**Practical Lab Plan: Advanced Facial Recognition with HDBSCAN**

**Title:** Advanced Facial Recognition Clustering using HDBSCAN

**Objective:**

1. Explore the application of **HDBSCAN** for clustering facial data in an unsupervised learning scenario.
2. Tackle real-world challenges like noisy datasets and unbalanced data distribution.
3. Visualize and evaluate clustering results for practical use cases.

**Lab Prerequisites:**

1. Basic knowledge of unsupervised learning and density-based clustering methods.
2. Familiarity with dimensionality reduction techniques such as PCA.
3. Proficiency in Python programming and related data analysis tools.

**Required Tools:**

* A Python development environment such as Jupyter Notebook or VS Code.
* Pre-installed libraries such as NumPy, Pandas, Scikit-learn, OpenCV, Matplotlib, and HDBSCAN.

**Lab Setup:**

**Dataset:**

Use a real-world dataset, such as:

* CelebA Dataset: Contains celebrity faces with multiple images per person.
* [Labeled](http://vis-www.cs.umass.edu/lfw/lfw.tgz) [Faces](http://vis-www.cs.umass.edu/lfw/lfw.tgz) [in](http://vis-www.cs.umass.edu/lfw/lfw.tgz) [the](http://vis-www.cs.umass.edu/lfw/lfw.tgz) [Wild](http://vis-www.cs.umass.edu/lfw/lfw.tgz) (LFW): A dataset of labeled facial images suitable for clustering tasks.

**Lab Steps:**

**Step 1: Load and Explore the Dataset**

* Load the dataset into the environment.
* Examine the shape and structure of the dataset to understand its features, including the number of images, dimensions, and labels (if available).
* Visualize a few images to familiarize students with the data format and potential challenges like varying image sizes or resolutions.

**Step 2: Data Preprocessing**

* Flatten the image data into a format suitable for machine learning models.
* Standardize the dataset to ensure uniform scaling, which is critical for clustering algorithms like HDBSCAN.

**Step 3: Dimensionality Reduction**

* Use PCA (Principal Component Analysis) to reduce the number of features while retaining essential variance in the data.
* Choose an appropriate number of components to optimize clustering performance without over-simplifying the dataset.

**Step 4: Apply HDBSCAN Clustering**

* Apply the HDBSCAN algorithm to the reduced data.
* Configure HDBSCAN parameters, such as minimum cluster size and distance metric, to achieve meaningful clusters.
* Extract and display the cluster labels assigned to each data point.

**Step 5: Introduce Real-Life Challenges**

1. **Noisy Data:**
   1. Simulate noisy data by adding random variations to images and analyze the algorithm’s performance on this altered dataset.
2. **Distance Metrics:**
   1. Experiment with different distance metrics, such as Manhattan or Cosine, to observe their effects on clustering results.

**Step 6: Visualize and Analyze Results**

* Create scatter plots or other visual representations to display the clustering of the reduced data.
* Identify noise points, which HDBSCAN labels as outliers, and analyze their distribution.
* Evaluate clustering quality using metrics like silhouette scores or other relevant techniques.

**Step 7: Real-Life Applications**

1. **Test on New Data:**
   1. Use HDBSCAN to assign a new image to one of the identified clusters. ○ Analyze how well the clustering algorithm generalizes to unseen data.
2. **Representative Images for Clusters:**
   1. Identify the most representative image in each cluster by finding the data point closest to the cluster center.

**Expected Outcomes:**

1. Hands-on experience in clustering facial data with unsupervised learning techniques.
2. Understanding the impact of dimensionality reduction and parameter tuning on clustering.
3. Insights into real-world challenges such as noise and imbalanced data.

**Additional Challenges:**

1. Work with a custom dataset by capturing images from a webcam or mobile camera.
2. Experiment with advanced feature extraction techniques like HOG or embeddings from pre-trained models.
3. Optimize HDBSCAN parameters for better performance in clustering tasks.