Student Name: Artem Kamov  
Student ID: 301220613

1. The Olivetti Faces Dataset was retrieved from the sklearn.datasets module using fetch\_olivetti\_faces. The dataset contains 400 grayscale images, each of size 64x64 pixels, corresponding to 40 different individuals (10 images per person). The data matrix X contains 4096 features per image (64x64=4096), and the target variable y contains the identity labels for each person.

  
  
Shape of X: (400, 4096) - 400 images and 4096 pixels per image.

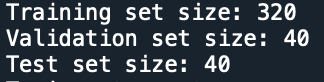
Shape of y: (400,) - 400 identity labels

2. Stratified sampling was applied to divide the dataset into training, validation, and test sets to ensuring an equal number of images per person in each set. This maintai the class balance across the splits which is important for ensuring consistent evaluation of classification models.s.

**Split Ratios**:

* **80% training set**: 320 images.
* **10% validation set**: 40 images.
* **10% test set**: 40 images.

**Rationale for Split**: 80/10/10 split was selected to provide sufficient data for training, while still retaining enough data for validation and testing since there are only 400 images in the dataset.



**Stratified Sampling Results**:

Train set counts per person: {0: 8, 1: 8, 2: 8, 3: 8, 4: 8, 5: 8, 6: 8, 7: 8, 8: 8, 9: 8, 10: 8, 11: 8, 12: 8, 13: 8, 14: 8, 15: 8, 16: 8, 17: 8, 18: 8, 19: 8, 20: 8, 21: 8, 22: 8, 23: 8, 24: 8, 25: 8, 26: 8, 27: 8, 28: 8, 29: 8, 30: 8, 31: 8, 32: 8, 33: 8, 34: 8, 35: 8, 36: 8, 37: 8, 38: 8, 39: 8}

Validation set counts per person: {0: 1, 1: 1, 2: 1, 3: 1, 4: 1, 5: 1, 6: 1, 7: 1, 8: 1, 9: 1, 10: 1, 11: 1, 12: 1, 13: 1, 14: 1, 15: 1, 16: 1, 17: 1, 18: 1, 19: 1, 20: 1, 21: 1, 22: 1, 23: 1, 24: 1, 25: 1, 26: 1, 27: 1, 28: 1, 29: 1, 30: 1, 31: 1, 32: 1, 33: 1, 34: 1, 35: 1, 36: 1, 37: 1, 38: 1, 39: 1}

Test set counts per person: {0: 1, 1: 1, 2: 1, 3: 1, 4: 1, 5: 1, 6: 1, 7: 1, 8: 1, 9: 1, 10: 1, 11: 1, 12: 1, 13: 1, 14: 1, 15: 1, 16: 1, 17: 1, 18: 1, 19: 1, 20: 1, 21: 1, 22: 1, 23: 1, 24: 1, 25: 1, 26: 1, 27: 1, 28: 1, 29: 1, 30: 1, 31: 1, 32: 1, 33: 1, 34: 1, 35: 1, 36: 1, 37: 1, 38: 1, 39: 1}

* **Training set counts per person**: 8 images per individual.
* **Validation set counts per person**: 1 image per individual.
* **Test set counts per person**: 1 image per individual.

3. Support Vector Machine (SVM) was chosen as the classifier because it is a common choice for image classification tasks since it works well for high-dimensional data like images. A Support Vector Machine (SVM) classifier with a linear kernel using 5-fold cross-validation on the training set was trained to predict the identity of individuals in the dataset. Cross-validation was used to assess the model's generalizability.

**Cross-validation Results**:

Cross-validation accuracy scores: [0.9375, 0.9375, 0.96875, 0.96875, 0.953125]

