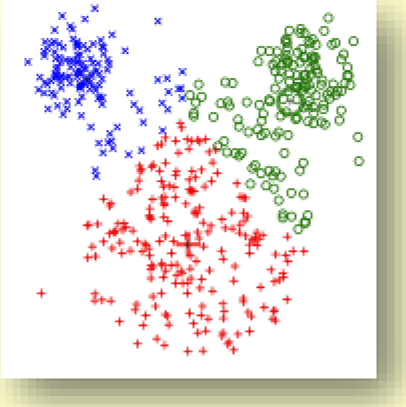


Cluster Analysis

Ali Ridho Barakbah, Entin Martiana

Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya

Variance

- Digunakan untuk mengukur nilai penyebaran dari data-data hasil clustering
- Dipakai untuk data yang bertipe unsupervised
- Variance pada clustering ada 2 macam:
  - Variance within cluster
  - Variance between clusters

Good cluster

is when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity)

Variance & homogeneity

internal homogeneity → Variance within cluster (V_w)

external homogeneity → Variance between clusters (V_b)

Ideal cluster

- The ideal cluster has minimum V_w to express internal homogeneity and maximum V_b to express external homogeneity.

minimum

$$V = \frac{V_w}{V_b}$$

Variance per Cluster

$$v_c{}^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(d_i - \overline{d}_i \right)^2$$

$v_c{}^2$ = variance pada cluster c
c = 1..k, dimana k = jumlah cluster
n_c = jumlah data pada cluster c
d_i = data ke-i pada suatu cluster
$\overline{d}_i$ = rata-rata dari data pada suatu cluster

Variance within cluster

$$v_w = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\cdot v_i^{\,2}$$

v_w = variance within cluster
N = jumlah semua data

Variance between clusters

$$v_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\overline{d_i} - \overline{d}\right)^2$$

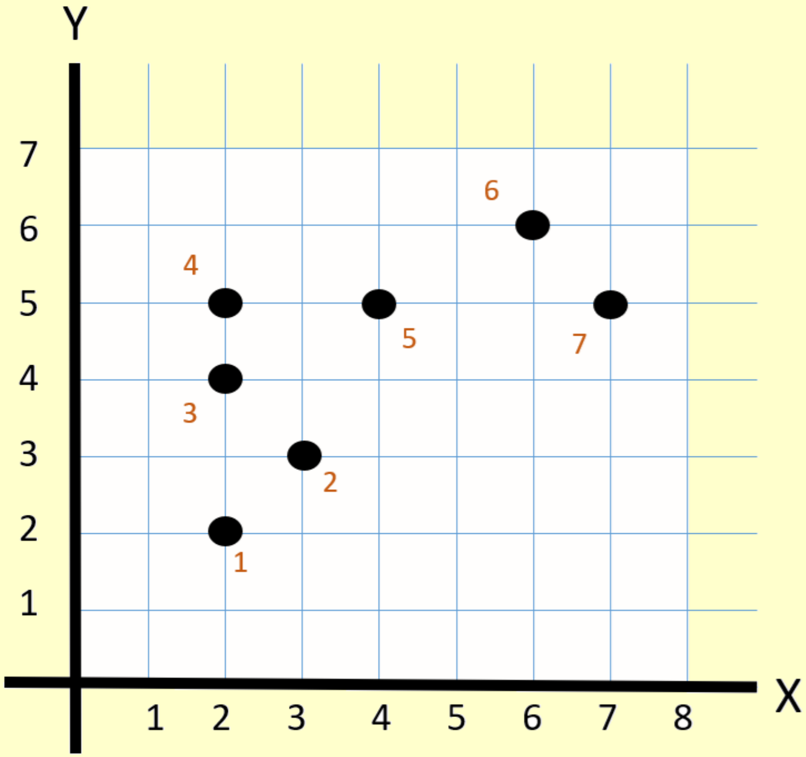
$\overline{d}$ = rata-rata dari $\overline{d_i}$

