Assignment 3 Part B Report

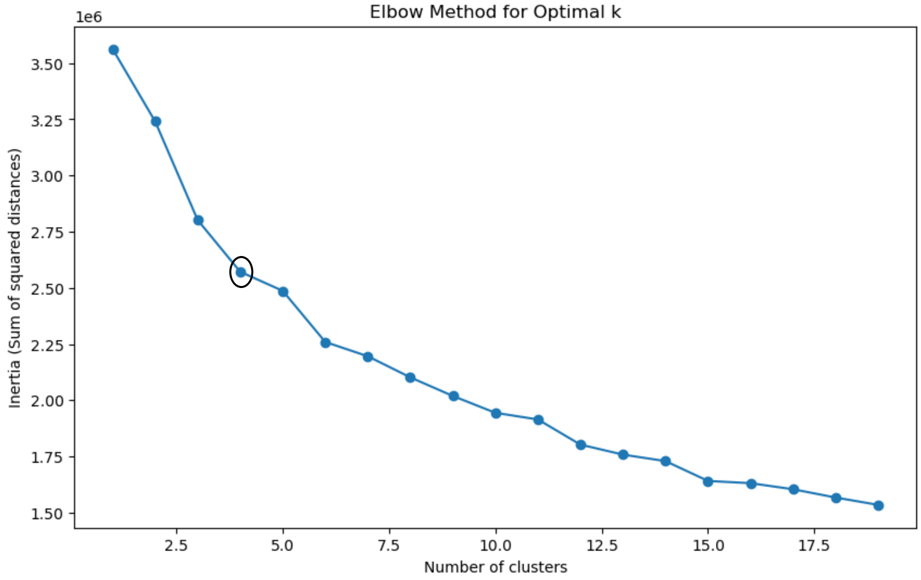
Sharanya Chakraborty

22CS10088

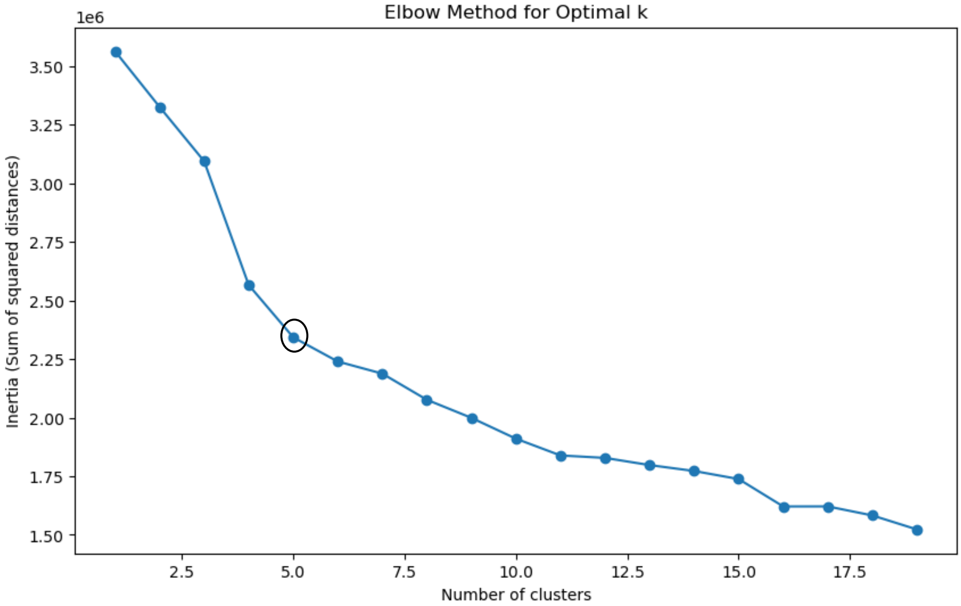
The following observations were made for the various steps followed in the assignment:

1. **Data Preprocessing:**
    - **Skipping Feature Engineering reduced the Silhouette Score** of the K-Means Clustering from its current value of 0.427095460133813 to around 0.3, hence feature engineering was implemented with PolynomialFeatures with degree 2. This **increased number of features from 22 to 275.**
    - Feature Correlation Analysis was done, but it didn't drop any features at the end.
2. **K-Means Clustering:**
    - The process followed can be summarised as follows:
        - Initialising centroids randomly as well as using K-Means++.
        - Assigning clusters to each data point based on Euclidean distance.
        - Updating centroids to the mean of all points assigned to each centroid.
        - Using Elbow Method to run K-Means on a range of K values to find the optimal K value.
        - Using the optimal K value and then calculating the Silhouette Score.
    - The **Elbow Method gave K=4** for Random Initialisation and **K=5** for K-Means++.

