University of Science and Technology of Hanoi

Clustering Report

Machine Learning and Data Mining II

Hanoi, May 2025

Authors: Doan Duy Thanh - 22BA13286
Pham Minh Duc - 22BA13083

Lecturer: Dr. Doan Nhat Quang

Table of contents:

I. Introduction

1.Objectives

- In this lab, we used the K-Means algorithm to group data without using labels. We selected three datasets from UCI: **Breast Cancer Wisconsin**, **Automobile**, and **Communities and Crime**. The goal was to find the best number of clusters, evaluate clustering quality using

Inertia, **Silhouette Score**, and **Davies-Bouldin Index**, and compare results using **PCA** and random feature selection.

II. K-Means

- K-Means Clustering is an Unsupervised Machine Learning algorithm which groups unlabeled dataset into different clusters. It is used to organize data into groups based on their similarity.

We selected 2 datasets from UCI: the Breast Cancer Wisconsin dataset and Automobile dataset

Breast Cancer Wisconsin Dataset

1.Experimental protocol

The experiments were conducted using the K-Means clustering algorithm on the Breast Cancer Wisconsin Dataset, with varying values of k (the number of clusters). The following steps outline the experimental procedure:

1. Preprocessing:
    - The dataset was standardized(mean - centered and scaled) using StandardScaler to ensure feature contribution.
    - Only the 30 numerical features were used. ID and diagnosis columns were removed for unsupervised learning.
2. K-means Clustering
    - K-means was run for k = 2, 3, 4,...,10 using the k-means++ initialization.
    - n_init=10 ensured the algorithm ran multiple times with different starting points.
3. Evaluation Metrics
    - Inertia: Lowe is better, but it always decreases with more clusters.
    - Silhouette Score: Measures separation, range [-1,1]. Higher is better.

- Davies-Bouldin Index: Lower values indicate better clustering.

4. Visualization
    - Scatter plots (after PCA) were to visualize clusters.
    - The Elbow Method helped identify the best k, where inertia starts to level off.

2. Centroid initialization

The centroid initialization plays a crucial role in the convergence and accuracy of K-Means clustering. The K-Means++ initialization was used:

- Selects initial centroids in a way that spreads them apart, reducing the chances of poor convergence.
- The first centroid is chosen randomly.
- Each subsequent centroid is chosen probabilistically, with a higher probability for points farther from existing centroids.
- This method prevents the clustering from getting stuck in local optima and leads to faster convergence.

