

DESCRIPTIVE MODELING

Subhashini Natarajan



JULY 4, 2022

SUBHASHINI NATARAJAN

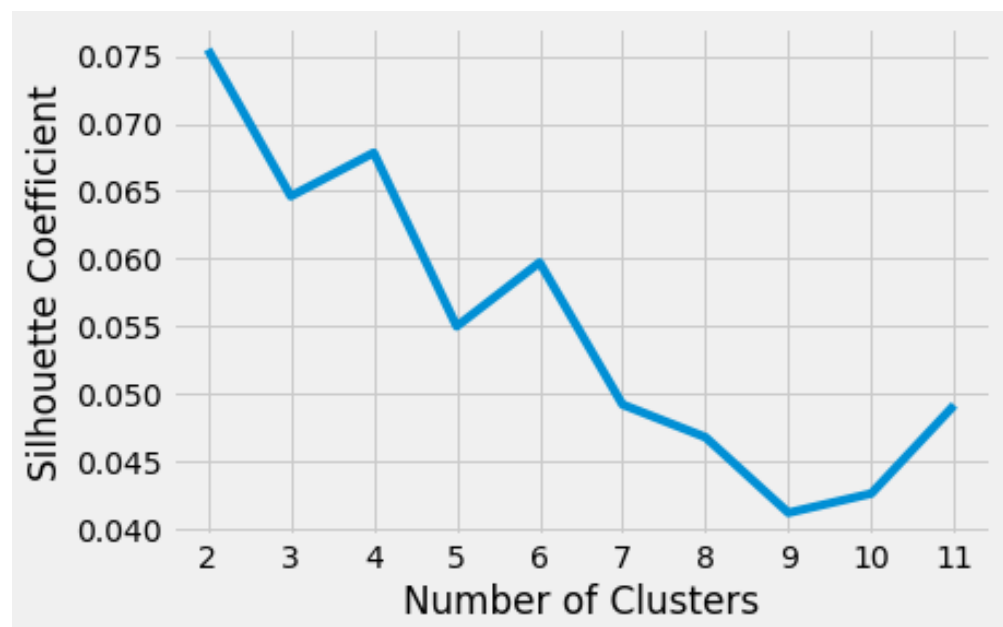
Objective:

The objective of the project is to apply clustering methods on dataset of [Amyotrophic Lateral Sclerosis \(ALS\)](#), patients.

Steps:

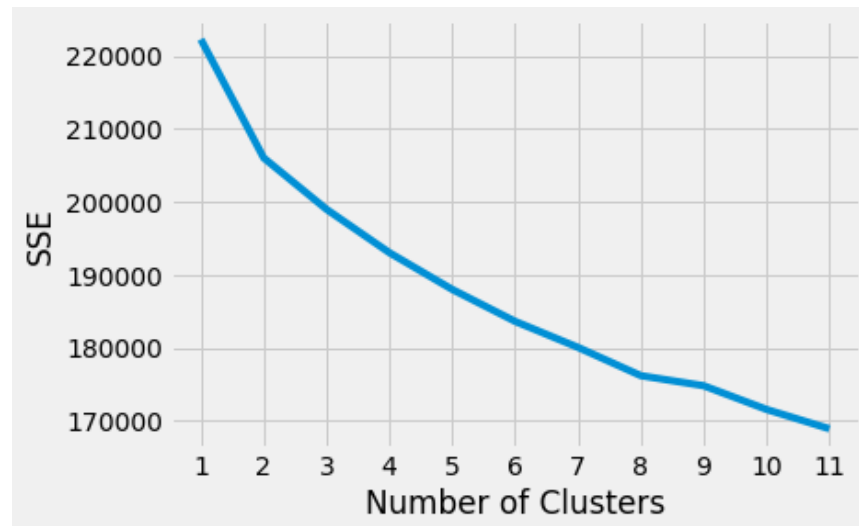
The following steps are performed as instructed.

1. Data irrelevant to patient's ALS condition to be removed -
As not an expert in ALS condition, removed only the id field and retained the other features pertaining to medical conditions.
2. Standard Scalar to be applied to the data -
Feature scaling is the preprocessing step for clustering. Standard scalar scales the data values so that the features have mean of 0 and standard deviation of 1. For the given dataset, the standard scalar method is applied and the scaled values are stored in x_std.
3. Determine and plot silhouette coefficient –
Silhouette coefficient measures the cluster cohesion and separation, quantifying how well the data point fits into the assigned cluster. For the dataset, silhouette coefficient is determined and plotted as below.



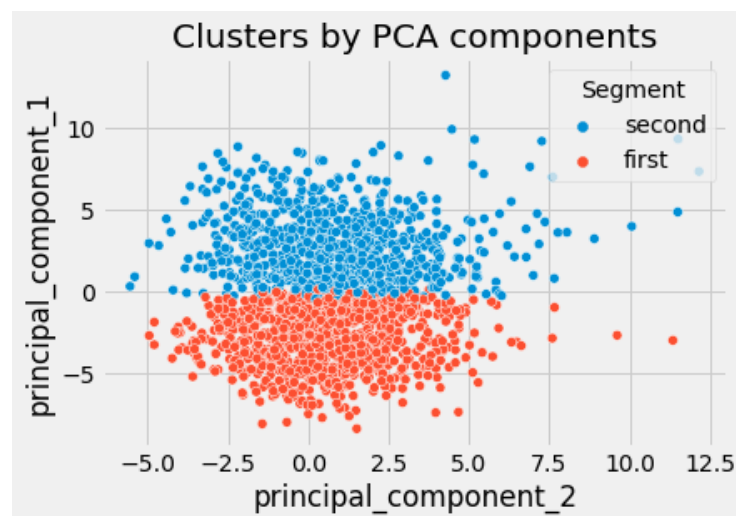
4. Optimal Number of clusters –

From the plot number of clusters vs SSE, after cluster 2, the plot continually drops down indicating overlap among the data values.



Also, the silhouette coefficient is maximum for the number of clusters = 2. Hence chose optimal number of clusters as 2.

5. K-means performed with optimal cluster number 2.
6. Principal Component Analysis performed to reduce the dimensionality, by combining variables into bigger, more meaningful features.
7. PCA components are plotted and the results are as below –



Summary:

The ALS patients can be classified into two different groups. ALSFRS-R progression slope that normalizes the score by the duration of symptoms is suggested as a predictor of survival. Based on this dataset and clustering, the patients are classified as one with faster progression of the disease and short survival duration, the other with slower progression of the disease and longer survival duration.

References:

https://en.wikipedia.org/wiki/ALS_Functional_Rating_Scale_-_Revised

<https://www.kaggle.com/competitions/als-progression>

<https://realpython.com/k-means-clustering-python/>