Variance pada keseluruhan cluster

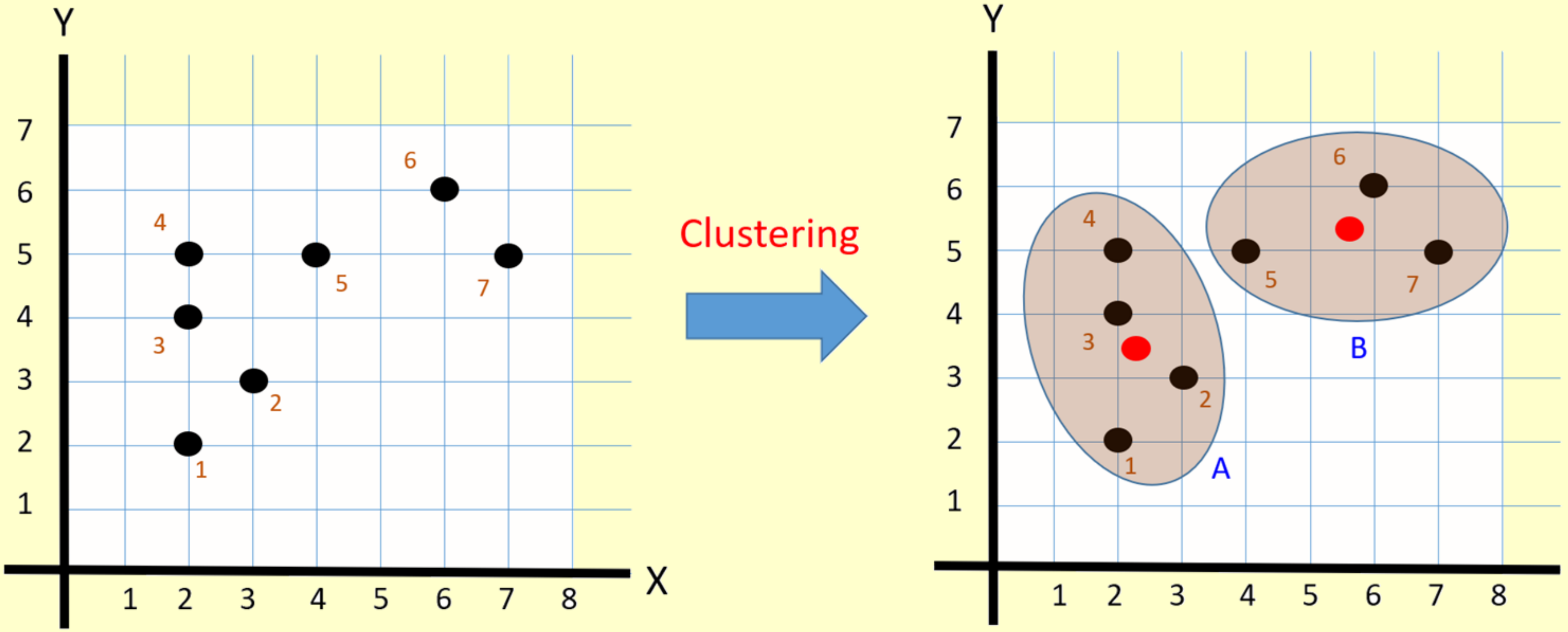
→ $$v = \frac{v_w}{v_b}$$

Contoh Kasus

	X	Y
Data1	2	2
Data2	3	3
Data3	2	4
Data4	2	5
Data5	4	5
Data6	6	6
Data7	7	5



Centroid$_A$ = (2.25, 3.5)
Centroid$_B$ = (5.67, 5.33)