      Plot for Random Initialisation:



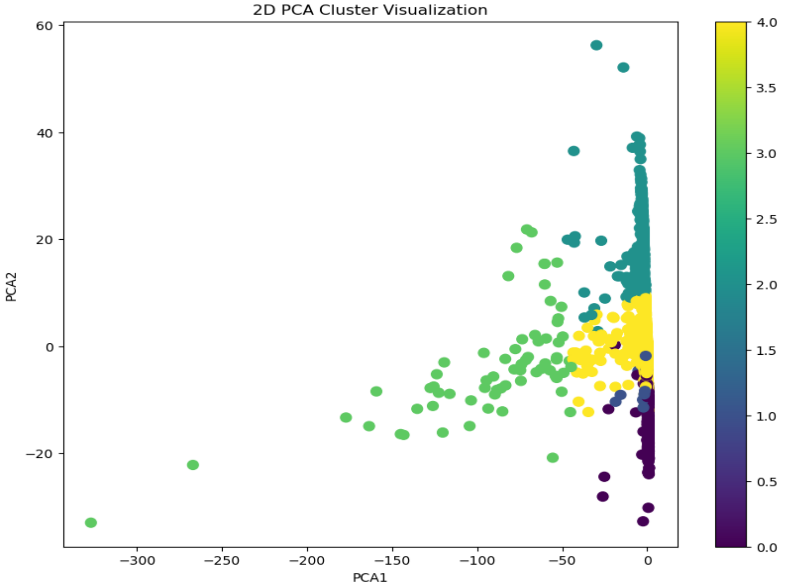
Plot for K-Means++:



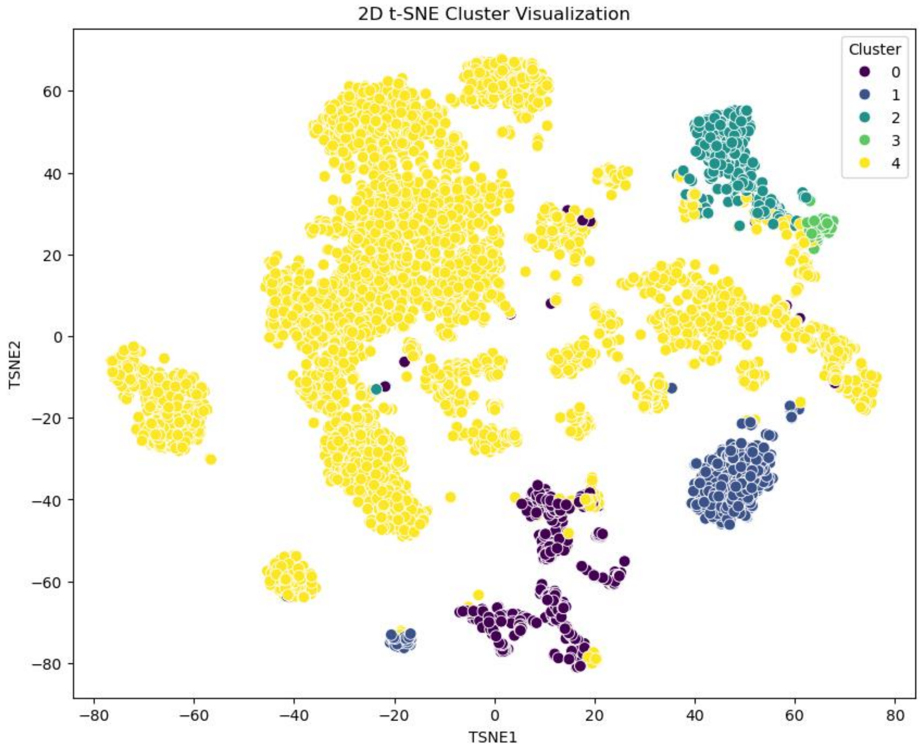
- Using **K-Means++** gave **a higher Silhouette Score** (0.427095460133813) than random initialisation (0.41179442410177447).
- Thus, **K-Means++** based K-Means Clustering was used for the further steps, as it **performed marginally better**.

3. Cluster Visualisation:
    - Both PCA and t-SNE were implemented. PCA was implemented from scratch, whereas sklearn's t-SNE library was used.
    - These were implemented since the number of features were 275, which is impossible to view on a 2D plane.

