**Assignment 5**

**Group: DWDM19G04**

**Roll Numbers: 2016BTECS00063, 2016BTECS00081, 2016BTECS00103**

**Batch: B7**

**Date: 5th October, 2019**

**Title: Clustering**

**Theory:**

It is a data mining technique used to place the data elements into their related groups. Clustering is the process of partitioning the data (or objects) into the same class, The data in one class is more similar to each other than to those in other cluster. The process of partitioning data objects into subclasses is called as cluster. A cluster consists of data object with high inter similarity and low intra similarity. The quality of cluster depends on the method used.

Clustering is also called as data segmentation, because it partitions large data sets into groups according to their similarity.

**Clustering can be helpful in many fields, such as:**

**1. Marketing:**

Clustering helps to find group of customers with similar behavior from a given data set customer record.

**2. Biology:**

Classification of plants and animal according to their features.

**3. Library:**

Clustering is very useful in book ordering.

**Types of clustering**

Clustering methods can be classified into the following categories:

**1. Partitioning**

In this approach, several partitions are created and then evaluated based on given criteria.

**2. Hierarchical method**

In this method, the set of data objects are decomposed (multilevel) hierarchically by using certain criteria.

**3. Density-based method**

This method is based on density (density reachability and density connectivity).

**4. Grid-based methods**

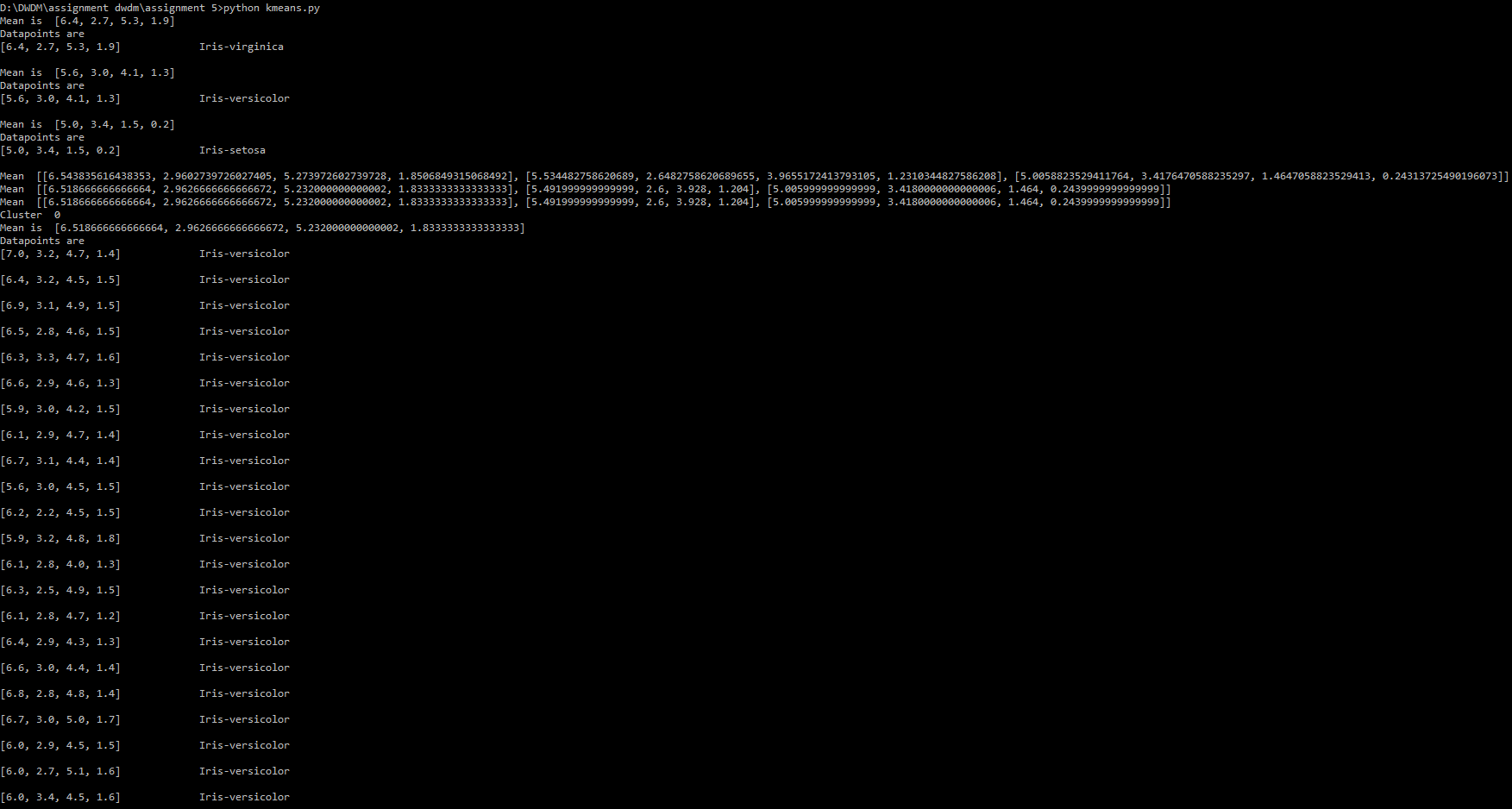
This approach is based on multi-resolution grid data structure.

