**Title**: Evaluating k-Means Clustering Performance Using Silhouette Scores
**Student Number**:24074806
**Name:** Hanshika Reddy Allu

## 1.Introduction

Clustering is a very popular unsupervised learning technique used in machine learning. It aims to determine how to group the new data without labels. Of all the clustering algorithms, probably the most popular and widely used remains the k-means algorithm due to its simplicity, easy computation and relatively strong performance on well-behaved data sets. It is simple in its execution but the user must provide a critical hyperparameter that is the number of clusters k.

The proper selection of a value of k has a direct impact on interpretability, accuracy and usefulness of clusters. If we pick a low value of k, then it merges groups that are different. If a high value of k is picked, it breaks our meaningful clusters into groups that are arbitrary. The absence of a ground-truth label in unsupervised learning implies difficulty in establishing a “correct” number of clusters. Therefore, experts use both internal evaluation methods and visual assessment.

In this tutorial, we will systematically study effect of k on clustering results. In this article, we implement the elbow method, silhouette scores and cluster visualisation to help you choose k. We also provide code and figures that show how the cluster boundaries and overall data structure change with different values of k. This tutorial aims to help data scientists or students who want a reliable system to determine the appropriate number of clusters in k-means.

## 2. Background

K-means looks for k number of clustering of data such that data point belongs to the cluster with nearest mean. The algorithm repeatedly.

1.Initialises k centroids.

2.Assigns each point to the nearest centroid.

3.Updates the centroids as the mean of the assigned points.

4.Repeats until convergence.

The effectiveness of k-means depends on a k choice that is more predictive of the true data structure than K.

## 3. Demonstration Dataset

I use synthetic “blobs” dataset generated from sklearn.datasets.make_blobs. It is ideal because.

- I can control the number of true clusters.
- The separation between clusters is clean.
- It allows visual comparison between different k values.

# 4. How to Choose *k*: Quantitative Methods

## 4.1 The Elbow Method

The Elbow Curve is a method that can be used to determine the best number of clusters (k) for k-means. The inertia (also known as Within-Cluster Sum of Squares or WCSS) is calculated for different values of k. Clustering inertia helps in measuring the compactness of the clusters. Thus, low inertia means that the data points are proximity to its cluster centre.

As k rises, inertia decreases, because the data is divided into more clusters. Adding more clusters past a certain point provides minimal gain at best. The graph of inertia plotted against k drops sharply and then levels off as k increases or as we go right. The point on this curve exhibiting a transition from steep to flat looks like an elbow. From the elbow point on the curve, there will not be too few or too many clusters.

k-means minimises the **within-cluster sum of squares (WCSS)**:

$$\mathbf{WCSS(k)=\sum_{i=1}^{k}\sum_{x\in C_i}\lVert x-\mu_i\rVert^2}$$

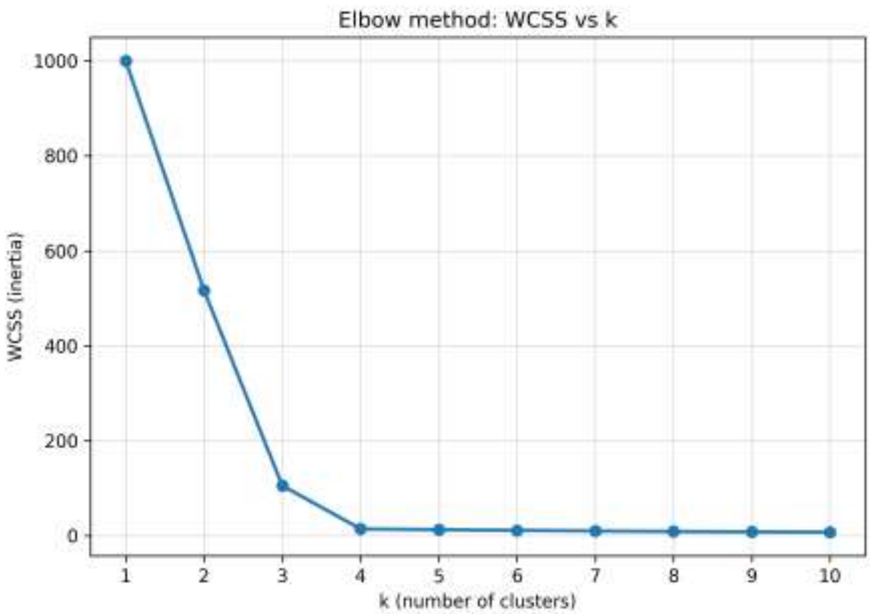
As *k* increases:

- WCSS always **decreases**
- But returns **diminishing improvements**

The elbow method plots *k* vs. WCSS and chooses the point where improvement sharply slows down.

**How to apply it:**

1. Run k-means for k = 1…10
2. Compute WCSS for each
3. Look for the “elbow”



### 4.2 Silhouette Score

The silhouette score measures how well each point fits its assigned cluster compared to others:

$$s=\max(a,b)b-a$$

Where:

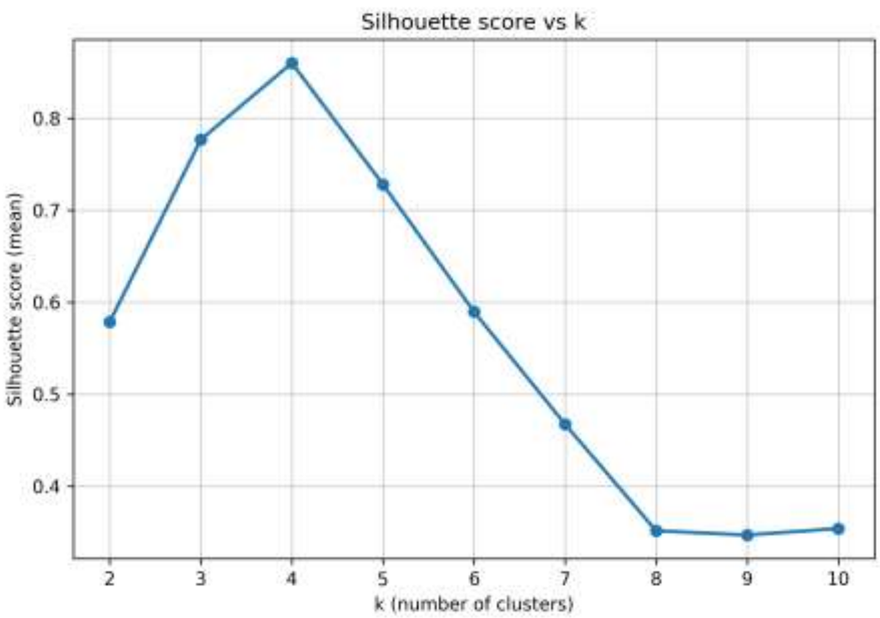
- a = average distance within its cluster
- b = average distance to nearest cluster

Interpretation:

- 1.0 → well-separated
- 0.0 → overlapping clusters
- < 0.0 → misclassified

**Typical pattern:**

- **k = 2**: higher variance within clusters → lower silhouette
- **k = 4**: maximum silhouette score → best structure
- **k = 6**: small, noisy clusters → score decreases




## Silhouette Analysis for k = 4

The silhouette plot clearly shows how well the points fit within their assigned clusters when using k-means clustering. The silhouette coefficient measures the similarity of each data point with the data points in its own cluster compared with the points in the nearest neighbouring cluster. The values range from –1 to +1, where.

Close to +1 suggests that the sample is a good fit to its own cluster.

