# Clustering

1. **About the model**

When you have unlabeled data, K-means clustering is a sort of unsupervised learning (i.e., data without defined categories or groups). The purpose of this technique is to locate groups in the data, with K representing the number of groups. Based on the attributes provided, the algorithm assigns each data point to one of K groups iteratively. Data points are grouped together based on how comparable their features are. The K-means clustering technique yields the following results:

1. The K clusters' centroids, which can be used to label new data.
2. Data labels for the training (each data point is assigned to a single cluster).

**Elbow Method**

The number of clusters ranges K from one to ten. We calculate WCSS for each value of K. ( Within-Cluster Sum of Square ). In a cluster, WCSS is the sum of squared distances between each point and the centroid. The plot appears like an Elbow when we plot the WCSS with the K value. The WCSS value will begin to fall as the number of clusters grows. When K = 1, the WCSS value is the highest. When we examine the graph, we can see that it will shift rapidly at a point, forming an elbow shape. The graph begins to travel practically parallel to the X-axis at this point. The ideal K value, or the optimal number of clusters, corresponds to this point.

1. Dataset

Amyotrophic lateral sclerosis, or ALS (also known in the US as Lou Gehrig’s Disease and as

Motor Neuron Disease in the UK) is a disease that involves the degeneration and death of the

nerve cells in the brain and spinal cord that control voluntary muscle movement. Death

typically occurs within 3 - 5 years of diagnosis. Only about 25% of patients survive for more

than 5 years after diagnosis.

The given data set is ALS, Training & Test data for the ALS where various factors are taken into account.

The given dataset names are :

Train :- ALS\_TrainingData\_2223.csv

Test :- ALS\_TestingData\_78.csv

For the train data there are 2223 rows and 101 columns,( ['ID', 'Age\_mean', 'Albumin\_max', 'Albumin\_median', 'Albumin\_min', 'Albumin\_range', 'ALSFRS\_slope', 'ALSFRS\_Total\_max', 'ALSFRS\_Total\_median', 'ALSFRS\_Total\_min', 'ALSFRS\_Total\_range', 'ALT.SGPT.\_max', 'ALT.SGPT.\_median', 'ALT.SGPT.\_min', 'ALT.SGPT.\_range', 'AST.SGOT.\_max', 'AST.SGOT.\_median', 'AST.SGOT.\_min', 'AST.SGOT.\_range', 'Bicarbonate\_max', 'Bicarbonate\_median', 'Bicarbonate\_min', 'Bicarbonate\_range', 'Blood.Urea.Nitrogen..BUN.\_max', 'Blood.Urea.Nitrogen..BUN.\_median', 'Blood.Urea.Nitrogen..BUN.\_min', 'Blood.Urea.Nitrogen..BUN.\_range', 'bp\_diastolic\_max', 'bp\_diastolic\_median', 'bp\_diastolic\_min', 'bp\_diastolic\_range', 'bp\_systolic\_max', 'bp\_systolic\_median', 'bp\_systolic\_min', 'bp\_systolic\_range', 'Calcium\_max', 'Calcium\_median', 'Calcium\_min', 'Calcium\_range', 'Chloride\_max', 'Chloride\_median', 'Chloride\_min', 'Chloride\_range', 'Creatinine\_max', 'Creatinine\_median', 'Creatinine\_min', 'Creatinine\_range', 'Gender\_mean', 'Glucose\_max', 'Glucose\_median', 'Glucose\_min', 'Glucose\_range', 'hands\_max', 'hands\_median', 'hands\_min', 'hands\_range', 'Hematocrit\_max', 'Hematocrit\_median', 'Hematocrit\_min', 'Hematocrit\_range', 'Hemoglobin\_max', 'Hemoglobin\_median', 'Hemoglobin\_min', 'Hemoglobin\_range', 'leg\_max', 'leg\_median', 'leg\_min', 'leg\_range', 'mouth\_max', 'mouth\_median', 'mouth\_min', 'mouth\_range', 'onset\_delta\_mean', 'onset\_site\_mean', 'Platelets\_max', 'Platelets\_median', 'Platelets\_min', 'Potassium\_max', 'Potassium\_median', 'Potassium\_min', 'Potassium\_range', 'pulse\_max', 'pulse\_median', 'pulse\_min', 'pulse\_range', 'respiratory\_max', 'respiratory\_median', 'respiratory\_min', 'respiratory\_range', 'Sodium\_max', 'Sodium\_median', 'Sodium\_min', 'Sodium\_range', 'SubjectID', 'trunk\_max', 'trunk\_median', 'trunk\_min', 'trunk\_range', 'Urine.Ph\_max', 'Urine.Ph\_median', 'Urine.Ph\_min'])

