University of Science and Technology of Hanoi

Machine Learning and Data Mining 2

Labwork 2 - Clustering

BA12-118 Nguyen Phuc Minh
BA12-066 Nguyen Hoang Ha

TABLE OF CONTENTS

I. K-means

1. Import necessary libraries

- Numpy, Pandas, Matplotlib, Seaborn, Sklearn, Yellowbrick, mpl_toolkits.

2. Data loading and cleaning

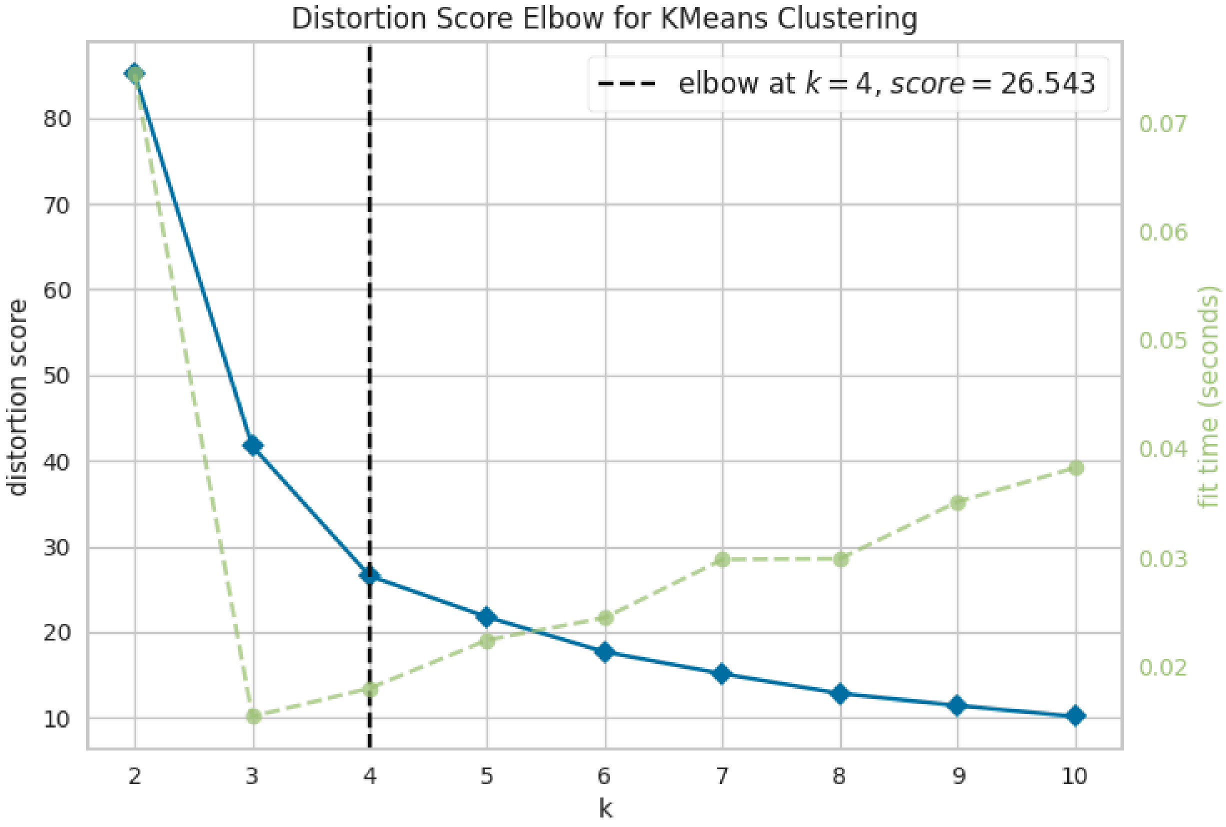
- Two datasets: Wholesale Customers and Heart Failure Clinical.
- Check the missing value.

3. Data preprocessing

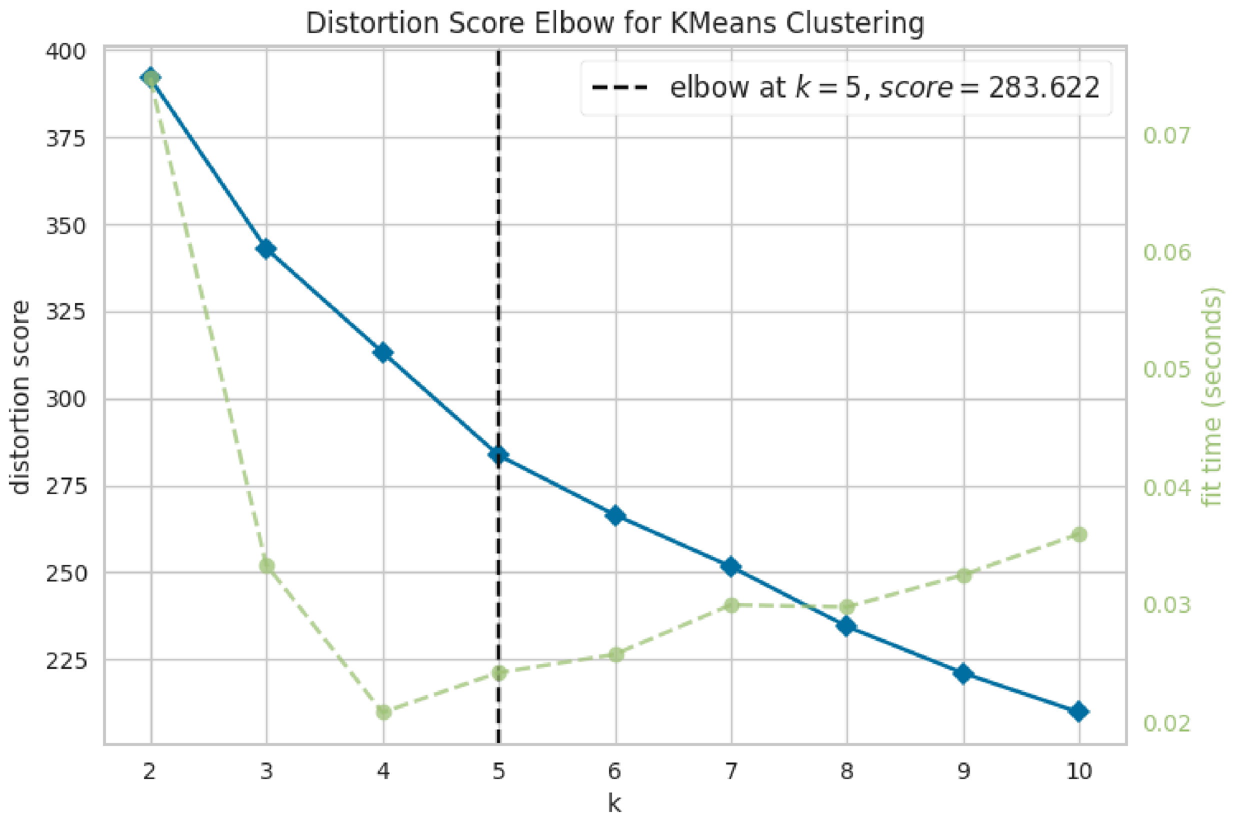
- Label encoding the categorical features.
- Scale the features using the standard scaler.

