SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

# ANALYSIS OF THE APPLICABILITY CRITERION FOR K MEANS CLUSTERING ALGORITHM RUN TEN NUMBER OF TIMES ON THE FIRST 25 NUMBERS OF THE FIBONACCI SERIES

**Guntuboyina Divya[1], R.Satya Ravindra Babu[2]**

[1]*Student-CSE Department,Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh.*

[2]*Associate Professor-CSE Department, Sanketika Vidya Parishad Engineering College, Visakhapatnam, AP*

**Article DOI: https://doi.org/10.36713/epra8497**
**DOI No: 10.36713/epra8497**

**ABSTRACT**

*In this research investigation Analysis Of The Applicability Criterion For K Means Clustering Algorithm Run Ten Number Of Times On The First 25 Numbers Of The Fibonacci Series is performed. For this analysis RCB Model Of Applicability Criterion For K Means Clustering Algorithm is used. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K- Means clustering algorithm is a scheme for clustering continuous and numeric data. As K-Means algorithm consists of scheme of random initialization of centroids, every time it is run, it gives different or slightly different results because it may reach some local optima. Quantification of such aforementioned variation is of some importance as this sheds light on the nature of the Discrete K-Means Objective function with regards its maxima and minima. The K-Means Clustering algorithm aims at minimizing the aforementioned Objective function. The RCB Model Of Applicability Criterion for K-Means Clustering aims at telling us if we can use the K-Means Clustering Algorithm on a given set of data within acceptable variation limits of the results of the K-Means Clustering Algorithm when it is run several times.*

**KEY WORDS**: *K-means clustering algorithm, RCB model and Cluster evaluation.*

## INTRODUCTION

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K clusters) fixed apriori. The main idea is to define K centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate K new centroids as barycentre of the clusters resulting from the previous step. After we have these K new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the K centres change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm aims at minimizing an objective function known as squared error function.

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

## OBJECTIVE OF THE PROJECT

Clusters the data into k groups where k is predefined. Select k points at random as cluster centres. Assign objects to their closest cluster centre according to the Euclidean distance function. Calculate the centroid or mean of all objects in each cluster.

## PURPOSE OF THE PROJECT

K-means clustering can be applied to machine learning or data mining Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation). Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization. K-means algorithm is useful for undirected knowledge discovery and is relatively simple. Kmeans has found wide spread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others.

## LITERATURE REVIEW

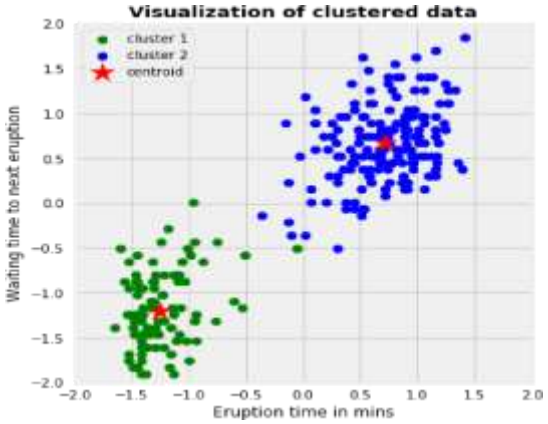
**K-means clustering:** K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. ... Data points are clustered based on feature similarity. Clusters the data into k groups where k is predefined. Select k points at random as cluster centres. Assign objects to their closest cluster centre according to the Euclidean distance function. Calculate the centroid or mean of all objects in each cluster

