Project

Wine Dataset : Unsupervised Learning with Dimensionality Reduction and Clustering

Student Name : Manolina Das

Institute : University of Kalyani

Course : B.Tech (CSE)

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 project explores the application of unsupervised learning techniques to analyze high-dimensional datasets. The work was carried out in two phases using two different datasets.

In the first phase, the **MNIST handwritten digits dataset** was used. K-Means clustering was applied directly to the 64-dimensional digit data, and the resulting cluster centers were found to resemble digit patterns, showing how unsupervised learning can capture meaningful structures even without labels. To improve interpretability and efficiency, Principal Component Analysis (PCA) was then applied to reduce the high-dimensional digit data to two dimensions, enabling clear visualization of clusters.

In the second phase, the focus shifted to the **Wine dataset** from scikit-learn, which contains 13 chemical features of 178 wine samples from three cultivars. The methodology involved data loading, feature scaling, applying PCA to reduce the feature space to two dimensions, and then performing K-Means clustering. The quality of the resulting clusters was evaluated using the silhouette score, which indicated strong separation and compactness of the clusters.

The findings from both datasets demonstrate that the combination of PCA and K-Means is highly effective in uncovering the underlying structure of data, successfully grouping samples into distinct and interpretable clusters. Visualizations using Matplotlib and OpenCV further reinforced the clarity of these results.

2 Introduction

Unsupervised learning is a critical branch of machine learning where algorithms learn patterns from untagged data. In many real-world scenarios, datasets are high-dimensional, making them difficult to analyze and visualize. This project addresses this challenge by combining two powerful unsupervised techniques: dimensionality reduction and clustering. The relevance lies in its application to exploratory data analysis, where the goal is to discover hidden structures, anomalies, or groups in data without prior knowledge. The primary technology used is Python, with its scientific computing libraries such as scikit-learn, pandas, and Matplotlib. By first reducing the 13-dimensional Wine dataset to a 2D space using PCA, we can effectively visualize the data and apply clustering algorithms like K-Means to identify natural groupings. This procedure is fundamental in fields like bioinformatics, market segmentation, and anomaly detection.

During the first two weeks of the internship, training was provided on the following foundational topics:

- Introduction to Python for Data Science (NumPy, Pandas)
- Data Visualization with Matplotlib and Seaborn
- Fundamentals of Machine Learning: Supervised vs. Unsupervised Learning
- In-depth study of Clustering Algorithms (K-Means, Hierarchical)
- Principles of Dimensionality Reduction (PCA)
- Model Evaluation Metrics for Unsupervised Learning (e.g., Silhouette Score)

3 Project Objective

The primary objectives of this project were to implement and evaluate a complete unsupervised learning pipeline. The specific goals are outlined below:

- To select and load a suitable high-dimensional dataset and perform basic exploratory data analysis (EDA) to understand its characteristics.
- To apply Principal Component Analysis (PCA) to reduce the dimensionality of the dataset from its original feature space to two principal components for visualization purposes.
- To implement the K-Means clustering algorithm on the dimensionally-reduced data to partition it into a predefined number of clusters (k=3).
- To evaluate the quality and separation of the identified clusters quantitatively using the silhouette score.
- To visualize the final clusters and their centroids on a 2D scatter plot to qualitatively assess the performance of the model.

4 Methodology

The project was executed following a structured pipeline, using Python 3 and its associated data science libraries (scikit-learn, pandas, Matplotlib, OpenCV). The entire process, from data acquisition to final visualization, is outlined below and illustrated in the flowchart in Figure 1.

1. **Data Collection:** The Wine dataset, a classic benchmark dataset, was loaded directly from the 'sklearn.datasets' module. This dataset contains 178 samples with 13 continuous features.
2. **Data Pre-processing:** Since PCA and K-Means are sensitive to feature scales, the data was standardized using 'StandardScaler' from scikit-learn. This process transforms each feature to have a mean of 0 and a standard deviation of 1.
3. **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to the scaled data. The number of components was set to 2 to project the 13-dimensional data onto a 2D plane, capturing the maximum possible variance.
4. **Clustering:** The K-Means algorithm was implemented on the 2D PCA-transformed data. The number of clusters was set to 3, corresponding to the known number of wine cultivars in the dataset.
5. **Model Evaluation:** The performance of the clustering was measured using the silhouette score, which provides a metric for how well-separated the clusters are.
6. **Visualization:** The final clusters were visualized using both Matplotlib and OpenCV to display the data points colored by their assigned cluster and to mark the cluster centroids.

