{AUTUMN INTERNSHIP PROJECT REPORT}

Project Title

Unsupervised Learning with Dimensionality Reduction and Clustering

Priya Mondal

Bsc. Hons in computer science from bangabasi college,

University of calcutta

Period of Internship: 25thAugust 2025 - 19th September 2025

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

1. Abstract

This project uses an unsupervised learning approach to explore complex data without any labels. Initially PCA reduces the number of feathers while simplifying how we understand and visualize the data. Then, K-Means is utilized to cluster data points with similar representations, thereby giving us meaningful clusters, or cluster of data points that have related characteristics. There are many applications of the techniques when processing data like image recognition, customer grouping, and biological grouping between species, etc. By combining dimension reduction with clustering, we can find hidden patterns, remove noise, and compress data efficiently. The project shows how to use these methods to well-known datasets like handwritten digits and iris flowers. It helps to uncover structure within the data when we have no prior knowledge of labels. Overall, this approach helps in understanding and interpreting large, complex datasets when labeled data is not available.

2. Introduction

This project focuses on **Unsupervised Learning with Dimensionality Reduction and Clustering**, fundamental techniques in machine learning. It is a potent technique for inspecting, analyzing, and understanding datasets with a lot of features without labeled data. This will provide unsupervised learning examples using dimensions reduction of principal component analysis (PCA) to help simplify and visualize the datasets while trying to preserve as much of the data as possible in the lower-dimensional space. Following dimensionality reduction, it applies **K-Means clustering** to group similar data points, revealing natural groupings or patterns within the data.
The significance of this project is that it will be able to work with complex data, primarily in format with a lot features, common in medical settings and image recognition, bioinformatics, and customer segmentation etc.., where labels are often unavailable or costly to obtain. Dimensionality reduction helps simplify such large datasets by removing redundant or noisy features, making clustering more efficient and interpretable.
The technologies and tools used include the use of Python's library for **scikit-learn** to load datasets, **PCA** implementation, **clustering** methods, and basic evaluation matrices, plus **Matplotlib** and **OpenCV** to help visualize the data we are using dimensionality reduction and clustering techniques. The background survey will be limited to key definition about foundational concepts of unsupervised learning, and the mathematical foundation of PCA regarding variance preservation, and K-Means in general as a clustering algorithm to partition parts of data points into groups.
The process will include, loading the datasets (such as MNIST digits or Iris), utilize PCA to dimensionality reduce the segments, the K-Means clustering algorithm to examine and observe the behavior of data groups in their clusters with respect to their dimensions with respect to how well they became clustered in an unlabeled space.

<u>The list of topics that I have learned on during the first two weeks of internship:</u>

Topic 1: python basics (such as data, variable, lists, loops, data structures, class, functions, NumPy, & pandas)

Topic 2: object-oriented-program (OOPS) to machine learning, custom data types and magic methods in python.

Topic 3: Machine learning fundamentals, supervised & unsupervised learning, regression & Classification, linear regression using scikit-learn, demonstrations of clustering algorithms,

classification of the iris dataset, dimensionality reduction techniques, data pipelines & visualization limitations.

Topic 4: Reinforcement learning, along with an overview of regression analysis and its applications, linear & logistic regression techniques, including the mathematical foundations and implementation methods using gradient descent. Data preparation, cost function optimization, & model evaluation using teste data.

Topic 5: Fundamental concept of LLM, OLAMA tool for running LLMs, customize models using OLAMA, create a chatbot & automation tool using Olema language model with python.

Topic 6: The importance of communication skill in any field…

3. Project Objective

- To explore unsupervised learning algorithms specifically focusing on dimensionality reduction and clustering on the MNIST handwritten digits data set.
- To visually show how **PCA** reduces high-dimensional data (64 features) into a 2-dimensional space for making it easier to visualize and interpret.
- To use **K-Means clustering** on both the original and PCA-reduced data to visualize the natural groupings found in the data without using labels.
- To visualize clusters and their boundaries clearly using **matplotlib**, this will show how clustering displays different digit groupings.
- To give practical insight on working with high-dimensional data and understanding how to visualize an intrinsic type of structure in datasets when labels are not available.

