Clustering Analysis Report: Anuran Calls Dataset (MFCCs)

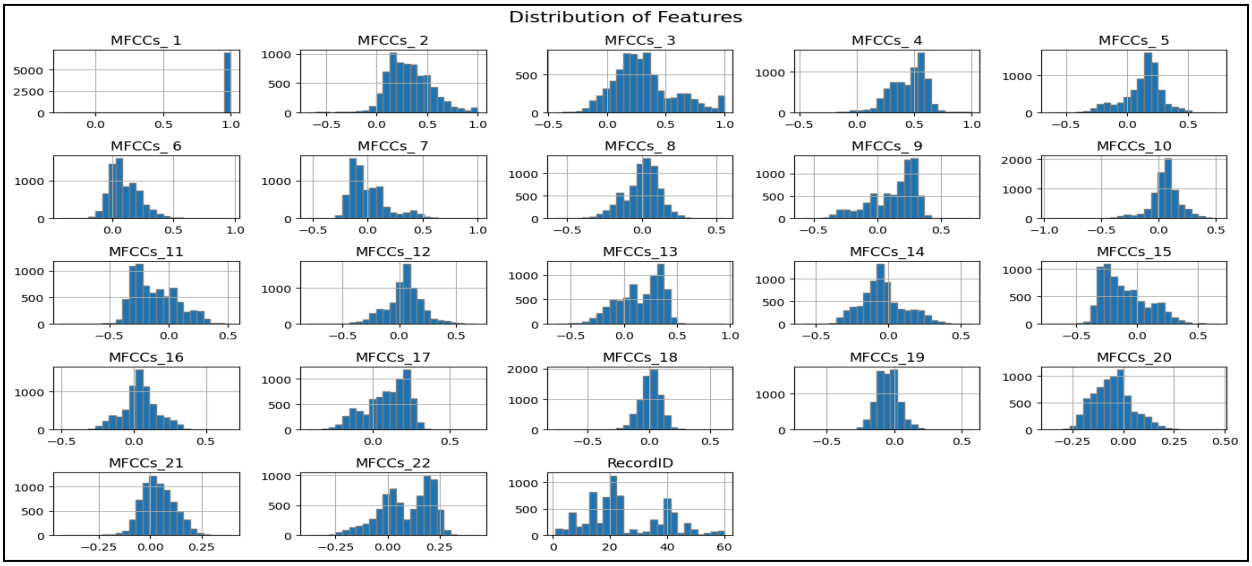
Overview

This report presents a comprehensive clustering analysis of the Anuran Calls Dataset, which contains Mel-Frequency Cepstral Coefficients (MFCCs) for frog call sounds. The primary objective is to group these frogs into clusters based on acoustic features. The clustering techniques employed include K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. The analysis also incorporates feature engineering, dimensionality reduction, and multiple evaluation metrics to assess clustering quality and performance.

