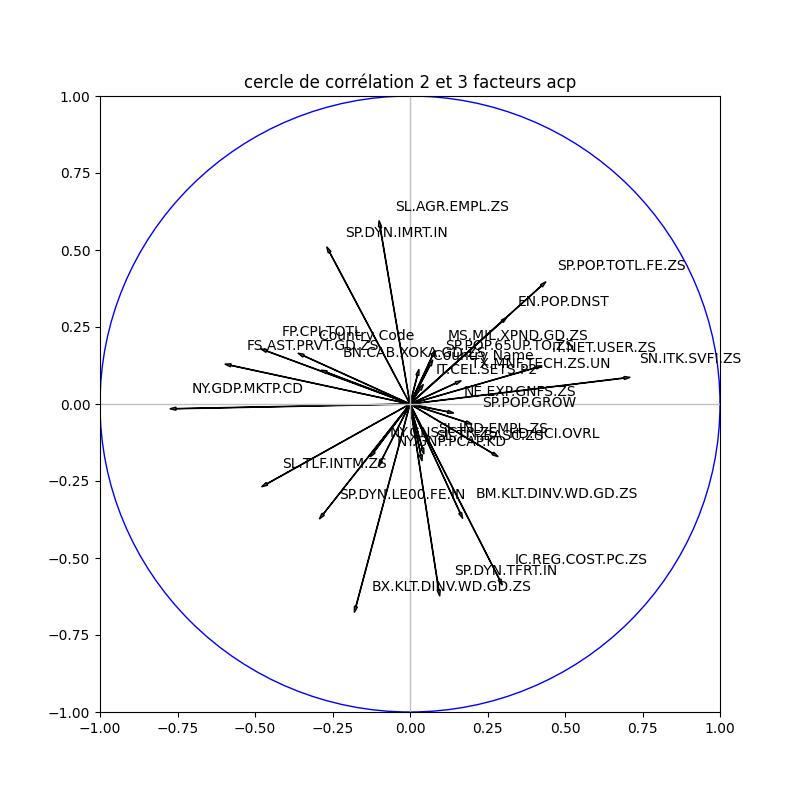


5. Que représentent les 4 premiers facteurs de l'ACP ? Vous pourrez répondre à cette question

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.





Les variables fortement corrélées aux premier facteur de l’ACP:

**ACP 0 :**

- SL.IND.EMPL.ZS (0.8068) : Pourcentage d'emploi dans l'industrie.

- BN.CAB.XOKA.GD.ZS (0.7079) : Solde de la balance des transactions courantes en pourcentage du PIB.

- BM.KLT.DINV.WD.GD.ZS (-0.6829) : Investissements directs étrangers nets en pourcentage du PIB.

- SL.AGR.EMPL.ZS (-0.4627) : Pourcentage d'emploi dans l'agriculture.

**ACP 1 :**

- SP.POP.65UP.TO.ZS (0.8702) : Pourcentage de la population âgée de 65 ans et plus.

- HD.HCI.OVRL (0.5573) : Indice global du capital humain.

- SL.AGR.EMPL.ZS (0.2515) : Pourcentage d'emploi dans l'agriculture.

- SL.TLF.INTM.ZS (-0.0181) : Pourcentage de la main-d'œuvre intermédiaire par rapport à la main-d'œuvre totale.

**ACP 2 :**

SN.ITK.SVFI.ZS (0.6906) : Indice de la sécurité financière des personnes âgées.

IT.NET.USER.ZS (0.4073) : Utilisateurs d'Internet (% de la population).

NY.GDP.MKTP.CD (-0.7549) : PIB (valeur totale en dollars US).

**ACP 3 :**

SP.DYN.TFRT.IN (-0.6042) : Taux de fécondité.

MS.MIL.XPND.GD.ZS (0.1568) : Dépenses militaires (% du PIB).

SP.DYN.IMRT.IN (0.4928) : Taux de mortalité infantile.

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.