4. K-means clustering for all dataset

- Elbow Method to determine the number of clusters to be formed (K)



Wholesale Customers



Heart Failure Clinical

As K increases, average distortion decreases, meaning clusters have fewer instances and they're closer to their centroids. However, the rate of improvement in distortion declines with higher K. The "elbow point" is where this decline is most significant, indicating where to stop dividing data into more clusters.

- Centroid initialization: K-means++
  Use the K-means++ algorithm for initializing the centroids to improve the original K-means algorithm's convergence rate and clustering quality by selecting initial centroids in a more intelligent way.

  - The algorithm starts by randomly selecting the first centroid from the data points.

  - For each data point, calculate its distance from the nearest centroid that has already been chosen. This distance is typically measured as the squared Euclidean distance.

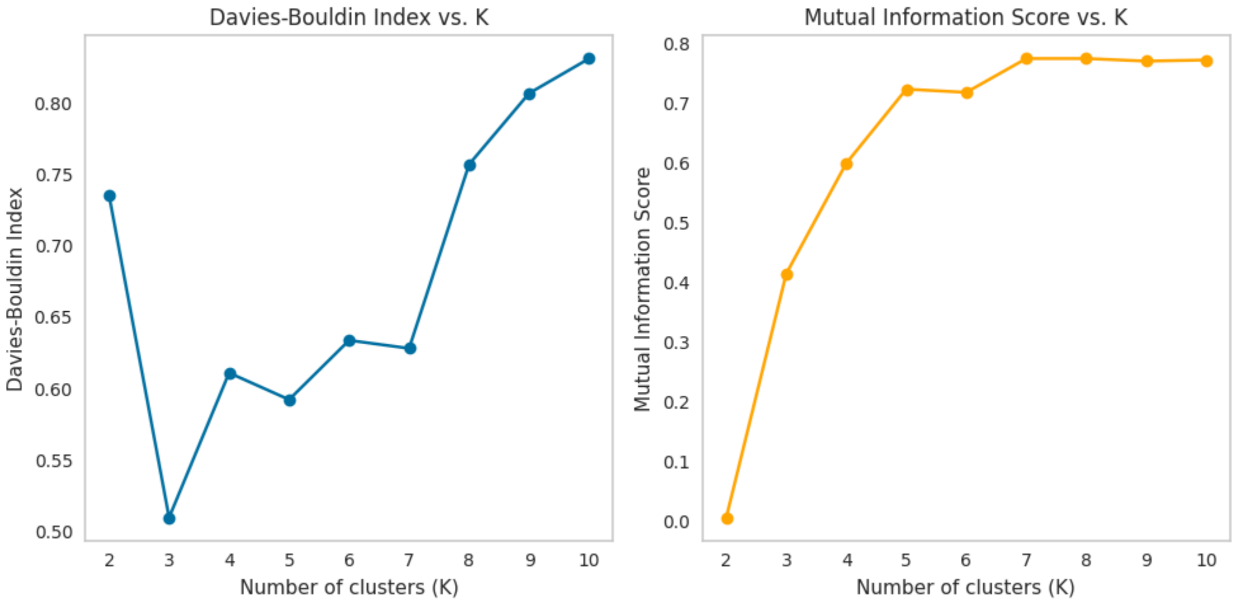
- The next centroid is chosen with a probability proportional to the square of the distance of each data point from the nearest centroid that has already been chosen. This means that data points that are further away from existing centroids are more likely to be selected as the next centroid.

Repeat: Steps 2 and 3 are repeated until k centroids have been selected.