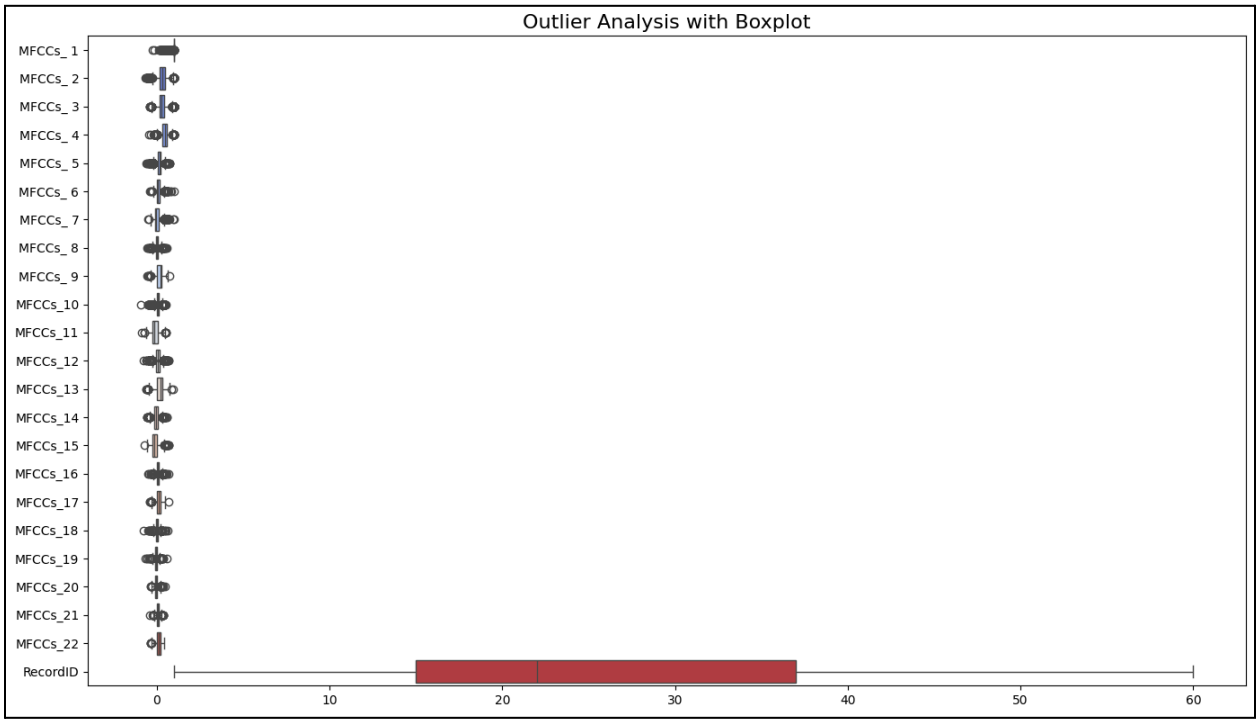
1. Data Preprocessing and Exploration

1.1 Exploratory Data Analysis (EDA)

- Dataset Overview: The dataset comprises **7,195** samples with 26 columns (22 MFCC features, 3 categorical labels, and 1 unique identifier).
- Missing Values: There were **no missing values** in the dataset.
- Feature Distribution: Each MFCC feature was visualized using histograms. Most features showed a roughly normal distribution, though some had long tails.



- Outliers: Box plots revealed the presence of potential outliers in multiple features, which might affect clustering results if left unaddressed.