Groupe de pays pour ACP 0 :

IND (Inde)

IDN (Indonésie)

SAU (Arabie Saoudite)

TUR (Turquie)

MEX (Mexique)

ARG (Argentine)

BGR (Bulgarie)

ZAF (Afrique du Sud)

Groupe de pays pour ACP 1 :

LUX (Luxembourg)

CYP (Chypre)

MLT (Malte)

ESP (Espagne)

GRC (Grèce)

EST (Estonie)

Groupe de pays pour ACP 2 :

FRA (France)

GBR (Royaume-Uni)

DEU (Allemagne)

ITA (Italie)

POL (Pologne)

SVN (Slovénie)

AUT (Autriche)

Groupe de pays pour ACP 3 :

AUS (Australie)

KOR (Corée du Sud)

JPN (Japon)

CAN (Canada)

FIN (Finlande)

SWE (Suède)

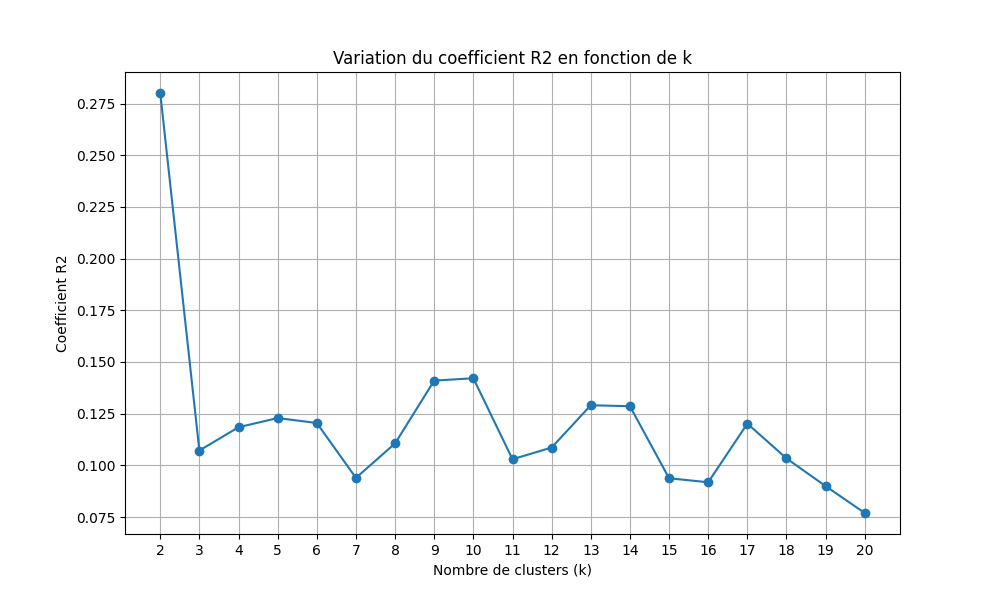
DNK (Danemark)

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.

France :

- Elle appartient à un groupe avec un indice moyen de la facilité de faire des affaires (IC.REG.COST.PC.ZS) de 5,71, ce qui est relativement faible.

- Le ratio du solde de la balance courante (% du PIB) (BN.CAB.XOKA.GD.ZS) est de 0,71, indiquant un équilibre ou un léger surplus.

- Le groupe a un pourcentage moyen d'emploi dans l'agriculture (SL.AGR.EMPL.ZS) de 3,89% et un pourcentage moyen d'emploi dans l'industrie (SL.IND.EMPL.ZS) de 22,22%.

- En termes de mortalité infantile (SP.DYN.IMRT.IN), elle a une moyenne de 3,54 décès pour 1000 naissances vivantes.

Mexique :

- Le Mexique se trouve dans un groupe avec un revenu national brut (NY.GNP.PCAP.KD) par habitant de 39655,76 dollars, en moyenne.

- Le pourcentage moyen de la population utilisant Internet (IT.NET.USER.ZS) est de 45,26%, ce qui peut indiquer un accès relativement bon à Internet dans ce groupe.

- Le Mexique a un pourcentage moyen de personnes âgées de plus de 65 ans (SP.POP.65UP.TO.ZS) de 7,94%, ce qui peut indiquer une population relativement jeune par rapport à d'autres groupes.

Bulgarie :

- La Bulgarie est dans un groupe avec un faible pourcentage moyen de chômage (SL.UEM.TOTL.ZS) de 7,94%.