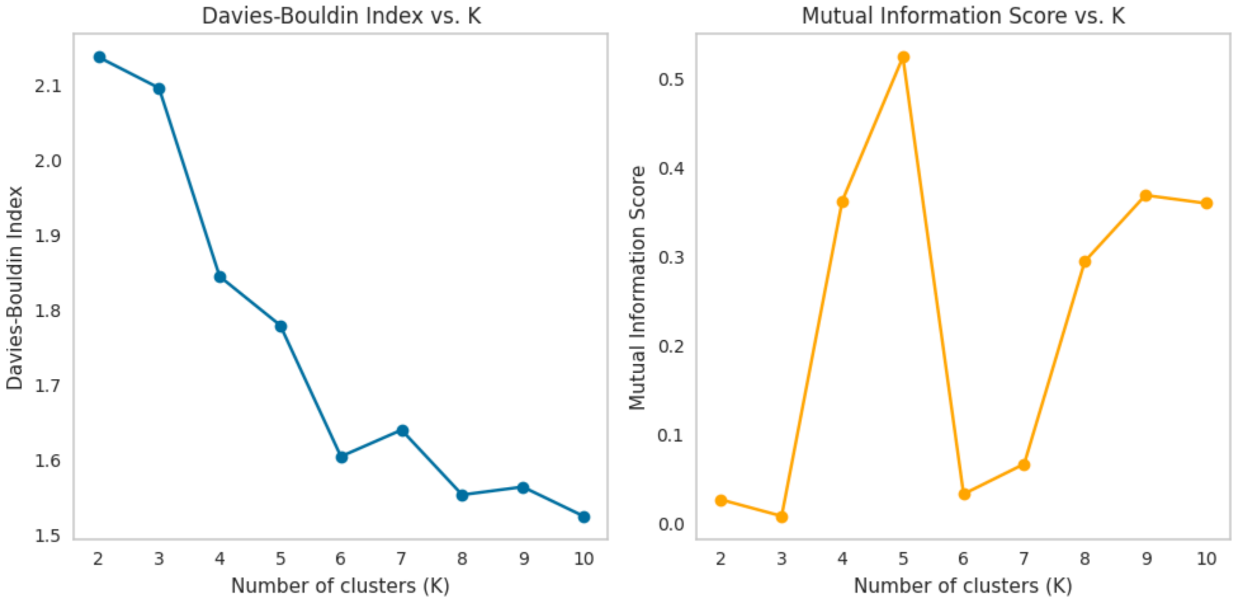
By using K-means++ initialization, the centroids are spread out more effectively across the dataset, which can lead to faster convergence and better clustering results compared to randomly selecting initial centroids.

- Implement and evaluate K-Means Clustering quality

Internal validation: Davies-Bouldin Index (DBI)	External validation: Mutual Information (MI)
The DBI measures the average similarity between each cluster and its most similar cluster, relative to the size of the clusters. A lower DBI indicates better clustering because it means that the clusters are well-separated from each other and compact within themselves.	The MI assesses how well the clustering algorithm assigns data points to clusters compared to the ground truth labels. The maximum value of MI depends on the number of clusters and the distribution of data points. Values closer to 1 or 0 indicate strong or weak agreement between the clustering and true labels. Negative values are possible but generally occur when the clustering algorithm performs worse than random labeling.