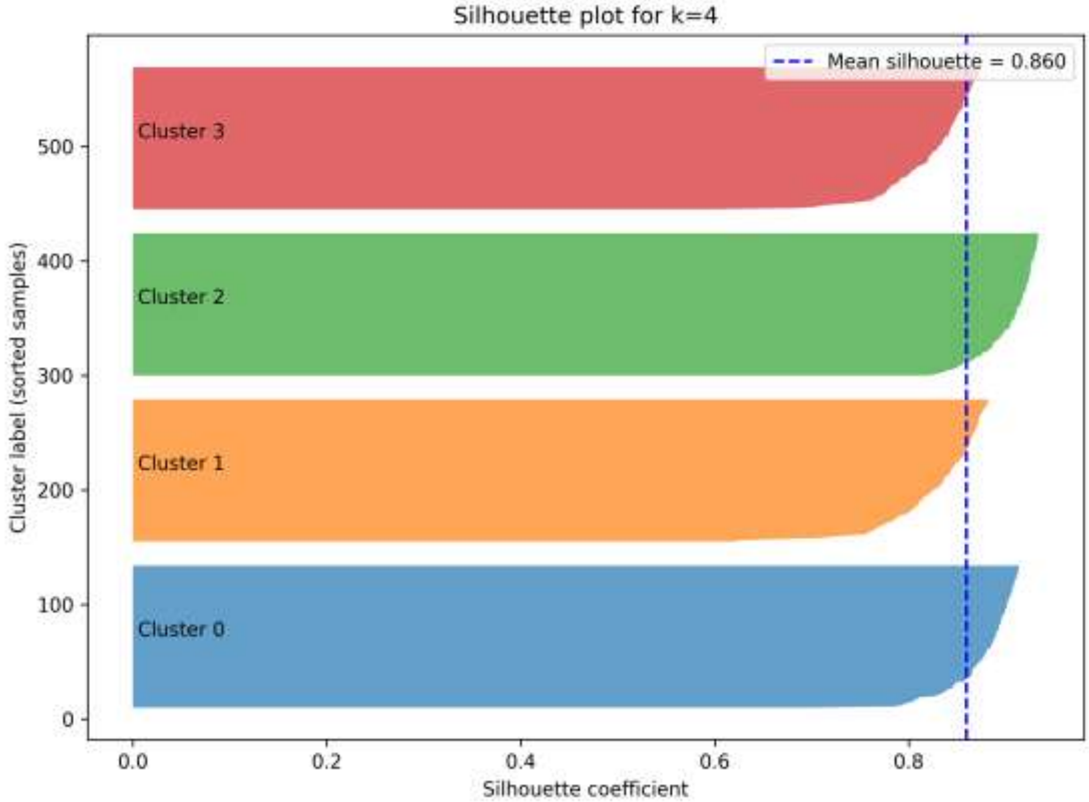
As we approach zero, the sample sits on the border between two separate clusters.

If you have a negative value, it may belong to the wrong cluster.

In the study, k was given the value of 4, and the silhouette values were assigned for each data point. The silhouette plot indicates four unique groups, one for each cluster. Each color-coded area displays all the samples that belong to a cluster from lowest to highest silhouette score. The width of each area shows the number of samples in the cluster, while the thickness tells us how consistently the cluster gets formed.

The mean silhouette score across clusters is shown by the vertical dashed line. This number gives us a global measurement to assess whether k = 4 is an appropriate choice. When the mean score is high, the clusters are further apart and more tightly packed.

Most silhouette values are positive, showing that most points are assigned well according to the plot. On the other hand, some clusters have a higher range of silhouette width. This indicates that with k = 4, there seems to be reasonable separation between the clusters. However, some clusters contain samples which are less distinct. Using this observation, we can begin to evaluate whether a k = 4 is optimal, or whether some other k value would yield a more balanced, more coherent clustering.