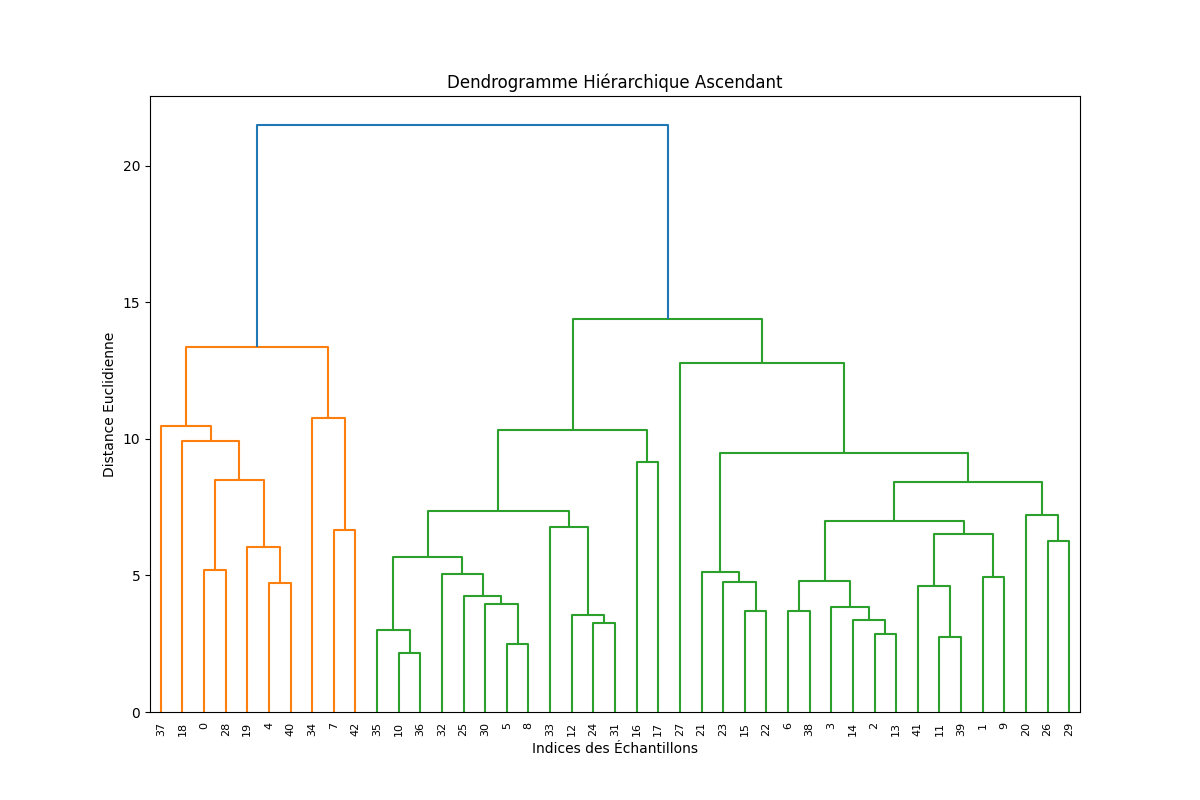
- Le taux de croissance de la population (SP.POP.GROW) est de 0,3%, ce qui peut indiquer une croissance démographique relativement faible.

- La Bulgarie a un pourcentage moyen de taxes sur le PIB (GC.TAX.TOTL.GD.ZS) de 50,90%, indiquant un niveau relativement élevé de prélèvements fiscaux.

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 ?

Les attributs sont des noms de produits.

Que représentent les valeurs présentes dans le fichier ?

Dans le fichier fourni, chaque ligne représente une transaction d'achat. Les valeurs présentes dans chaque ligne sont les noms des produits achetés dans cette transaction. Ces valeurs indiquent quels produits ont été achetés lors de chaque transaction. Si un produit est acheté dans une transaction, son nom est présent dans la liste de produits correspondante à cette transaction. Si un produit n'est pas acheté, la cellule reste généralement vide.

Comment agit les commandes du script fourni, permettant la discrétisation ?

Les commandes du script fournies permettent la discrétisation des données en utilisant la méthode de one-hot encoding. La bibliothèque mlxtend est utilisée pour effectuer cette opération. Plus précisément, la classe TransactionEncoder de mlxtend.preprocessing est utilisée pour convertir les données transactionnelles en un format adapté pour l'application d'algorithmes d'association tels que l'algorithme Apriori. Cette conversion transforme les données en une matrice binaire où chaque colonne représente un item possible et chaque ligne représente une transaction. Si un item est présent dans une transaction, la valeur correspondante dans la matrice est définie à 1 ; sinon, elle est définie à 0.