Wholesale Customers



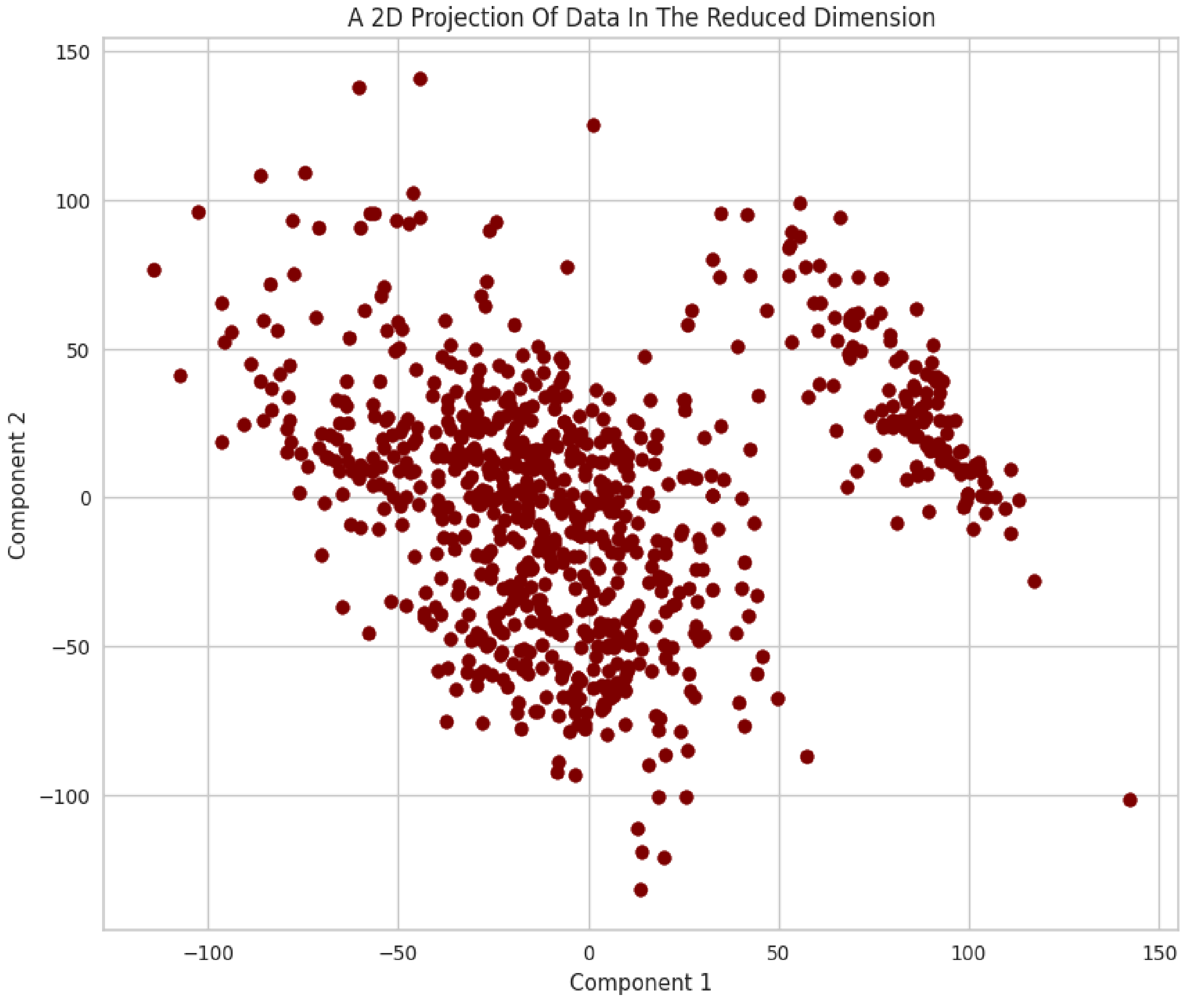
Heart Failure Clinical

I. Subspace clustering

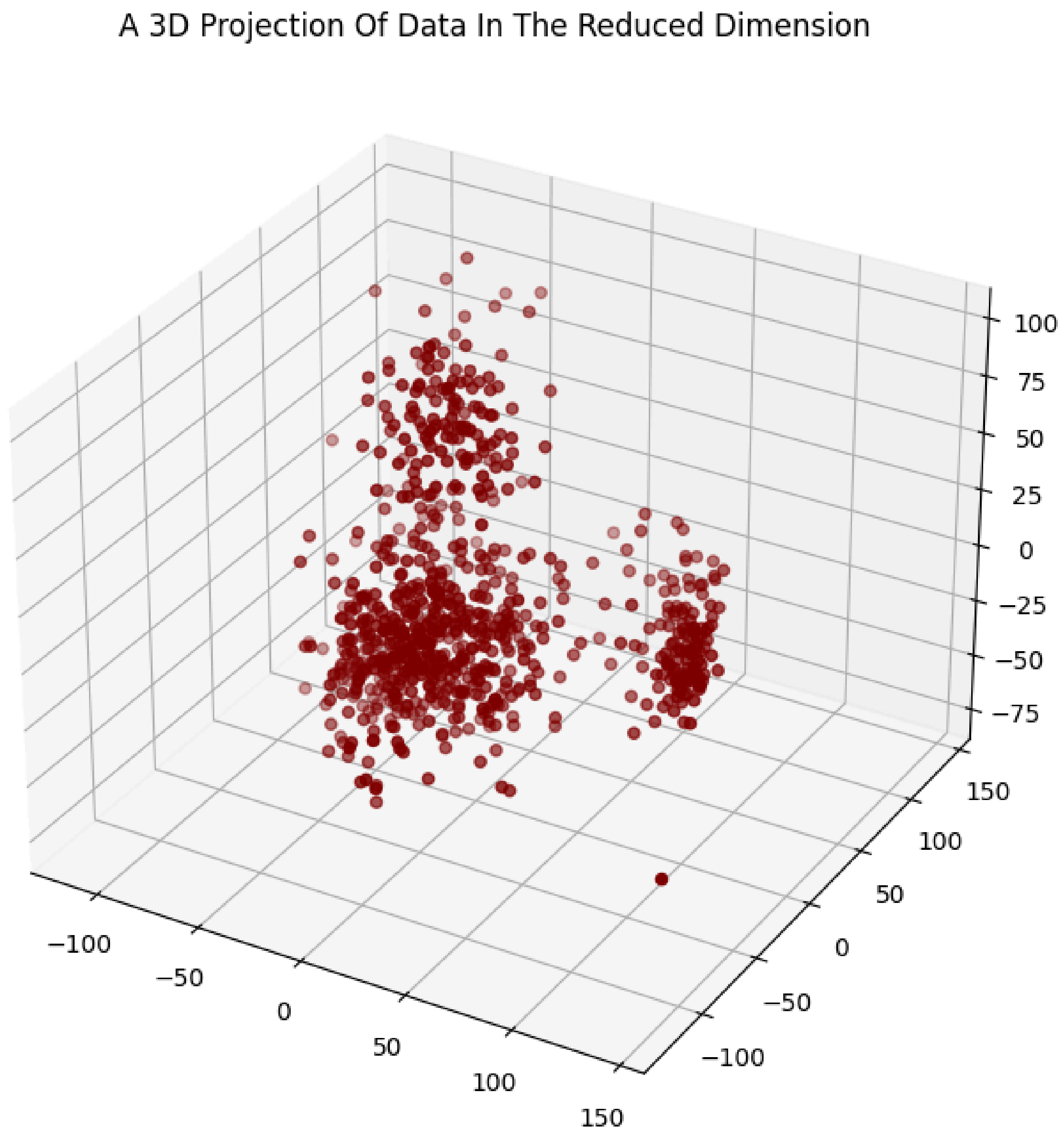
1. Dataset

In this section, we use a dataset containing 802 samples for 802 corresponding people who have been found to have different types of cancer. Each sample contains expression values of more than 20K genes.

