Unsupervised Learning with Dimensionality Reduction and Clustering

NAME: Rumela Dasgupta

COURSE/BATCH: BSc Computer Science(Hons.)

SECTION 1: Foundational Internship

INSTITUTE NAME: ST. XAVIER'S COLLEGE(AUTONOMOUS), KOLKATA

Period of Internship: 21ST January 2026 – 17TH February 2026

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI, Kolkata

1. Abstract

This project was based on an unsupervised machine learning approach to understand handwritten digit data using clustering techniques. The sklearn digits dataset was used, which contains 1797 digit images (0–9), where each image is an 8×8 grayscale representation. The main model used in this project was K-Means clustering with 10 clusters. Since clustering algorithms do not use true labels during training, an additional mapping step was required to convert the cluster IDs into actual digit predictions. PCA was also applied to reduce the original 64-dimensional feature space into 2 dimensions, which made clustering and visualization easier. The final predicted labels were evaluated using accuracy score, confusion matrix, and classification report. The overall accuracy achieved was approximately 56%. The report also highlights which digits were predicted well and which digits were commonly confused. This project helped in building a clear understanding of unsupervised learning, PCA, and evaluation techniques in machine learning.

2. Introduction

In this project, an unsupervised approach was explored to understand how clustering algorithms can group similar images even without being given the actual digit labels.

The dataset used in this project is the sklearn digits dataset. It contains 1797 handwritten digit samples, and each digit image is represented as an 8×8 pixel grayscale image. For machine learning, each image is flattened into a 64-dimensional feature vector.

K-Means clustering was selected as the main model because it is one of the simplest and most widely used clustering algorithms. It works by dividing the dataset into clusters such that the distance between data points and their assigned cluster center is minimized. Since K-Means assigns cluster labels such as 0–9, these labels do not directly represent the digit values. Therefore, a label mapping strategy was used. In this method, the most frequent true digit label inside each cluster was selected using the mode.

In addition to clustering on the original dataset, PCA (Principal Component Analysis) was applied to reduce the 64-dimensional data into 2 dimensions. This was done to make the dataset easier to visualize and also to see how clustering behaves after dimensionality reduction.

Relevance of the Project

- Demonstrates unsupervised learning on image-based data.
- Shows how PCA can reduce dimensions for clustering.
- Explains label mapping to interpret clustering output.
- Uses evaluation metrics for performance analysis.

Technology and Tools Used

- Python
- Google Colab / Jupyter Notebook
- NumPy
- Scikit-learn
- SciPy
- Matplotlib

Topics Covered During First Two Weeks of Internship Training

The following topics were covered during the first two weeks of internship training:

Week 1

1. Introduction – Welcome Note – What to expect from the internship
2. Python Basics – 1 (Data, Variables, Lists, Loop)
3. Python Basics – 2 (Data Structures)
4. Python Basics – 3 (Class, Functions, OOPS)
5. Python Basics – 4 (NumPy, Pandas)

Week 2

6. Machine Learning Overview
7. Machine Learning 1 (Regression)
8. Machine Learning 2 (Classification)
9. LLM Fundamentals
10. Communication Skills

3. Project Objective

The objectives of this project are:

- To load and understand the sklearn handwritten digits dataset.
- To apply K-Means clustering on digit image data.
- To reduce the dataset dimensions using PCA.
- To perform clustering on PCA-reduced 2D data.
- To map cluster labels to true digit labels using mode-based mapping.
- To evaluate clustering performance using accuracy score, confusion matrix, and classification report.

4. Methodology

This project was implemented using Python in Google Colab. The complete workflow was carried out step-by-step as described below.

Step 1: Loading the Dataset

The dataset was loaded using the load_digits() function from sklearn.
The shape of the dataset was printed and it was found that:

- The dataset contains **1797 samples**
- Each sample has **64 features**, representing an 8×8 image flattened into a vector

This confirmed that the dataset was properly loaded and ready for analysis.