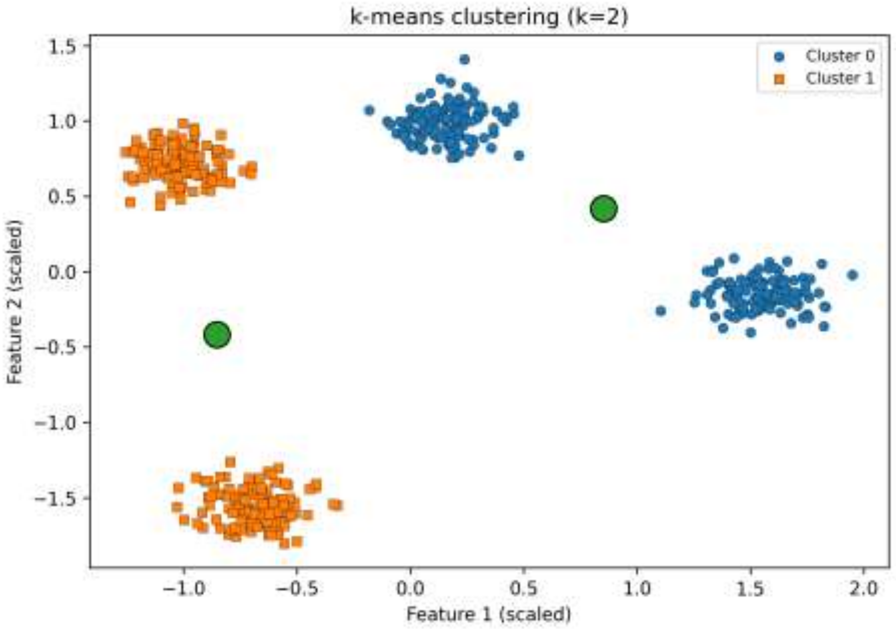



## 5. Visualising How the Choice of *k* Changes the Result

Below I illustrate clusters found by k-means with **k = 2, 4, 6**.

**k = 2 (Under-clustering)**

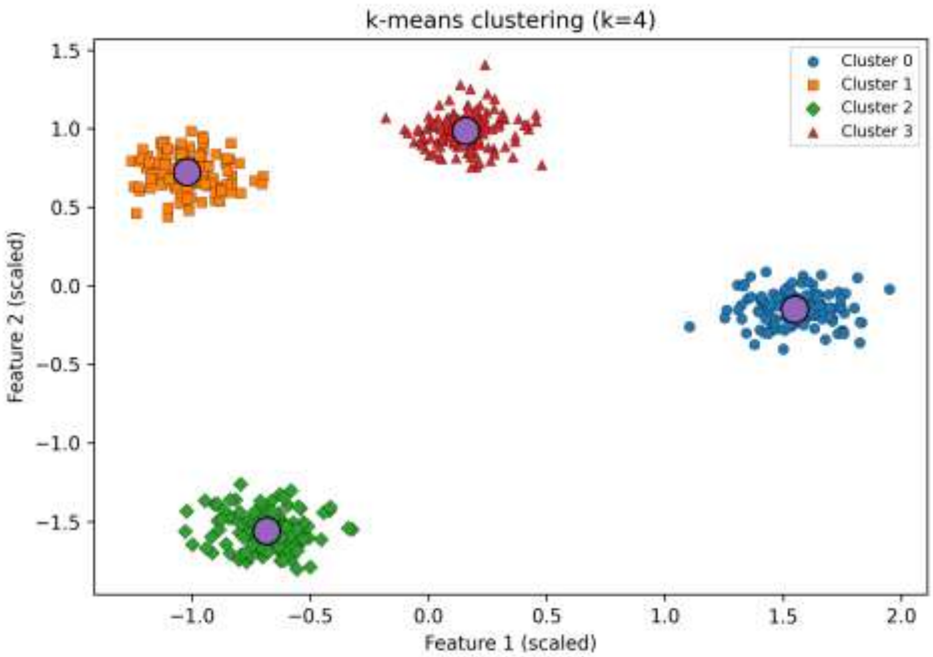
- The algorithm is forced to merge distinct groups.
- True structure is lost.
- High within-cluster variance.