PCA Plot:

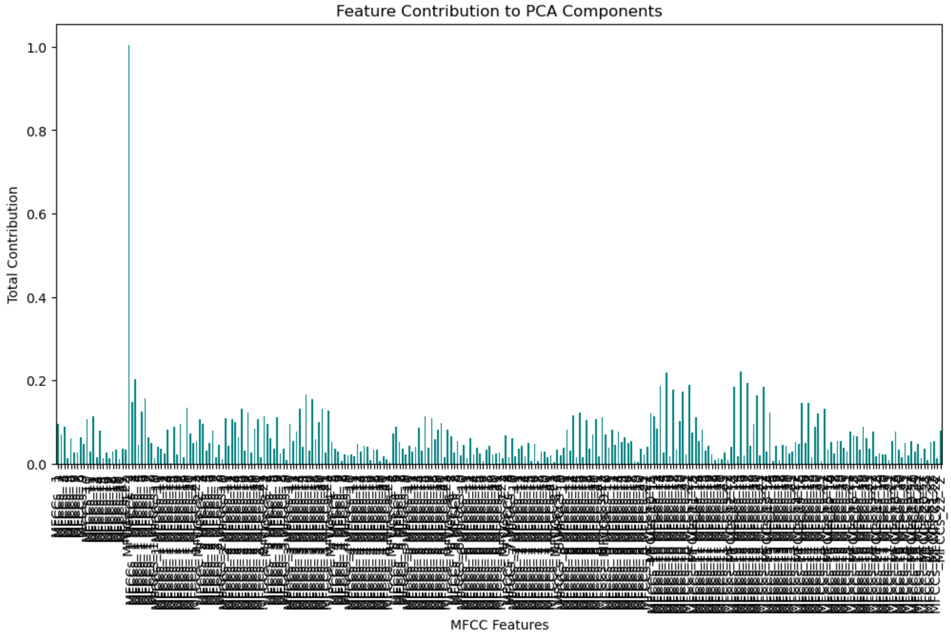


t-SNE Plot:

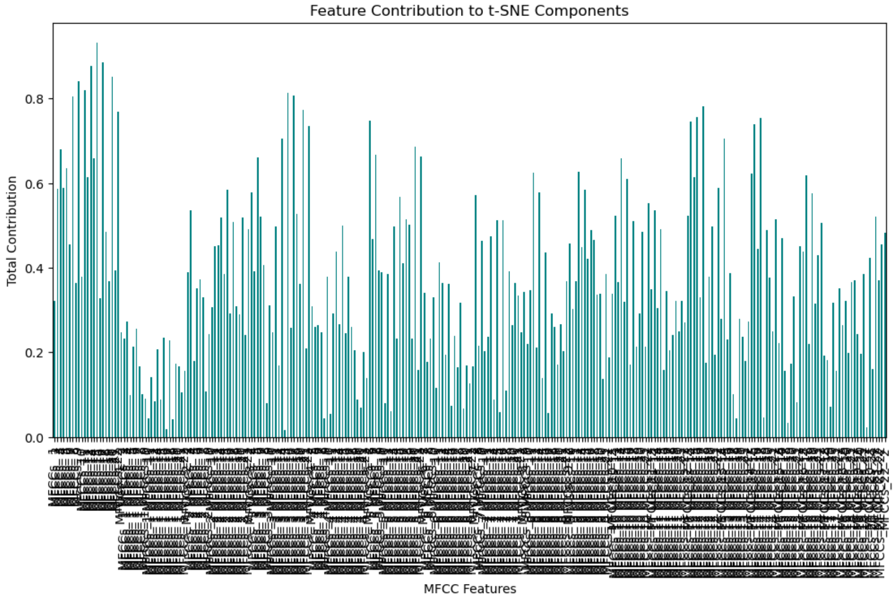


- t-SNE managed to capture the clustering better than PCA. This verifies the fact the PCA focuses on preserving global variance and dimensionality reduction rather than preserving local clusters.
- **Feature Contribution:** This was plotted for both PCA and t-SNE.

