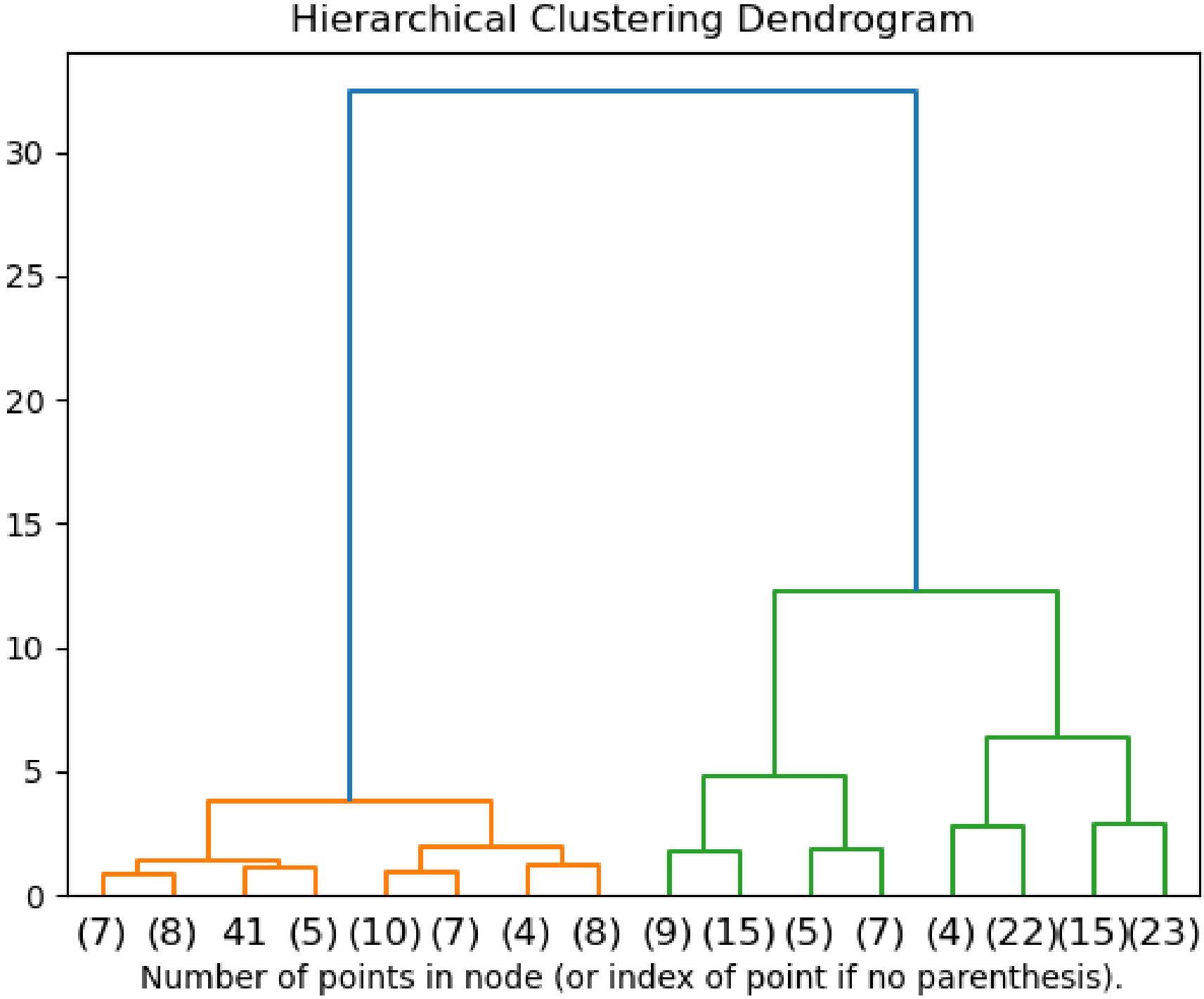
Analytics Vidhya

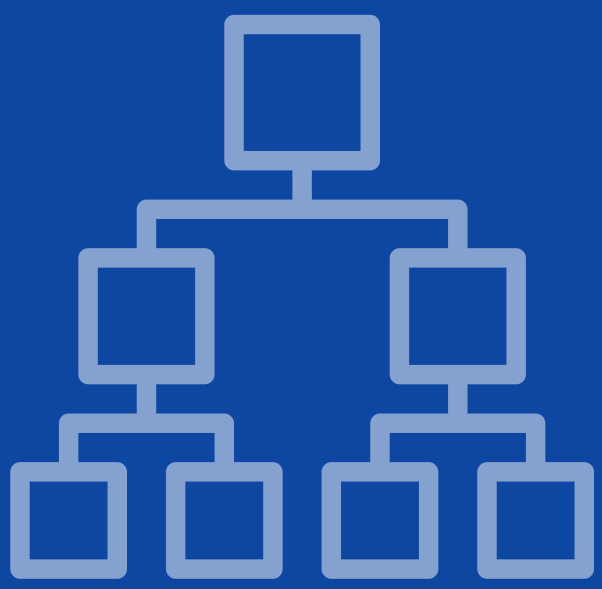
A Guide to Hierarchical Clustering



Analytics Vidhya

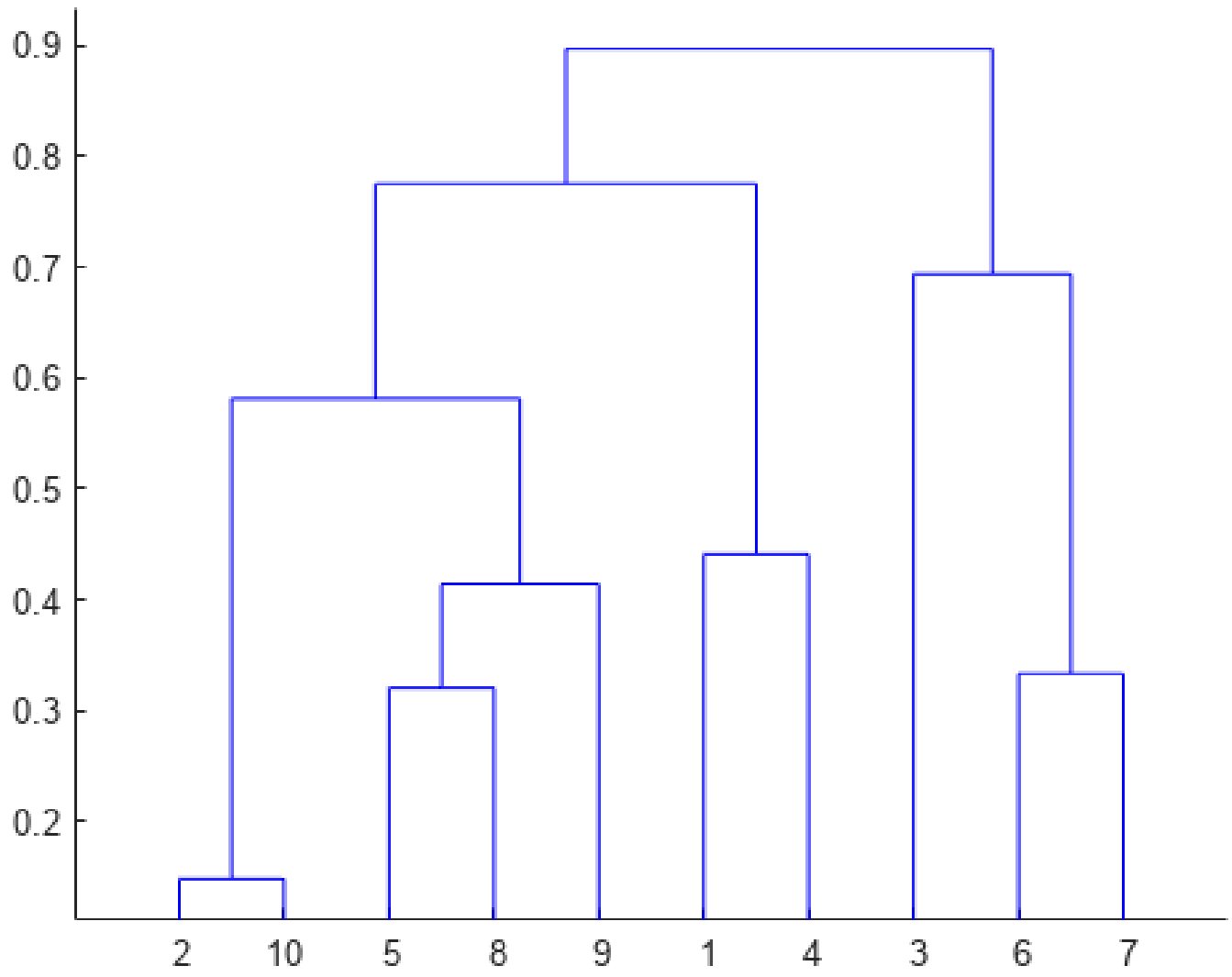
Imagine, you find yourself in a land of data, where customers roam freely and patterns hide in plain sight. As a brave data scientist you embark on a quest to uncover the secrets of customer segmentation.

Fear not, for the answer lies in the enchanting realm of H**ierarchical Clustering**!



What is Hierarchical Clustering?

Hierarchical clustering merges similar clusters iteratively, starting with each data point as a separate cluster. This creates a tree-like structure that shows the relationships between clusters and their hierarchy. This is known as **dendrogram**.