For the test data we have 78 rows and 131 columns, (['ID', 'Age\_mean', 'Albumin\_max', 'Albumin\_median', 'Albumin\_min', 'Albumin\_range', 'ALSFRS\_slope', 'ALSFRS\_Total\_max', 'ALSFRS\_Total\_median', 'ALSFRS\_Total\_min', 'ALSFRS\_Total\_range', 'ALT.SGPT.\_max', 'ALT.SGPT.\_median', 'ALT.SGPT.\_min', 'ALT.SGPT.\_range', 'AST.SGOT.\_max', 'AST.SGOT.\_median', 'AST.SGOT.\_min', 'AST.SGOT.\_range', 'Basophils\_max', 'Basophils\_median', 'Basophils\_min', 'Basophils\_range', 'Bicarbonate\_max', 'Bicarbonate\_median', 'Bicarbonate\_min', 'Bicarbonate\_range', 'Bilirubin..total.\_max', 'Bilirubin..total.\_median', 'Bilirubin..total.\_min', 'Bilirubin..total.\_range', 'Blood.Urea.Nitrogen..BUN.\_max', 'Blood.Urea.Nitrogen..BUN.\_median', 'Blood.Urea.Nitrogen..BUN.\_min', 'Blood.Urea.Nitrogen..BUN.\_range', 'BMI\_max', 'bp\_diastolic\_max', 'bp\_diastolic\_median', 'bp\_diastolic\_min', 'bp\_diastolic\_range', 'bp\_systolic\_max', 'bp\_systolic\_median', 'bp\_systolic\_min', 'bp\_systolic\_range', 'Calcium\_max', 'Calcium\_median', 'Calcium\_min', 'Calcium\_range', 'Chloride\_max', 'Chloride\_median', 'Chloride\_min', 'Chloride\_range', 'Creatinine\_max', 'Creatinine\_median', 'Creatinine\_min', 'Creatinine\_range', 'Eosinophils\_max', 'Eosinophils\_median', 'Eosinophils\_min', 'Eosinophils\_range', 'Gender\_mean', 'Glucose\_max', 'Glucose\_median', 'Glucose\_min', 'Glucose\_range', 'hands\_max', 'hands\_median', 'hands\_min', 'hands\_range', 'Hematocrit\_max', 'Hematocrit\_median', 'Hematocrit\_min', 'Hematocrit\_range', 'Hemoglobin\_max', 'Hemoglobin\_median', 'Hemoglobin\_min', 'Hemoglobin\_range', 'leg\_max', 'leg\_median', 'leg\_min', 'leg\_range', 'Lymphocytes\_max', 'Lymphocytes\_median', 'Lymphocytes\_min', 'Lymphocytes\_range', 'Monocytes\_max', 'Monocytes\_median', 'Monocytes\_min', 'Monocytes\_range', 'mouth\_max', 'mouth\_median', 'mouth\_min', 'mouth\_range', 'onset\_delta\_mean', 'onset\_site\_mean', 'Platelets\_max', 'Platelets\_median', 'Platelets\_min', 'Potassium\_max', 'Potassium\_median', 'Potassium\_min', 'Potassium\_range', 'pulse\_max', 'pulse\_median', 'pulse\_min', 'pulse\_range', 'Red.Blood.Cells..RBC.\_max', 'Red.Blood.Cells..RBC.\_median', 'Red.Blood.Cells..RBC.\_min', 'Red.Blood.Cells..RBC.\_range', 'respiratory\_max', 'respiratory\_median', 'respiratory\_min', 'respiratory\_range', 'Sodium\_max', 'Sodium\_median', 'Sodium\_min', 'Sodium\_range', 'SubjectID', 'trunk\_max', 'trunk\_median', 'trunk\_min', 'trunk\_range', 'Urine.Ph\_max', 'Urine.Ph\_median', 'Urine.Ph\_min', 'Urine.Ph\_range', 'White.Blood.Cell..WBC.\_max', 'White.Blood.Cell..WBC.\_median', 'White.Blood.Cell..WBC.\_min', 'White.Blood.Cell..WBC.\_range'])

Where we use few columns for the predictions by using **k-Means** algorithm, with different K vales and clusters.

Loading the data and importing required packages.

A screenshot of a computer

Description automatically generated with medium confidence

1. Data pre-processing

The train dataset & test data set is imported using pandas with df\_train & df\_test data frame.