Cluster Variance

$$v_c{}^2 = \frac{1}{n_c - 1}\sum_{i=1}^{n_c}\left(d_i - \overline{d_i}\right)^2$$

$$v_A{}^2 = \tfrac{1}{4-1}\left(\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2.25\\3.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}3\\3\end{bmatrix} - \begin{bmatrix}2.25\\3.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}2\\4\end{bmatrix} - \begin{bmatrix}2.25\\3.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}2\\5\end{bmatrix} - \begin{bmatrix}2.25\\3.5\end{bmatrix}\right)^2\right)$$

$$= \tfrac{1}{3}\left(\left(\begin{bmatrix}-0.25\\-1.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}0.75\\-0.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}-0.25\\0.5\end{bmatrix}\right)^2 + \left(\begin{bmatrix}-0.25\\1.5\end{bmatrix}\right)^2\right)$$

$$= \tfrac{1}{3}\left(\begin{bmatrix}-0.25 & -1.5\end{bmatrix}\begin{bmatrix}-0.25\\-1.5\end{bmatrix} + \begin{bmatrix}0.75 & -0.5\end{bmatrix}\begin{bmatrix}0.75\\-0.5\end{bmatrix} + \begin{bmatrix}-0.25 & 0.5\end{bmatrix}\begin{bmatrix}-0.25\\0.5\end{bmatrix} + \begin{bmatrix}-0.25 & 1.5\end{bmatrix}\begin{bmatrix}-0.25\\1.5\end{bmatrix}\right)$$

$$= \tfrac{1}{3}\ (2.3125 + 0.8125 + 0.3125 + 2.3125) = 1.9167$$

$$v_B{}^2 = \tfrac{1}{3-1}\left(\left(\begin{bmatrix}4\\5\end{bmatrix} - \begin{bmatrix}5.67\\5.33\end{bmatrix}\right)^2 + \left(\begin{bmatrix}6\\6\end{bmatrix} - \begin{bmatrix}5.67\\5.33\end{bmatrix}\right)^2 + \left(\begin{bmatrix}7\\5\end{bmatrix} - \begin{bmatrix}5.67\\5.33\end{bmatrix}\right)^2\right)$$

$$= \tfrac{1}{2}\left(\left(\begin{bmatrix}-1.67\\-0.33\end{bmatrix}\right)^2 + \left(\begin{bmatrix}0.33\\0.67\end{bmatrix}\right)^2 + \left(\begin{bmatrix}1.33\\-0.33\end{bmatrix}\right)^2\right)$$

$$= \tfrac{1}{2}\left(\begin{bmatrix}-1.67 & -0.33\end{bmatrix}\begin{bmatrix}-1.67\\-0.33\end{bmatrix} + \begin{bmatrix}0.33 & 0.67\end{bmatrix}\begin{bmatrix}0.33\\0.67\end{bmatrix} + \begin{bmatrix}1.33 & -0.33\end{bmatrix}\begin{bmatrix}1.33\\-0.33\end{bmatrix}\right)$$

$$= \tfrac{1}{2}\ (2.8978 + 0.5578 + 1.8778) = 2.6667$$

Variance within cluster & Variance between clusters

$$v_w = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\cdot v_i^{\,2}$$

$$v_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\overline{d_i} - \overline{d}\right)^2$$

$$\begin{aligned} v_w &= \frac{1}{7-2}\big((4-1)\ 1.9167 + (3-1)2.6667\big) \\ &= \frac{1}{5}(5.7501 + 5.3334) = 2.2167 \end{aligned}$$

$$\begin{aligned} v_b &= \frac{1}{2-1}\left(4\left(\begin{bmatrix}2.25\\3.5\end{bmatrix} - \begin{bmatrix}3.96\\4.415\end{bmatrix}\right)^2 + 3\left(\begin{bmatrix}5.67\\5.33\end{bmatrix} - \begin{bmatrix}3.96\\4.415\end{bmatrix}\right)^2\right) \\ &= \left(4\left(\begin{bmatrix}-1.71\\-0.915\end{bmatrix}\right)^2 + 3\left(\begin{bmatrix}1.71\\0.915\end{bmatrix}\right)^2\right) \\ &= \left(4\left(\begin{bmatrix}-1.71 & -0.915\end{bmatrix}\begin{bmatrix}-1.71\\-0.915\end{bmatrix}\right) + 3\left(\begin{bmatrix}1.71 & 0.915\end{bmatrix}\begin{bmatrix}1.71\\0.915\end{bmatrix}\right)\right) \\ &= \big(4(3.761325) + 3((3.761325))\big) = 26.329275 \end{aligned}$$