## Classification vs Clustering

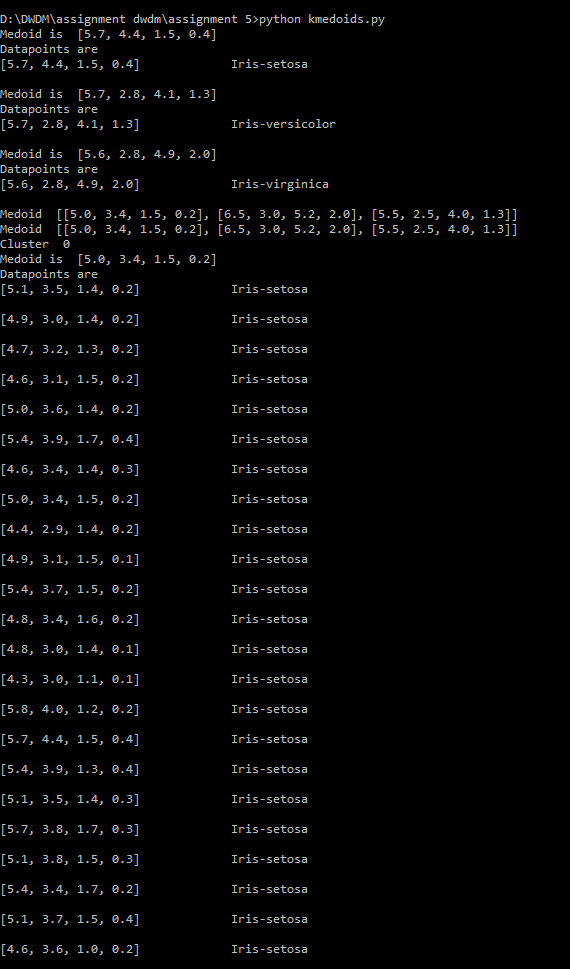
|  |  |
| --- | --- |
| **Classification** | **Clustering** |
| It is supervised learning. | It is unsupervised learning. |
| Classification contains previously categorized training set. | In clustering, the characteristics of similarity of data is not known. |
| Decision tree is used to partition and segment record. | There are a variety of algorithms for clustering, which generally share the same property of interactively assigning records to a cluster. |

**Output:**

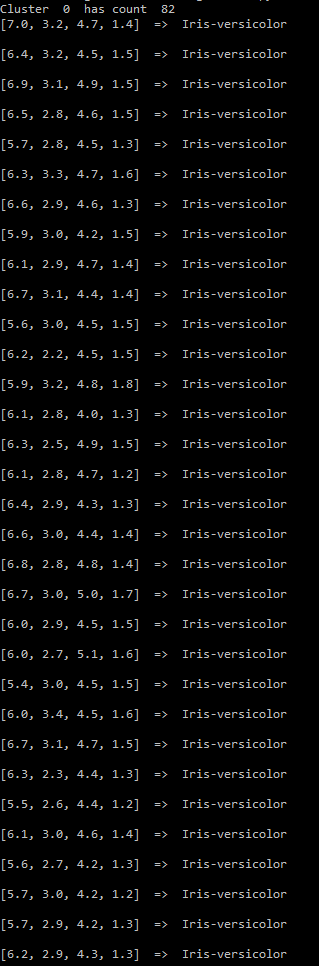
1. **K-means:**



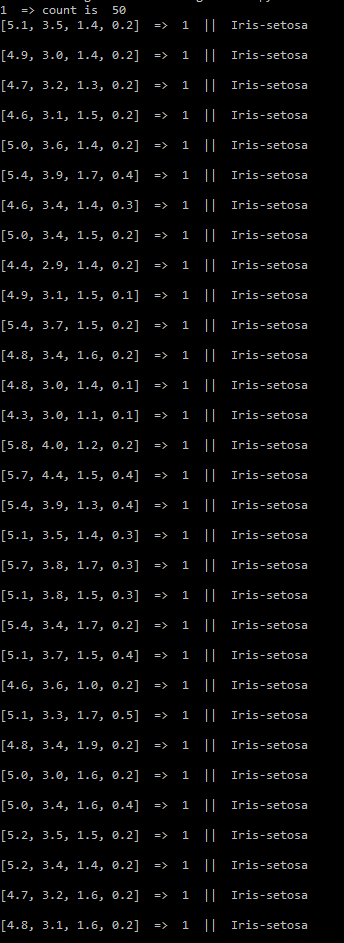
1. **K-medoids:**



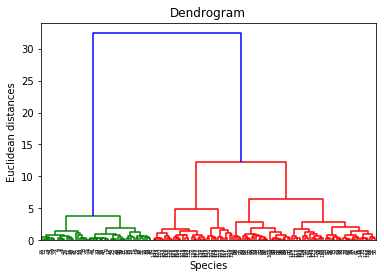
1. **Diana:**



1. **DBScan:**



**Dendogram:**



**Tabular Result:**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | Data Samples | Correctly classified | Accuracy |
| K-Mean | 150 | 102 | 68% |
| K-Medoid | 150 | 77 | 51.33% |
| DB-Scan | 150 | 121 | 80.66% |

**Conclusion:**

We studied and implemented different clustering algorithms on iris dataset. From the observations, we can conclude that DBScan clustering gives the highest accuracy as it prevents overfitting.