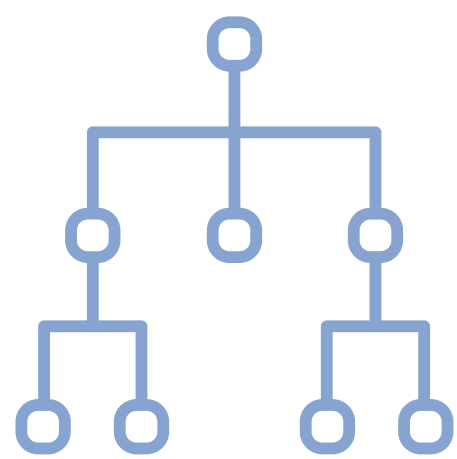


Why Hierarchical Clustering?

- Unlike some other algorithms, hierarchical clustering does not require specifying the number of clusters at the beginning.

- Starting with individual items, it gradually creates clusters so you can explore different granularities.

- It is particularly useful when the optimal number of clusters is unknown or when you want to analyze data at different levels of detail.



Types of Hierarchical Clustering

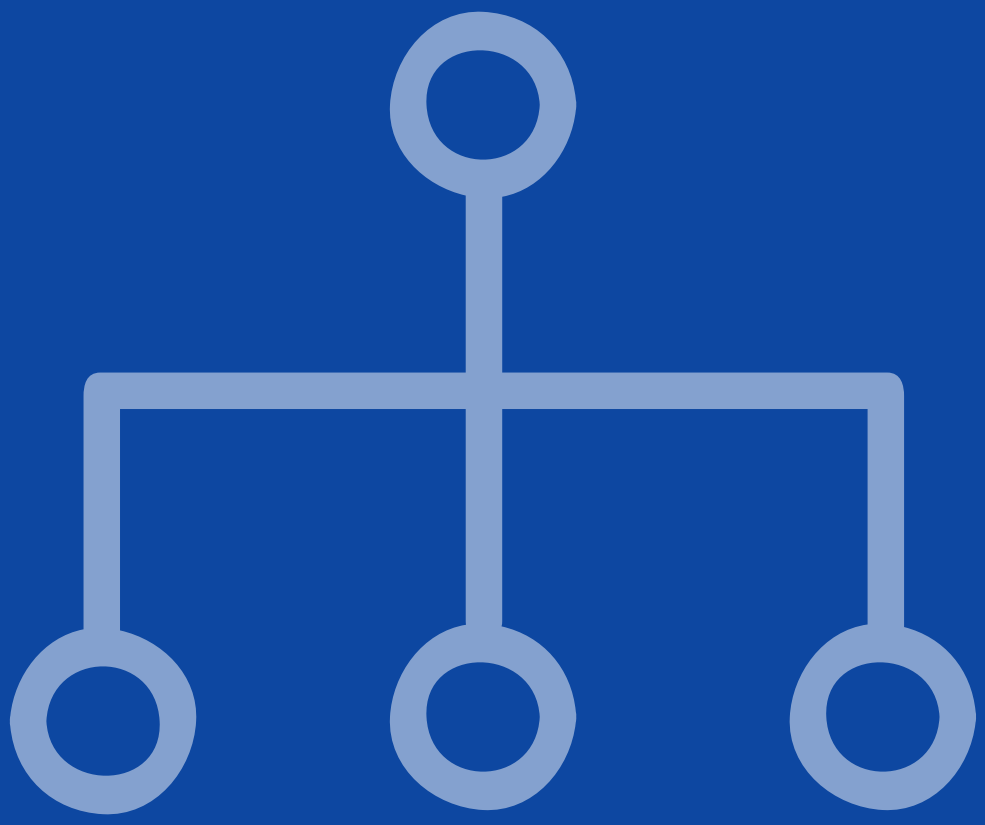
There are mainly two types of hierarchical clustering:

- **Agglomerative Hierarchical Clustering**
- **Divisive Hierarchical Clustering**

