BIG DATA PROGRAMING

ASSIGNMENT-6

As the primary step the spark context and session are created, after this import the necessary models like in this assignment Kmeans from the MLib library in spark and also the required libraries likes cluster evaluation.

The next step is to read the given text file with svmlib in spark as a data frame, which consists of x and y coordinates and their labels implies whether a point belongs to a class.

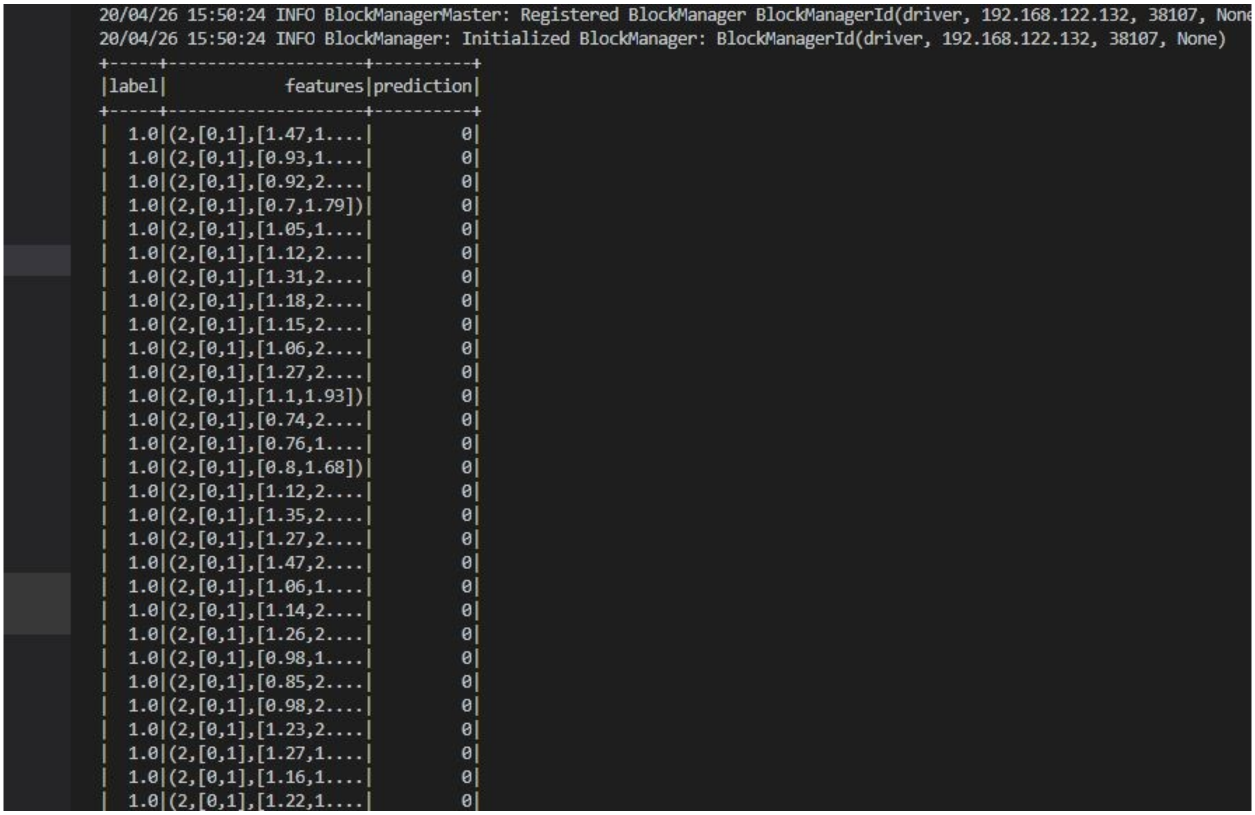
```
dataset = spark.read.format("libsvm").load("/home/vmalapati1/data/kmeans_input.txt")
```

Bisecting k-means is a hybrid approach between Divisive Hierarchical Clustering (top down clustering). It uses K means to each and every split. And this process continuous recursively.