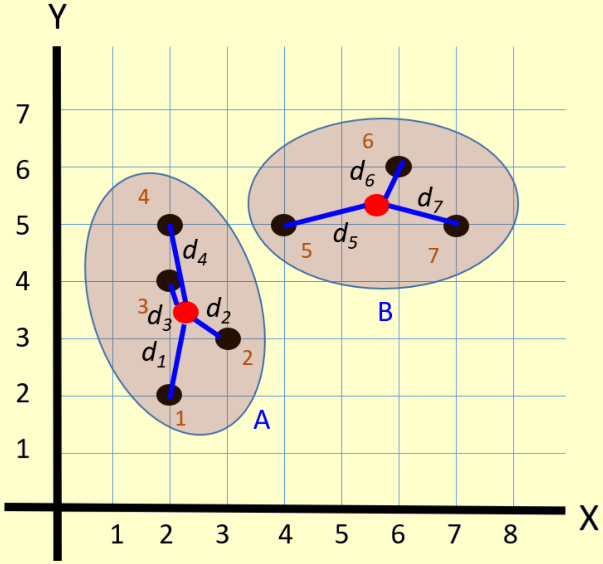
$$\begin{aligned} v &= \frac{v_W}{v_b} = \frac{2.2167}{26.329275} \\ &= 0.0842 \end{aligned}$$

Sum of Squared Error

The most widely used criterion to quantify cluster homogeneity is the Sum of Squared Error (SSE) criterion

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} \left\| m_{ij} - \bar{s}_i \right\|^2$$

$$\boxed{\text{SSE} = \frac{d_1+d_2+d_3+d_4+d_5+d_6+d_7}{7}}$$

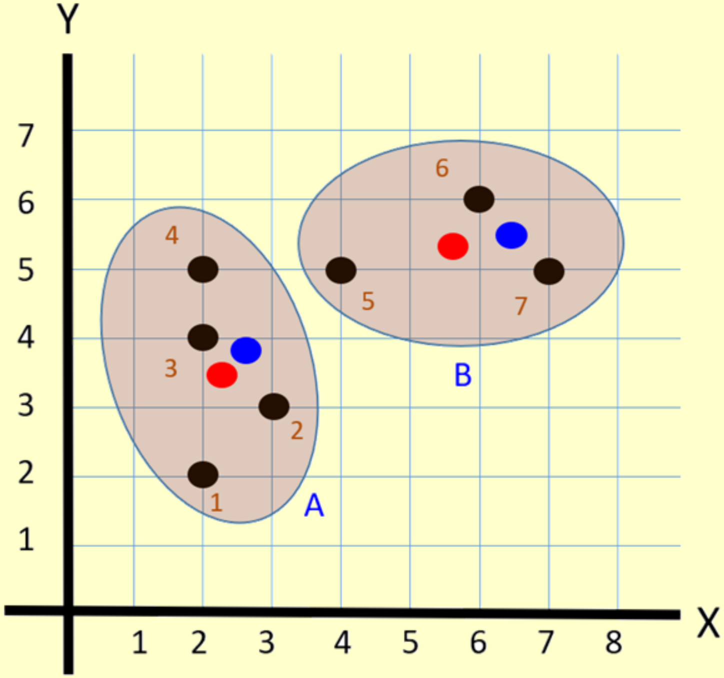


Error ratio

- Dipakai jika dataset yang digunakan adalah supervised
- Biasanya digunakan untuk mengukur tingkat presisi dari metode clustering
- Rumus:

$$Error = \frac{missclassified}{jumlahdata} \times 100\%$$

	X	Y	Class
Data1	2	2	1
Data2	3	3	1
Data3	2	4	1
Data4	2	5	1
Data5	4	5	1
Data6	6	6	2
Data7	7	5	2