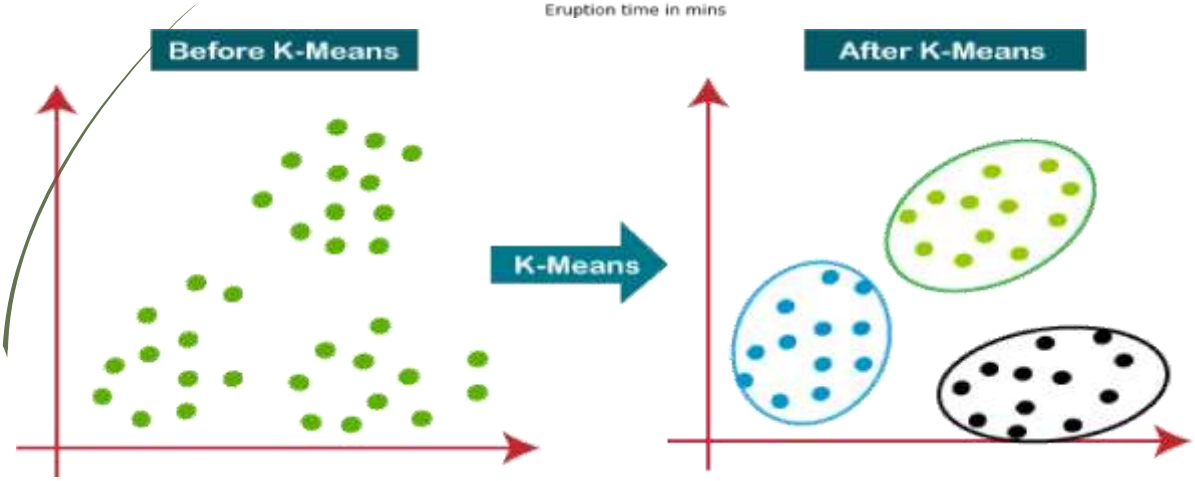
**How do you interpret k-means clustering?**

It calculates the sum of the square of the points and calculates the average distance. When the value of k is 1, the within-cluster sum of the square will be high. As the value of k increases, the within-cluster sum of square value will decrease.

**RCB model**: The RCB Model Of Applicability Criterion for K-Means Clustering

The RCB Model Of Applicability Criterion for K-Means Clustering is detailed in the following lines: The



SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

Applicability Criterion This analysis is presented for the univariate case of dataset. Let the data points be represented by xi = 1 to n

This analysis is presented for the univariate case of dataset.

Let the data points be represented by $x_i = 1 \text{ to } n$.

Let $x_{ij-Cavg}$ be the Cluster Average of the Cluster to which $x_j$ belongs to in the $j^{th}$ Run of the K-Means Clustering Algorithm.

Let $x_{ij-C\max}$ be the Cluster Maximum of the Cluster to which $x_i$ belongs to in the $j^{th}$ Run of the K-Means Clustering Algorithm.

Let $x_{ij-C\min}$ be the Cluster Minimum of the Cluster to which $x_i$ belongs to in the $j^{th}$ Run of the K-Means Clustering Algorithm.

Let $x_{ij-CSS}$ be the Cluster Silhouette Score of the Cluster to which $x_i$ belongs to in the $j^{th}$ Run of the K-Means Clustering Algorithm.

Let $x_{ij-CSSW}$ be the Cluster Sum Of Squares Within of the Cluster to which $x_i$ belongs to in the $j^{th}$

We now compute the Deviations

$$\delta_{ij-Cavg} = \left\{\frac{\sum_j x_{ij-Cavg}}{N}\right\} - \left(x_{ij-Cavg}\right)$$ where $N$ is the number of Runs of the K-Means Clustering Algorithm.

Similarly, we compute

$$\delta_{ij-C\max} = \left\{\frac{\sum_j x_{ij-C\max}}{N}\right\} - \left(x_{ij-C\max}\right)$$

$$\delta_{ij-C\min} = \left\{\frac{\sum_j x_{ij-C\min}}{N}\right\} - \left(x_{ij-C\min}\right)$$

$$\delta_{ij-CSS} = \left\{\frac{\sum_j x_{ij-CSS}}{N}\right\} - \left(x_{ij-CSS}\right)$$

$$\delta_{ij-CSSW} = \left\{\frac{\sum_j x_{ij-CSSW}}{N}\right\} - \left(x_{ij-CSSW}\right)$$

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

We now Min-Max Normalize in the Range [0, 1] all the above sets of values $\delta_{ij-Cavg}$, $\delta_{ij-C\max}$, $\delta_{ij-C\min}$, $\delta_{ij-CSS}$, $\delta_{ij-CSSW}$ separately (such aforementioned normalization done separately for each set of values). Let these thusly normalized values be represented by $\tilde{\delta}_{ij-Cavg}$, $\tilde{\delta}_{ij-C\max}$, $\tilde{\delta}_{ij-C\min}$, $\tilde{\delta}_{ij-CSS}$, $\tilde{\delta}_{ij-CSSW}$.