2. Use PCA to visualize the data distribution in 2D and 3D.



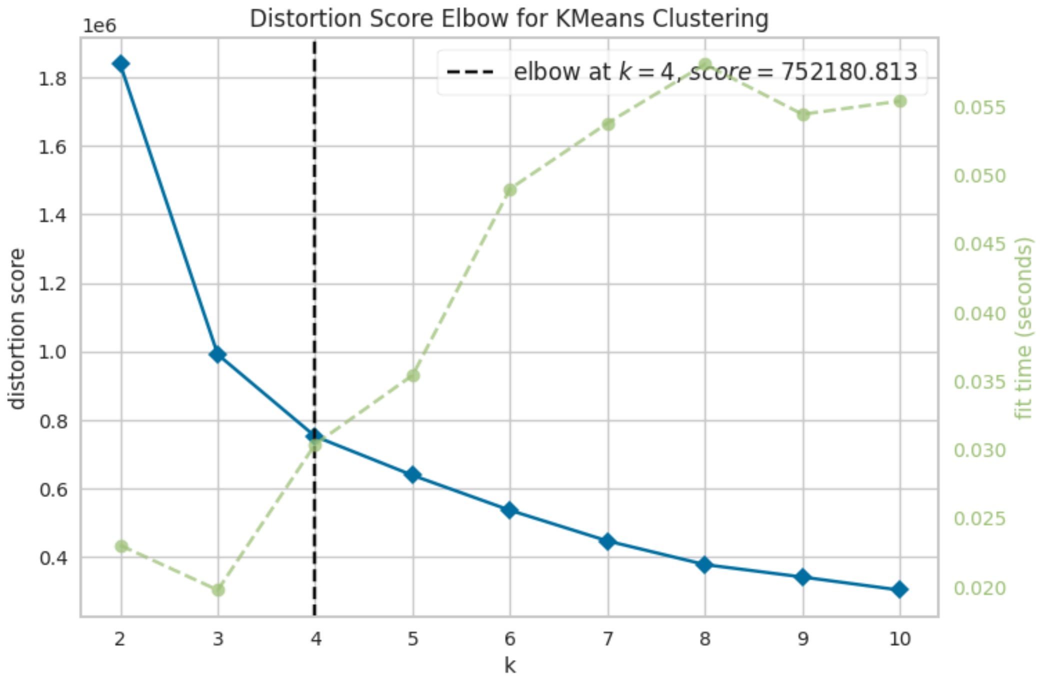
2 Principal Components



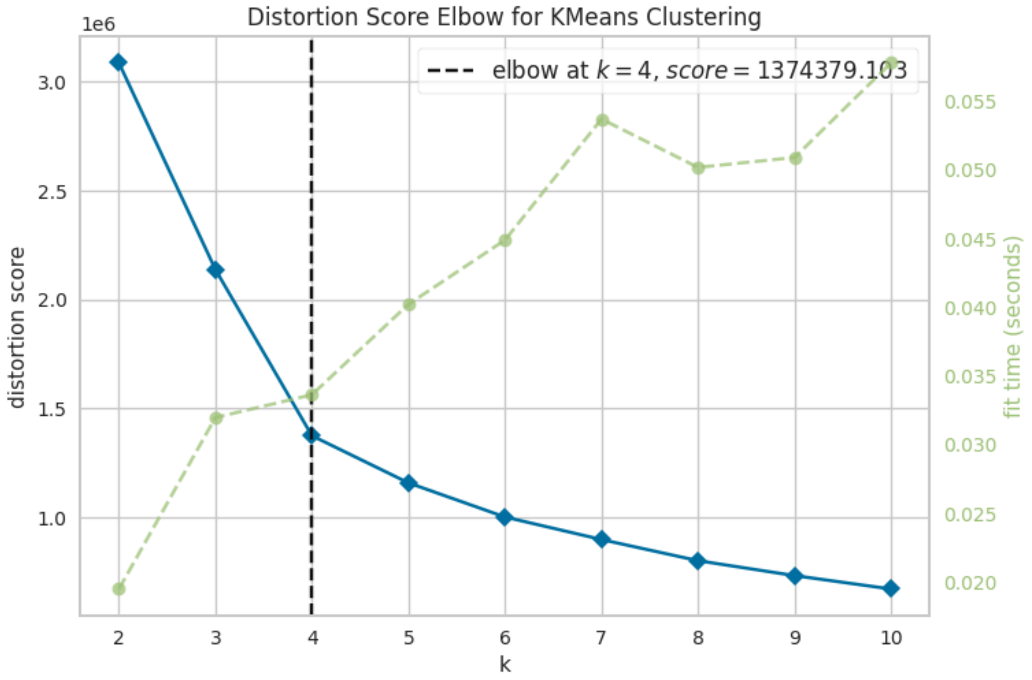
3 Principal Components

3. Apply Clustering after PCA

- Elbow Method to determine the number of clusters to be formed (K)