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

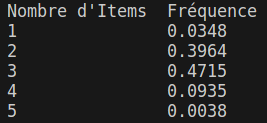
fréquents à l'aide de la fonction apriori() de mlxtend en fixant un support minimal à 0,002.

Expliquez ce que signifie cette valeur de 0,002. Combien d'itemsets obtenez-vous ?

Dans le contexte donné, fixer un support minimal à 0,002 signifie que nous recherchons des itemsets fréquents qui apparaissent dans au moins 0,2% des transactions.

J’obtiens 4223 d'itemsets.

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

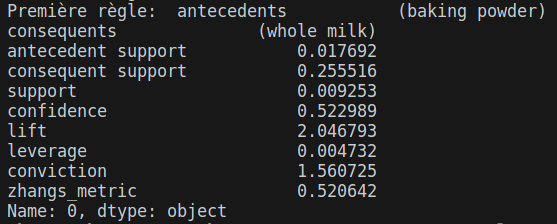


3. Construisez les règles à partir de l'ensemble des itemsets précédemment obtenus en recourant à la fonction association\_rules() de mlxtend. Pour ne pas avoir un nombre trop important de règles générées, fixez un seuil minimal de 0,5 pour la valeur de confiance.

Combien de règles obtenez-vous ?



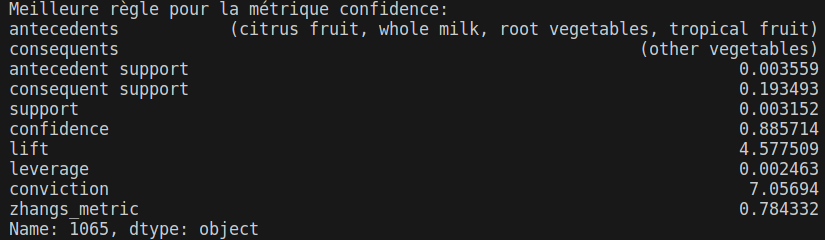
Commentez la première règle obtenue.

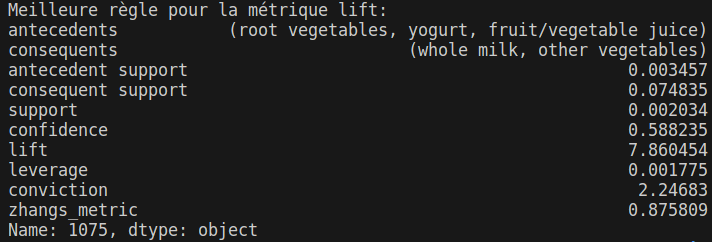


Cette règle indique une association assez forte entre "baking powder" et "whole milk", avec une confiance de 52.3%. Cela signifie que lorsqu'un client achète "baking powder", il y a une probabilité assez élevée (environ 52.3%) qu'il achète également "whole milk". Le lift supérieur à 1 suggère une association positive significative entre les deux items, ce qui indique que les clients qui achètent "baking powder" ont une probabilité plus élevée que la moyenne d'acheter également "whole milk".

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

confidence et lift. Quelle est la meilleure règle pour l'ensemble grocery pour chacune des métriques confidence et lift ?





Commentez les scores et règles obtenus.

**Pour la métrique de confidence :**

La meilleure règle a les antécédents (citrus fruit, whole milk, root vegetables, tropical fruit) et les conséquents (other vegetables). Voici une analyse détaillée des scores et de la règle :

- Support : Le support de cette règle est de 0.003152, ce qui indique qu'elle s'applique à environ 0.315% des transactions dans l'ensemble de données.

- Confidence : La confiance de la règle est très élevée, à 0.885714, ce qui signifie que 88.6% des transactions contenant tous les antécédents contiennent également les conséquents. Cela indique une forte association entre les antécédents et les conséquents.

- Lift : Le lift est également élevé, à 4.577509, ce qui suggère une association significative entre les antécédents et les conséquents. Un lift supérieur à 1 indique une association positive entre les items.

- Interprétation : Cette règle montre qu'il y a une forte probabilité (88.6%) que les clients qui achètent citrus fruit, whole milk, root vegetables et tropical fruit achètent également other vegetables. Le lift élevé (4.577509) confirme cette forte association entre ces items.