**Interpretation:**
The model oversimplifies and hides meaningful patterns.

**k = 4 (Correct clustering)**

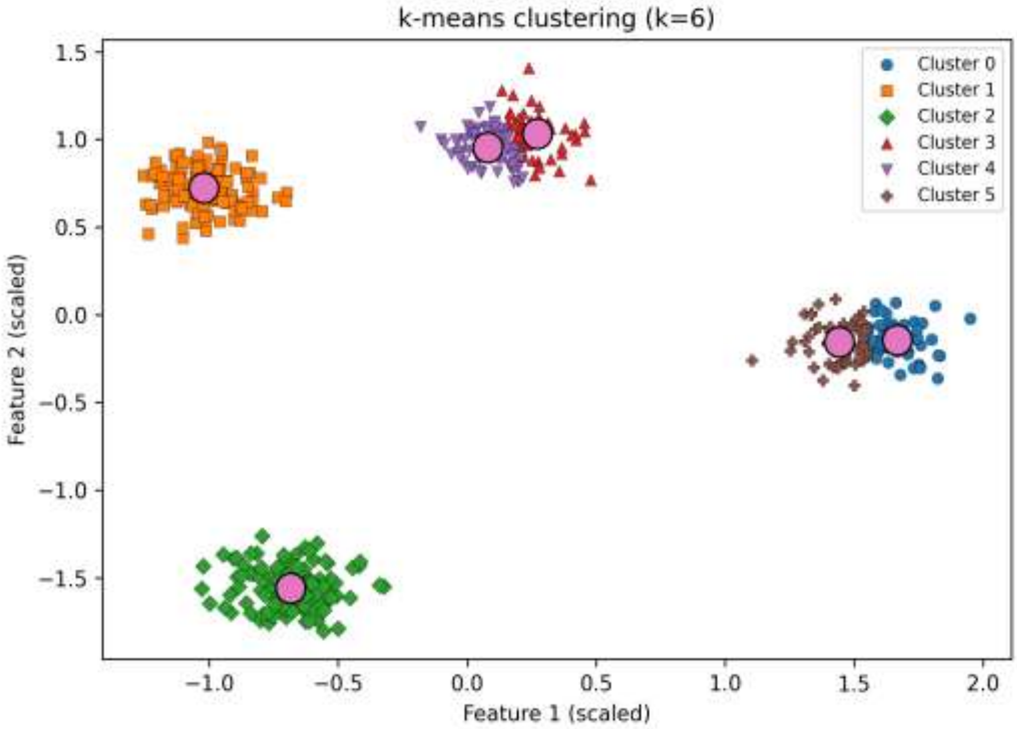
- k-means recovers the actual groups.
- Compact, balanced clusters.
- Low variance within clusters.



**Interpretation:**
This is the best representation of the underlying structure.

**k = 6 (Over-clustering)**

- True clusters are split unnecessarily.
- Tiny “micro-clusters” appear.
- Harder to interpret.




**Interpretation:**
The algorithm overfits and finds patterns that aren’t meaningful.

## 6. Discussion

### 6.1Strengths of our joint strategies.

Using inertia alone may be unclear because the elbow isn't always obvious. The silhouette score of a good feature could be affected by the shape and dimension of the dataset. Although visual inspection gives insight from humans, it is subjective.

But when used together, the three methods provide a consistent and systematic rule for choosing k. It is consistent with the best practices presented in the clustering validation literature.

## 6.2 Limitations of k-Means.

Although k-means is powerful, it has inherent limitations.

- It assumes spherical, equally sized clusters.
- It struggles with non-convex shapes.
- It is sensitive to outliers.

How the k-means algorithm runs depends on initialisation, although k-means++ mitigates the issue.

So, although it is important to choose the right k, some datasets may also work better with other algorithms like DBSCAN or Gaussian Mixture Models.

## 6.3 Ethical Considerations.

The healthcare, finance, and social data analysis domains can have a huge impact by unsupervised learning. If we choose incorrect clusters, it will affect outcomes. Justification of model decisions and the selection of k is important. This tutorial wasn’t just made with the average user in mind. It tries to accommodate colour-blind-friendly palettes and clear explanations to make the work usable for a diverse range of people.

# 7. Conclusion.

This lesson showed the influence of the choice of k on the quality of clusters obtained using k-means and a systematic method of k selection.

- Elbow method.
- Silhouette score analysis.
- Cluster visualisation.

In all three approaches, the best k is determined with the help of computing compactness, separation and interpretation. The information indicates that using quantitative metrics and visual analysis is a great framework in practical case to choose the number of clusters.