{AUTUMN INTERNSHIP PROJECT REPORT FORMAT}

Unsupervised Learning and Dimensionality Reduction on the Wine Dataset

Sumedha Sarkar,

M.Sc. in Statistics, University of Delhi

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

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

Abstract :

This project explores the Wine dataset using Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), and clustering methods to uncover hidden structures within the data. The dataset consists of 178 wine samples with 13 chemical features. After conducting descriptive and visual analysis, PCA was applied to reduce dimensionality and visualize separability of wine classes. Clustering techniques such as KMeans were implemented to validate natural groupings of the data. Comparative results highlight the effectiveness of PCA in improving interpretability of clustering. The study demonstrates the application of machine learning techniques in identifying patterns in multivariate datasets and offers insights for classification and dimensionality reduction.

Introduction :

The classification of wines based on their chemical composition is an important problem in both agriculture and food science. By leveraging machine learning, we can analyze patterns that distinguish different wine cultivars. The dataset used in this project, the Wine dataset from UCI Machine Learning Repository, contains continuous variables such as alcohol, flavanoids, phenols, and color intensity.

The relevance of this project lies in:

- Applying modern data analysis tools to a real-world dataset.
- Demonstrating dimensionality reduction and clustering in practice.
- Exploring relationships between variables using EDA.

Background material survey: Previous studies show that PCA is effective in reducing redundancy in high-dimensional datasets, and clustering is often applied to validate group structures without labels.

Procedure used:

1. Data loading and cleaning.
2. Exploratory Data Analysis (EDA) to understand distributions, correlations, and group patterns.
3. PCA for dimensionality reduction and visualization.
4. KMeans clustering to discover inherent patterns.

Purpose: To understand the data, test the usefulness of PCA for dimensionality reduction, and examine clustering performance on reduced dimensions.

Training topics during first two weeks of internship:

The internship primarily focussed on self-paced learning, and mentor-guided project work covering the following topics:

- Data Structures & Algorithms

- Python Programming
- Basic Statistics
- Machine Learning (ML)
- Artificial Intelligence (AI)
- Large Language Models (LLMs)
- Communication Skills

Project Objective :

The main objectives of this project are:

- To perform exploratory data analysis on the Wine dataset and identify feature patterns.
- To apply PCA and evaluate its effectiveness in dimensionality reduction.
- To perform clustering using KMeans and interpret the results.
- To visualize class separability in reduced dimensions.
- To compare clustering performance on original and PCA-reduced features.

Methodology :

The project followed a structured data analysis workflow using the Wine dataset from sklearn.datasets. The methodology can be summarized as follows:

1. **Library Import**
    - Essential libraries such as Pandas, Seaborn, Matplotlib, and Scikit-learn were imported to support data analysis, clustering, and visualization tasks.
2. **Dataset Loading**
    - The Wine dataset was loaded using load_wine(as_frame=True) from Scikit-learn.
    - A DataFrame was constructed containing all chemical features of wine samples along with the target wine class labels.
3. **Data Preparation**
    - Features (chemical properties) were separated from the target column (WineClass).
    - The target column was dropped for unsupervised learning since clustering does not use labels.
4. **Data Scaling**
    - To ensure comparability among features with different units, data standardization was applied using **StandardScaler**, which transforms features into zero mean and unit variance.
5. **KMeans Clustering**
    - The **Elbow Method** was employed by plotting Within-Cluster Sum of Squares (WCSS) for cluster numbers ranging from 1 to 10.
    - From the elbow plot, the optimal cluster number was chosen as **3**, aligning with the known classes in the dataset.

    - KMeans was fitted with n_clusters=3, and cluster labels were assigned to each sample.

6. **Dimensionality Reduction (PCA)**
    - Principal Component Analysis (PCA) with 2 components was applied to reduce the high-dimensional dataset to two dimensions for effective visualization.
    - Both wine samples and cluster centers were projected into PCA space.
7. **Visualization**
    - A scatter plot was generated with PCA1 and PCA2 as axes, where points were colored based on cluster labels.
    - Cluster centroids were highlighted with red "x" markers, visually demonstrating separation of clusters.

Methodology Flowchart

```
Data Collection
      ↓
Data Preprocessing (Scaling, Cleaning)
      ↓
Exploratory Data Analysis (EDA)
      ↓
Dimensionality Reduction (PCA)
      ↓
Clustering (KMeans)
      ↓
Result Evaluation & Visualization
```

Github Link :

https://github.com/SumedhaSarkarStat/Project/blob/main/Unsupervised_Learning_With_Dimensionality_Reduction%20(1).ipynb

Data Analysis and Results :

Descriptive Analysis

- The dataset consisted of multiple numerical features representing chemical compositions of wine.
- Standardization ensured that all variables contributed equally to clustering.

Clustering Results

- The **Elbow Method** indicated that 3 clusters were appropriate.
- KMeans with 3 clusters successfully grouped the wine samples into distinct clusters.

PCA Visualization

- PCA reduced the 13-dimensional dataset to **2 principal components**, capturing maximum variance for visualization.
- A scatter plot clearly showed three clusters with reasonable separation.

- Cluster centroids in PCA space confirmed the stability of the clustering solution.

➢ **Summary of Results:**

- The dataset could be meaningfully grouped into 3 clusters.
- PCA allowed effective 2D visualization of high-dimensional clustering results.
- Visual inspection showed clusters were fairly distinct, validating the choice of 3 clusters.

Conclusion :

The project successfully demonstrated the use of **unsupervised learning (KMeans clustering)** combined with **dimensionality reduction (PCA)** on the Wine dataset. Key conclusions include:

- The Elbow Method suggested 3 clusters, which is consistent with the actual number of wine classes.
- Standardization played a crucial role in achieving balanced clustering across all features.
- PCA effectively reduced dimensionality while retaining structure, enabling clear visualization of clusters.
- The results indicate that clustering algorithms, when combined with dimensionality reduction, can provide meaningful insights even without class labels.

APPENDICES :

- **References**

- Scikit-learn Documentation: https://scikit-learn.org/stable/datasets/toy_dataset.html
- Seaborn Documentation: https://seaborn.pydata.org/
- Matplotlib Documentation: https://matplotlib.org/

- **Github Link**

https://github.com/SumedhaSarkarStat/Project