We also compute the (Sample) Standard Deviations of these aforementioned normalized sets of values, separately for each set, as:

$$\tilde{\delta}_{ij-Cavg} \rightarrow \sigma_{(ij-Cavg)S} = \left\{ \frac{\left\{ \left( \frac{\sum_j \tilde{\delta}_{ij-Cavg}}{N} \right) - \tilde{\delta}_{ij-Cavg} \right\}^2}{N-1} \right\}^{1/2}$$

$$\tilde{\delta}_{ij-C\max} \rightarrow \sigma_{(ij-C\max)S} = \left\{ \frac{\left\{ \left( \frac{\sum_j \tilde{\delta}_{ij-C\max}}{N} \right) - \tilde{\delta}_{ij-C\max} \right\}^2}{N-1} \right\}^{1/2}$$

$$\tilde{\delta}_{ij-C\min} \rightarrow \sigma_{(ij-C\min)S} = \left\{ \frac{\left\{ \left( \frac{\sum_j \tilde{\delta}_{ij-C\min}}{N} \right) - \tilde{\delta}_{ij-C\min} \right\}^2}{N-1} \right\}^{1/2}$$

$$\tilde{\delta}_{ij-CSS} \rightarrow \sigma_{(ij-CSS)S} = \left\{ \frac{\left\{ \left( \frac{\sum_j \tilde{\delta}_{ij-CSS}}{N} \right) - \tilde{\delta}_{ij-CSS} \right\}^2}{N-1} \right\}^{1/2}$$

$$\tilde{\delta}_{ij-CSSW} \rightarrow \sigma_{(ij-CSSW)S} = \left\{ \frac{\left\{ \left( \frac{\sum_j \tilde{\delta}_{ij-CSSW}}{N} \right) - \tilde{\delta}_{ij-CSSW} \right\}^2}{N-1} \right\}^{1/2}$$

As the total number of unique results possible in a K-Means Clustering Algorithm for making K Clusters with $n$ data points is given by $m = {}^nC_K$

the Population Standard Deviations of the above Sample Standard Deviations are given by

$$\sigma_{(ij-Cavg)Pop} = \sqrt{m} \cdot \sigma_{(ij-Cavg)S}$$

$$\sigma_{(ij-C\max)Pop} = \sqrt{m} \cdot \sigma_{(ij-C\max)S}$$

$$\sigma_{(ij-C\min)Pop} = \sqrt{m} \cdot \sigma_{(ij-C\min)S}$$

$$\sigma_{(ij-CSS)Pop} = \sqrt{m} \cdot \sigma_{(ij-CSS)S}$$

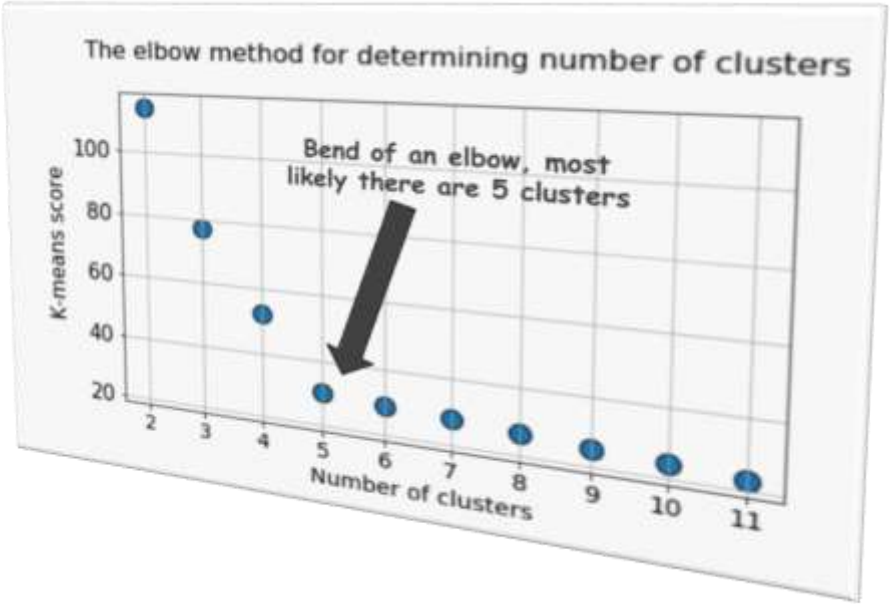
$$\sigma_{(ij-CSSW)Pop} = \sqrt{m} \cdot \sigma_{(ij-CSSW)S}$$

