Dans le cadre du projet final, j'ai écrit un code Python pour analyser une banque de données appelée Donnees_libertes.csv (fausses données dedans donc aucune conclusion réelle possible), que j'ai importé sous forme de pandas DataFrame. Ce DataFrame contient divers indices et indicateurs qui pourraient être liés au bonheur des individus. Voici une brève explication de mon code et des hypothèses que j'ai émises en examinant les relations entre les différentes variables.

1. J'ai commencé par vérifier et installer les bibliothèques requises à partir d'un fichier **requirements.txt**. Ensuite, j'ai importé les bibliothèques nécessaires et mes fonctions personnalisées.
2. J'ai chargé les données à partir d'un fichier CSV et les ai stockées dans une base de données SQLite pour faciliter leur manipulation.
3. Pour mieux comprendre les données, j'ai effectué une analyse descriptive de la colonne **Indice_liberte_expression**. J'ai calculé des statistiques et tracé un histogramme pour visualiser la distribution des données.

D'après les résultats du test de normalité de Shapiro-Wilk et les autres statistiques descriptives concernant la colonne Indice_liberte_expression, voici mon interprétation :

- Le test de normalité de Shapiro-Wilk a donné une valeur de W = 0.9543 et une p-value très faible (0.0000). Étant donné que la p-value est inférieure à un seuil communément accepté de 0.05, on rejette l'hypothèse nulle selon laquelle les données suivent une distribution normale. En d'autres termes, les données de la colonne Indice_liberte_expression ne semblent pas suivre une distribution normale.
- La moyenne de la colonne Indice_liberte_expression est de 0.5032. Cela indique que, en moyenne, les valeurs de cette colonne sont légèrement supérieures à 0.5.
- L'écart-type de la colonne Indice_liberte_expression est de 0.2892, ce qui montre une certaine dispersion des données autour de la moyenne. Cela signifie que les valeurs de cette colonne varient dans une plage relativement large.
- La médiane de la colonne Indice_liberte_expression est de 0.5051, ce qui est très proche de la moyenne. Cela suggère que les données sont assez symétriques, bien qu'elles ne suivent pas une distribution normale.
- La plage interquartile (IQR) pour la colonne Indice_liberte_expression est de 0.5020, ce qui indique la dispersion des données entre le premier et le troisième quartile.

En résumé, les données de la colonne Indice_liberte_expression ne semblent pas suivre une distribution normale, avec une moyenne légèrement supérieure à 0.5 et une dispersion modérée. Les valeurs de cette colonne varient dans une plage relativement large, ce qui indique une diversité dans l'indice de liberté d'expression parmi les individus ou les groupes étudiés.

4. J'ai utilisé un algorithme de tri par insertion pour trier la colonne **Indice_moyen** et enregistrer la liste triée dans un fichier texte.
5. Pour examiner les relations entre les différentes variables et la colonne cible **Heureux_ou_non**, j'ai effectué une classification KNN et tracé la précision en fonction du nombre de voisins. J'ai également utilisé la méthode des machines à vecteurs de support (SVM) pour explorer davantage les relations entre les variables.

Les résultats du SVM montrent que la précision moyenne pour la classification de la colonne cible Heureux_ou_non est de 0.98 avec un écart-type de 0.01. Cela signifie que le modèle SVM est capable de prédire correctement le bonheur (ou non) des individus dans environ 98% des cas, ce qui indique une performance très élevée. Il est important de noter que la performance d'un modèle sur des données d'apprentissage ne garantit pas nécessairement une performance similaire sur de nouvelles données, il est donc essentiel de valider le modèle sur un ensemble de données de test indépendant pour évaluer sa capacité à généraliser.
Le nombre de vecteurs de support pour chaque classe est [506, 490], ce qui montre que le modèle SVM a utilisé 506 vecteurs de support pour la classe "heureux" et 490 vecteurs de support pour la classe "non heureux". Ces vecteurs de support sont les points de données les plus influents pour déterminer la frontière de décision entre les deux classes.
En conclusion, il semble que les variables présentes dans le DataFrame ont une capacité significative à prédire si un individu est heureux ou non, comme en témoigne la haute précision du modèle SVM.
Le modèle KNN avait lui une précision plus faible de 0.91 concernant la prédiction de Heureux_ou_non.

6. J'ai appliqué une analyse en composantes principales (ACP) pour réduire la dimensionnalité des données et visualiser l'importance relative des différentes variables.
7. J’ai créé deux fonctions anonymes pour calculer la moyenne des indices de liberté et l'indice de bien-être pour chaque individu. J'ai ensuite ajouté ces nouvelles variables au DataFrame et les ai enregistrées dans un fichier CSV.
8. Enfin, j'ai appliqué un codage récursif aux données et j’ai enregistré le data frame avec le nouveau code.