1.2 Data Scaling

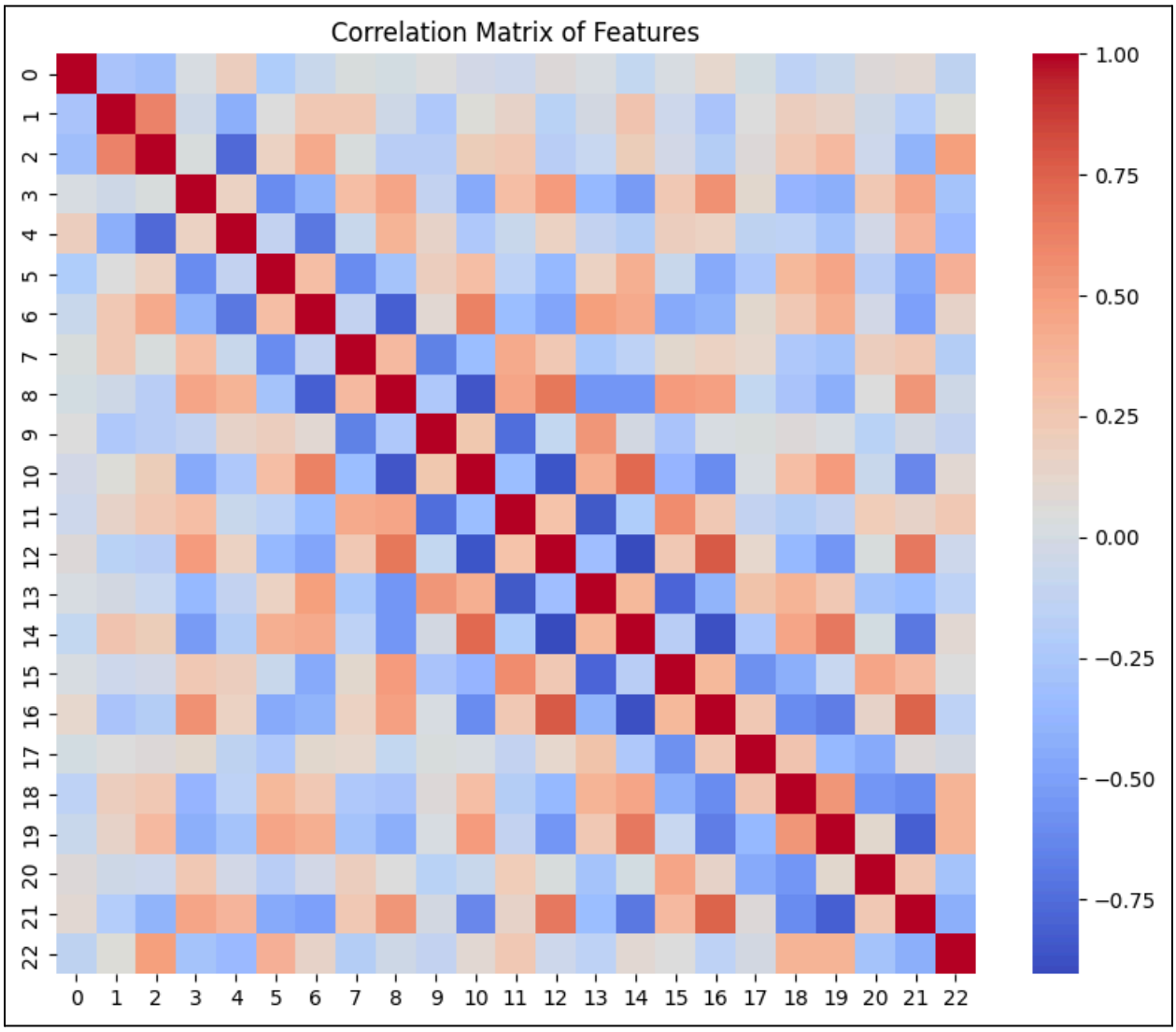
- Standardization: All numerical features were scaled using "StandardScaler", bringing them to a similar range and standardizing variances, essential for distance-based clustering algorithms.

1.3 Feature Engineering

- Polynomial Features: To explore higher-dimensional interactions, polynomial features of degree 2 (interaction only) were generated, resulting in a total of **276 features**. However, these features were later reduced based on correlations to retain only the most informative ones.

1.4 Feature Correlation Analysis

- Correlation Matrix: A heatmap of feature correlations indicated that several features were highly correlated (correlation > 0.9).

