**Unsupervised Learning with Dimensionality Reduction and Clustering**

Sayan Pal

B.Sc. cs (Honours) Batch 1

Bangabasi College

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

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

1. **Abstract**

This internship project explores the application of unsupervised learning techniques to high-dimensional datasets, focusing on dimensionality reduction and clustering. Principal Component Analysis (PCA) was employed to reduce the feature space, enabling effective visualization and improving clustering performance. K-Means clustering was then applied to both the original and reduced datasets to identify inherent groupings without labeled data. The project utilized the Digits and Wine datasets to demonstrate the methodology, with results visualized using Matplotlib and OpenCV. An object-oriented approach was implemented to streamline the workflow, making the system modular and reusable. Evaluation metrics such as silhouette score were used to assess clustering quality. The findings highlight the power of combining PCA with clustering algorithms to uncover meaningful patterns in complex data.

1. **Introduction**

Unsupervised learning is a core area of machine learning that focuses on discovering hidden patterns in data without predefined labels. This project explores the synergy between dimensionality reduction techniques and clustering algorithms to analyze high-dimensional datasets. By reducing the number of features and grouping similar data points, we gain insights into the structure of complex datasets.

1. **Project Objective**

 Understand the principles of unsupervised learning and its applications in data analysis.

 Apply dimensionality reduction techniques, specifically PCA, to simplify high-dimensional datasets.

 Implement K-Means clustering to identify natural groupings within unlabeled data.

 Evaluate clustering performance using metrics such as silhouette score.

 Visualize reduced data and clustering results using Matplotlib and OpenCV.

 Develop a modular, object-oriented pipeline for clustering and visualization.

 Analyse and interpret the structure of clusters to gain insights from complex datasets.

1. **Methodology**

**4.1 Dataset Selection**

Two datasets were used:

**Digits Dataset**: Contains 64 features representing 8×8-pixel images of handwritten digits.

**Wine Dataset**: A high-dimensional dataset with 13 chemical features of wine samples from three cultivars.

**4.2 Dimensionality Reduction**

Principal Component Analysis (PCA) was used to reduce the dimensionality of both datasets to 2 components. This enabled effective visualization and improved clustering performance.

**4.3 Clustering**

K-Means clustering was applied to both the original and PCA-reduced datasets. The number of clusters was set to:

10 for the Digits dataset (representing digits 0–9).

3 for the Wine dataset (representing three wine classes).

**4.4 Visualization**

Matplotlib was used to visualize clusters and centroids.

Decision boundaries were plotted using mesh grids.

OpenCV was used for alternate visualization, adapted for Google Colab using cv2\_imshow.

**5.Implementation**

**5.1 Digits Dataset**

* Loaded using sklearn.datasets.load\_digits.
* PCA reduced the 64-dimensional data to 2D.
* K-Means clustered the data into 10 groups.
* Cluster centers were visualized as digit images.
* Decision boundaries were plotted to show cluster separation.

**5.2 Wine Dataset**

* Loaded using sklearn.datasets.load\_wine.
* Data was standardized using StandardScaler.
* PCA reduced the 13-dimensional data to 2D.
* K-Means clustered the data into 3 groups.
* Silhouette score was used to evaluate clustering quality.

**5.3 Object-Oriented Clustering**

A class YourDataClustering was created with methods for:

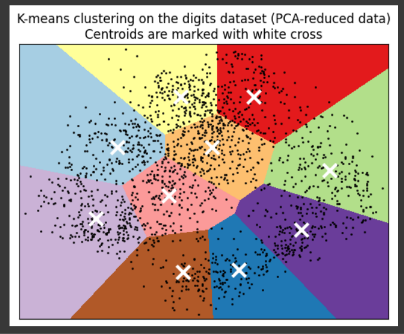
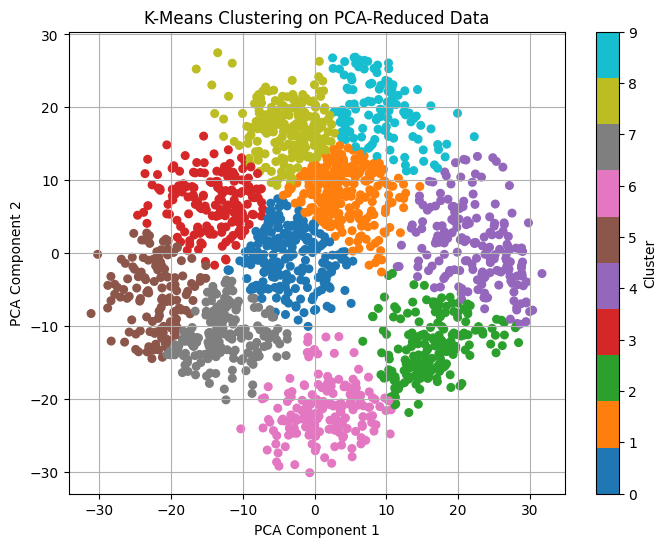
* Loading and preprocessing data.
* Applying PCA and KMeans.
* Evaluating clusters using silhouette score.
* Visualizing results with Matplotlib and OpenCV.

**6. Data Analysis and Results**

 Digits **Dataset**: PCA + KMeans revealed clear digit-like cluster centers. Visualization showed distinct groupings.

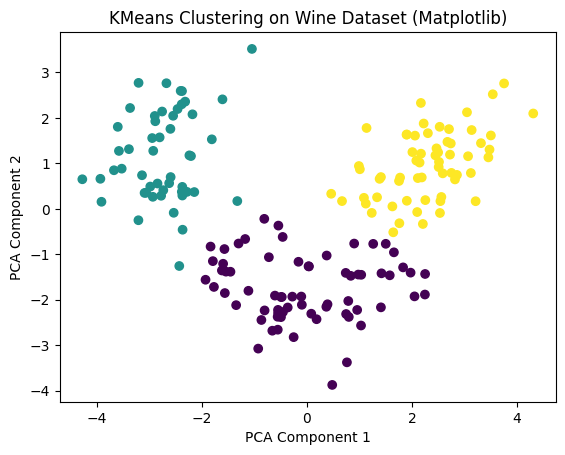
 Wine **Dataset**: Achieved a silhouette score of approximately 0.3–0.5, indicating moderate clustering quality. PCA helped separate the wine classes effectively.

 Object-oriented implementation made the pipeline modular and reusable.



Output figure1: K-Means clustering on PCA-Reduced Data.

Output figure 2: - for K-Means Clustering on the Digits dataset (PCA-Reduced Data) centroids are marked with white cross.



Output figure 3: - for K-Means Clustering on Wine Dataset(matpotlid).

**7.Conclusion**

This project demonstrated the effectiveness of combining dimensionality reduction with clustering in unsupervised learning. PCA improved both visualization and clustering performance. The object-oriented approach provided a clean and extensible framework for future experimentation.

**8. APPENDICES**

 Jolliffe**, I. T., & Cadima, J. (2016).** Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

 MacQueen**, J. (1967).** Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281–297.

 Pedregosa**, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).** Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

 **Rousseeuw, P. J. (1987).** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

 van **der Maaten, L., & Hinton, G. (2008).** Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579–2605.

 Bishop**, C. M. (2006).** Pattern Recognition and Machine Learning. Springer.

 Hastie**, T., Tibshirani, R., & Friedman, J. (2009).** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.