The df\_train data set with top 5 rows using head function

A screenshot of a computer

Description automatically generated with medium confidence

The shape of the train data set

Graphical user interface, application

Description automatically generated

Checking any null values in the training data

Text

Description automatically generated

Description of the train data

A screen shot of a computer

Description automatically generated with low confidence

Same as for test data

A screenshot of a computer screen

Description automatically generated with medium confidence

Shape of test data set

Graphical user interface, application, website

Description automatically generated

Checking any null values in the test data set

Text

Description automatically generated

The description of the test data

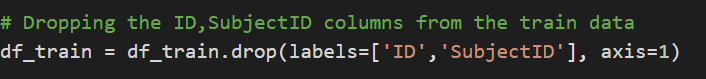
A screen shot of a computer

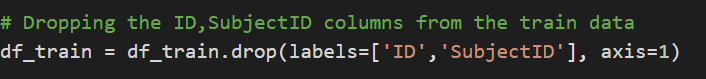
Description automatically generated with low confidence

1. Data visualizations

Creating graph for the train data, which are effective for the prediction in devastating disease.

For this removing the columns, doesn’t play a major account, which are **ID,Subject ID**

****

****

from the train data, axis=1 indicate 1st column where columns are present, in result the new data is stored in the df\_train

Creating histograph for the train data, few graphs are below

Graphical user interface

Description automatically generated

Graphs are created for each column present in the train data.

The density plots for the train data set, which we can understand the observation in data.

Chart, line chart

Description automatically generated

Getting log values for the train data to each column

The natural logarithm log is the inverse of the exponential function

Text

Description automatically generated

Creating density graph for the log train data set

A picture containing chart

Description automatically generated

Heat map:

The correlation for the train data is calculated by using corr() function



Creating heatmap for the corr train data

A screenshot of a computer

Description automatically generated with medium confidence



1. The data fitting

By using Principal component analysis clustering from the sklearn package.

Graphical user interface, text, application

Description automatically generated

It can be used for feature selection and visualization of higher-dimensional data when the feature is large. It is an unsupervised learning technique that can be used to identify patterns, clusters, and possibly any latent information.

Fitting the df\_train with PCA, transforming into the x\_train data

The idea behind StandardScaler is that it will modify your data so that its distribution has a mean value of 0 and a standard deviation of 1.

In the case of multivariate data, this is done feature by feature (in other words independently for each column of the data).

Given the distribution of the data, each value in the dataset will have the mean value subtracted and then divided by the standard deviation of the entire dataset (or feature in the multivariate case).

Graphical user interface, text

Description automatically generated

For the test data

Graphical user interface, text

Description automatically generated

1. Applying the model

The ELBOW METHOD is for assisting data scientists in determining the ideal number of clusters by fitting the model with a range of K values.

Text

Description automatically generated

The result graph is:

Chart

Description automatically generated with medium confidence

The maximum accuracy occurs from the K value is 2(K=2) to the K value is 4(K=4) from the graph above.

1. The results and test score

Creating K\_M\_model function, for K vales

Text

Description automatically generated

This function returns cluster lables and the center values

Testing with value K=2

Graphical user interface, application

Description automatically generated

For K=3

Graphical user interface, application

Description automatically generated

For K=4 is also created as shown above

Visualizing the cluster results, where K=4 cluster label added to the train data

A screenshot of a computer

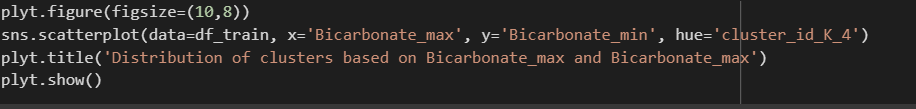
Description automatically generated with medium confidence

Creating inverse log transform for the above train data

A screenshot of a computer

Description automatically generated

TheDistribution of Clusters with Bicarbonate\_max and Bicarbonate\_min, with hue = cluster\_id\_K\_4(k=4)



The graphs shows:

Chart, scatter chart

Description automatically generated

TheDistribution of Clusters with Hemoglobin\_median & Hemoglobin\_range with hue = cluster\_id\_K\_4(k=4)

Chart, scatter chart

Description automatically generated

And created for few variables above graphs with clusters o-4

1. Metrics

Calculating the silhouette score value for different cluster

When **K=2**

Graphical user interface, text, website

Description automatically generated

When K=3

Graphical user interface, text, website

Description automatically generated

When k=4

Text

Description automatically generated

1. Conclusion

The silhouette coefficient, often known as the silhouette score, is a statistic used to assess the effectiveness of a clustering procedure. Its value ranges from -1 to 1.

1: Clusters are well separated and different from one another.

0 denotes that the clusters are unrelated or that the distance between them is insignificant.

-1: Clusters were wrongly assigned.

So, from the above silhouette coefficient value for different K values

When K=2, the silhouette coefficient value is 0.64

When k=3, the silhouette coefficient value is 0.59

When k=4, the silhouette coefficient value is 0.56

From the 3 different K values, we can conclude that when **K=2** is the best fit for the model.