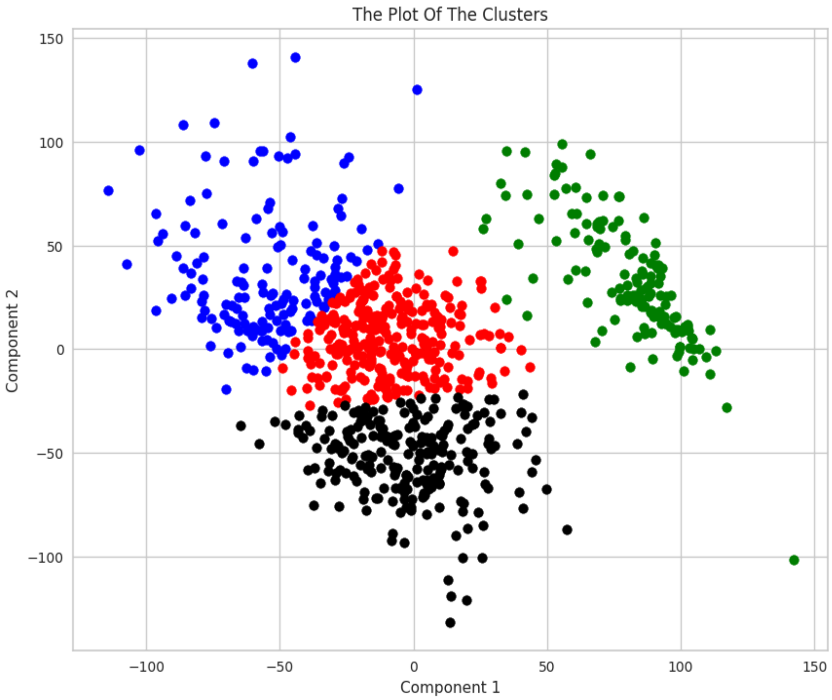


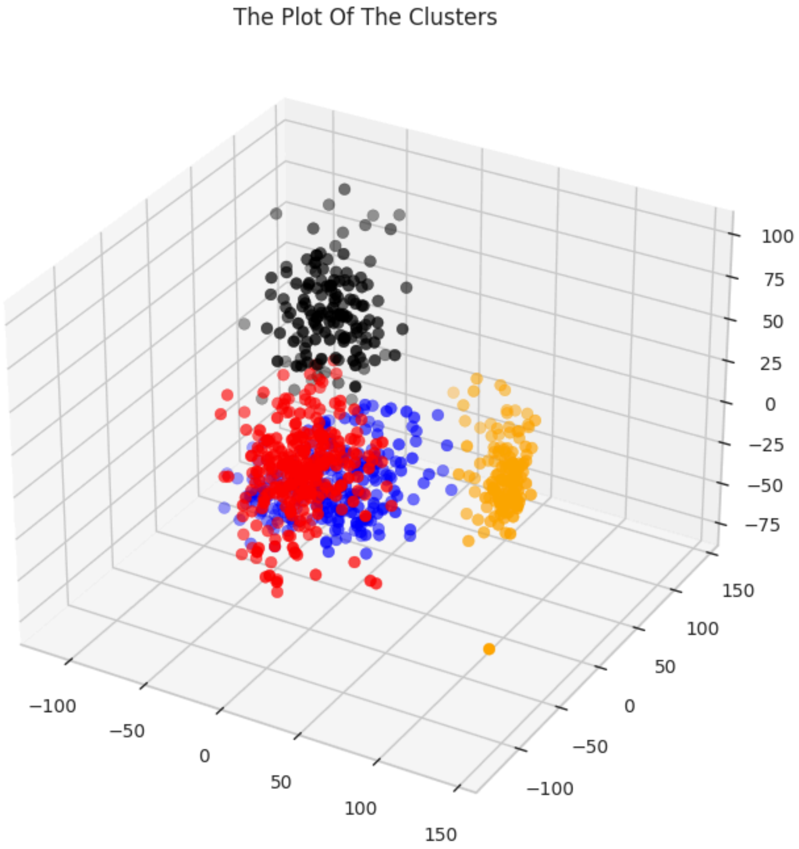
Elbow for PCA2



Elbow for PCA3

- Clustering distribution in 2D and 3D for K = 4





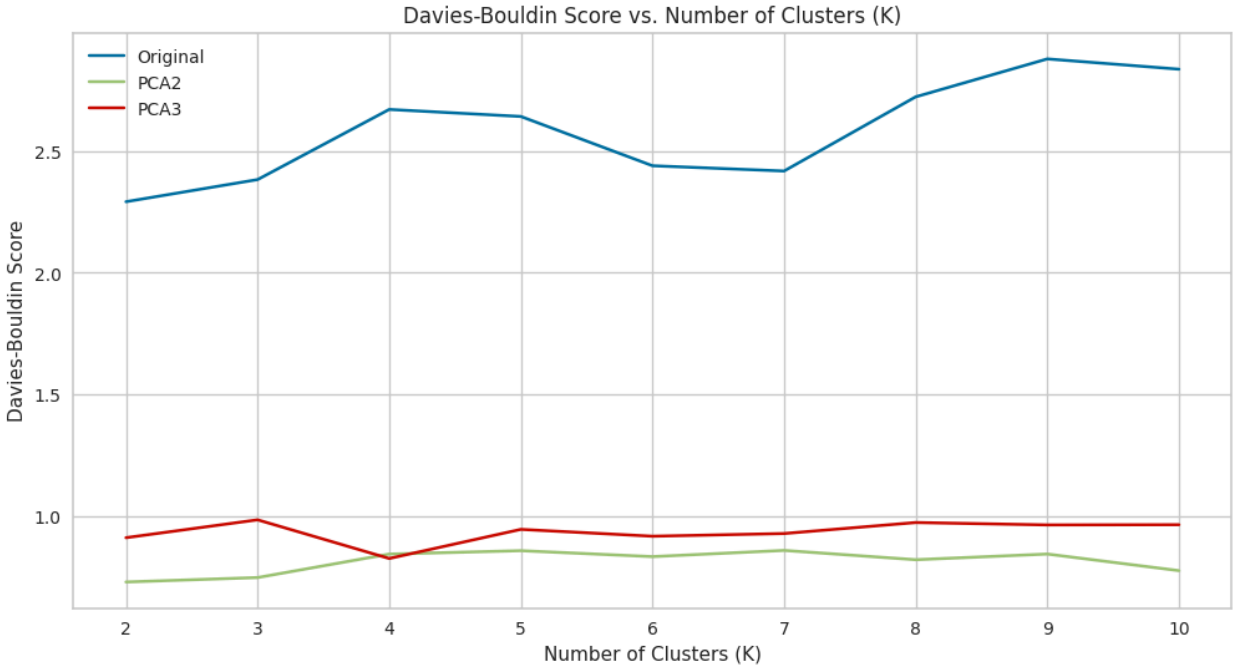
- Compare the performance before and after the dimensionality reduction

➢ Davies-Bouldin Index:

- The Davies-Bouldin score measures the average similarity ratio of each cluster with its most similar cluster. A lower Davies-Bouldin score indicates better cluster separation and compactness.

