**Name - Jatin Mahawar**
**Roll.No. - 22CS30032**

# Machine Learning Assignment 3 Part B Report

# Clustering Process

- **Data Preprocessing**:
  - **Handling Missing Values**: No missing values were found in the dataset.
  - **Categorical Encoding**: The categorical columns (`Family`, `Genus`, `Species`) were encoded using Label Encoding to convert them into a numeric format.
  - **Data Standardization**: StandardScaler was used to normalize the MFCC features, ensuring all features contributed equally to the clustering process.
- **Dimensionality Reduction**:
  - **PCA (Principal Component Analysis)** was applied to reduce the dimensionality of the data. This step helped remove redundancy among the MFCC features, which had been identified as correlated in the correlation matrix.
- **K-Means Clustering**:
  - **Choosing Optimal K**: Using the Elbow Method and Silhouette Score, the optimal number of clusters was identified as **K = 2**. Silhouette's Score for **k=2** comes **0.33**.
  - **Clustering Execution**: K-Means was then applied with `n_clusters=2`, creating four clusters based on the MFCC feature space.
- **Evaluation Metrics**:
  - **Silhouette Score**: A score of **0.329** was obtained, indicating reasonably good clustering with compact and well-separated clusters.
  - **Davies-Bouldin Index**: The index was **1.308**, which is relatively low, suggesting that clusters are compact and distinct.
  - **Calinski-Harabasz Score**: A score of **3284.605** was achieved, which is relatively high and further indicates that clusters are well-separated and dense.
- **Visualization**:
  - Using PCA, the clusters were visualized in 2D space. The visualization showed two clusters, validating that K-Means could capture separable groupings in the MFCC feature space. This visualization also supported the choice of `K=2` based on the observed data distribution.

# Analysis of Clustering Evaluation Metrics

- **Silhouette Score**: A score of **0.329** suggests moderately well-separated clusters. Since the silhouette score ranges from -1 to 1, a score above 0.5 indicates that the clusters are distinct, though there may still be some overlap or ambiguity between them.
- **Davies-Bouldin Index**: The relatively low value of **1.308** indicates that the clusters have a high degree of separation and are compact, which is desirable in clustering.
- **Calinski-Harabasz Score**: The high score of **3284.605** reflects well-defined clusters with clear boundaries. This high value indicates that the within-cluster variance is low while the between-cluster variance is high, confirming that the clusters are meaningful and distinct.

## Limitations of K-Means and Alternative Clustering Techniques

- **K-Means Limitations**:
  - **Sensitivity to Initial Centroids**: The K-Means algorithm is sensitive to the starting points of the centroids, potentially leading to different cluster assignments on other runs. The notebook may have addressed this using multiple initializations (e.g., `k-means++`), but this still introduces variability.
  - **Spherical Cluster Assumption**: K-Means assumes that clusters are spherical and evenly sized, which may not be appropriate for frog acoustic data if the clusters have complex or elongated shapes. The assumption of Euclidean distance as the similarity metric may not fully capture the true structure of the frog species' sound profiles.
  - **Scalability**: K-Means is generally efficient but can only work with very large datasets if there are many dimensions or samples. PCA helped mitigate this, but it remains a limitation when dealing with complex datasets.
- **Alternative Clustering Methods**:
  - **Hierarchical Clustering**: This approach could explore nested groupings, potentially providing more insights into the hierarchical relationships between frog species. However, it may be computationally expensive for larger datasets.
  - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: DBSCAN does not assume spherical clusters and can identify clusters of arbitrary shapes. It may work well if frog species exhibit varied sound patterns, but it requires careful tuning of parameters like `epsilon` and `min_samples`.
  - **Gaussian Mixture Models (GMM)**: GMM is a probabilistic model that assumes data is generated from a mixture of Gaussian distributions. It can handle elliptical clusters better than K-Means. However, it is computationally more expensive and requires the assumption that clusters are Gaussian-distributed.