We now Min-Max Normalize in the Range [0, 1] all the above sets of values $\sigma_{(ij-Cavg)Pop}$, $\sigma_{(ij-C\max)Pop}$, $\sigma_{(ij-C\min)Pop}$, $\sigma_{(ij-CSS)Pop}$, $\sigma_{(ij-CSSW)Pop}$ separately (such aforementioned normalization done separately for each set of values). Let these thusly normalized values be represented by $\bar{\sigma}_{(ij-Cavg)Pop}$, $\bar{\sigma}_{(ij-C\max)Pop}$, $\bar{\sigma}_{(ij-C\min)Pop}$, $\bar{\sigma}_{(ij-CSS)Pop}$, $\bar{\sigma}_{(ij-CSSW)Pop}$

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

We now find a weighted average v.We can then say r=(1-v) as the Coefficient Of Robustness of the results of the K-means Clustering Algorithm for a given data set.

**Elbow method**: The most common approach for answering this question is the so-called elbow method. It involves running the algorithm multiple times over a loop, with an increasing number of cluster choice and then plotting a clustering score as a function of the number of clusters. The score is, in general, a measure of the input data on the k-means objective function i.e, Some form of intra-cluster distance relative to inner-cluster distance.Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (sse) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow.




**Sample Screenshots**:

Elbow Curve Method:

The elbow method runs k-means clustering on the dataset for a range of values of k (say 1 to 10).

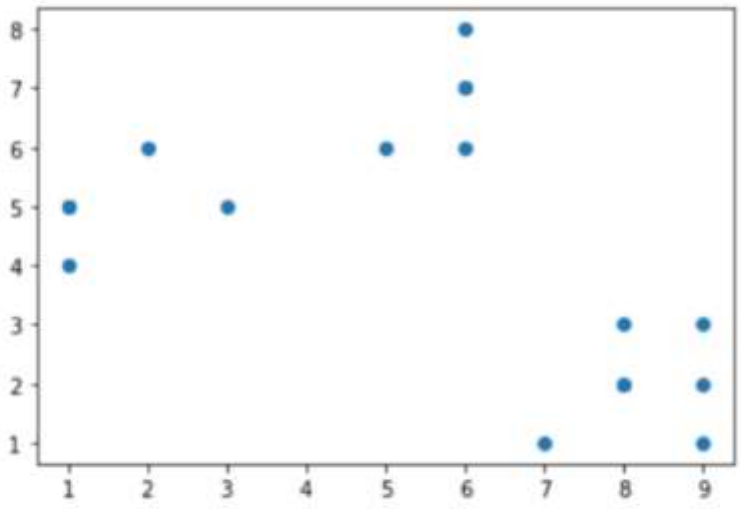
- Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
- Plot these points and find the point where the average distance from the centroid falls suddenly ("Elbow").