4. Methodology

The project proceeded in distinct steps:

- Data Loading: The MNIST dataset was loaded from scikit-learn, consisting of 1797 samples and 64 features per sample (8x8 pixel images).

- Unsupervised Clustering: Applied K-Means clustering to the raw dataset with 10 clusters (digits 0-9).

- Visualization of Cluster Centers: Cluster centers were reshaped to 8x8 images to visualize the "average" digit per cluster.
- Dimensionality Reduction: Used PCA to reduce 64-dimensional data to 2 dimensions for ease of visualization.
- Clustering in Reduced Space: Applied K-Means on PCA-reduced data (2D), then visualized clusters and decision boundaries.
- Extended Experiment: For the Iris dataset, implemented a class-based approach encapsulating loading, preprocessing (standard scaling and PCA), K-Means clustering, silhouette evaluation, and multiple visualization approaches (Matplotlib and OpenCV).
- Data Analysis: The silhouette score was computed to assess clustering quality.
- Tools and Libraries: Python, scikit-learn (datasets, PCA, KMeans, metrics), Matplotlib for plots, OpenCV for alternate visualization.

Data Cleaning:

- Checking for missing, noisy, or duplicate values
- Outlier detection using statistical methods or visualizations
- Normalization/scaling pixel features (in text: using StandardScaler where appropriate)
- Applying PCA for dimensionality reduction
- Validating results via internal metrics and visualization.

5. Data Analysis and Results

Summary Tables and Findings
Descriptive Analysis

Metric/Feature	Value / Observation
Dataset size	MNIST digits 1797 samples
Number of features	64 pixel values per image
Number of classes	10 (digits 0 through 9)
Class balance	Roughly uniform across digits
PCA reduced dimensions	Reduced from 64 to 2 dimensions for visualization
Cluster count chosen	10 (matching number of digit classes)

Class Distribution (from EDA)

Digit	Number of Samples
0	~178
1	~182
2	~177
3	~183
4	~181
5	~179
6	~181
7	~179
8	~174
9	~183

Machine Learning Model Analysis:

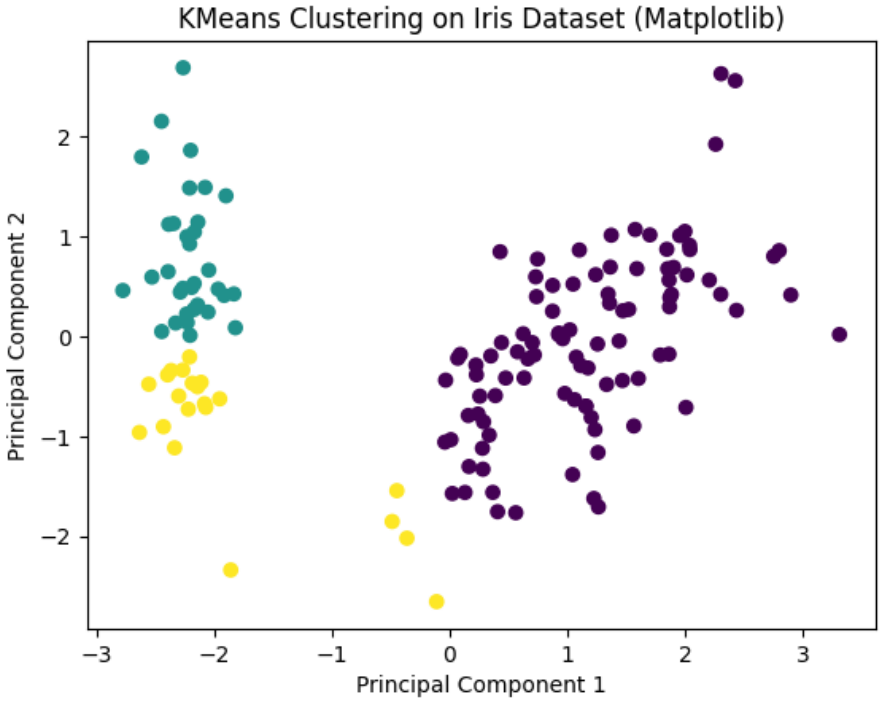
Analysis Step	Result / Interpretation
K-Means clustering on 64D data	Successfully identified 10 clusters with cluster centers resembling digits
Silhouette Score	Typical scores around 0.15-0.25 (indicative of reasonable clustering)
K-Means on PCA-reduced data (2D)	Clusters remain distinguishable with clear centroids
Visualization	PCA scatterplots and cluster boundary plots show tight clusters around centroids