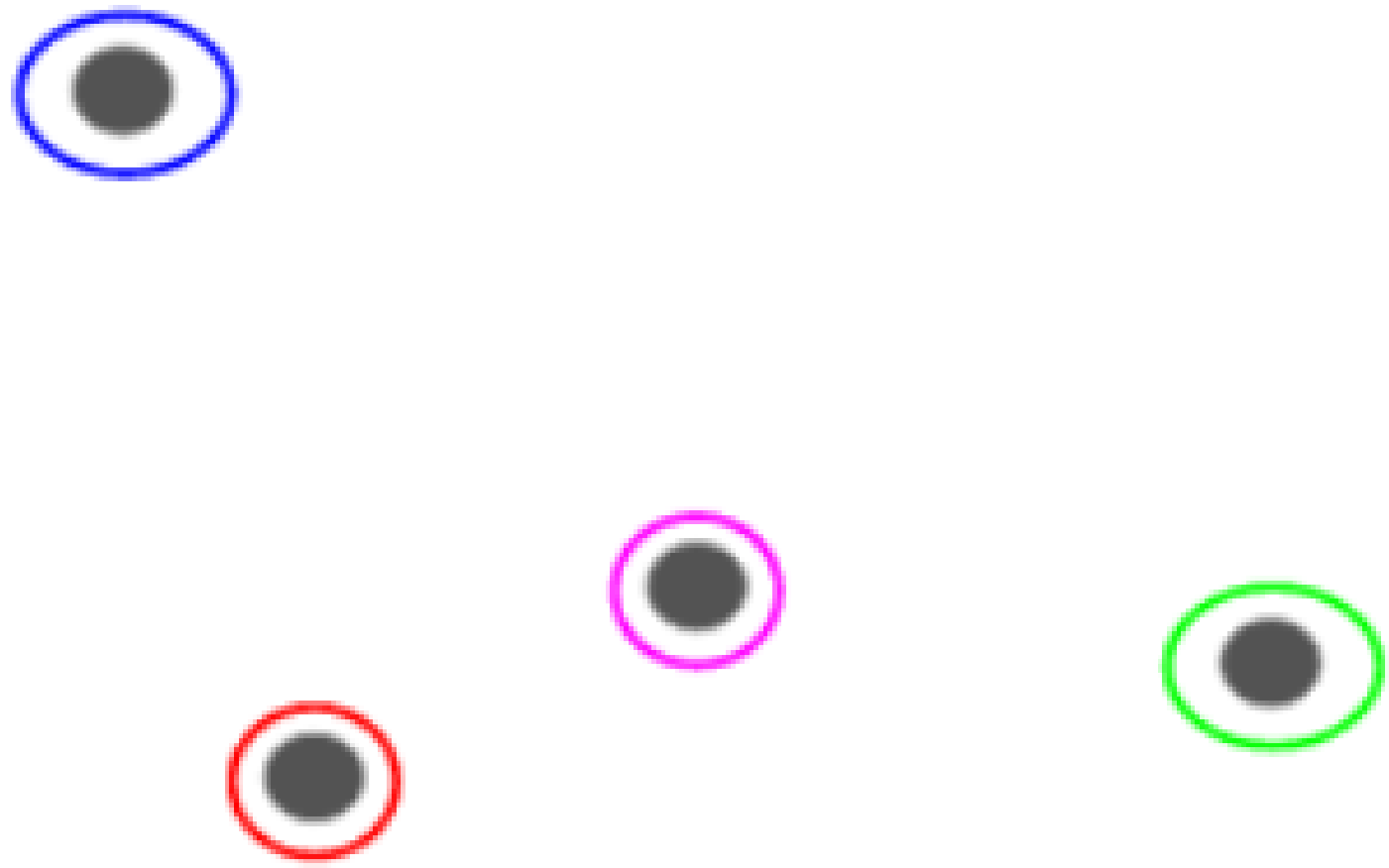
Let’s understand each type in detail.