Step 2: Applying K-Means Clustering

K-Means clustering was applied on the original 64-dimensional dataset with:

- n_clusters = 10
- random_state = 0

The model produced:

- Cluster labels for each digit sample
- Cluster centers with shape **(10, 64)**

Step 3: Reshaping Cluster Centers

To understand what each cluster center represents, the cluster centers were reshaped into image form:

- From **(10, 64)** to **(10, 8, 8)**

This helped in interpreting each cluster center as an “average digit image”.

Step 4: Dimensionality Reduction using PCA

PCA was applied to reduce the feature space from 64 dimensions to 2 dimensions.
This step produced:

- reduced_data.shape = (1797, 2)

Reducing the dimensions made the dataset easier to visualize and also allowed clustering to be tested on a simplified representation.

Step 5: Clustering on Reduced Data

K-Means clustering was applied again, but this time using the PCA-reduced dataset.
This produced a new set of cluster labels for each digit.

Step 6: Mapping Cluster Labels to True Digit Labels

Since K-Means does not use the digit labels during clustering, the cluster IDs do not directly match digits.
To solve this, a mapping dictionary (labels_map) was created:

- For each cluster (0 to 9), all samples belonging to that cluster were selected.
- The most common true digit label among those samples was found using the mode.
- That digit label was assigned as the meaning of the cluster.

This mapping step was necessary to convert clustering results into meaningful predictions.

Step 7: Generating Final Predictions

After label mapping, final predicted labels were generated by converting each cluster label into the mapped digit label.
The predictions were stored in pred_labels.

Step 8: Model Evaluation

To measure how well the unsupervised clustering worked, the predicted labels were compared with the true labels using:

- Accuracy Score
- Confusion Matrix
- Classification Report

These metrics provided both overall performance and digit-wise performance.

5. Data Analysis and Results

Dataset Summary

- Total images: **1797**
- Features per image: **64**

- Classes: **0 to 9**

K-Means Cluster Center Shape

- digits.data.shape → **(1797, 64)**
- kmeans.cluster_centers_.shape → **(10, 64)**
- After reshaping → **(10, 8, 8)**

PCA Output

- reduced_data.shape → **(1797, 2)**

Cluster-to-Label Mapping

A mapping dictionary was generated where each cluster was assigned the most frequent digit label found in that cluster.

Accuracy was computed using:

accuracy_score(digits.target, pred_labels)

The output was a floating point value representing overall accuracy.

Accuracy Score

The overall accuracy obtained after mapping cluster labels to digit labels was:

Accuracy ≈ 0.5665 (56.65%)

This accuracy shows that even though the model was unsupervised, it was still able to group and predict many digits correctly.

Confusion Matrix

A **10×10 confusion matrix** was generated to identify which digits were most commonly confused with others.

Classification Report

A full classification report was printed showing:

- Precision
- Recall
- F1-score
  for each digit class (0–9)

6. Conclusion

This project successfully demonstrated the use of unsupervised learning for handwritten digit recognition. K-Means clustering was applied on the sklearn digits dataset, and PCA was used to reduce the feature space for clustering and visualization. Since clustering does not use true labels, a mode-based label mapping method was used to interpret cluster labels as digit labels. After evaluation, the overall accuracy achieved was around 56.65%.

Although this accuracy is lower than supervised learning models, it is still meaningful because the model learned patterns without using digit labels. The confusion matrix and classification report showed that certain digits were predicted well, while others were confused due to similarity in handwriting styles. Future improvements could include increasing PCA components, using better clustering algorithms, or extracting stronger features before clustering.

7. APPENDICES

Appendix A: References

- Scikit-learn documentation
- SciPy documentation
- sklearn digits dataset documentation

GITHUB Repository:

rumela-dasgupta/-Unsupervised-Learning-with-Dimensionality-Reduction-and-Clustering-Project-Notebook