- Higher scores for the original dataset suggest that the clusters are less compact and more spread out in the high-dimensional space.

- Lower scores after dimensionality reduction imply that the clusters are more distinct and well-defined in the reduced dimensional space.

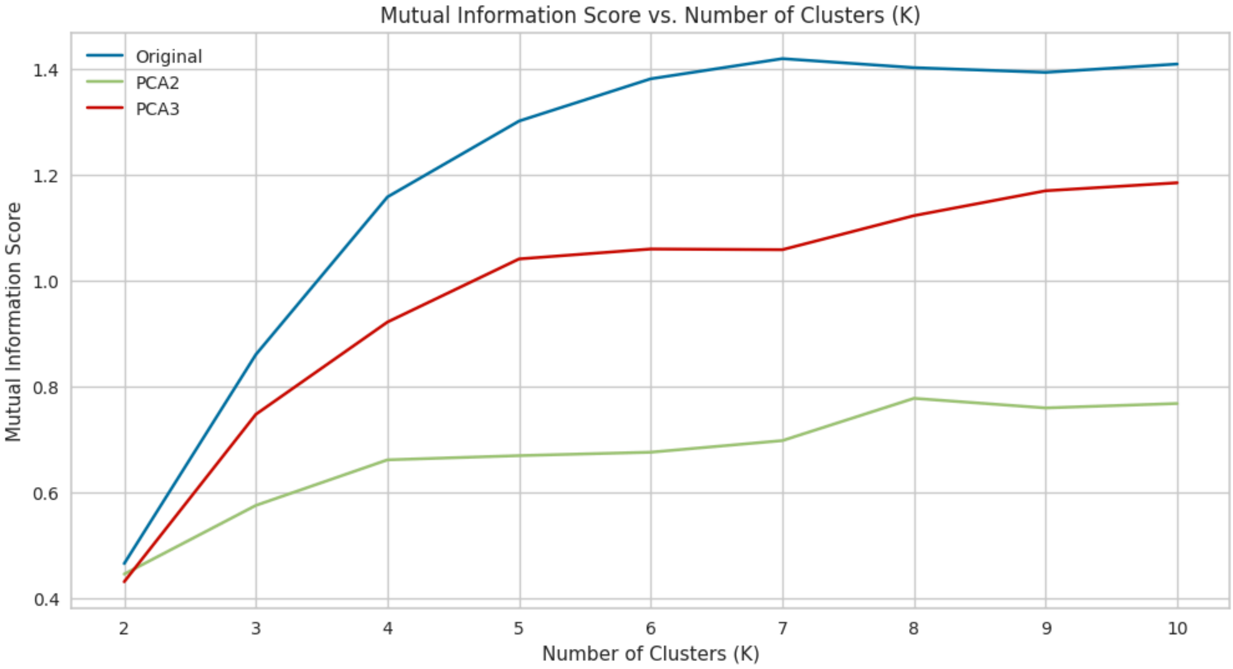


➢ Mutual Information:

- As dimensionality decreases, some information inherent in the original high-dimensional space is inevitably lost because fewer components cannot capture all the variability and nuances present in the original data.

- Clustering algorithms operate on a dataset with less information, potentially leading to clusters that are less representative of the true underlying structure, resulting in lower Mutual Information scores.

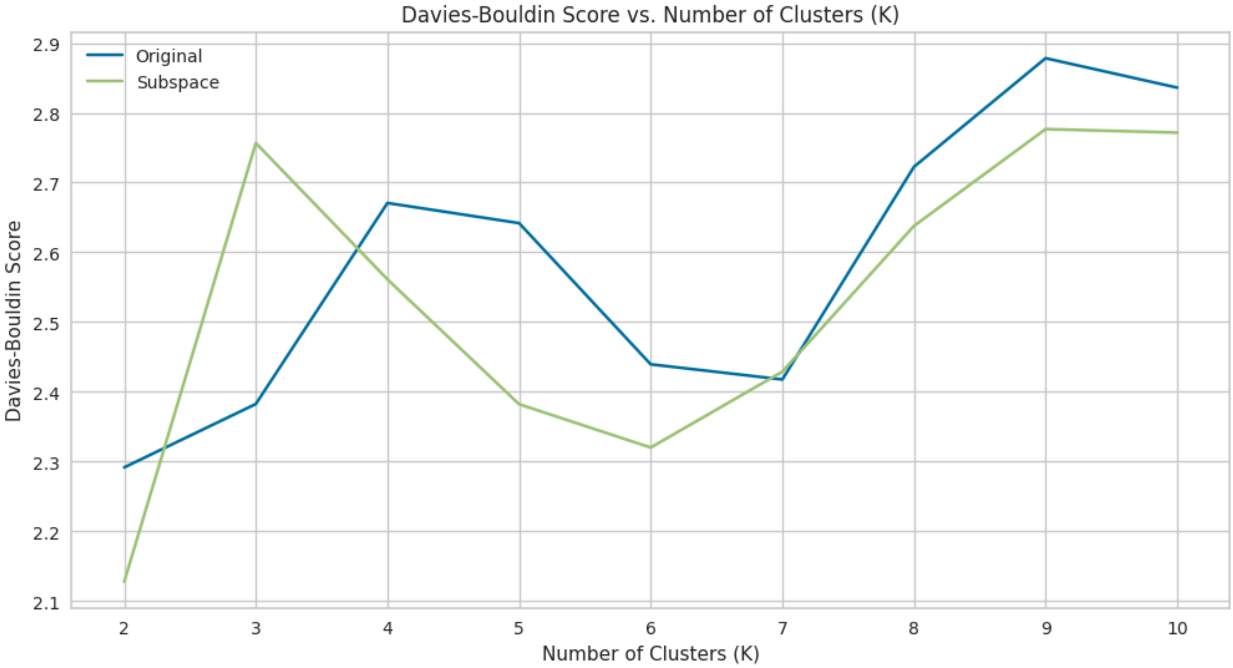
- Although PCA2 and PCA3 datasets typically retain a significant portion of the variance, some detailed information is still lost, affecting the clusters' alignment with the original data label