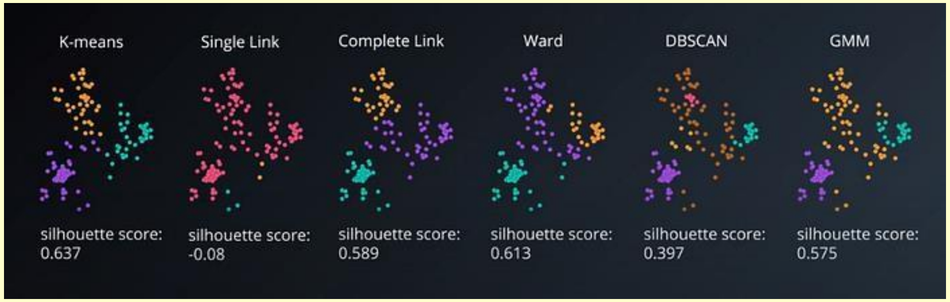


	Label	A → 1 B → 2	A → 2 B → 1
Data 1	1	1	2
Data 2	1	1	2
Data 3	1	1	2
Data 4	1	1	2
Data 5	1	2	1
Data6	2	2	1
Data7	2	2	1
Misclassified		1	6
Error ratio		14.3%	

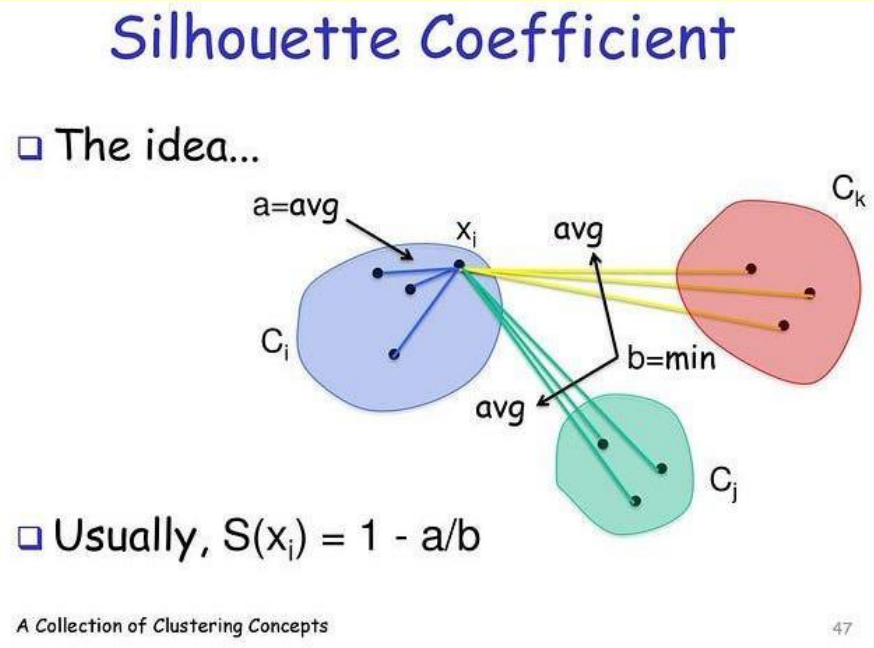
Silhoutte Score

The silhouette value is a measure of **how similar** an object is to its own cluster (**cohesion**) compared to other clusters (**separation**). The Silhouette coefficient is a value between **-1 and 1**, where **higher values indicate a better clustering**. This index is especially useful for high-dimensional datasets where visualizing the clustering's is not possible.

Cluster Validation is the process by which clustering's are scored after they are executed. This provides a means of comparing different clustering algorithms and their results on a certain dataset.



Silhoutte Score



Silhouette Coefficient

❑ The idea...