Analytics Vidhya

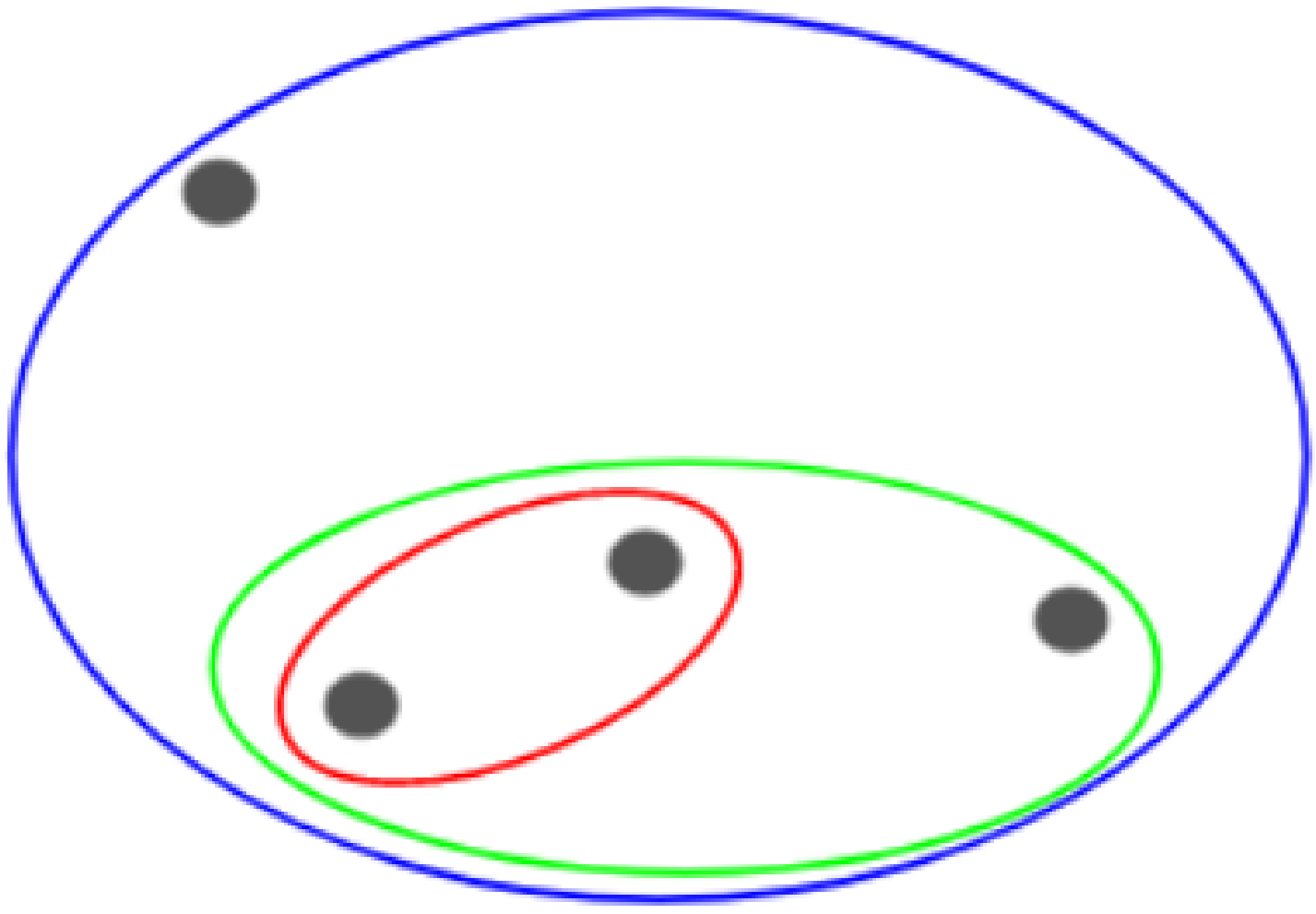
Agglomerative Hierarchical Clustering



We assign each point to an individual cluster in this technique. Suppose there are 4 data points. We will assign each of these points to a cluster and hence will have 4 clusters in the beginning.



Then, at each iteration, we merge the closest pair of clusters and repeat this step until only a single cluster is left:

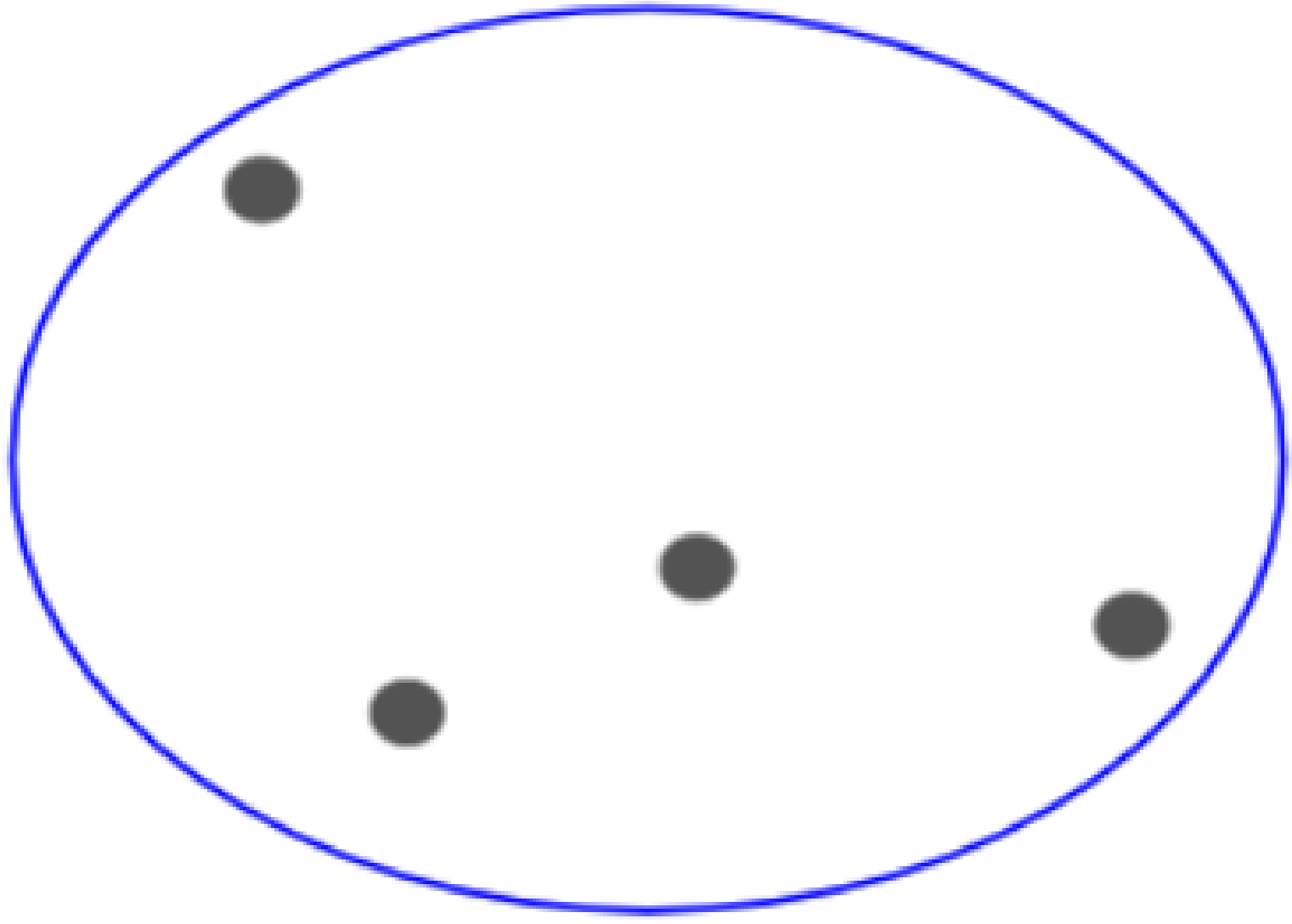


Since we are merging (or adding) the clusters at each step this type of clustering is also known as **additive hierarchical clustering**.

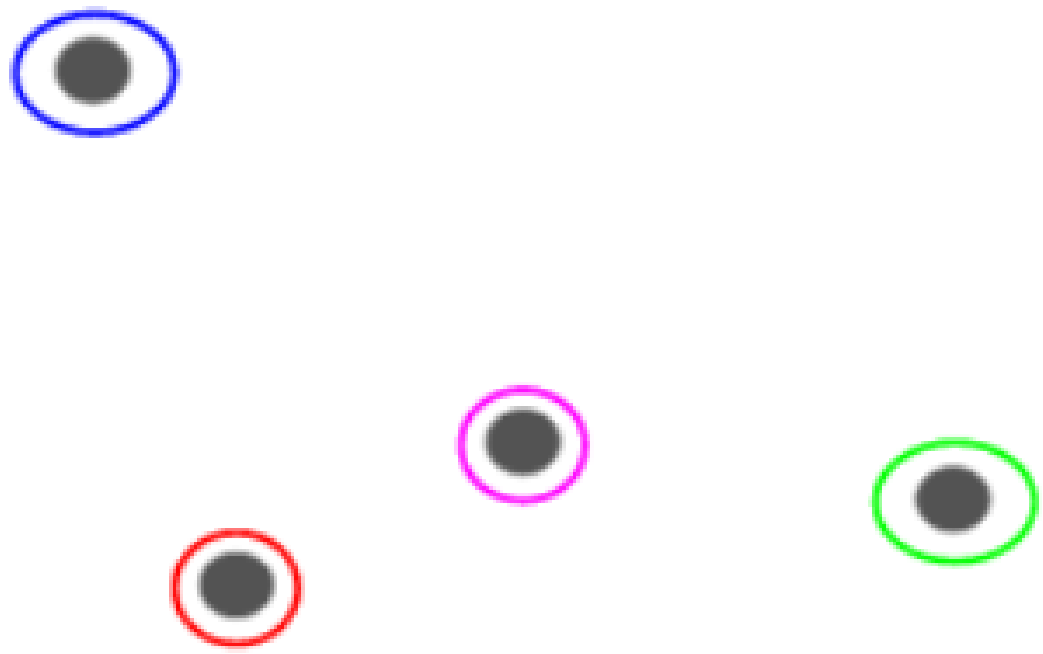
Analytics Vidhya

Divisive Hierarchical Clustering

We start with a single cluster and assign all the points to that cluster. So, it doesn’t matter if we have 10 or 1000 data points. All these points will belong to the same cluster at the beginning:



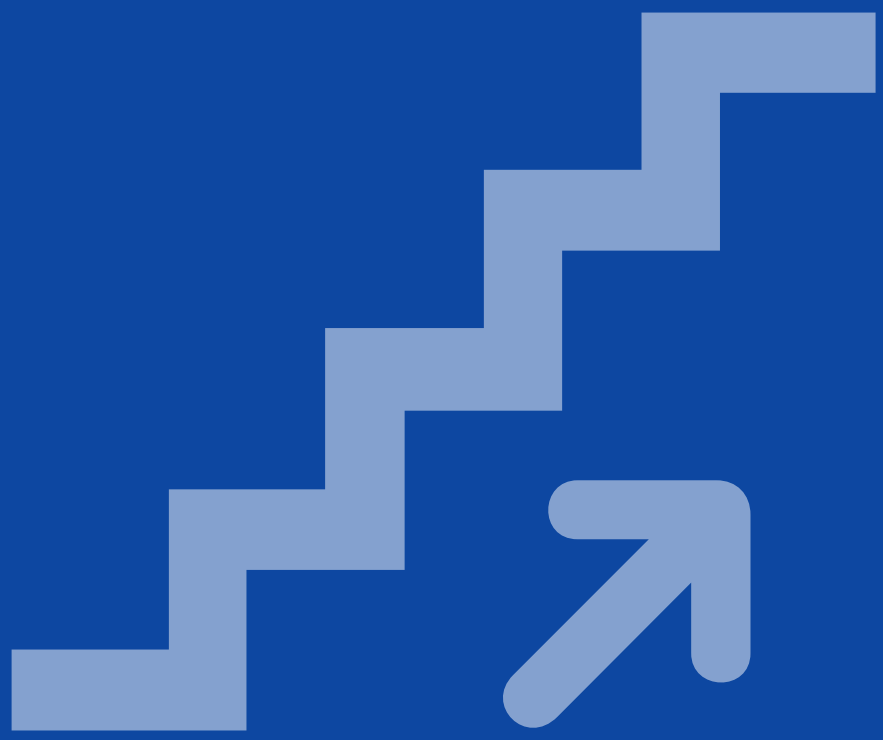
Now, at each iteration, we split the farthest point in the cluster and repeat this process until each cluster only contains a single point:



We are splitting (or dividing) the clusters at each step, hence the name **divisive hierarchical clustering**.

Analytics Vidhya

Steps to Perform Hierarchical Clustering



Lets understand with a simple example.

We have five students and their scores in two subjects: **Math and English.**

We want to cluster them based on their academic performance using **hierarchical clustering**.

Step 1: Preprocess the data:

Assume that the scores are already in a comparable format, so we don't need to perform any preprocessing in this case.

Student	Maths	English
Student 1	85	70
Student 2	75	65
Student 3	90	95
Student 4	80	85
Student 5	95	90

Step 2: Select a distance metric:

We'll use **Euclidean distance** as our distance metric to measure the similarity between student performance.

Step 3: Choose a linkage method:

We will use the **complete linkage method**, which calculates the distance between clusters as the maximum distance between any two points in each cluster.

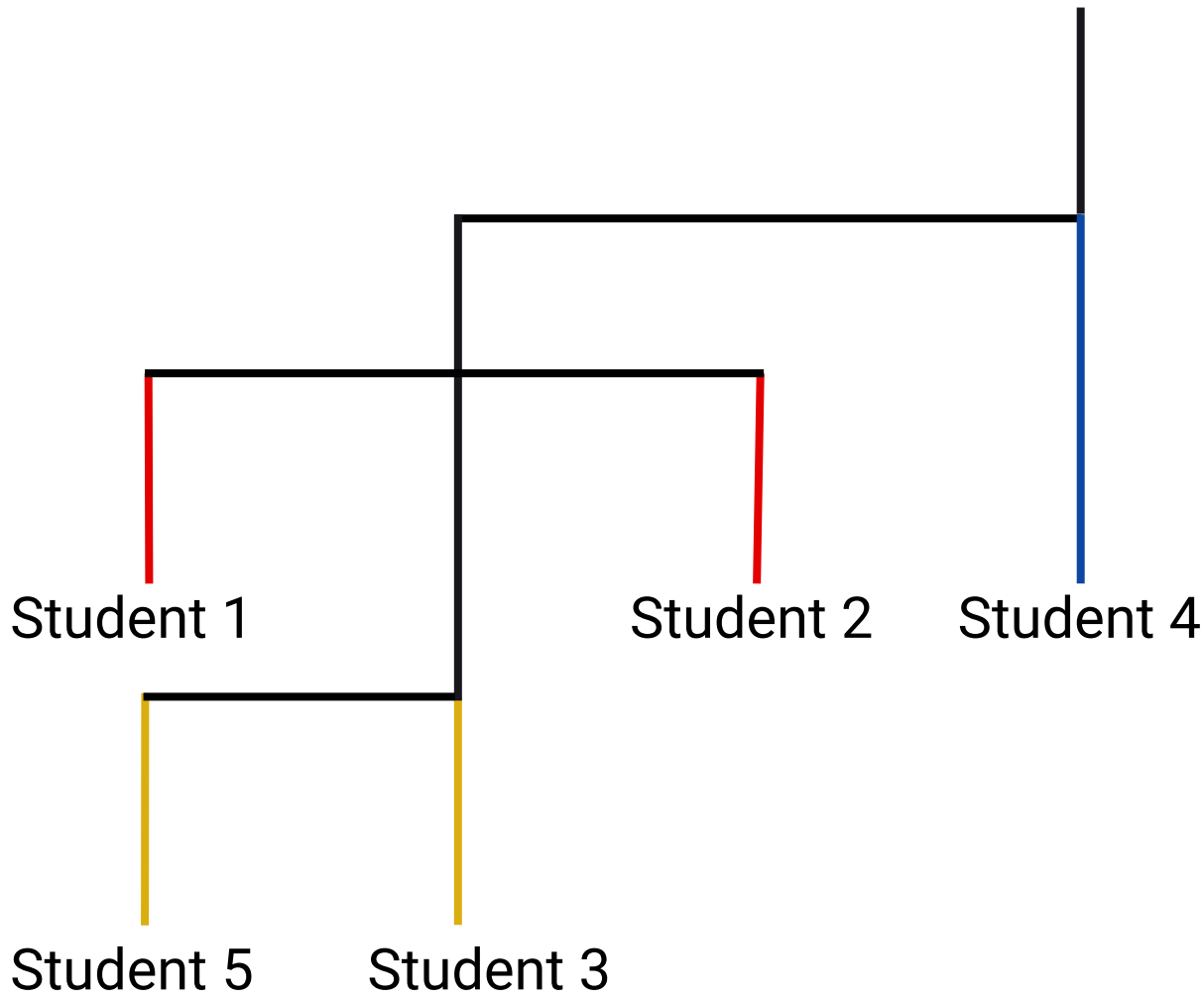
Step 4: Create a distance or similarity matrix:

We can calculate the pairwise Euclidean distances between each pair of students based on their scores. The resulting distance matrix would look like this:

	Student 1	Student 2	Student 3	Student 4	Student 5
Student 1	**0**	15.81	35.44	19.21	18.03
Student 2	15.81	**0**	37.42	20.40	21.54
Student 3	35.44	37.42	**0**	10.00	10.00
Student 4	19.21	20.40	10.00	**0**	16.28
Student 5	18.03	21.54	10.00	16.28	**0**

Step 5: Build the dendrogram:

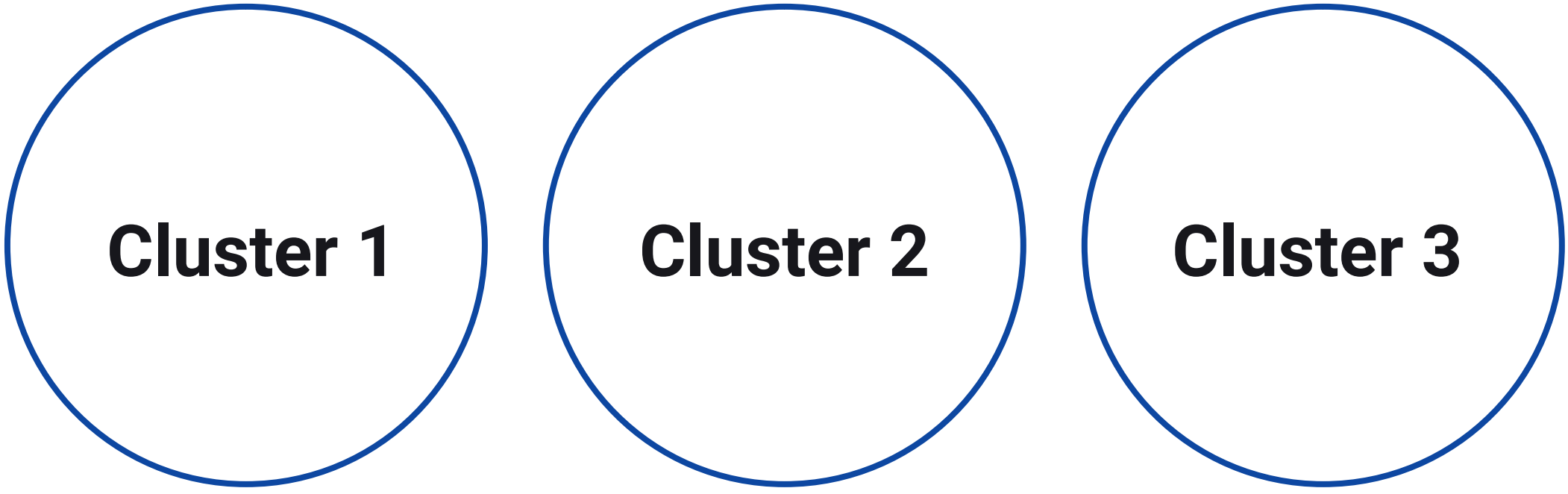
Using the distance matrix, we can construct the dendrogram that represents the hierarchical clustering process. Here's an example of a dendrogram for our student marks:



The dendrogram shows the merging of clusters as we move up the hierarchy. Initially, each student is represented as an individual cluster, and then they are gradually merged based on their similarities.

Step 6: Determine the number of clusters

We observe the heights of the dendrogram branches to determine the appropriate number of clusters. Let's say we decide to have three clusters.



Step 7: Assign data points to clusters

Based on the determined number of clusters, we assign each student to their respective cluster based on the dendrogram. Let's assume the cluster assignments are as follows:

Cluster 1: Student 1, Student 2

Cluster 2: Student 5, Student 3

Cluster 3: Student 4

Analytics Vidhya

That's a wrap. Was this post Helpful?

Follow us for more!

Analytics Vidhya