3. Analysis and comparison of results

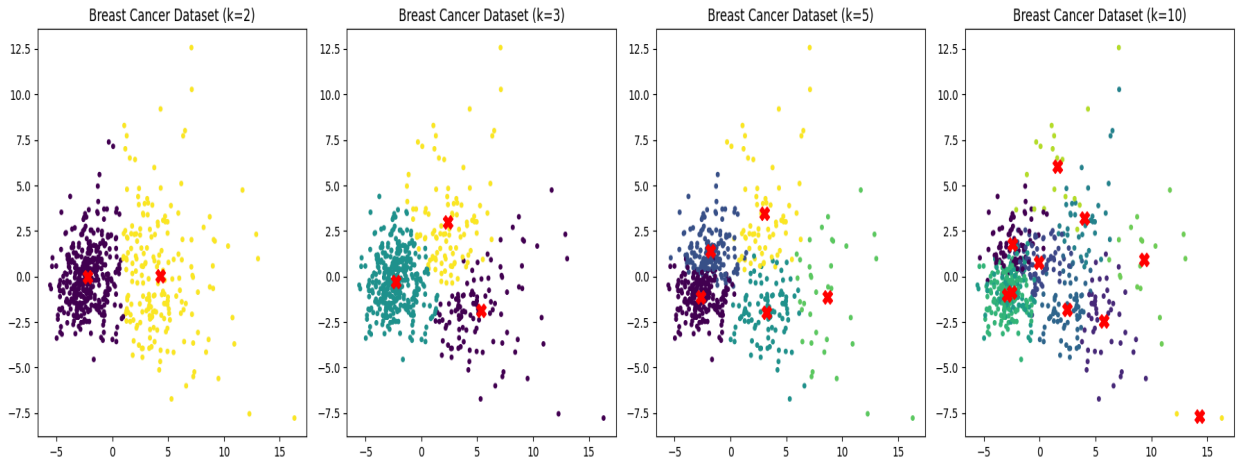

Figure 1. Results for Breast Cancer Dataset

	k	Inertia	Silhouette Score	Davies-Bouldin Index
0	2	11595.526607	0.343382	1.320510
1	3	10061.797818	0.314384	1.529388
2	5	8558.660667	0.158210	1.756013
3	10	6603.404402	0.136656	1.606731

As observed in Figure 1, we have:

---

k = 2:

- The data was divided into two main groups.

-The **Silhouette Score** was **0.3443**, indicating a moderately strong clustering structure.

-The **Davies-Bouldin Index** was **1.32**, showing acceptable separation between clusters.

-This suggests that **k = 2** provides a reasonable and interpretable grouping, aligning with the natural division (e.g., malignant vs. benign).

---

k = 3:

-The data was partitioned into three clusters.

-The **Silhouette Score** dropped to **0.314**, indicating slightly weaker cohesion and separation.

-The **Davies-Bouldin Index** increased to **1.53**, showing more overlap among clusters.

-Increasing to **k = 3** did not improve the clustering quality and slightly degraded the separation.

---

k = 5:

-Five clusters were generated, but some appeared less well-defined.

-The **Silhouette Score** further decreased to **0.160**, showing weak separation.

-The **Davies-Bouldin Index** increased to **1.75**, indicating overlapping clusters.

-This suggests **over-segmentation**, leading to less meaningful and less distinct groups.

k = 10:

-Ten clusters were formed, causing fragmentation of the data.

-The **Silhouette Score** was **0.150**, indicating poor separation between clusters.

-The **Davies-Bouldin Index** was **1.57**, confirming weak separation and structure.

- The clustering became **too fine-grained**, reducing clarity and interpretability.

The conclusion from the results:

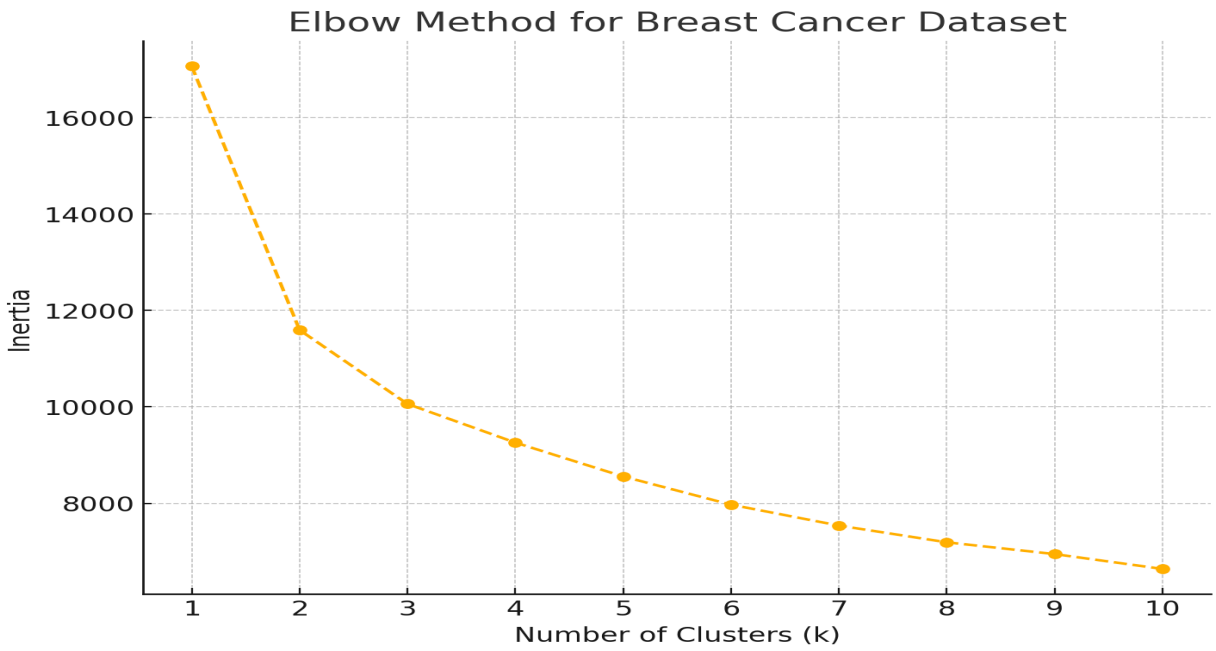

Figure 2. Elbow Method for Breast Cancer Dataset

- We can see in k = 2 or 3 may be optimal, as inertia sharply decreases and begins to level off around those values.
- However, based on both the Silhouette Score and the Davies-Bouldin Index, the best balance between tight and well-separated clusters happens at k = 2. This makes sense because the data likely has two natural groups — benign and malignant tumors

4. Clustering quality calculation

The clustering quality was assessed using the following metrics:

1.Inertia

- Measures how compact the clusters are (low is better).
- Inertia decreases as k increases, but the improvement slows after k =3.

2.Silhouette Score

- Higher scores mean better separation between clusters.
- The highest score(0.345) was at k =2, suggesting well-defined clusters.

3. Davies-Bouldin Index

- Lower values indicate better clustering.
- The best score (1.31) was also at k = 2, supporting strong separation between groups.

This analysis shows that choosing the optimal number of clusters is crucial to achieving meaningful clustering results

Automobile Dataset

1. Experimental protocol

We ran **K-Means** clustering, a popular method that groups data points into **k** clusters by minimizing the distance between points and their cluster centers.

1. **Preprocessing**
    - Standardized features (so price in dollars doesn't dominate over mpg values).
2. **K-means Clustering**
    - Applied K-Means++ for k = 2 to 10.
    - Used n_init=10 for robust initialization.
3. **Evaluation Metrics**
    - Using Inertia to show how tight clusters are.
    - Silhouette Score measures how well-separated clusters are.