Mean cross-validation accuracy: **95.31%**

**precision recall f1-score support**

0 1.00 1.00 1.00 1

1 1.00 1.00 1.00 1

2 1.00 1.00 1.00 1

3 1.00 1.00 1.00 1

4 1.00 1.00 1.00 1

5 1.00 1.00 1.00 1

6 1.00 1.00 1.00 1

7 0.50 1.00 0.67 1

8 1.00 1.00 1.00 1

9 0.00 0.00 0.00 1

10 1.00 1.00 1.00 1

11 1.00 1.00 1.00 1

12 1.00 1.00 1.00 1

13 1.00 1.00 1.00 1

14 0.50 1.00 0.67 1

15 1.00 1.00 1.00 1

16 1.00 1.00 1.00 1

17 1.00 1.00 1.00 1

18 1.00 1.00 1.00 1

19 1.00 1.00 1.00 1

20 1.00 1.00 1.00 1

21 1.00 1.00 1.00 1

22 1.00 1.00 1.00 1

23 1.00 1.00 1.00 1

24 1.00 1.00 1.00 1

25 1.00 1.00 1.00 1

26 1.00 1.00 1.00 1

27 1.00 1.00 1.00 1

28 1.00 1.00 1.00 1

29 1.00 1.00 1.00 1

30 1.00 1.00 1.00 1

31 0.00 0.00 0.00 1

32 1.00 1.00 1.00 1

33 1.00 1.00 1.00 1

34 1.00 1.00 1.00 1

35 1.00 1.00 1.00 1

36 1.00 1.00 1.00 1

37 1.00 1.00 1.00 1

38 1.00 1.00 1.00 1

39 1.00 1.00 1.00 1

accuracy 0.95 40

macro avg 0.93 0.95 0.93 40

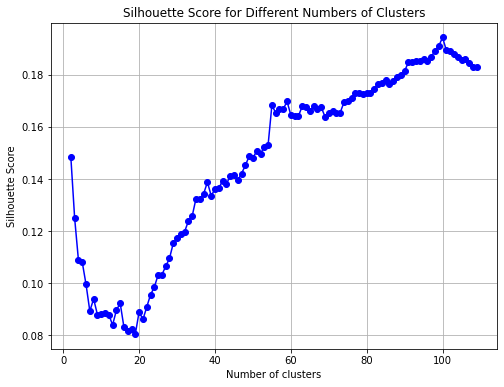
weighted avg 0.93 0.95 0.93 40

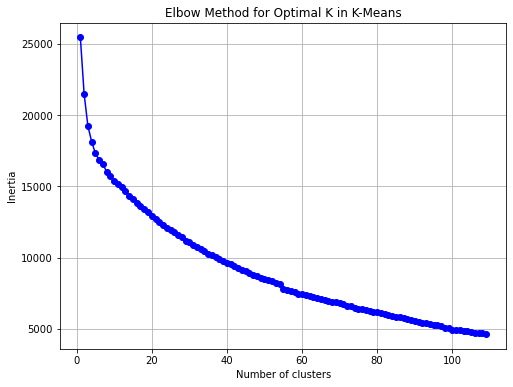
The high cross-validation accuracy indicates that the classifier performs well on unseen data, with minimal variation in accuracy across folds.

* **Accuracy** on the validation set: 95%.
* Most classes achieved perfect precision, recall, and F1-scores, except for a few classes like 7, 9, 14, and 31, which had lower precision and recall.
  + Notably, classes 7, 9, 14, and 31 had difficulty, with some having zero recall, which could be due to either noise or challenging visual differences in the dataset.

The overall performance on the validation set is consistent with the high accuracy achieved during cross-validation indicating that the model generalizes well to unseen data.

4.





The best number of clusters is 100 with a silhouette score of 0.194.

The silhouette score of 0.1942795217037201 for 100 clusters indicates that increasing the number of clusters improves the cohesion and separation to some extent, although it is not a particularly high score which suggests some overlapping clusters.

Based on the elbow method plot, it looks like the best number of clusters is 5. Howverer, it also should be noted that inerinertia curve flattens out after 100 clusters meaning that tional clusters don't significantly improve the model.

Since it was required to use the the silhouette score approach to choose the number of clusters, the number of clusters of 100 was chosen to train the clasifier once again.

5.