For PCA:

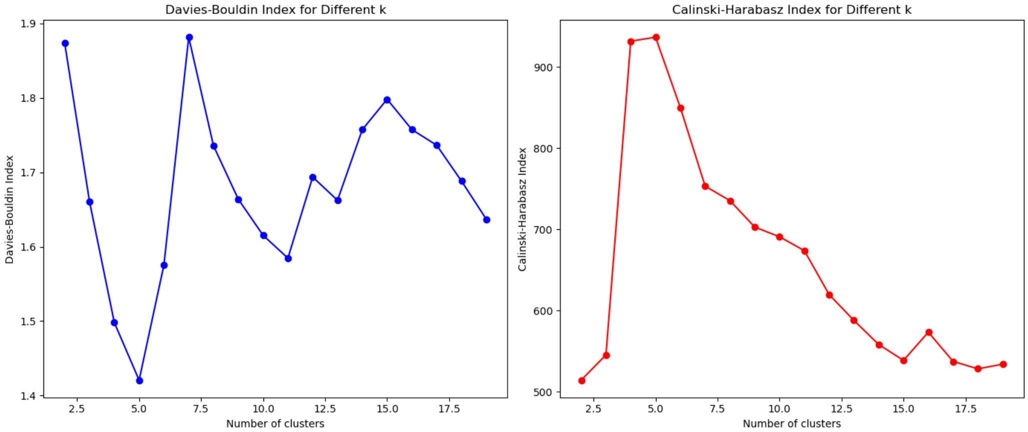


For t-SNE:



- Both the plots show that **some features contributed higher than others**, and the peaks in the bar graphs are approximately at the same positions.

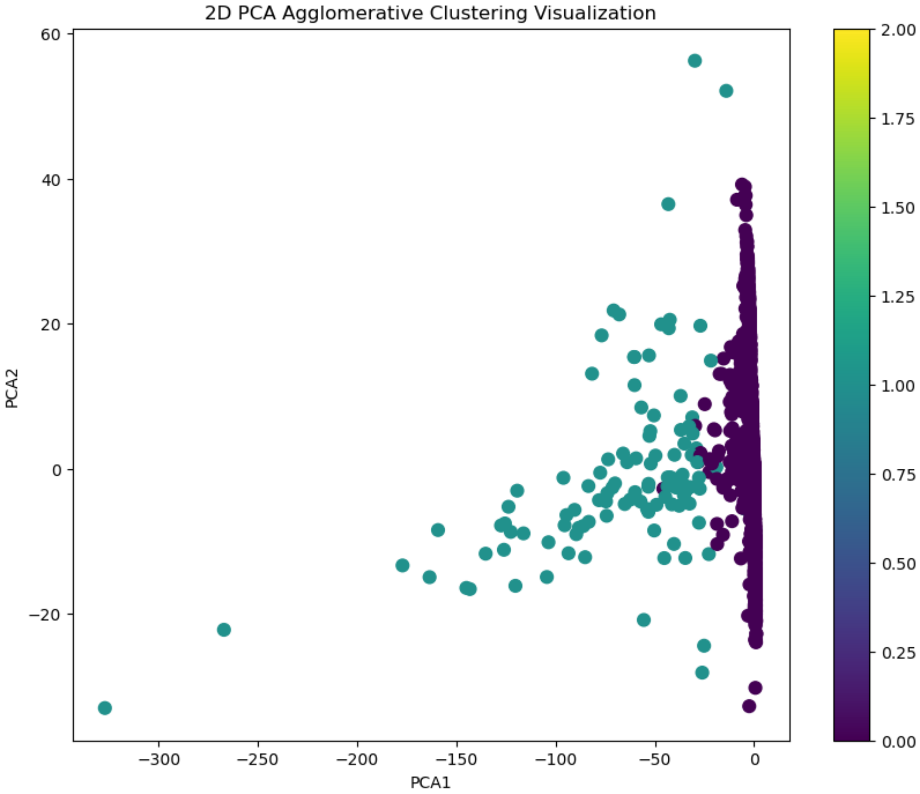
4. **Cluster Evaluation Metrics:**
   - **Davies-Bouldin Index** and **Calinski-Harabasz Index** were used to assess the quality of clusters over a range of K-Values (2 to 20).
   - The following plots were obtained:

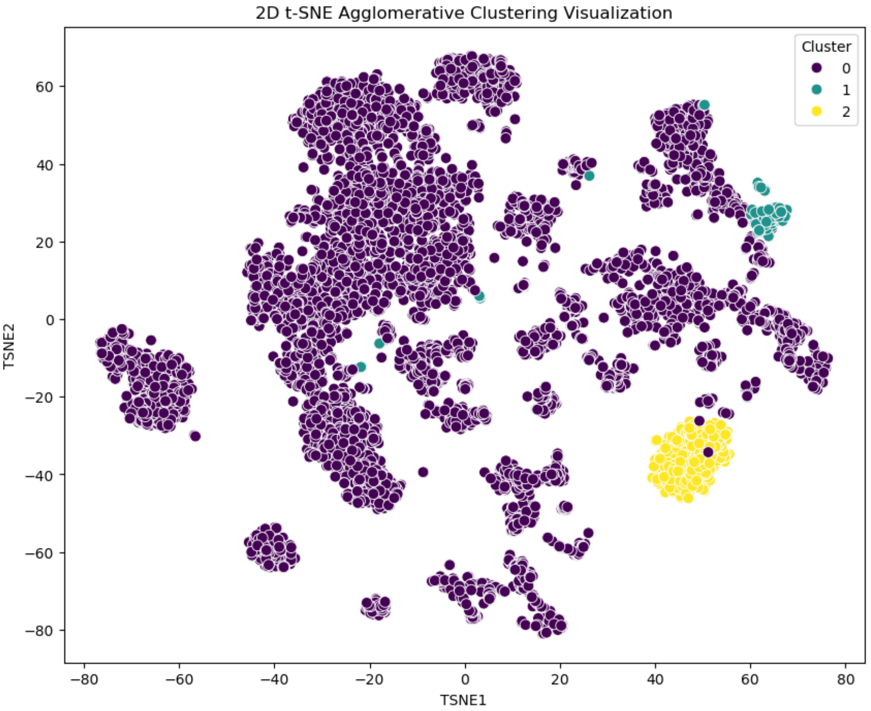


   - The Davies-Bouldin Index evaluates the average similarity ratio of each cluster with respect to its most similar cluster. **A lower value** indicates better clustering.
   - The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the ratio of the sum of between-cluster dispersion and within-cluster dispersion. **A higher value** indicates better clustering.
   - It is evident from the plots that the lowest value for DB index (1.4207948297923072) and highest value for CH index (936.374266306559) are obtained **for K=5**, which also **validates the Elbow Method and Silhouette Score.**

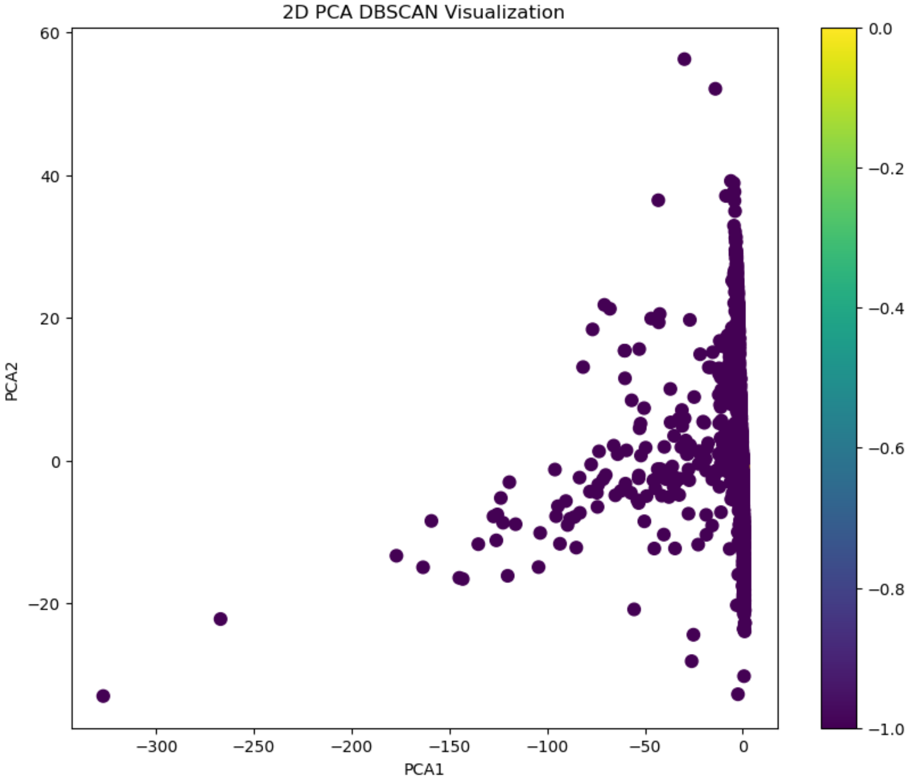
5. **Comparison with Other Clustering Algorithms:**