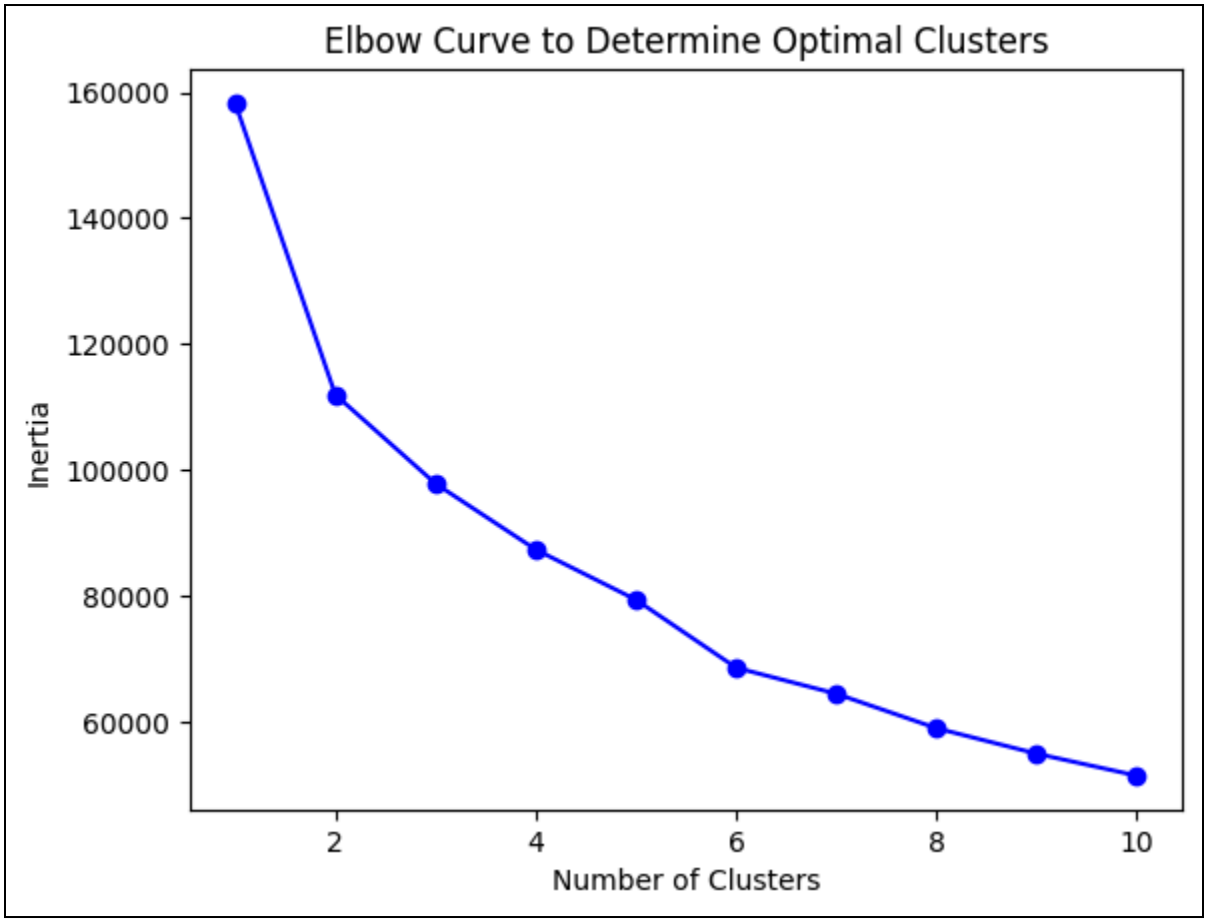


- **Removal of Redundant Features**: Features with a high correlation were removed to reduce redundancy, resulting in a reduced feature set of 22 dimensions.