5 Data Analysis and Results

The analysis was performed in two stages: initial exploratory analysis and the results from the machine learning pipeline.

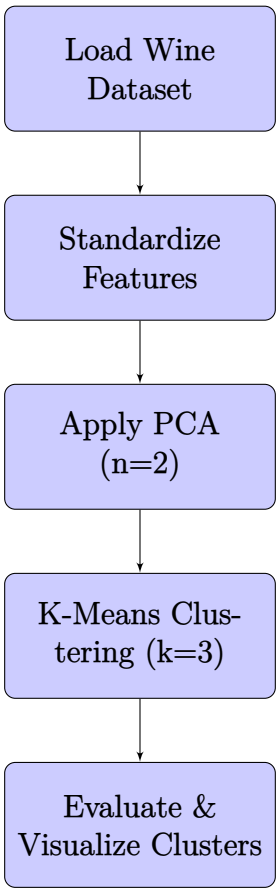

Figure 1: Project Methodology Flowchart.

5.1 Descriptive Analysis

Initial exploration of the Wine dataset revealed 178 samples and 13 numeric features with no missing values. A correlation heatmap (Figure 2) was generated to understand the relationships between features. It showed several highly correlated features, such as 'flavanoids' and 'total_phenols', justifying the use of PCA to reduce multicollinearity and simplify the model. In addition, histograms were plotted for each feature to examine their individual distributions (Figure 3) .

5.2 Inferential Analysis and Model Results

After applying the PCA and K-Means pipeline, the model achieved a **Silhouette Score of 0.565**. A silhouette score ranges from -1 to 1, where a value closer to 1 indicates that the clusters are dense and well-separated. A score of 0.565 suggests that the clusters are reasonably distinct.

The visualization of the clusters on the two principal components (plotted using MatplotLib, Figure 4 and OpenCV, Figure 5) provides a clear qualitative confirmation of this result. The three clusters are visually distinct with minimal overlap, and the data points are grouped tightly around their respective centroids. This confirms that the unsupervised pipeline successfully identified the underlying structure in the data, corresponding to the three different cultivars of wine.

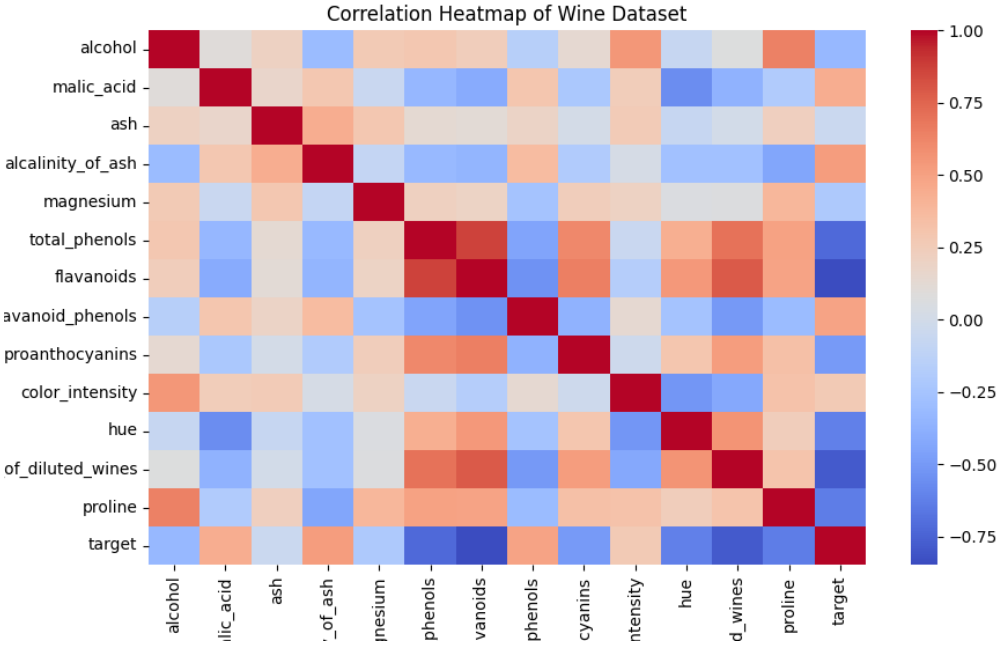

Figure 2: Correlation Heatmap of the Wine Dataset Features.