**Pour la métrique de lift :**

La meilleure règle a les antécédents (root vegetables, yogurt, fruit/vegetable juice) et les conséquents (whole milk, other vegetables). Voici une analyse détaillée des scores et de la règle :

- Support : Le support de cette règle est de 0.002034, ce qui signifie qu'elle s'applique à environ 0.203% des transactions dans l'ensemble de données.

- Confidence : La confiance de la règle est de 0.588235, ce qui indique que 58.8% des transactions contenant les antécédents contiennent également les conséquents. Cela indique une association modérée entre les antécédents et les conséquents.

- Lift : Le lift est élevé, à 7.860454, ce qui suggère une association significative entre les antécédents et les conséquents. Un lift élevé indique une forte association entre les items.

- Interprétation : Cette règle montre qu'il y a une probabilité relativement élevée (58.8%) que les clients qui achètent root vegetables, yogurt et fruit/vegetable juice achètent également whole milk et other vegetables. Le lift élevé (7.860454) indique une forte association entre ces items, ce qui suggère que les clients qui achètent ces items ont une probabilité plus élevée que la moyenne d'acheter également whole milk et other vegetables.

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 ?



Le lait entier est le produit le plus acheté avec le yaourt et le café.

### 2 Analyse prédictive

**2.1 Classement**

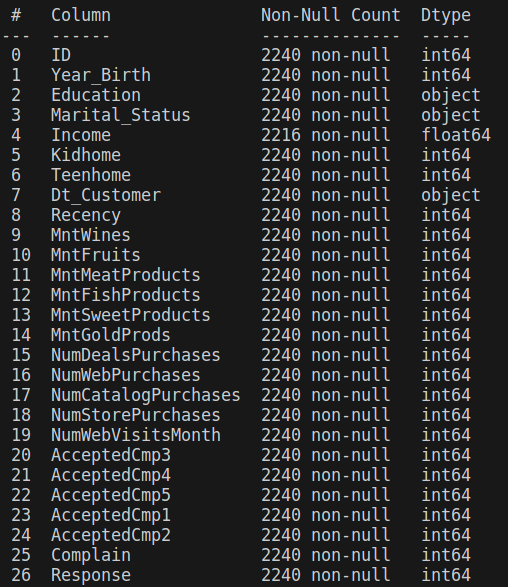
1. Ouvrez les données à l'aide de la bibliothèque pandas. Combien y a-t-il d'attributs dans ce

fichier ?



Il y a 27 attributs pour le fichier marketing\_campaign.

Quels sont leurs types ?

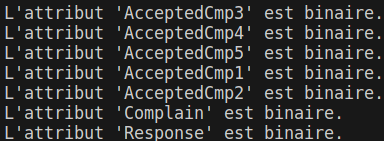


Il y a 23 attributs de type entier (int64), 3 attributs de type object et 1 attribut de type nombre flottant (float64).

Quelle est la variable à prédire ?

La variable à prédire dans ce contexte est Response. Cet attribut représente si le client a accepté l'offre dans la dernière campagne promotionnelle ou non.

Quels attributs, indiqués comme numériques, prennent deux valeurs et sont donc en réalité binaires ?



Les attributs binaire sont :

AcceptedCmp3

AcceptedCmp4

AcceptedCmp5

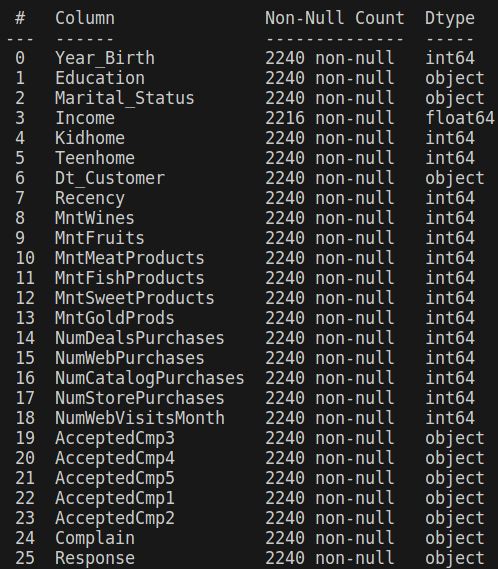
AcceptedCmp1

AcceptedCmp2

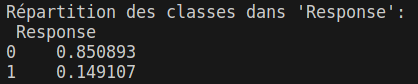
Complain

Response

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.