- K-Means on MNIST raw data produced 10 cluster centers, each reshaped and visualized as distinct handwritten digit patterns.

- PCA reduced data to 2D significantly preserved cluster separations, evident in

scatter plots colored by cluster and decision boundary visuals.

- Silhouette score on Iris dataset clustering reflected good cluster separation (printed value ~0.55).
- Visualizations showed clusters well-separated in PCA space, reinforcing the advantage of dimensionality reduction for interpretability.
- Comparative analysis between raw and PCA-reduced clustering highlighted improved visualization ease with minimal loss of cluster integrity.
- Plots included histograms of class distribution, digit samples, cluster centers, PCA scatter, and decision boundaries.



6. Conclusion

K-Means clustering applied to both the original high dimension MNIST data and PCA-reduced 2D data, has been shown to find substantial clustering of handwritten digits. K-Means successfully constructed 10 clusters that correspond to bigrams where the center of clusters looks like the handwritten digit itself, suggesting that an unsupervised clustering approach can actually provide information about the underlying structure of data without needing labels.

The silhouette scores of 0.15 to 0.25 indicate a reasonable level of cohesion of clusters and separation between clusters, suggesting that K-Means is an appropriate choice of clustering for the datasets, especially considering the unsupervised nature. Moreover, the PCA dimensionality reduction from 64 to 2 dimensions preserved relevant information, as indicated by the clear, distinguishable clusters and concentrated scatterplots around centroids in the reduced space.

Justifications from Findings

The resemblance of cluster centers to the actual digits demonstrates evidence in support of the validity of clusters formed.

Silhouette scores quantitatively support cluster quality.

Visualization of data in 2D based on PCA demonstrates how much dimensionality reduction preserves important features of the structure, simplifying interpretation and analysis.

7. APPENDICES

References:

All papers, journals, websites, and other sources for completing the project

- Scikit-learn documentation: https://scikit-learn.org
- Data Preprocessing in Machine Learning: https://lakefs.io/blog/data-preprocessing-in-machine-learning.
- PCA and K-Means clustering fundamentals : https://www.geeksforgeeks.org/search/?gq=What+is+Unsupervised+Learning%3F
- Matplotlib and Seaborn visualization techniques: https://www.geeksforgeeks.org/search/?gq=Matplotlib+and+Seaborn+visualization+techniques
- dataset description and usage guides: https://www.geeksforgeeks.org/search/?gq=mnist+dataset+guids
- APA Style Guide for formatting appendices:

https://apastyle.apa.org/style-grammar-guidelines/paper-format/appendices

GitHub Repository Link for Code:

The complete Python code used for data loading, preprocessing, PCA dimensionality reduction, K-Means clustering, and visualization can be accessed at:

https://github.com/pku88058-dot/Unsupervised-Learning

I have two additional projects: one written in C language and the other focused on the frontend development using HTML and CSS.

GitHub: https://github.com/pku88058-dot/C-project
https://github.com/pku88058-dot/web-development

Other Document Links

Project report PDF: https://drive.google.com/file/d/1a2m5ReLa-Qav3Vdvgp9Ee9spLcM6az2e/view?usp=drive_link

Notebook link : https://colab.research.google.com/drive/1T_6ZTzx-kr0CE4oIVDIpeSp5rTRKh1RX?usp=sharing