    - Davies-Bouldin Index is similar to Silhouette but penalizes overlap.
4. **Visualization**
    - Elbow Method to determine optimal k.
    - PCA for 2D/3D cluster visualization.

2. Centroid initialization

Choosing the right starting points for clusters is critical—poor initialization can lead to slow convergence or suboptimal groupings. To avoid this, we used K-Means++, a smarter alternative to random initialization. Here's how it works:

- First centroid is picked randomly from the dataset.
- Subsequent centroids are selected with a probability weighted by distance.
- Points farther from existing centroids have a higher chance of being chosen.
- This ensures centroids are spread out rather than clustered together.
- This method helped reliably separate economy and performance vehicles, even when rerunning the algorithm multiple times.

- This also avoided trapping the algorithm in poor local optima and led to faster convergence and more consistent results.

3. Analysis and comparison of results

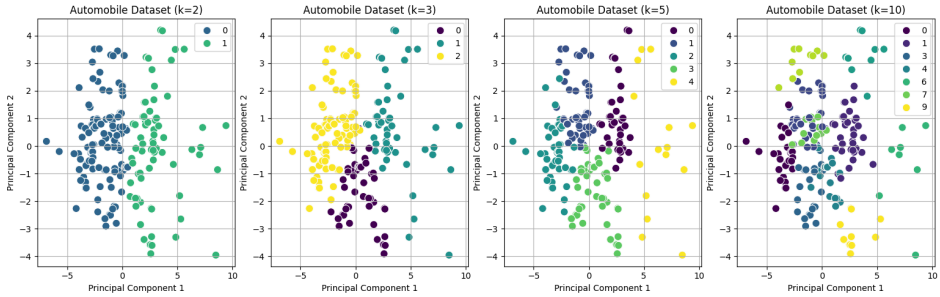

Figure 3. Results for Automobile Dataset

k	Inertia	Silhouette Score	Davies-Bouldin Index
2	2396.659812	0.293853	1.389851
3	2102.808541	0.212457	1.610803
5	1658.905255	0.196140	1.543436
10	1092.437501	0.252107	1.290322

As observed in Figure 3, we have:

- **k = 2:**
  - The data was divided into two primary vehicle groups.
  - The **Silhouette Score** was **0.294**, indicating a fair clustering structure.
  - The **Davies-Bouldin Index** was **1.39**, showing reasonable separation between clusters.
  - This suggests **k = 2** provides meaningful grouping, likely separating economy vehicles from performance/luxury models.