1. Les classes sont-elles équilibrées dans les fichiers ?



Nous pouvons voir que la classe 0 est beaucoup plus fréquente que la classe 1. Cette répartition indique un déséquilibre des classes dans la variable 'Response'. En effet, la classe 0 représente environ 85% des données, tandis que la classe 1 représente seulement environ 15% des données.

Y a-t-il des valeurs manquantes dans les données ?



Comment ces valeurs sont-elles représentées dans le ficher ?

Par ‘,,’ qui indique qu’il n’y a pas de valeurs sur la case.

Quels sont les attributs concernés par des valeurs manquantes ?

Il y a 24 valeurs manquantes pour l’attribut Income.

1. Expliquez en une phrase en quoi consiste la méthode élémentaire de classement DummyClassifier utilisant la stratégie most\_frequent.

La méthode élémentaire de classement DummyClassifier utilisant la stratégie "most\_frequent" consiste à prédire toujours la classe la plus fréquente dans l'ensemble de données, ignorant ainsi les caractéristiques des variables d'entrée.

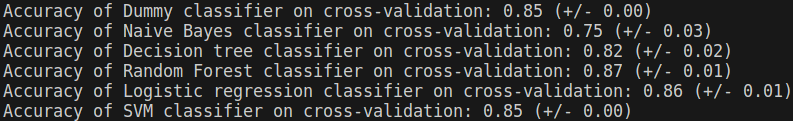
Dans quel cas ce classifieur pourrait donner les meilleurs résultats parmi ceux testés ?

Le classifieur DummyClassifier avec la stratégie "most\_frequent" pourrait donner les meilleurs résultats lorsque les classes sont déséquilibrées et que la classe majoritaire est très prédominante dans l'ensemble de données.

1. 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 ?



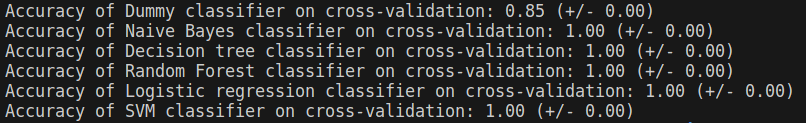
La méthode Random Forest classifier est la plus performante par rapport au taux de classification, avec une précision moyenne de 0.87 et un écart type de 0.01 lors de la validation croisée en 5 blocs. Cela signifie que le modèle Random Forest a produit des prédictions plus précises par rapport aux autres méthodes testées dans votre ensemble de données.

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

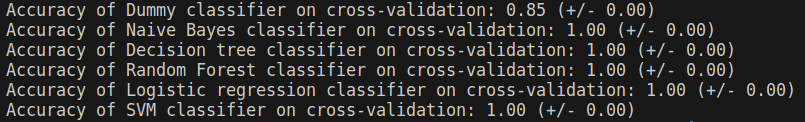
la question précédente ?



Les performances sont bien meilleure qu’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.



Le résultat est le même qu’à la question précédente

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

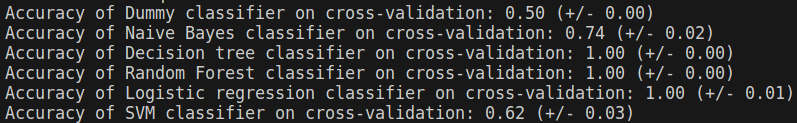
classification (accuracy ). Pourquoi cette métrique ne permet-elle pas d'évaluer correctement

les méthodes pour des classes déséquilibréesire.

Le taux de classification (accuracy) mesure simplement la proportion de prédictions correctes parmi toutes les prédictions effectuées. Cependant, lorsque les classes sont déséquilibrées, c'est-à-dire lorsqu'il y a une grande différence dans le nombre d'instances de chaque classe, l'accuracy peut être trompeuse.

Prenons un exemple où 90 % des instances appartiennent à la classe 1 et seulement 10 % à la classe 0. Un classificateur pourrait simplement prédire toutes les instances comme appartenant à la classe 1 et obtenir une accuracy de 90 %. Cependant, ce classificateur n'a aucune capacité à distinguer les instances de la classe 0, ce qui le rend inefficace dans la plupart des situations réelles.

Construisez un nouveau jeu de données équilibré 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.



Les taux de classification ont baissé par rapport à la question précédente, ce qui montre bien que lorsque les classes sont déséquilibrées, l'accuracy peut être trompeuse.