SIG720 Task 1P

Solve the following set of problems using Python and submit the code file with extension .ipynb.

Part A: Data Preprocessing (Microclimate Data)

1. Load the "Microclimate sensors data" dataset and print the feature name with numbers of missing entries.
2. Fill in the missing entries. For filling any feature, you can use either the mean or median value of the feature values from observed entries. Explain the reason behind your choice and print replacement value of each feature.
3. Split the “LatLong” column into separate Latitude and Longitude columns using appropriate encoding.
4. Apply Min-Max scaling on the continuous features.
5. Plot the distribution of scaled features before and after scaling. Comment on any changes observed.

Part B: Clustering and PCA

6. Load the “Obesity” dataset and remove the class label "NObeyesdad".
7. Encode the categorical features using appropriate techniques.
8. Use the Silhouette Coefficient to determine the optimal number of clusters (k).
9. Apply KMeans and KMeans++ clustering algorithms using the optimal k. Compare and report the clustering performance.
10. Load the “gene expression” dataset and apply PCA to reduce the data to 3 principal components. Report the variance explained by these components.
11. Apply KMeans on the original features of the gene dataset and the first three components returned by PCA. Compare the results using the given labels.