**What is K-Modes Clustering?**

K-Modes clustering is an unsupervised machine learning technique. It is a partition clustering algorithm used to group a dataset into K clusters.  K-Modes clustering can be used in machine learning applications that need to partition data having categorical variables.

K-Modes clustering is an iterative algorithm that starts by selecting k initial data points as centroids of the cluster. After that, each data point in the dataset is assigned to a cluster based on its similarity with the centroids. After creating clusters for the first time, we select a new centroid in each cluster using the mod of each feature in the data. After selecting new clusters, we calculate their dissimilarity from each data point and regroup the clusters. This process continues until the process converges and there is no change to the clusters in two consecutive iterations.

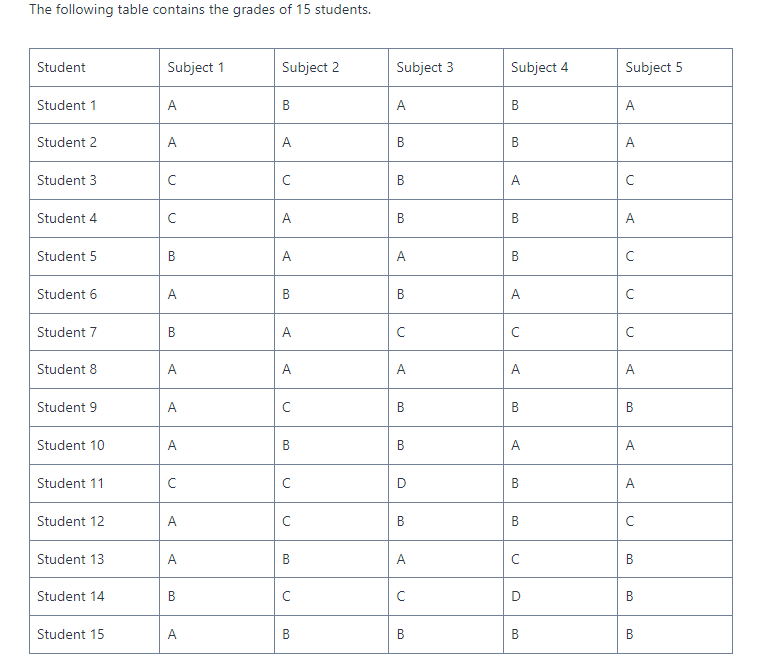
## Why Cannot We Use K-means Clustering on Categorical Data?

To calculate the distance or dissimilarity between two data points in the dataset, the K-means clustering algorithm uses euclidean distance or Manhattan distance. Similarly, the k-means algorithm uses the mean of the features of the data points in a cluster to calculate centroids for new clusters.

We are able to calculate the Euclidean distance, Manhattan distance, or mean of the features in a cluster because K-means clustering is primarily used for numeric data. These operations cannot be defined for a dataset having categorical features.

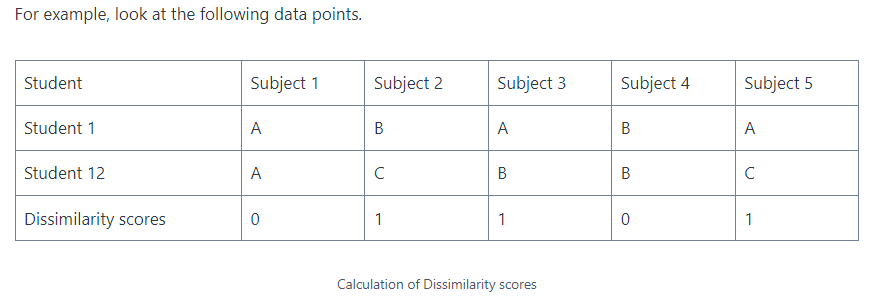
## Calculation of Dissimilarity Score Between Two Data Points in K-Modes Clustering

In K-Modes clustering, we cannot use Euclidean or Manhattan distance as a dissimilarity measure for the data points. Hence, we will use a different measure as discussed here. For calculating the dissimilarity measure, consider the following dataset.



Suppose that we have to partition the above table into 3 clusters. For this, we need to define a dissimilarity score. For this, we will use the following procedure.

* First, we will compare each feature of the data points. If the values in a feature are equal, we will assign a dissimilarity value of 0 to the feature.
* If the values in the corresponding features are different, we will assign a dissimilarity value of 1 to the feature.
* The final dissimilarity score between two data points will be the sum of the dissimilarity value of each feature.



Here, Subject 1 and Subject 4 have the same values for student 1 as well as student 12. Hence, these features have been assigned a dissimilarity value of 0.