2. K-Means Clustering

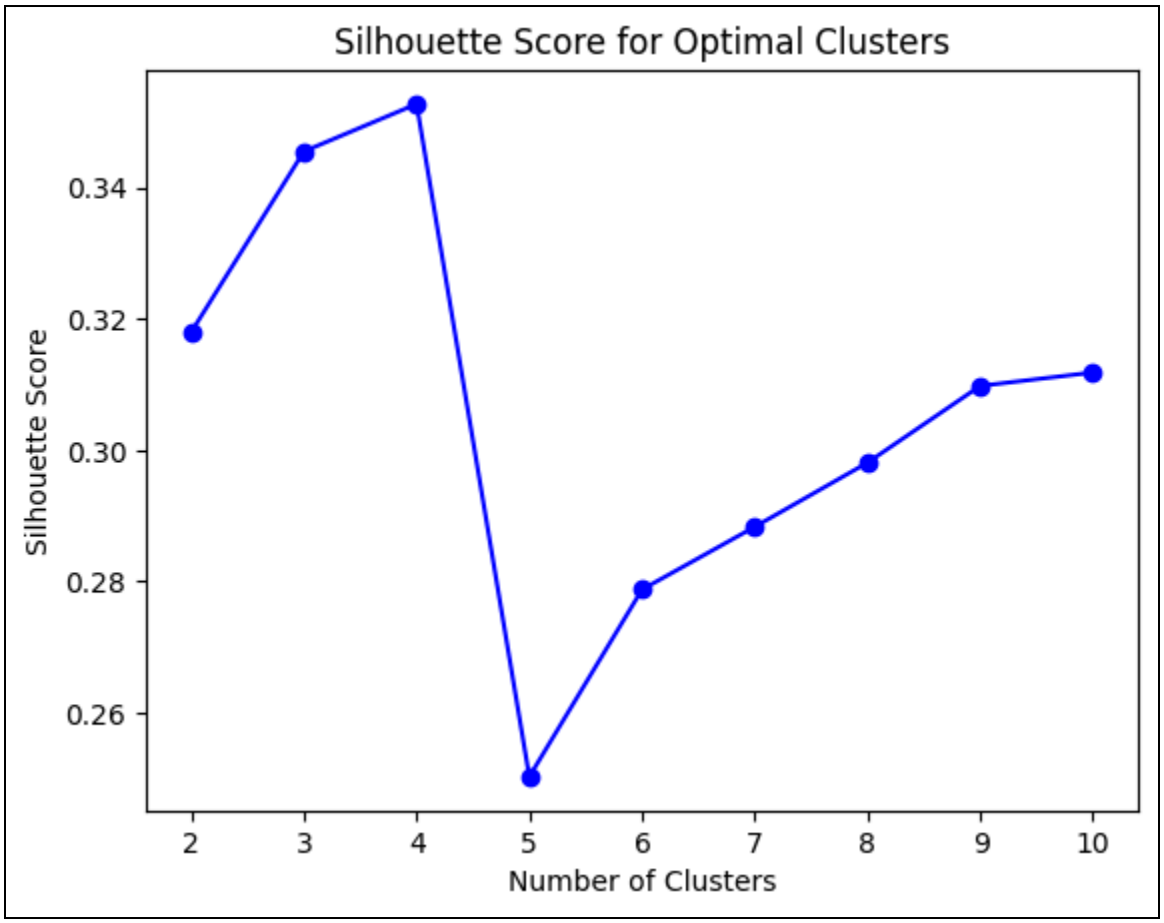
2.1 Elbow Method

- Inertia Plot: The Elbow Method was used to determine the optimal number of clusters. A plot of inertia values showed a notable decrease around 4 clusters, suggesting that this number could be optimal.



2.2 Silhouette Score Evaluation

- Silhouette Score: The silhouette score was calculated for a range of cluster numbers (2 to 10). The maximum silhouette score was observed at **4 clusters**, confirming the result from the Elbow Method.



2.3 K-Means Clustering Implementation

- K-Means Clustering: Using the optimal cluster count of 4, K-Means was implemented on the reduced feature set, resulting in 4 distinct clusters.

2.4 Cluster Initialization

- Comparison of Initialization Methods:

- The 'random' initialization method achieved an inertia of 86,464.57, while 'k-means++' had an inertia of 87,213.49. Although close, 'k-means++' provided slightly better clustering stability and was thus chosen for further analysis.