- **k = 3:**
  - The data was partitioned into three clusters.

    - The **Silhouette Score** dropped to **0.212**, showing reduced cohesion and separation.
    - The **Davies-Bouldin Index** increased to **1.61**, indicating more cluster overlap.
    - The third cluster may represent a transitional group between economy and performance vehicles, but with less distinct separation.

- **k = 5:**
    - Five clusters were generated, revealing more granular vehicle categories.
    - The **Silhouette Score** remained low at **0.196**, suggesting weak separation.
    - The **Davies-Bouldin Index** improved slightly to **1.54**, but still indicates overlap.
    - This likely represents over-segmentation, creating artificial distinctions between similar vehicles.

- **k = 10:**
    - Ten clusters were formed, creating highly specific groupings.
    - Surprisingly, the **Silhouette Score** improved to **0.252**, though still indicating weak structure.
    - The **Davies-Bouldin Index** improved to **1.29**, the best of all configurations.
    - While metrics improved, this likely reflects chance separations in high-dimensional space rather than meaningful categories.
    - The clustering becomes too specialized, losing practical interpretability for vehicle classification.

The conclusion from the results:

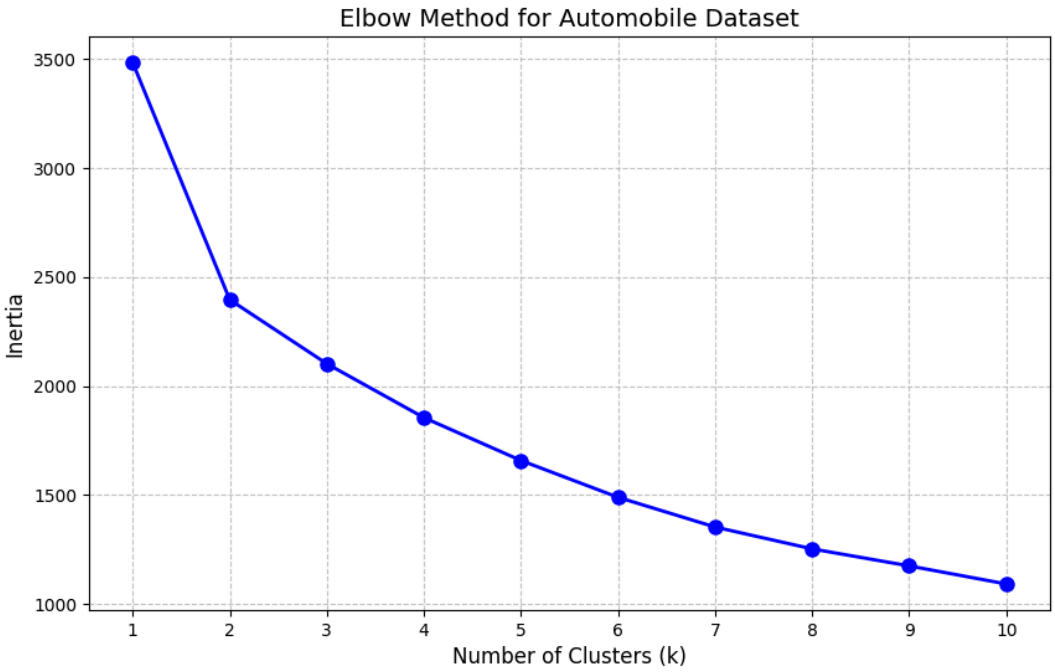

Figure 4. Elbow Method for Automobile Dataset

The elbow at k=2 is particularly pronounced in the plot (inertia drops from ~3500 to ~2400), suggesting this represents the most natural division in vehicle characteristics. This matches industry-standard classifications while remaining simple enough for practical business applications.