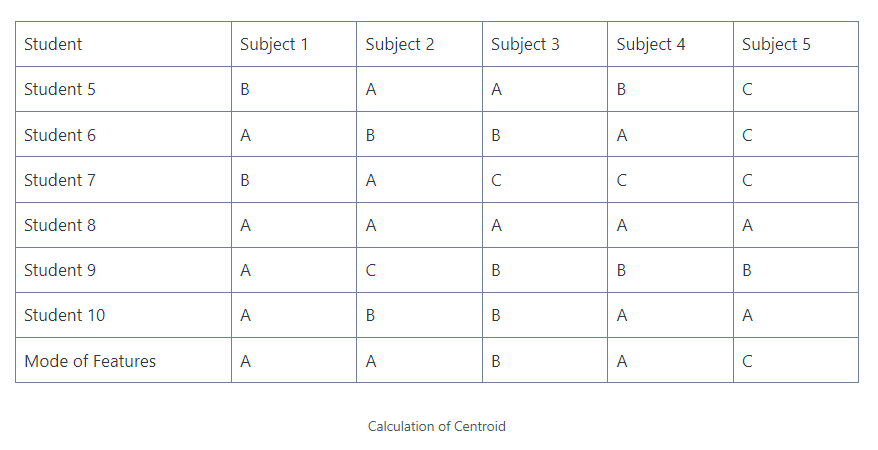
Subject2, Subject 3, and Subject 5 have different values for both students. Hence, we have assigned the value 1 as the dissimilarity value for each of the features.

The final dissimilarity score between both the students is the sum of dissimilarity values of all the features i.e. 0+1+1+0+1 =3.

Now that we have discussed the calculation of dissimilarity score for the K-Modes clustering algorithm, let us now discuss how to calculate a new centroid for any given cluster.

## Calculation of New Centroid in a Cluster

Suppose that the following data points belong to a single cluster.



To calculate the new centroid for the cluster, we will calculate the mod of the values in each feature of the data points. Here, the mod is the value with the most frequency.

In the above cluster, A is the mode for feature Subject 1. Similarly, A, B, A, and C are the mod of features of Subject 2, Subject 3, and Subject 4 respectively.

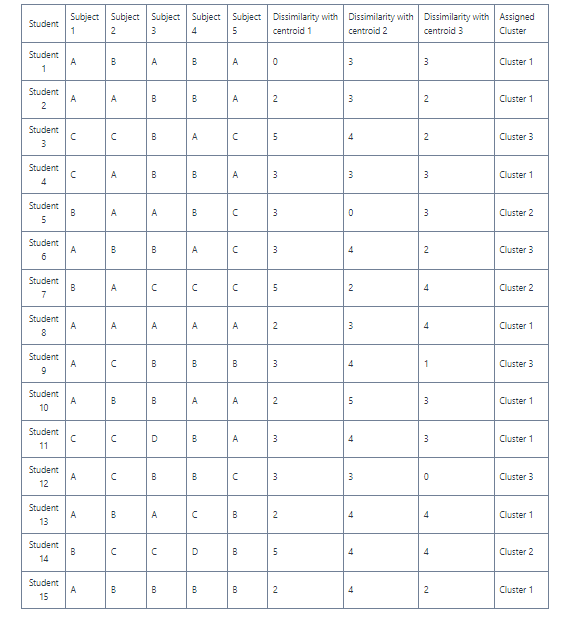
Hence, the new centroid for the cluster will be (A, A, B, A, C).

## K-Modes Clustering Numerical Example

### Iteration 1

First, we will select three student records as centroids for the clusters. Let it be Student 1, Student 5, and Student 12.

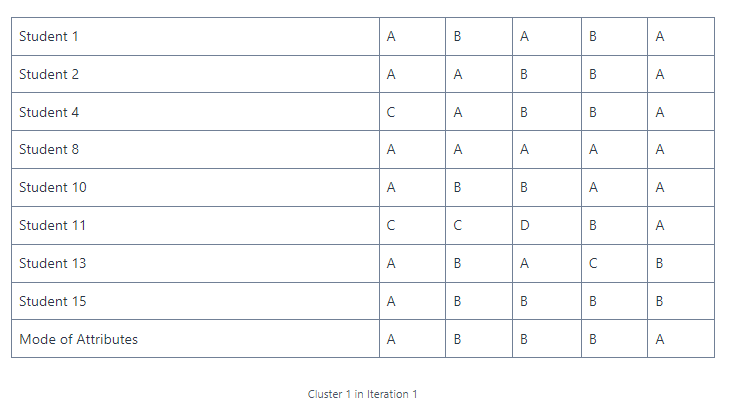
* Centroid 1 (Student 1) = (A, B, A, B, A)
* Centroid 2 (Student 5) = (B, A, A, B, C)
* Centroid 3 (Student 12) = (A, C, B, B, C)



In the above table, we have assigned each data point to a cluster based on the dissimilarity score. If a data point has the same dissimilarity value compared to the two centroids, we have assigned the data points to the cluster that comes first.

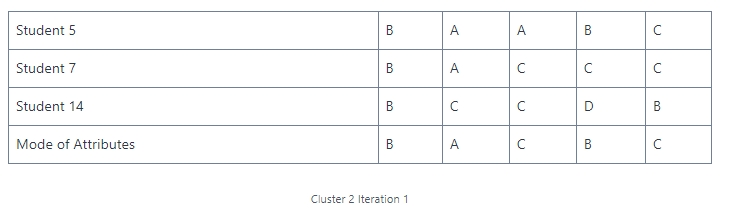
Now, we will calculate new centroids for each cluster.