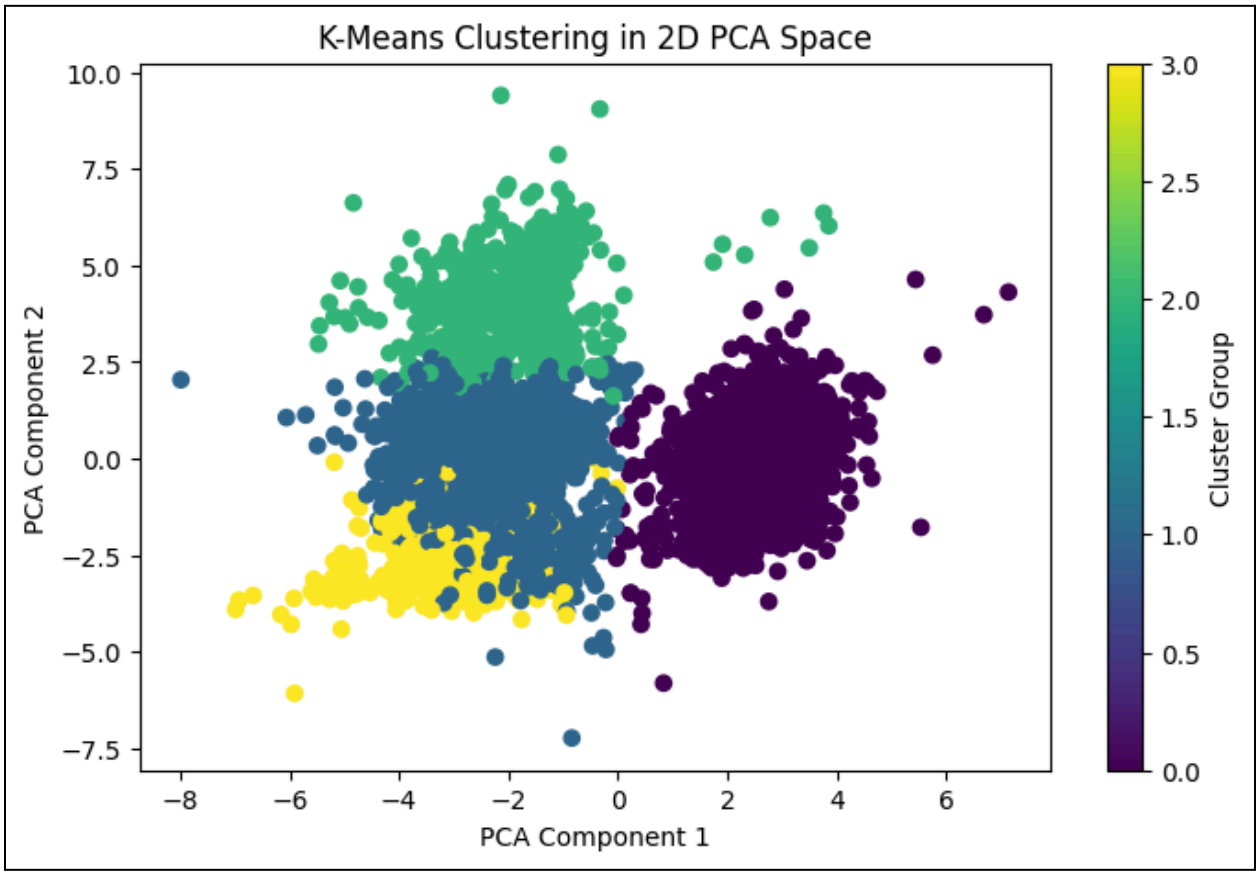
3. Cluster Visualization

3.1 Dimensionality Reduction

- PCA for Visualization: PCA reduced the features to 2 principal components for visualization. This simplified 2D space allowed for a clear display of cluster separation.