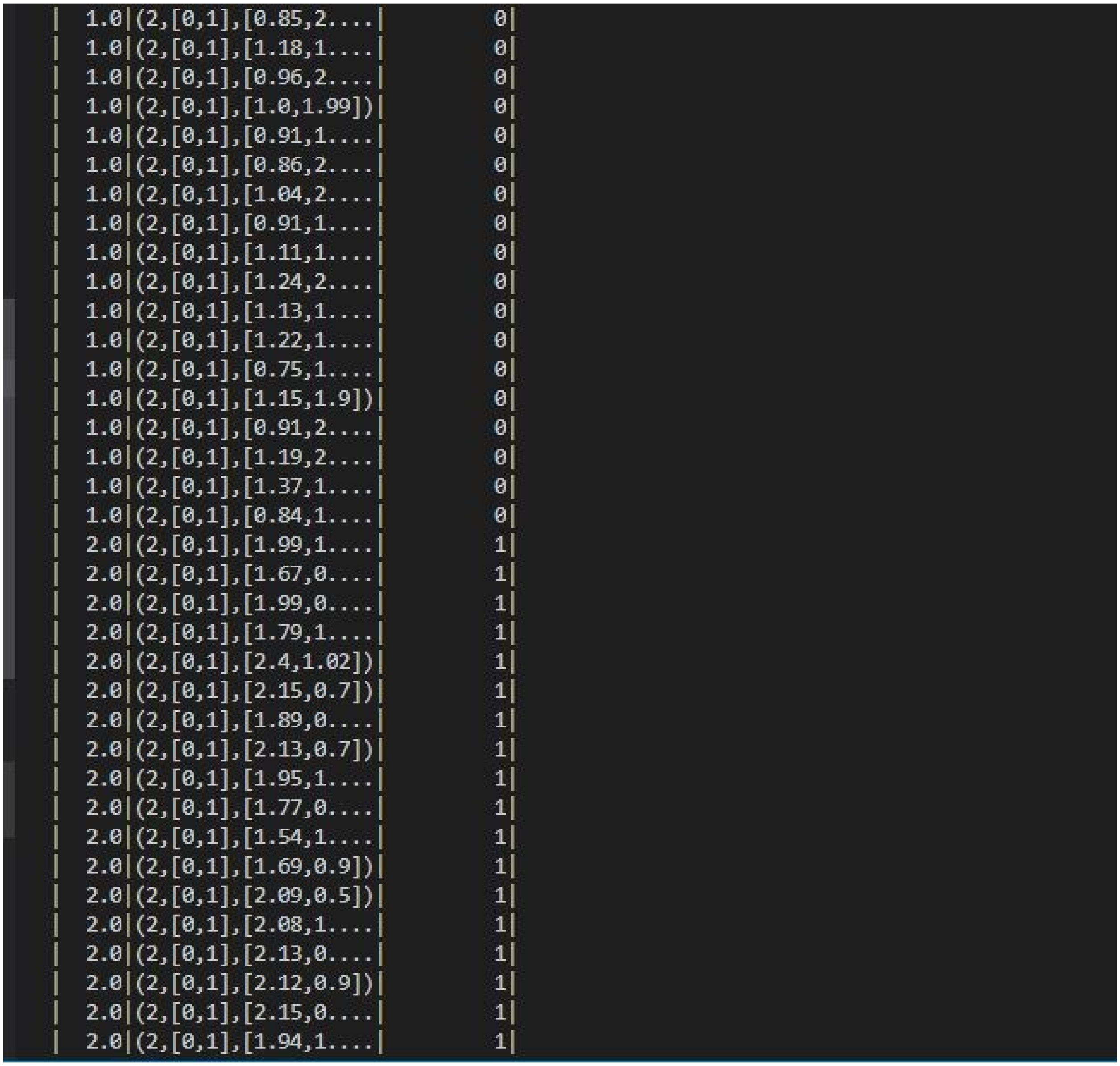
This data frame is given as input to the model for training, but before that a model is initialized with the required parameters likes number of classes and etc.. After this the model is trained using model.fit method with dataset as a input parameter. Using transformed operation we can predict the results that are given by the model for the given dataset as input. And next we evaluate the model with sum of squared errors. And further centers of the clusters are computed using the clusters centers method of the model that we trained in above

```
#INTIALIZING THE KMMEANS ALGO
kmeans = BisectingKMeans(k=2, seed=1)  # 2 clusters here
# TRIANING THE MODEL WITH ABOVE DATA FRAME
model = kmeans.fit(dataset)
# PREDICTING THE RESULTS BASED ON THE INPUT OF THE MODEL
transformed = model.transform(dataset)
```