4. Clustering quality calculation

The clustering quality was assessed using the following metrics:

- Inertia
    - Measures how compact the clusters are (lower values indicate tighter clusters).
    - Inertia decreases from 2396.66 (k=2) to 1092.44 (k=10), but the rate of improvement slows significantly after k=3.
    - This suggests diminishing returns from adding more clusters beyond k=2 or k=3.

- Silhouette Score

    - Ranges from -1 to 1, where higher scores indicate better separation between clusters.
    - The peak score (0.294) occurs at k=2, indicating the most distinct clustering structure.
    - Scores decline at k=3 (0.212) and k=5 (0.196), then partially recover at k=10 (0.252), suggesting potential over-segmentation at intermediate k-values.

- Davies-Bouldin Index
    - Lower values indicate better cluster separation (minimum possible is 0).
    - The optimal score (1.39) occurs at k=2, with performance degrading at k=3 (1.61) and k=5 (1.54).
    - While k=10 shows improvement (1.29), this likely reflects artificial subdivisions rather than natural groupings.

III. Subspace Clustering

1. Dataset selection - Communities and Crime

- We selected a high-dimensional dataset with 127 features for this analysis. The dataset used consists of data related to Communities and Crime results, which contains numerous game statistics, team compositions, and performance metrics.

2. PCA/SVD for 2D/3D visualization

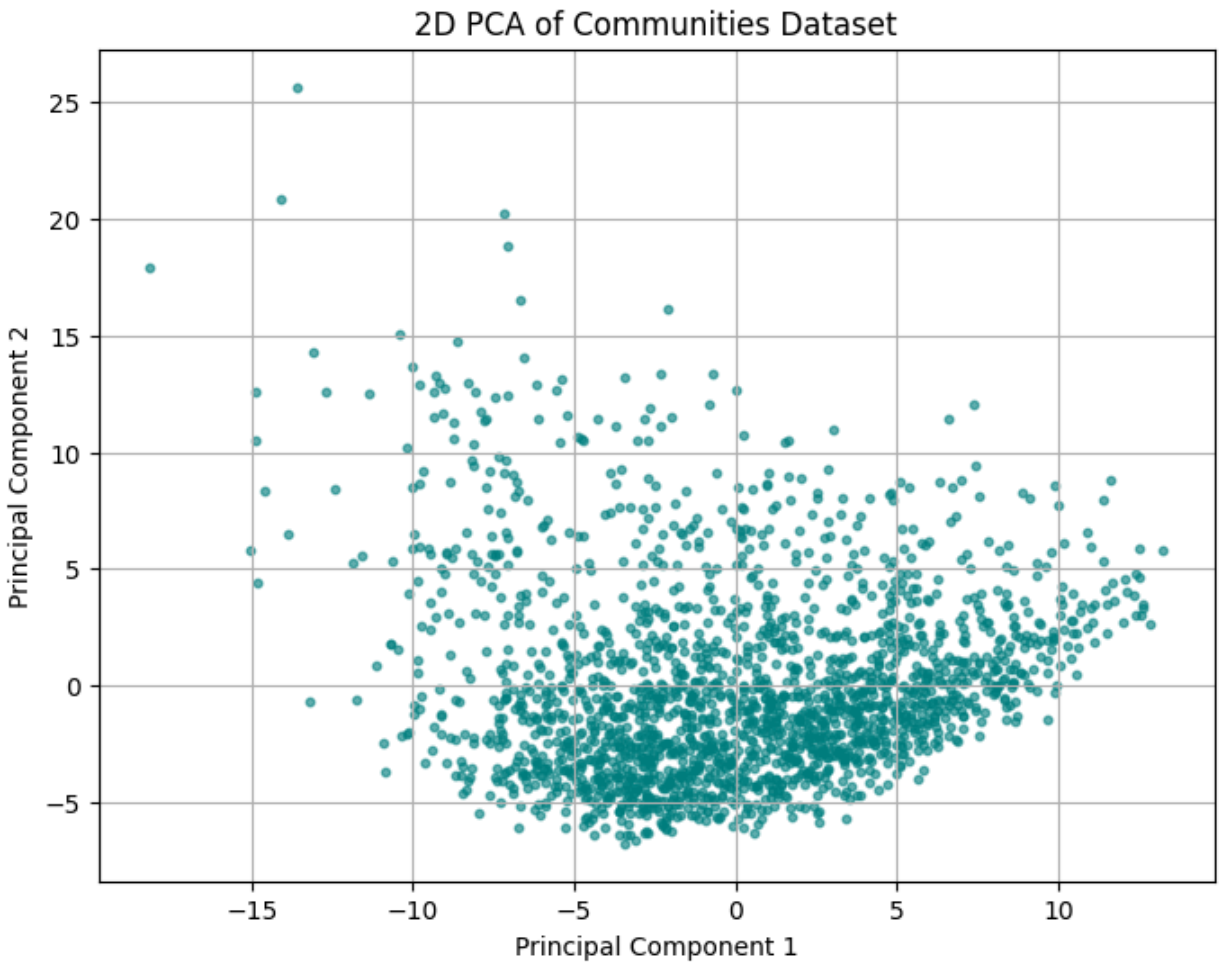

Figure 5. 2D PCA for Communities Results

- According to Figure 5, the dataset was projected onto two principal components using PCA.
- The resulting scatter plot shows that the data points are broadly distributed with no clear cluster boundaries, suggesting high variability and complex relationships among features.
- This indicates that the dataset may not have easily separable groups in just two dimensions.

