MLPR LAB-5 REPORT

1. What are the common distance metrics used in distance-based classification algorithms?

In this lab, the Euclidean distance is the default metric used by Kmeans to calculate the straight-line distance between face features (Hue and Saturation). Other common metrics include Manhattan distance, which sums absolute differences, and Cosine similarity, which focuses on the angle between feature vectors rather than their magnitude.

2. What are some real-world applications of distance-based classification algorithms?

In this lab , with face detection, a primary application is facial recognition and grouping, where an unknown face (like the template image) is assigned to a known identity cluster. It is also used in image segmenting and recommendation systems to group users with similar aesthetic or behavioral profiles.

3. Explain various distance metrics.

- **Euclidean Distance:** Calculated as $\sqrt{\sum(x_i - y_i)^2}$, which you used to find the nearest centroid for Dr. Tharoor's face.
- **Manhattan Distance:** The sum of horizontal and vertical components, useful when features (like Hue vs. Saturation) should not be squared or weighted quadratically.

4. What is the role of cross-validation in model performance?

Prevents Overfitting: It ensures your color boundaries aren't tuned only to the specific lighting of one image, allowing the model to work on different photos.

Tests Generalization: By testing on data the model hasn't seen, it confirms that the template face is being classified based on real patterns rather than coincidental noise.

5. Explain variance and bias in terms of KNN?

In a distance-based context like KNN, High Bias occurs if k is too large, leading to an oversimplified model that ignores local face-color differences. High Variance occurs if k is too small (ex: $k = 1$), making the classification overly sensitive to "noise" or a single outlier pixel in a face's saturation, causing the model to change drastically with slight data changes.