**MACHINE LEARNING LAB**

Name – Manish Kumar Singh

Class – CSE-3

Roll No. - 40876802717

**EXPERIMENT 8**

**AIM :** Select two datasets. Each dataset should contain examples from multiple classes. For training purposes assume that the class label of example is unknown. Apply the k-means algorithm and apply it to the selected data. Evaluate the process by measuring the sum of euclidean distance of each example from its class center. Test the performance of the algorithm as a function of the parameter k.

**Github Link:-**

<https://github.com/sngmanish/ML_Assignments/tree/master/LabProgram8>

**Algorithm**

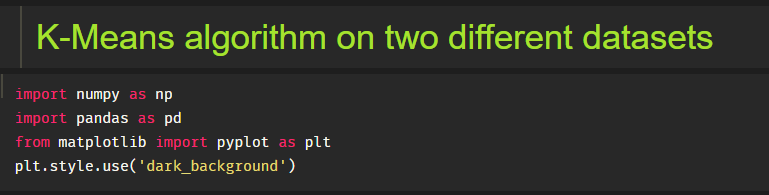
Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

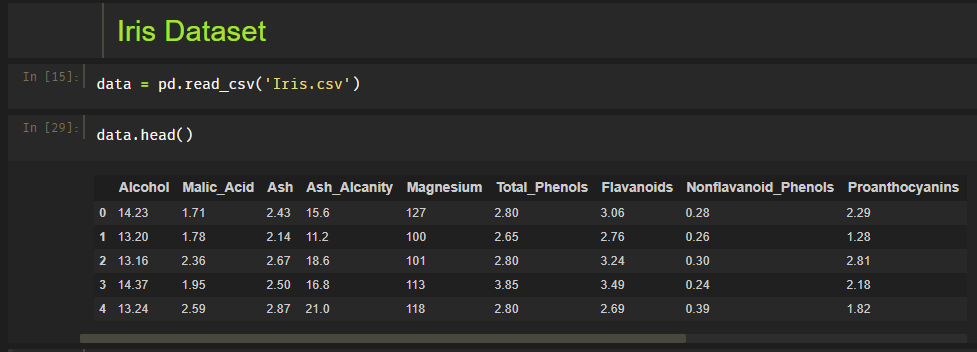
* Specify number of clusters K.
* Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
* Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