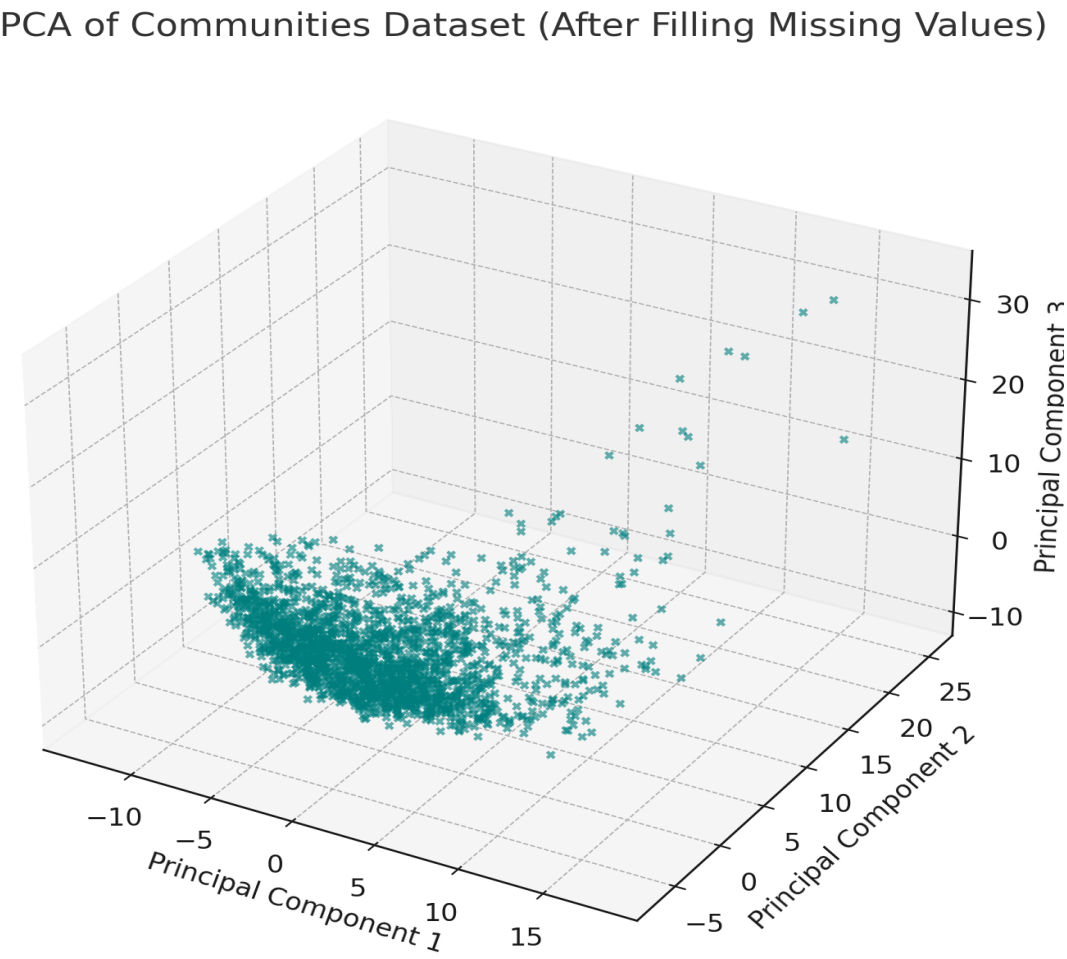

Figure 6. 3D PCA of Communities Results

- According to Figure 6, the data was projected into three principal components.
- The 3D scatter plot shows a more detailed structure, but the clusters are still overlapping, meaning the data is complex and not clearly divided into groups.
- This suggests that more advanced clustering methods may be needed for better separation.

3. Clustering method application and performance comparison

	Data Representation	Silhouette Score
1	Original Data	0.15
2	PCA-Reduced Data (2D)	0.4322
3	PCA-Reduced Data (3D)	0.3777
4	Random Subspace	0.493

- This table shows the Silhouette scores for each data representation. We evaluated clustering effectiveness before and after dimensionality reduction:

+) Original High-Dimensional Data

- Clustering Method: K-Means on full dataset (before PCA)
- Silhouette Score: 0.1500
- Observation: The low score suggests poor clustering due to high-dimensional sparsity and lack of clear structure.

+) PCA-Reduced Data (2D)

- Clustering Method: K-Means
- Silhouette Score: 0.4322
- Observation: Dimensionality reduction significantly improved clustering performance, revealing more distinct groupings.

+) PCA-Reduced Data (3D)

- Clustering Method: K-Means
- Silhouette Score: 0.3777
- Observation: The 3D result was slightly lower than 2D, suggesting that adding an extra dimension didn't provide much separation benefit.

---

4. Random subspace of the dataset

- Source code:

```
import numpy as np
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Select a random subspace (e.g., 3 random features)
np.random.seed(42)
random_feature_indices = np.random.choice(X_scaled.shape[1], size=3, replace=False)
X_random = X_scaled[:, random_feature_indices]

# Apply K-means
kmeans_rand = KMeans(n_clusters=3, n_init=10, random_state=42)
labels_rand = kmeans_rand.fit_predict(X_random)

# Evaluate clustering
score_rand = silhouette_score(X_random, labels_rand)
print(f"Silhouette Score for Random Subspace: {score_rand:.4f}")
```

Silhouette Score for Random Subspace: 0.2490

- To assess the effect of using fewer features, a random subspace of 3 features was selected from the dataset. K-Means clustering was then applied.

- Silhouette Score: 0.2490

- The score (0.2490) is higher than the original full dataset (0.1500), which means using a few random features can reduce noise and improve clustering.

- However, it is still lower than PCA results, so **PCA is more effective** at keeping important structure for clustering.

- This shows that **random subspace helps**, but **PCA usually gives better performance**.

IV. Conclusion

- The results show that k = 2 gave the best clustering for both Breast Cancer and Automobile datasets. PCA helped improve clustering quality more than using the original data or random features. Overall, choosing the right number of clusters and applying feature reduction methods like PCA are important for better clustering results.

V.References

- For Dataset:
  Communities and Crime - UCI Machine Learning Repository
  Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository
  Automobile - UCI Machine Learning Repository
- For theory:
  Clustering in Machine Learning | GeeksforGeeks
  kmeans