❑ Usually, $S(x_i) = 1 - a/b$

A Collection of Clustering Concepts 47

Sumber: Link

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

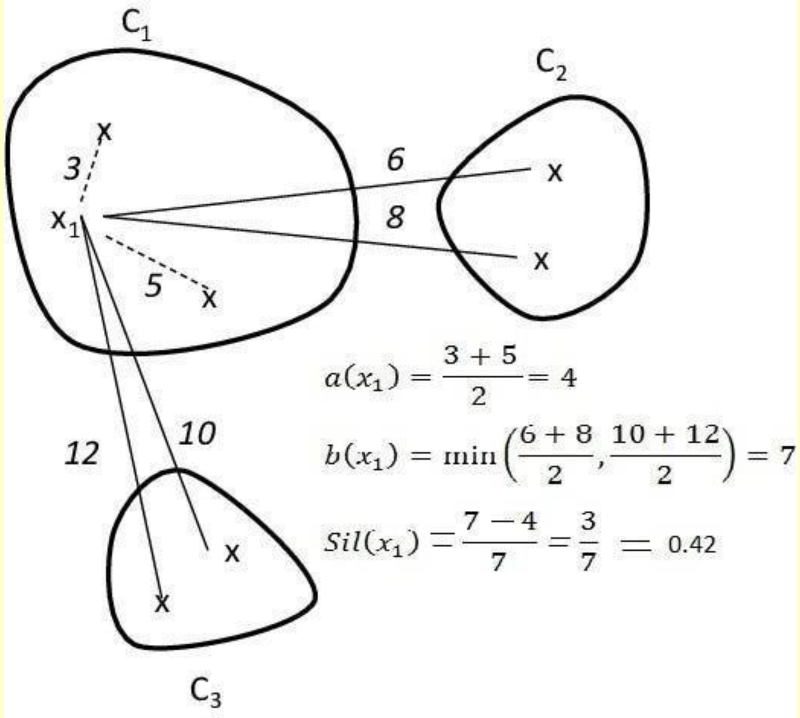
a(i) = avg distance of i to other point same cluster
C(i) = the number of points belonging to cluster i
d(i,j) = jarak data i ke j dalam cluster C(i)

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

b(i) = avg distance to nearest other cluster
C(k) = the number of points belonging to cluster k
d(i,j) = jarak data i ke j, dimana j anggota cluster k

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

Silhoutte Score



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$Sil(x_1) = \frac{7-4}{7} = \frac{3}{7} = 0.42$$

Overall Silhouette score for **the complete dataset can be calculated as the mean of silhouette score for all data points in the dataset.** As can be seen from the formula **silhouette score** would always lie between **-1 to 1** representing better clustering.

Davies and Bouldin

The Davies-Bouldin Index is a validation metric that is used to evaluate clustering models. It is calculated as **the average similarity measure of each cluster with the cluster most similar to it.** In this context, **similarity is defined as the ratio between inter-cluster and intra-cluster distances**. As such, this index ranks well-separated clusters with less dispersion as having a better score.

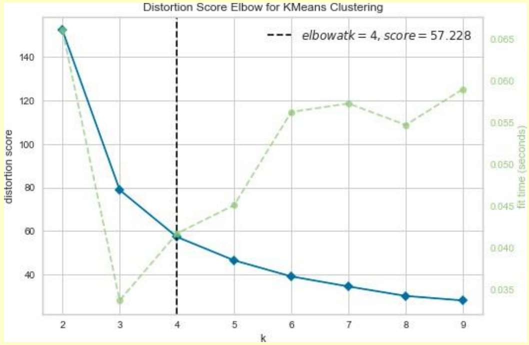
$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{s_i + s_j}{d_{ij}}\right)$$

where:

- k is the number of clusters
- s(i) is the average within-cluster distance of cluster i
- s(j) is the average within-cluster distance of cluster j
- d(i,j) is the between-cluster distance of cluster i and j

Finding optimal number of clusters in clustering

1. The Elbow method/ SSE Plot is used to find the elbow in the elbow plot. The elbow is found when the dataset becomes flat or linear after applying the cluster analysis algorithm. The elbow plot shows the elbow at the point where the number of clusters starts increasing.
2. The silhouette plot of a point measures how close that point lies to its nearest neighbor points, across all clusters. It provides information about clustering quality which can be used to determine whether further refinement by clustering should be performed on the current clustering.


Distortion Score Elbow for KMeans Clustering
elbow at k = 4, score = 57.228
distortion score
fit time (seconds)
k