Let us see the python code with the help of an example.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
X1 = [3, 1, 1, 2, 1, 6, 6, 6, 5, 6, 7, 8, 9, 8, 9, 9, 8]
X2 = [5, 4, 5, 6, 5, 8, 6, 7, 6, 7, 1, 2, 1, 2, 3, 2, 3]
plt.scatter(X1,X2)
plt.show()
```

Getting Optimal Clusters:

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal



Scatter Plot between X1 and X2

Visually we can see the optimal number of clusters should be 3.
Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
Plot these points and find the point where the average distance from the centroid falls suddenly .

Elbow Method:

```
Sum_of_squared_distances = []
K = range(1,10)
for num_clusters in K :
 kmeans = KMeans(n_clusters=num_clusters)
 kmeans.fit(data_frame)
 Sum_of_squared_distances.append(kmeans.inertia_)
plt.plot(K,Sum_of_squared_distances,'bx-')
plt.xlabel('Values of K')
plt.ylabel('Sum of squared distances/Inertia')
plt.title('Elbow Method For Optimal k')
plt.show()
```

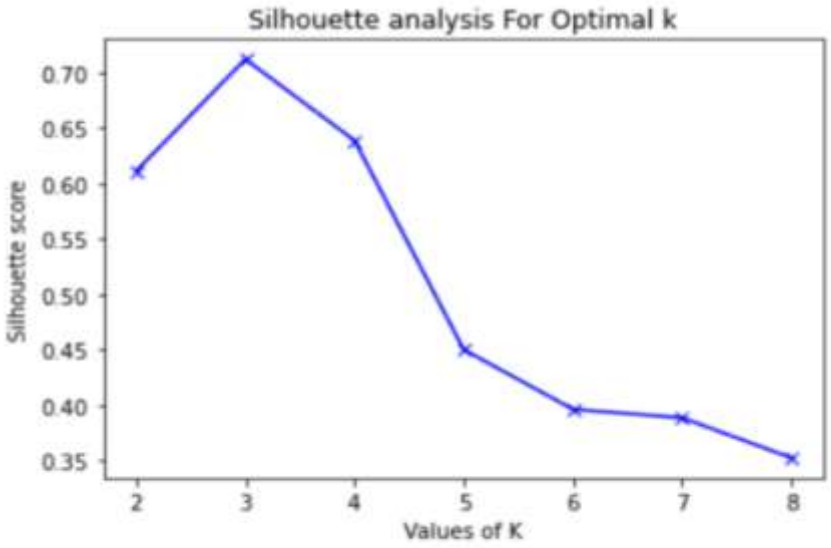
SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

Silhouette Score:

```
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
silhouette_avg = []
for num_clusters in range_n_clusters:

 # initialise kmeans
 kmeans = KMeans(n_clusters=num_clusters)
 kmeans.fit(data_frame)
 cluster_labels = kmeans.labels_

 # silhouette score
 silhouette_avg.append(silhouette_score(data_frame, cluster_labels))plt.plot(ran
plt.xlabel('Values of K')
plt.ylabel('Silhouette score')
plt.title('Silhouette analysis For Optimal k')
plt.show()
```




Line plot between K and Silhouette score

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 9 | September 2021 - Peer Reviewed Journal

## REFERENCES

[1] *Lloyd, Stuart P., "Least squares quantization in PCM". Bell Telephone Laboratories Paper.* (1957)

[2] *Lloyd, Stuart P.,* "Least squares quantization in PCM", *IEEE Transactions on Information Theory, Vol 28 No 2, pp 129–137. (1982)*

[3] *MacQueen, J. B.,* Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. (1967).*

[4] Shraddha Shukla and Naganna S., A Review On K-means Data Clustering Approach, *International Journal of Information & Computation Technology*. Vol 4, No 17, pp. 1847-1860 (2014)

[5] Giri, Chandini., Rallabhandi, Satya Ravindra Babu., *Quantification Of Variation Of Results Of The K-Means Clustering Algorithm Run Ten Times On The First Twenty Five Primes – A Criterion For Applicability Of The K-Means Clustering.*

[6] *Bagadi, R. C.*, R B Ideas Journal, Volume 1, Issue 5, Twelfth Edition, October 2021. *Applicability Criterion For K Means Clustering Algorithm*. Independently Published by Amazon – Kindle Direct Publishing, USA. July 23rd 2021
ISBN: 9798542222288

[7]https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[8] *Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems.* ***52*** *(2): 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241.*