In cluster 1, we have the following rows.



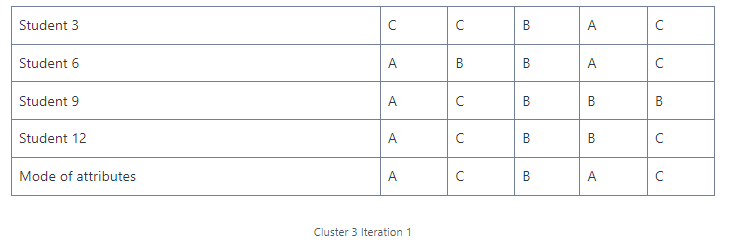
After calculating the mode of attributes in the cluster, the new centroid of cluster 1 is (A, B, B, B, A).

Now let us calculate the centroid of cluster 2.



The new centroid for cluster 2 is  (B, A, C, B, C).

Now, let us calculate the centroid for cluster 3.

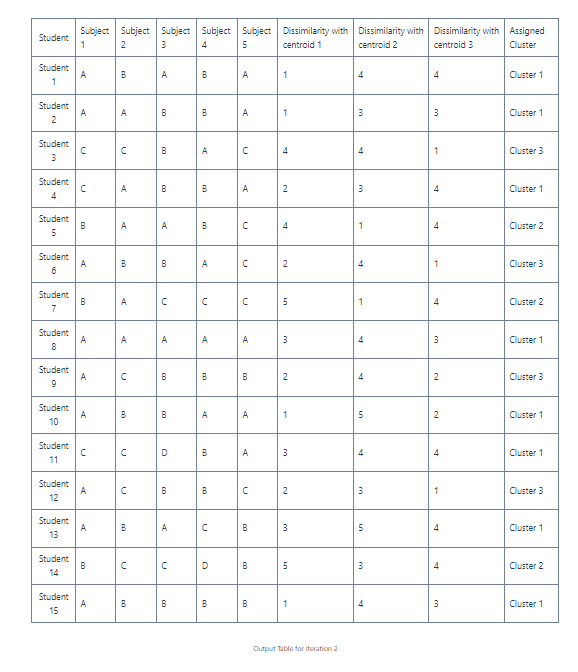


New Centroid for Cluster 3  is (A, C, B, A, C ).

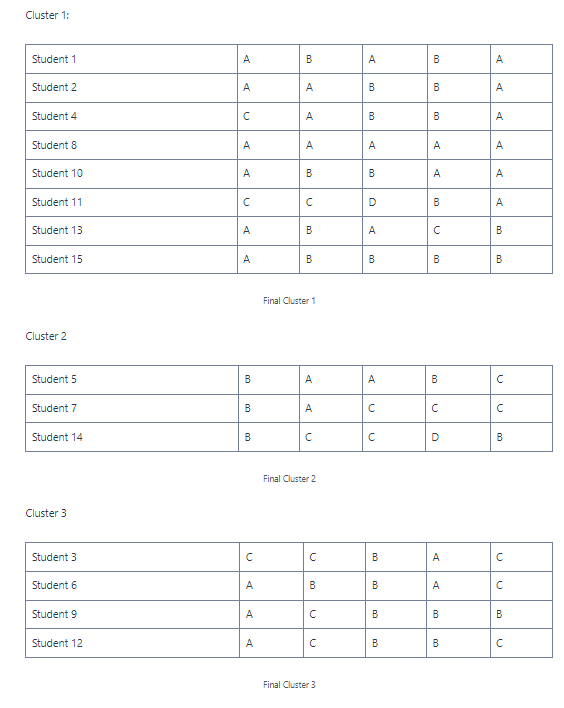
### Iteration 2

So, the new centroids calculated for cluster 1, cluster 2, and cluster 3 are (A, B, B, B, A), (B, A, C, B, C), (A, C, B, A, C ) respectively.

Now, we will calculate the dissimilarity of each data point from the new centroids and reassign the data points to the clusters. The results have been tabulated below.



You can observe that despite changing the centroids, no changes have happened in the clusters. Hence, the following are the final clusters calculated using K-Modes clustering.

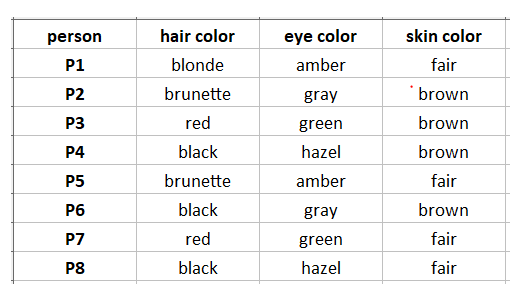


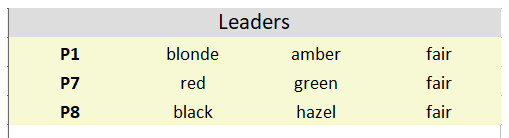
References :

https://codinginfinite.com/k-modes-clustering-algorithm-with-numerical-example/

https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

Practice Problem:





Final Result :