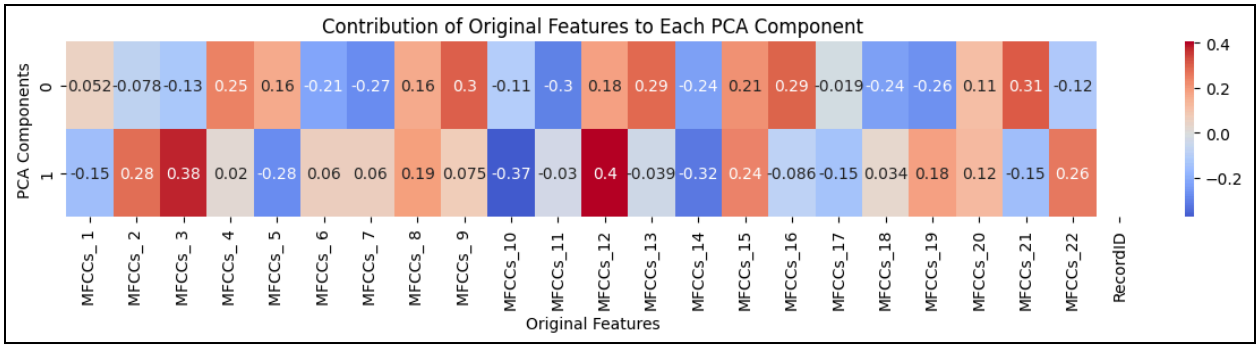
3.2 Cluster Plots

- Scatter Plot: Each cluster was visualized in a 2D scatter plot. The 4 clusters were reasonably well-separated, indicating that clustering was effective.



3.3 Feature Contribution to Clustering

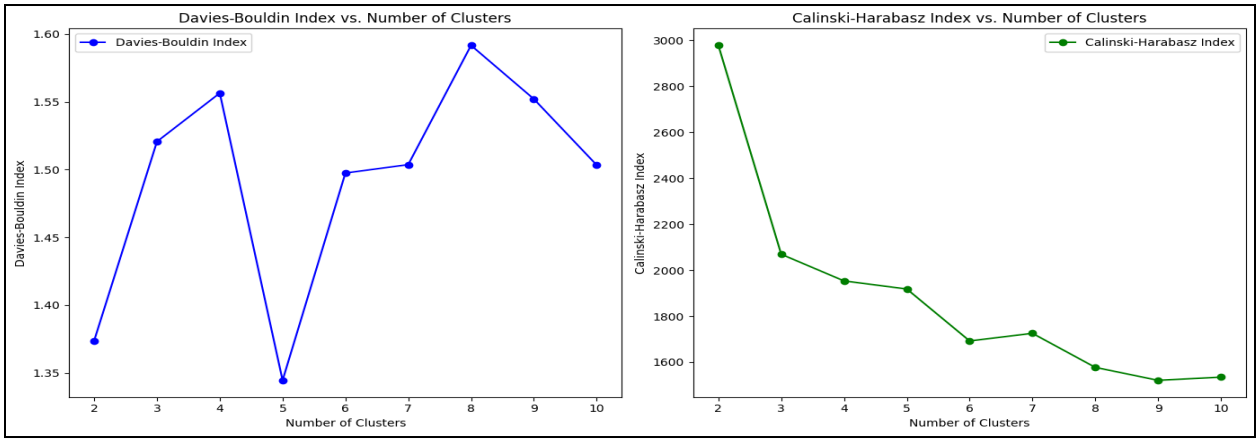
- Feature Contribution via PCA: The contributions of the original MFCC features to each PCA component were visualized. Certain MFCCs contributed more significantly to the principal components, suggesting their importance in cluster differentiation.



