Assignment 4

Problem Statement

To build an unsupervised learning model for male and female voices in Hindi

Data:

https://www.kaggle.com/datasets/vivmankar/hindi-speech-classification

Task:

1. Read train.csv file and train folder for audios
2. Select 2 samples for each female and male voices, plot zero crossing rate. Comment your inferences.
3. Extract zero crossing rate features for all audios.
4. Impute the null values with two options,
    a. zero impute
    b. KNN imputation. (https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html)
5. Apply following,
    a. PCA on zero impute and generate the clusters using KMeans.
    b. PCA on KNN imputation with KMeans, Agglomerative, DBSCAN, and mean shift.
    c. T-SNE on zero impute with KMeans.
6. Apply silhouette score metric on the labels. Comment on the metric results. (https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/)
7. Publish your notebook in that Kaggle dataset.