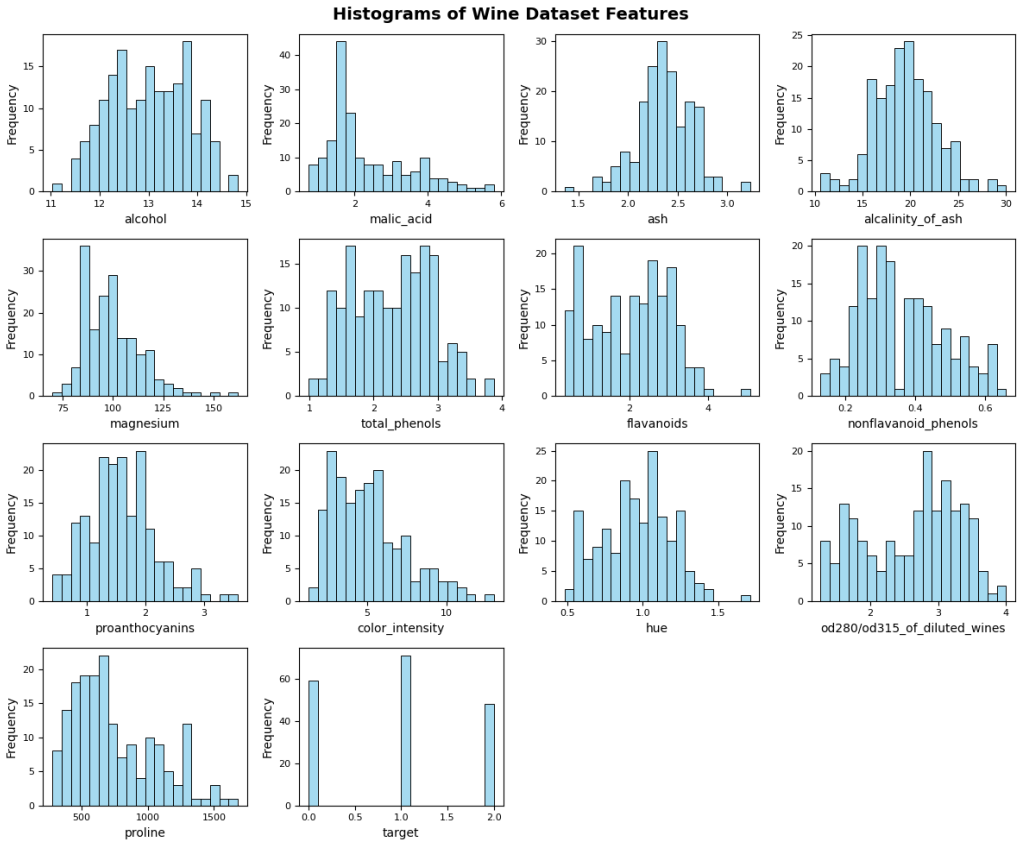

Figure 3: Histograms plotted for individual features.

6 Conclusion

This project successfully demonstrated a complete unsupervised learning workflow for analyzing high-dimensional data. By integrating data preprocessing, PCA for dimensionality reduction, and K-Means for clustering, we were able to uncover and visualize the intrinsic groupings within

