PROJECT 4

Questions :

* Variables catégorielles dans le CE1 ? On parle de la catégorie RFM ?
* Devrais-je utiliser le log pour avoir qqc de plus gaussien ?
* La plupart des clients n’ont qu’un seul produit, c’est un problème pendant la division du nombre de produits (frequency) par quartile. Ce que je fais n’a pas l’air terrible, il faudrait que je fasse un cluster séparé avec nb\_produits > 1 ?
* Comment afficher mes données correctement ? Quelles variables choisir ? Devrais-je reduire mes features ? Ce que j’ai n’est pas parfait.
* Comment comparer les parametres des DBScan/optics et hierarchical clusterings ? Comparaisons manuelles des clusters ? CE4. Je pourrais utiliser le RFM\_Score as true label et utiliser ARI ?
* Should I display the reachability\_plot for the OPTICS clustering ?
* Forme et stabilite des clusters ? CE 2-3
* Il serait intéressant également de regarder l’évolution de la la distribution de vos features numériques dans le temps, en utilisant des méthodes graphiques ou un test de Kolmogorov-Smirnov ?

##### Définir la stratégie d’élaboration d’un modèle d’apprentissage non supervisé, sélectionner et entraîner des modèles adaptés à une problématique métier afin de réaliser une segmentation ou une réduction de données.

* CE1 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder).
* CE2 Vous avez créé de nouvelles variables à partir de variables existantes.
* CE3 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.
* CE4 Vous avez normalisé les variables lorsque c'est requis.
* CE5 Vous avez défini la stratégie d’élaboration d’un modèle pour répondre à un besoin métier.
* CE6 Vous avez proposé un nombre de segments et une répartition adaptés à la problématique métier.
* CE7 Vous avez explicité la stratégie d'ajout de nouveaux clients.
* CE8 Vous avez pris en compte la nature des variables d'entrée dans le choix des algorithmes et de la distance.
* CE9 Vous avez testé et comparé plusieurs algorithmes.

##### Évaluer les performances des modèles d’apprentissage non supervisé selon différents critères (scores, temps d'entraînement, etc.) en adaptant les paramètres afin de choisir le modèle le plus performant pour la problématique métier.

* CE1 Vous avez choisi des métriques adaptées pour évaluer les valeurs candidates pour le nombre de segments (selon algorithme, par exemple : coefficient silhouette / elbow pour le KMeans, eps/min samples pour le DBScan).
* CE2 Vous avez évalué la forme des clusters.
* CE3 Vous avez évalué la stabilité des clusters à l’initialisation.
* CE4 Vous avez optimisé les hyper-paramètres pertinents dans les différents algorithmes.
* CE5 Vous avez justifié le choix final d'algorithme et d'hyper-paramètres.
* CE6 Vous avez analysé la stabilité du modèle dans le temps afin d’en assurer sa maintenance
* CE7 La convention PEP8 est respectée. Le code est commenté (commentaires réguliers, docstrings dans les fonctions) .