4. Re-apply clustering in a random subspace of the dataset and comment on the results.

   Evaluation of clustering quality of the original dataset and its subspace

➢ Davies-Bouldin Index:



The erratic increase and decrease in the Davies-Bouldin Index (DBI) for both the original dataset clustering and the subspace of dataset clustering could be attributed to the following reasons:

- Sensitivity to Cluster Structure: The DBI is sensitive to the cluster structure and the distances between clusters. If the clusters are not well-separated or if there are outliers present, the DBI values can fluctuate erratically as the number of clusters changes.

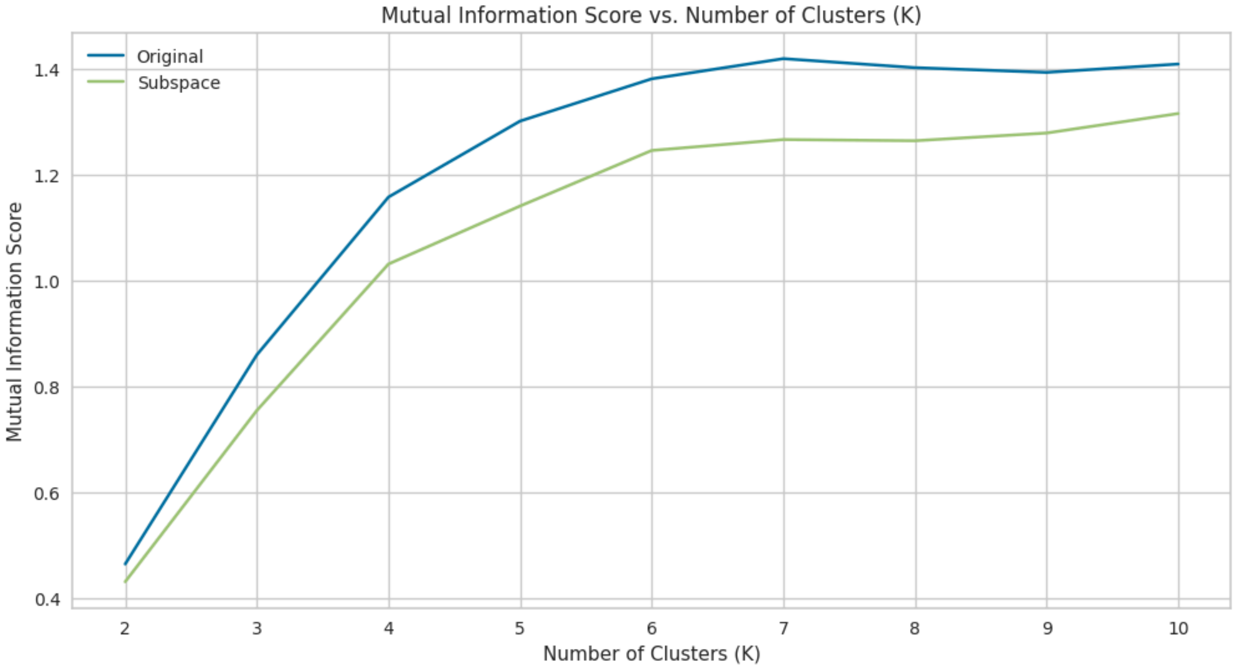
- Complex Data Patterns: In datasets with complex patterns or overlapping clusters, the DBI values may not follow a smooth decreasing trend with an increase in the number of clusters. This can lead to erratic fluctuations in the DBI values.

- Noise and Outliers: The presence of noise or outliers in the data can also impact the DBI values. Outliers can significantly affect the clustering results and lead to erratic changes in the DBI values.

- Suboptimal Clustering: If the clustering algorithm used is not able to effectively capture the underlying structure of the data, it can result in erratic changes in the DBI values as the number of clusters varies.

Overall, the erratic behavior of the DBI values in both scenarios may be indicative of the complexity and variability in the data, as well as the sensitivity of the DBI metric to different aspects of the clustering process.

➢ Mutual Information:



When the subspace is randomly chosen from the original data for clustering, the lower Mutual Information (MI) Scores in the subspace clustering compared to the original dataset clustering can be explained by the following reasons:

- Randomness of Subspace Selection: Since the subspace is randomly chosen from the original data, there is no guarantee that the selected subset of features captures the most relevant information for clustering. This randomness can lead to a subset that may not fully represent the underlying structure of the data, resulting in lower MI Scores.

- Information Loss: Randomly selecting a subspace from the original data can result in information loss, as the selected subset may not contain all the essential features necessary for accurate clustering. This loss of information can lead to lower mutual information values in the subspace clustering.

- Reduced Discriminative Power: The randomly chosen subspace may not preserve the discriminative power present in the full original dataset. As a result, the clustering algorithm may struggle to differentiate between clusters effectively, leading to lower MI Scores in the subspace clustering.

- Impact on Cluster Separation: The randomly selected subspace may not maintain the same level of cluster separability as the full dataset. This can affect the clustering performance and result in lower mutual information values for the subspace clustering.

In summary, when a subspace is randomly chosen from the original data for clustering, the lower MI Scores in the subspace clustering compared to the original dataset clustering can be attributed to the randomness of subspace selection, information loss, reduced discriminative power, and impact on cluster separation.