***Clustering***

***Exercice 1 : Classement des clients***

* *Lire les données à partir du fichier marketing.csv.*
* *Quel est la taille du DataFrame, de quel type sont les features.*
* *Combien y-a-t-il de valeurs manquantes dans les différentes features. Supprimez les samples ayant des valeurs manquantes.*
* *La colonne Duration correspond à la date d’enregistrement du client dans la base. Nous allons transformer ces données en une différence en nombre de jour avec le client le plus ancien. Dans un premier temps vous transformez la feature Duration en objet de type date grâce à la fonction to\_datetime de pandas, puis vous récupérerez la data minimum.*
* *Vous pourrez alors utiliser la fonction apply avec une lambda expression pour transformer les dates de la feature Duration en différence. Pour cela vous devez effectuer la différence avec la data minimum et récupérer grâce à l’attribut days le nombre de jours.*
* *Combien y-a-t-il de valeurs différentes dans la feature Education. Utilisez ensuite un LabelEncoder de la bibliothèque preprocessing de sklearn pour coder les données. Vous pourrez utiliser la fonction fit\_transform pour effectuer l’opération.*
* *Remplacez les valeurs de la feature Status par les entiers 1 et 2. Vous utiliserez pour cela la fonction apply. Cela permettra de compter le nombre de personnes impliquées dans l’enregistrement.*
* *On souhaite maintenant ajouter quelques informations qui pourraient être utiles lors de la classification des clients.*
* *Ajouter une feature Children qui compte le nombre d’enfants de la famille, cette feature correspond à la somme des petits kid et des ados Teen.*
* *Ajouter également un feature Family qui compte le nombre de personnes de la famille. Cette feature regroupe les données de Status et de Children.*
* *Ajouter une feature is\_parent qui indique s’il y a au moins un enfant dans la famille (valeur 1 ou 0).*
* *Ajouter une feature Total qui regroupe toutes les dépenses vin, fruits, viande, poison, bonbons, or.*
* *Effectuer une analyse statistique sur les données. Que constatez-vous.*
* *Utilisez maintenant la fonction scatterplot de seaborn pour afficher les relations entre le revenu et l’age en fonction du la taille de la famille.*
* *Identifiez à quel niveau se trouvent les outliers dans ces deux features et supprimez-les.*
* *Affichez la matrice des corrélations en utilisant la fonction heatmap de seaborn. Que constatez-vous.*
* *Utiliser maintenant un modèle StandardScaler pour normaliser les données.*
* *Avec l’algorithme du PCA affichez la courbe d’évolution du taux d’explicabilité en fonction du nombre de dimensions. Combien faudrait-il retenir de dimensions si l’on souhaite pourvoir conserver 94% de l’information.*
* *Créez le nouveau jeu de données (Data\_acp) composé des données standardisées et avec les dimensions réduites sur 94% de l’information.*
* *Effectuer maintenant dix clustering sur Data\_acp en utilisant le model KMeans de sklearn avec le nombre de cluster variant de 1 à 10, et affichez l’inertie. Quel est le nombre de cluster qu’il faudrait retenir.*
* *Utilisez le modèle* *AgglomerativeClustering pour classer les individus de Data\_acp.*
* *Enregistrez les résultats de la classification dans une feature Clusters dans le Data d’origine.*
* *Effectuez un affichage 3d à partir des données de Data\_acp. Pour cela vous commencerez par récupérerez (Dt) seulement les samples qui ont des revenus Income < 100000. Vous utiliserez pour les variables x, y et z de la fonction plot les features de Income, Total et Family de Dt, vous y ajouterez une couleur c égal à Clusters. Que constatez-vous. P.S. Pour le code d’un affichage 3d vous pourrez trouver des exemples sur le net.*
* *Affichez maintenant comment se répartissent les classes grâce à la fonction countplot de seaborn.*
* *Grâce aux fonctions groupeby et mean affichez les valeurs moyennes des samples en fonction des différents clusters. Vous ne ferez ces moyennes que sur les features : Income, Children, Family, Total, is\_parent. Que peut-on dire sur le Clusters.*
* *On souhaite maintenant regarder plus en détail l’influence de ces paramètres sur les différents clusters.*
* *Affichez le nombre de données dans chaque cluster, vous pouvez utiliser la fonction countplot de seaborn. Que voit-on.*
* *Afficher la répartition des revenus dans chaque cluster. Vous utiliserez pour cela la fonction histplot avec un bins de 30 et un affichage du kde. Comment peut-on caractérisé les clusters en fonction du revenu.*
* *Comment pour chaque cluster se répartissent les dépenses totales Total en fonction du revenu Income. Vous pourrez utiliser la fonction scatterplot de seaborn. Comment peut-on affiner les caractéristiques des clusters si l’on prend en compte la somme totale dépensée.*
* *Comment la variable Family impacte la classification. Vous utiliserez ici la fonction countplot.*
* *Résumez comment sont organisé les différents clusters.*

***Exercice 2 : Classification des champignons***

* *Lire les données sur les champignons dans le fichier Champignons.csv.*
* *Combiens y-a-t-il des données dans le DataFrame.*
* *Y-a-t-il des données manquantes. Dans ce cas supprimés ces données.*
* *Affichez les pourcentages de champignons toxiques et comestibles.*
* *Effectuez une copie de la base d’origine.*
* *Nous voulons encoder les données catégorielles autres que la target Classe. Pour cela vous utiliserez l’encodeur LabelEncoder de la bibliothèque preprocessing de sklearn. Pour cela vous devez donc créer un modèle de type LabelEncoder et ensuite l’entrainer (fit) sur une colonne et transformer les données (transform) sur cette même colonne. La fonction fit\_transform effectue l’opération en un coup.*
* *Vous devrez donc parcourir toutes les colonnes du DataFrame et les transformer. Attention vous ne devez pas transformer la target.*
* *Découpez maintenant le DataFrame en un jeu de teste et un jeu d’entrainement avec une découpe de 0.33.*
* *Cette découpe respecte-t-elle la répartition en pourcentage entre les individus toxiques et comestibles. Vous afficherez donc ces pourcentages pour les y test et train.*
* *Utilisez un modèle de support vector machine pour entrainer et estimer la toxicité des champignons. Quels sont les résultats obtenus.*
* *Grâce à la matrice de confusion affichez les différences entre les valeurs prédites et celles attendues.*
* *Recherchez quels sont les individus qui ont été mal classés et affichez-les.*