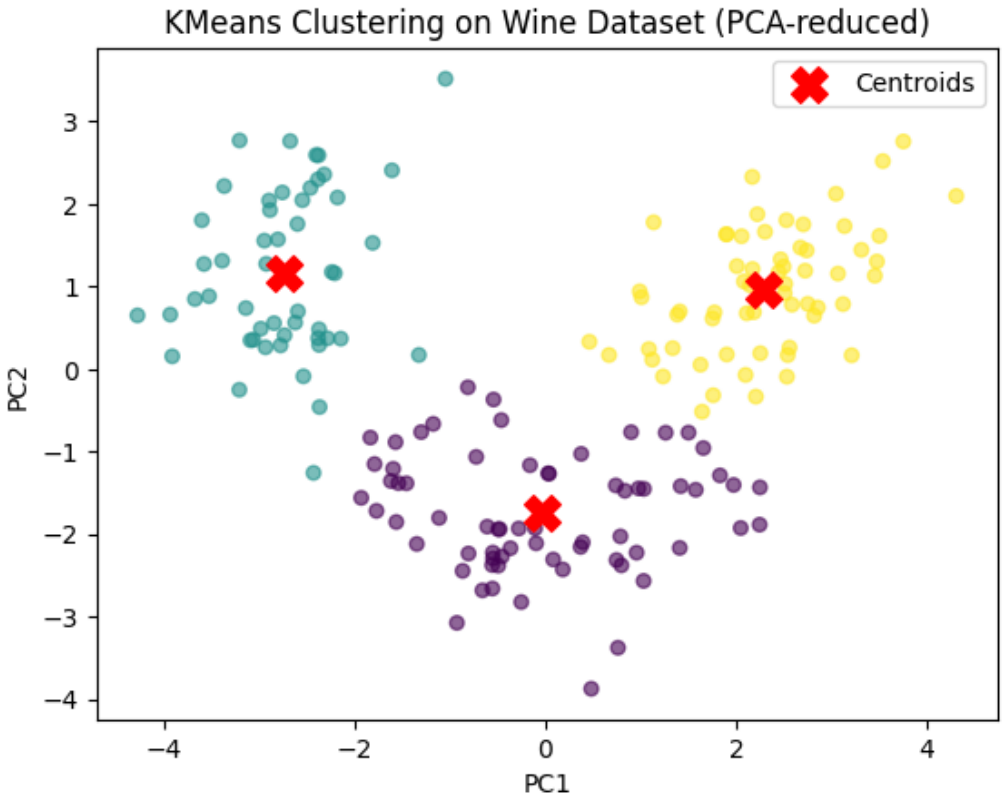

Figure 4: K-Means Clustering on PCA-Reduced Wine Dataset, using MatplotLib.

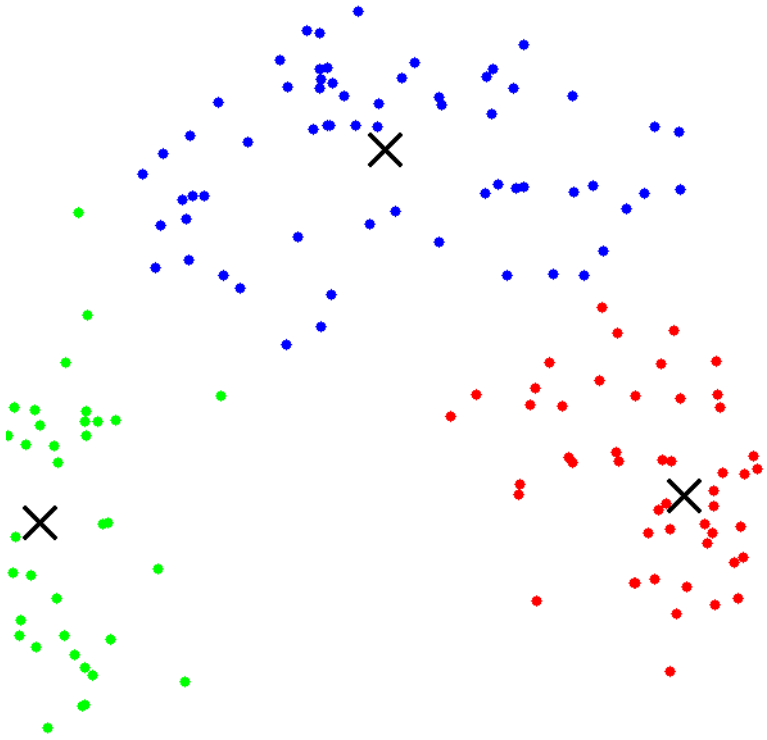

Figure 5: K-Means Clustering on PCA-Reduced Wine Dataset, using OpenCV.

the Wine dataset without relying on any labels. The resulting silhouette score of 0.565 and the clear visual separation of clusters validate the effectiveness of this approach. The findings confirm that unsupervised methods are powerful tools for exploratory data analysis, capable

of revealing hidden patterns that can inform further investigation.

For future work, this pipeline could be extended by exploring other clustering algorithms (e.g., DBSCAN, Agglomerative Clustering) or dimensionality reduction techniques (e.g., t-SNE) to compare their performance on this dataset. Additionally, applying this methodology to more complex, real-world datasets could yield valuable insights in various domains.

Appendices

A. References

- Scikit-learn: Machine Learning in Python - scikit-learn.org
- Wine Dataset - scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html

B. GitHub link for the Jupyter Notebook

The Jupyter Notebook developed for this project is available at the following GitHub repository: github.com/manolina-13/ProjectIdeas_TIH_ISI

C. Presentation Video Link

The required presentation screen recording has been uploaded to YouTube and is available on : YouTube