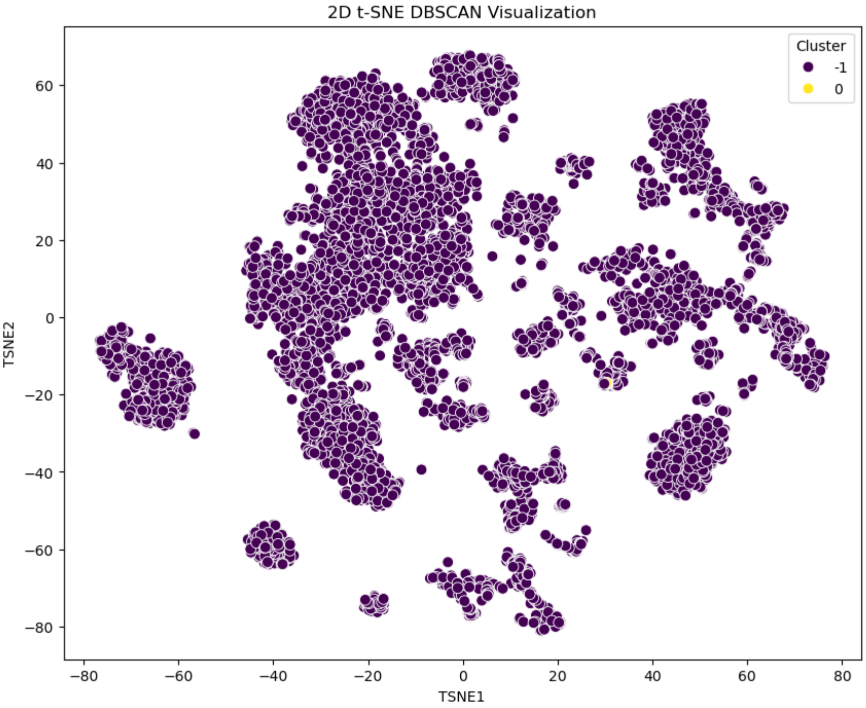
- Both **Agglomerative Hierarchical Clustering** and **DBSCAN** were implemented with the help of the sklearn libraries.
- By testing on a few cluster number values, **n = 3** was chosen as the optimal clustering for Agglomerative Hierarchical Clustering.





- DB Index of **1.1931521544499344** and CH Index of **911.68915206132** were obtained. This suggests that since the DB Index is lower, Hierarchical Clustering performed somewhat better than K-Means on this dataset. (n = 2 clusters gave a higher CH index but a higher DB Index, so that too signifies its better performance).
- For DBSCAN, setting eps as low as 1e-9 and as high as 1e9 gave only 1 cluster. Since we weren't taught about DBSCAN in class, I can't explain why this happened.





- But one thing can be said with certainty- DBSCAN performed poorer since one cluster prevented calculation of DB Index and CH Index.

Key Insights:

1. Data Preprocessing and Exploration:
   - Initial data exploration and preprocessing were critical in understanding the dataset and ensuring it was ready for clustering. **Scaling the features and feature engineering helped in achieving better clustering performance.**
2. K-Means Clustering:
   - The Elbow Method suggested k=5 as the optimal number of clusters. This was supported by the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. K-Means **provided a good balance between simplicity and performance**.
   - PCA and t-SNE visualizations confirmed well-separated clusters, demonstrating the effectiveness of K-Means in this context.
3. Feature Contribution:
   - Analysing feature contributions using PCA showed that **certain MFCC features significantly influenced cluster separation**, offering insights into which acoustic features were most informative.
4. Agglomerative Hierarchical Clustering:
   - Agglomerative Clustering was effective in capturing complex cluster structures with **lesser number of clusters**. However, it was computationally more expensive compared to K-Means.
5. DBSCAN:
   - DBSCAN created only **1 cluster**.

Conclusion

In conclusion, this assignment effectively demonstrated the use of various clustering techniques and evaluation metrics to analyse and visualize the Anuran Calls Dataset (MFCCs).

- K-Means emerged as a reliable method for this dataset, balancing performance and simplicity.
- Agglomerative Clustering added depth to our analysis by capturing complex structures.
- DBSCAN offered robustness against outliers and flexibility in cluster shapes.

Each algorithm showed distinct strengths and highlighted the importance of choosing the right clustering method based on the dataset characteristics and the analysis objectives. The insights gained from feature contributions further enhanced our understanding of the dataset, guiding us towards more informed and effective clustering.