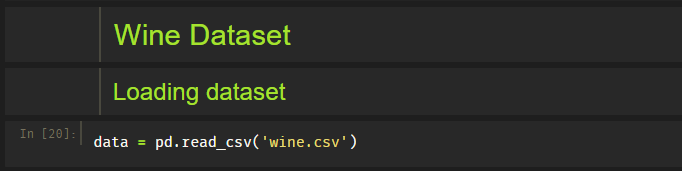
**Program Snippets :-**

1. **Importing Modules**



1. **Loading dataset :-**





1. **Data Visualization**



* **For Iris dataset**

****

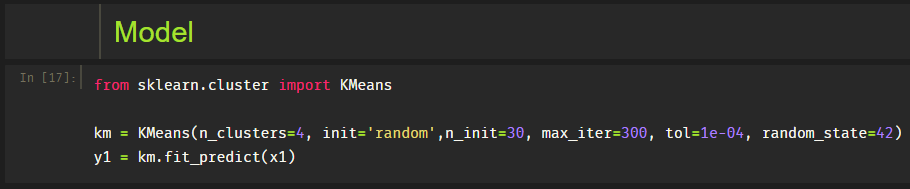
* **For wine dataset**

****

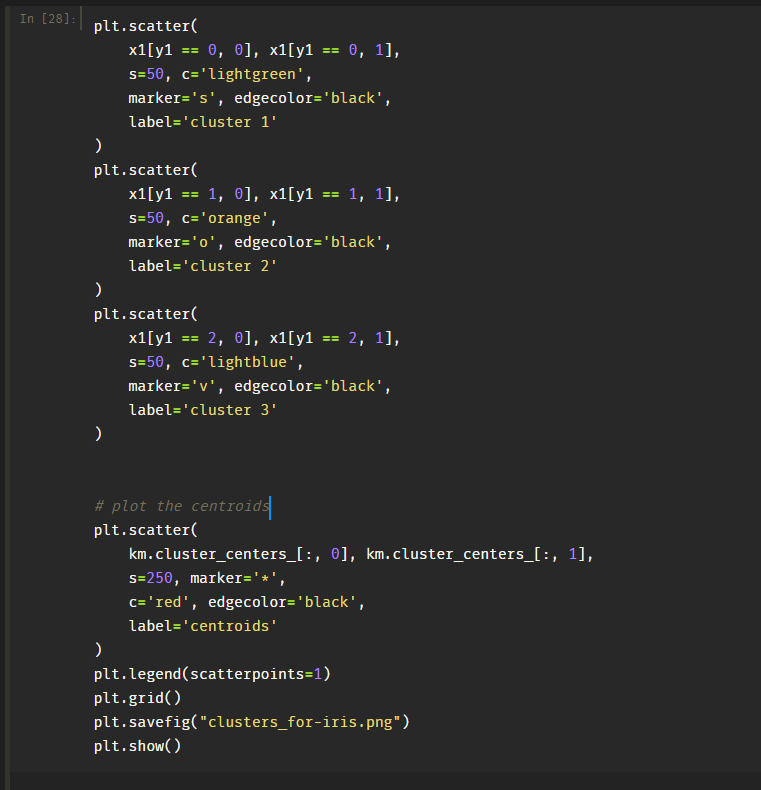
1. **Converting data into X**



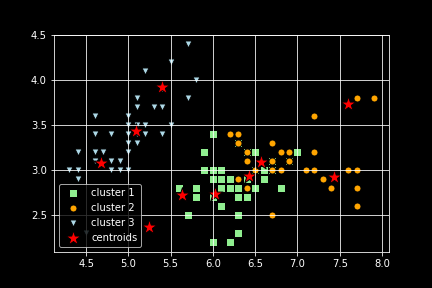
1. **Model Kmeans**



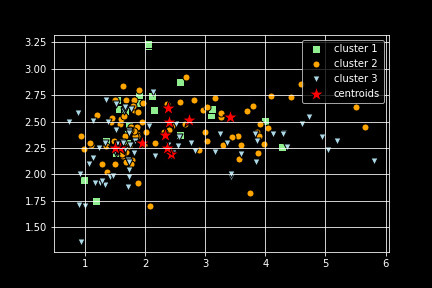
1. **New points introduced to check k-means Model**



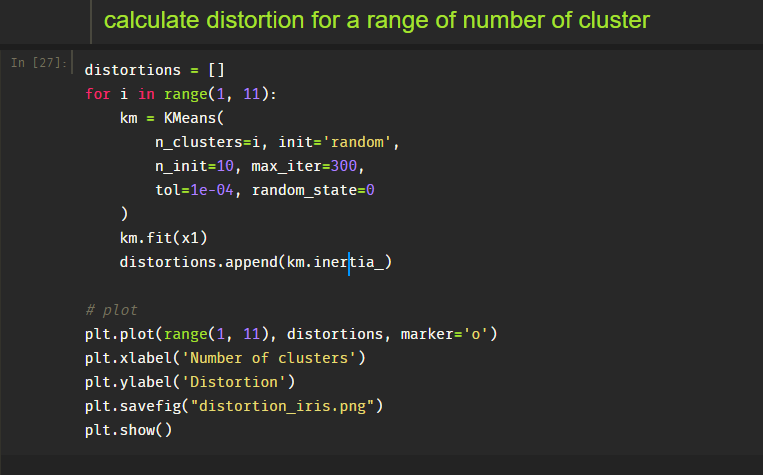
* **For Iris Dataset**

****

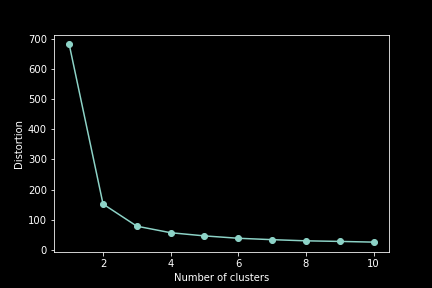
* **For Wine Dataset**

****

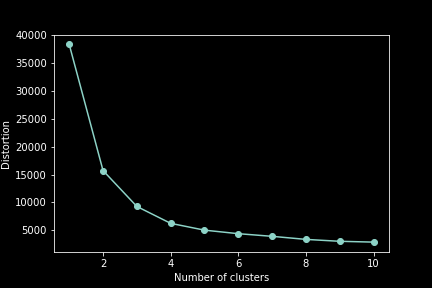
1. **Calculate distortion for a range of number of cluster**



* **For Iris Dataset**



* **For wine Dataset**

****