README

Brief Description:

The **Analysis** directory contains the following files and folders:

- **1) Analysis for Optimal Number Of Clusters** : This folder contains the following graphs which help us determine the optimal number of clusters for the Liver Disorders Dataset:
  - **Calinski_Harabasz_Index_versus_Number_Of_Clusters.png** : This graph plots the **Calinski Harabasz Index (CH Index)** versus the number of clusters. We vary the number of clusters for the KMeans Clustering model and then compute the CH index to measure its performance. CH index does not use ground truth labels and thus is a good metric for measuring the clustering.
  - **Silhouette_Index_versus_Number_Of_Clusters.png** : This graph plots the **Silhouette Index** versus the number of clusters. We vary the number of clusters for the KMeans Clustering model and then compute the Silhouette index to measure its performance. Silhouette index does not use ground truth labels and thus is a good metric for measuring the clustering.
  - **Wangs_Method_Of_Cross_Validation.png** : This graph plots the average number of disagreements obtained from **Wangs_Method_Of_Cross_Validation** versus the number of clusters. We vary the number of clusters for the KMeans Clustering model and then compute the average number of disagreements to measure its performance. The lesser the number of disagreements the better the clustering.
- **2) Test_A**: This folder contains the following graphs obtained after performing the Test_A as described in the assignment:
  - **Test_A_Adjusted_Rand_Score_Analysis.png** : This graph plots the **Adjusted Rand Index (ARI)** versus the Iteration Number. This graph shows the stability of the ARI upon repeatedly performing KMeans Clustering using k random initial centres.
  - **Test_A_Fowlkes_Mallows_Score_Analysis.png** : This graph plots the **Fowlkes Mallows Index (FM Index)** versus the Iteration Number. This graph shows the stability of the FM Index upon repeatedly performing KMeans Clustering using k random initial centres.
  - **Test_A_Homogeneity_Score_Analysis.png** : This graph plots the **Homogeneity Index** versus the Iteration Number. This graph shows the stability of the Homogeneity Index upon repeatedly performing KMeans Clustering using k random initial centres.
  - **Test_A_Normalized_Mutual_Info_Score_Analysis.png** : This graph plots the **Normalized Mutual Info Index (NMI Index)** versus the Iteration Number. This graph shows the stability of the NMI Index upon repeatedly performing KMeans Clustering using k random initial centres.
- **3) Improved_Test_A**: This folder contains the following graphs obtained after performing the Improved_Test_A. In Improved_Test_A instead of using k random initial centres, we use the **kmeans++ heuristic** to initialize the centres:
  - **Improved_Test_A_Adjusted_Rand_Score_Analysis.png** : This graph plots the **Adjusted Rand Index (ARI)** versus the Iteration Number. This graph shows

the stability of the ARI upon repeatedly performing KMeans Clustering using k random initial centres.
    - **Improved_Test_A_Fowlkes_Mallows_Score_Analysis.png** : This graph plots the **Fowlkes Mallows Index (FM Index)** versus the Iteration Number. This graph shows the stability of the FM Index upon repeatedly performing KMeans Clustering using k random initial centres.
    - **Improved_Test_A_Homogeneity_Score_Analysis.png** : This graph plots the **Homogeneity Index** versus the Iteration Number. This graph shows the stability of the Homogeneity Index upon repeatedly performing KMeans Clustering using k random initial centres.
    - **Improved_Test_A_Normalized_Mutual_Info_Score_Analysis.png** : This graph plots the **Normalized Mutual Info Index (NMI Index)** versus the Iteration Number. This graph shows the stability of the NMI Index upon repeatedly performing KMeans Clustering using k random initial centres.
- **Sample_Run.txt**: Contains output of the sample run of our program.
- **Report.pdf**: Contains a detailed analysis report of our KMeans Clustering Model, different initialization methods, different performance metrics, finding the optimal number of clusters for the Liver Disorders Dataset, perfoming Test A and Improved Test A and a section of extra analysis that we have performed to have better understanding of our models.

The **Project** directory contains the following files and folders:

- **Analysis** : Analysis is a folder which gets populated when analysis.py is run
- **bupa.DATA**:Contains the data used for training the K Means Clustering Model. The data is comprised of 345 samples with 6 continuous valued attributes (of which 1 is used to calculate the target label) and 1 selector attribute (which is discarded)
- **requirements.txt**: Contains all the necessary dependencies and their versions
- **utility.py**: Contains all the helper functions such as wangs_method_cross_validation etc. (if any)
- **models.py** : Contains the implementation of the KMeans class
- **analysis.py**: Contains the python code we implemented to perform analysis on the Data and KMeans Clustering Model.

Directions to use the code

---

1. Download the **Project** directory in into your local machine.
2. Ensure all the necessary dependencies with required version and latest version of Python3 are available (verify with requirements.txt)

    ```
    pip3 install -r requirements.txt
    ```

3. Run the analysis.py file.
    - Initially, it will output some information regarding the project such as author etc.
    - Next, it will **prompt you to enter the value of k**, which is the number of clusters to be formed when the KMeans model is constructed. The value of k has to be greater than 1. Please enter an integer value (k>1).
    - The next **prompt will query you for the mehtod of initialization of the initial cluster centres**. We have implemented two methods, one is called **lloyd** and the other is called **kmeans++**. Please enter either **'lloyd'** or

'kmeans++'. Note, that the method of initilization entered here will be used throughout the analysis.

- **lloyd cluster centre initialization** : The initial cluster centres are k randomly selected data points from the dataset.
- **kmeans++ cluster centre initialization** : The initial cluster centres are selected according to the kmeans++ heuristic.

- Now, a model is built with the above user given k and user given menthod of initialization, and then the performance of the model is measured with and without ground truth labels using various metrics and the same is given as output. (More details in the Report.pdf)
- After that, analysis is performed to determine the optimal number of clusters. In this step several graphs are created and saved in appropriate folders and the user is **prompted to enter the value of c, the number of permutations for which Wang's Mehtod of Cross Validation is to be run**. (We suggest the value of 20)
- After that, the Test_A and Improved_Test_A are performed, whose graphs are again created and saved in appropriate folders

Final Remarks

The dataset is taken from the following link.
https://archive.ics.uci.edu/ml/datasets/liver+disorders