Validation accuracy with reduced dimensionality: 0.875

Classification Report on Validation Set:

precision recall f1-score support

0 0.50 1.00 0.67 1

1 1.00 1.00 1.00 1

2 0.00 0.00 0.00 1

3 0.33 1.00 0.50 1

4 1.00 1.00 1.00 1

5 1.00 1.00 1.00 1

6 1.00 1.00 1.00 1

7 1.00 1.00 1.00 1

8 1.00 1.00 1.00 1

9 0.00 0.00 0.00 1

10 1.00 1.00 1.00 1

11 1.00 1.00 1.00 1

12 0.00 0.00 0.00 1

13 1.00 1.00 1.00 1

14 1.00 1.00 1.00 1

15 1.00 1.00 1.00 1

16 1.00 1.00 1.00 1

17 0.00 0.00 0.00 1

18 1.00 1.00 1.00 1

19 1.00 1.00 1.00 1

20 0.50 1.00 0.67 1

21 1.00 1.00 1.00 1

22 1.00 1.00 1.00 1

23 1.00 1.00 1.00 1

24 1.00 1.00 1.00 1

25 1.00 1.00 1.00 1

26 1.00 1.00 1.00 1

27 1.00 1.00 1.00 1

28 0.00 0.00 0.00 1

29 1.00 1.00 1.00 1

30 1.00 1.00 1.00 1

31 1.00 1.00 1.00 1

32 1.00 1.00 1.00 1

33 1.00 1.00 1.00 1

34 1.00 1.00 1.00 1

35 1.00 1.00 1.00 1

36 1.00 1.00 1.00 1

37 1.00 1.00 1.00 1

38 1.00 1.00 1.00 1

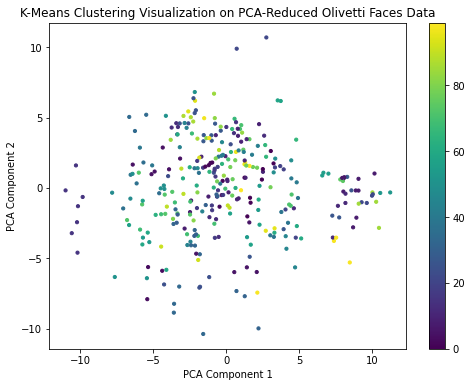
39 0.50 1.00 0.67 1

accuracy 0.88 40

macro avg 0.82 0.88 0.84 40

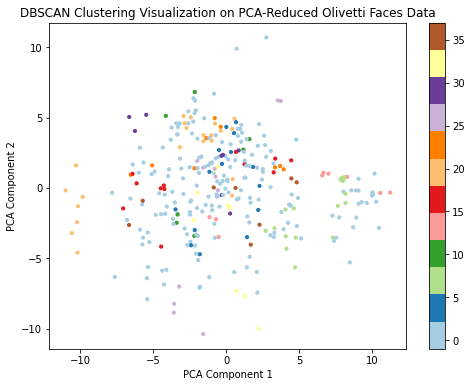
weighted avg 0.82 0.88 0.84 40

* **Validation accuracy** of 0.875 is quite good, given the dimensionality reduction applied through PCA.
* The **classification report** shows that many of the classes are performing well with precision and recall values of 1.0 for most, though there are a few classes (like 2, 9, 17, and 28) with performance issues, likely due to misclassifications or inherent similarities in the data.



The **K-Means clustering visualization** on the PCA-reduced Olivetti faces data shows a clear scatter with distinct clusters, though some points may be overlapping due to dimensionality reduction from 4096 features to just 2 components.

6. After testing different parameter values, eps=7.5 and min\_samples=2 were chosen as the optimal settings for the Olivetti faces dataset.



Number of clusters found by DBSCAN: 38

Number of noise points: 31

The chosen parameters resulted in a lot of small clusters (38 clusters), indicating that the faces in the dataset are highly varied, and DBSCAN can detect subtle groupings based on facial features.

Only 31 noise points were identified, which suggests that most of the faces were successfully grouped into clusters. A small amount of noise is a good outcome as it indicates that DBSCAN effectively assigned most faces to meaningful clusters.

Visual inspection of the plot shows a large number of small clusters scattered around, which aligns well with DBSCAN's ability to find dense regions in data.

Overall Conclusion:

DBSCAN was used effectively to explore clustering on the Olivetti dataset.

Meanwhile, K-Means showed that 100 clusters achieved the best silhouette score of 0.194, reflecting a overseparation of faces.

DBSCAN, with eps=7.5 and min\_samples=2, found 38 clusters with minimal noise points, effectively capturing dense regions of face data.

In classification, the model achieved 95% cross-validation accuracy, and with PCA for dimensionality reduction, the accuracy slightly decreased to 87.5%, indicating some loss of information in the reduced feature space.