4. Cluster Evaluation Metrics

Evaluation Using Multiple Metrics

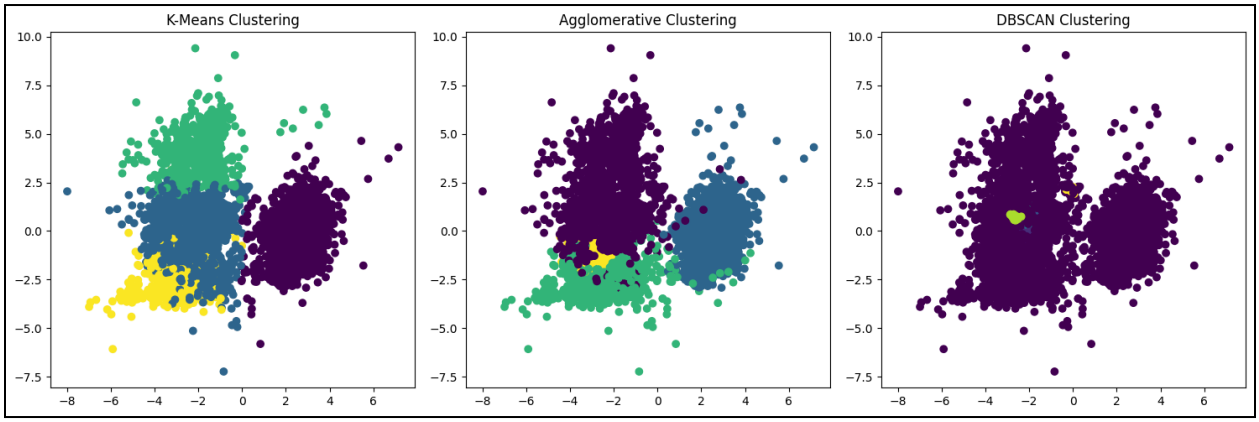
- **Davies-Bouldin Index and Calinski-Harabasz Index**:
  - Davies-Bouldin Index ranged from 1.34 (for 5 clusters) to 1.56 (for 4 clusters), suggesting that 5 clusters might yield slightly better intra-cluster similarity.
  - Calinski-Harabasz Index had the highest score for 2 clusters (2977.43), but considering other metrics, 4 clusters were still optimal.
- Conclusion: While both metrics provided useful insights, the silhouette score and Elbow Method supported 4 clusters as the optimal number, balancing intra-cluster cohesion and inter-cluster separation.



5. Comparison with Other Clustering Algorithms

Algorithm Comparison

- K-Means: Provided well-separated clusters, especially in reduced PCA space. This method worked well for spherical clusters but might struggle with varying density and shape.
- Agglomerative Clustering: Showed comparable results to K-Means but struggled with clear separation at higher dimensions.
- DBSCAN: DBSCAN identified several noise points, suggesting its effectiveness in handling outliers. However, the number of clusters was sensitive to "eps" and "min_samples" parameters.



Strengths and Weaknesses:
- K-Means: Efficient but assumes spherical clusters and struggles with irregular cluster shapes.
- Agglomerative Clustering: Effective for hierarchical relationships but computationally intensive for large datasets.
- DBSCAN: Great for density-based clusters but may fail if clusters have varying density.

Summary of Clustering Process

The clustering process revealed 4 distinct clusters based on the MFCC features. The clusters were validated using multiple metrics, including the silhouette score and Davies-Bouldin Index, both supporting the choice of 4 clusters. PCA allowed for visual separation of clusters and highlighted the contribution of various MFCC features to cluster differentiation.

Key Insights

- MFCC features effectively grouped frogs based on acoustic characteristics.
- Dimensionality reduction and outlier handling improved clustering quality and visual interpretability.
- K-Means performed best overall, while DBSCAN showed promise for handling noise.

Limitations of Clustering Algorithms

- K-Means: Assumes spherical clusters, making it less effective for non-spherical distributions.
- Agglomerative Clustering: Computationally expensive for large datasets.
- DBSCAN: Sensitive to parameter settings and less suitable for clusters with varying densities.

Conclusion

This analysis provided meaningful clustering of frog species based on MFCC features, with the optimal number of clusters identified as 4. The results demonstrated the applicability of K-Means for this dataset while highlighting the strengths of DBSCAN for handling noise. Future work could explore additional clustering algorithms like spectral clustering and alternative feature engineering techniques to enhance clustering performance further.