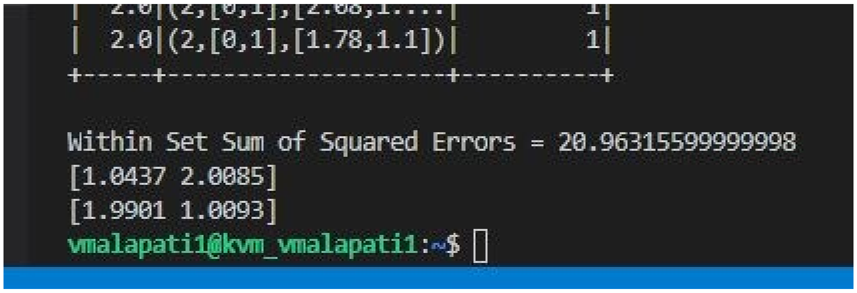
The final output of the model is as shown below.



```
20/04/26 15:50:24 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.122.132, 38107, None
20/04/26 15:50:24 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.122.132, 38107, None)
+-----+--------------------+----------+
|label|            features|prediction|
+-----+--------------------+----------+
|  1.0|(2,[0,1],[1.47,1....|         0|
|  1.0|(2,[0,1],[0.93,1....|         0|
|  1.0|(2,[0,1],[0.92,2....|         0|
|  1.0|(2,[0,1],[0.7,1.79])|         0|
|  1.0|(2,[0,1],[1.05,1....|         0|
|  1.0|(2,[0,1],[1.12,2....|         0|
|  1.0|(2,[0,1],[1.31,2....|         0|
|  1.0|(2,[0,1],[1.18,2....|         0|
|  1.0|(2,[0,1],[1.15,2....|         0|
|  1.0|(2,[0,1],[1.06,2....|         0|
|  1.0|(2,[0,1],[1.27,2....|         0|
|  1.0|(2,[0,1],[1.1,1.93])|         0|
|  1.0|(2,[0,1],[0.74,2....|         0|
|  1.0|(2,[0,1],[0.76,1....|         0|
|  1.0|(2,[0,1],[0.8,1.68])|         0|
|  1.0|(2,[0,1],[1.12,2....|         0|
|  1.0|(2,[0,1],[1.35,2....|         0|
|  1.0|(2,[0,1],[1.27,2....|         0|
|  1.0|(2,[0,1],[1.47,2....|         0|
|  1.0|(2,[0,1],[1.06,1....|         0|
|  1.0|(2,[0,1],[1.14,2....|         0|
|  1.0|(2,[0,1],[1.26,2....|         0|
|  1.0|(2,[0,1],[0.98,1....|         0|
|  1.0|(2,[0,1],[0.85,2....|         0|
|  1.0|(2,[0,1],[0.98,2....|         0|
|  1.0|(2,[0,1],[1.23,2....|         0|
|  1.0|(2,[0,1],[1.27,1....|         0|
|  1.0|(2,[0,1],[1.16,1....|         0|
|  1.0|(2,[0,1],[1.22,1....|         0|
```

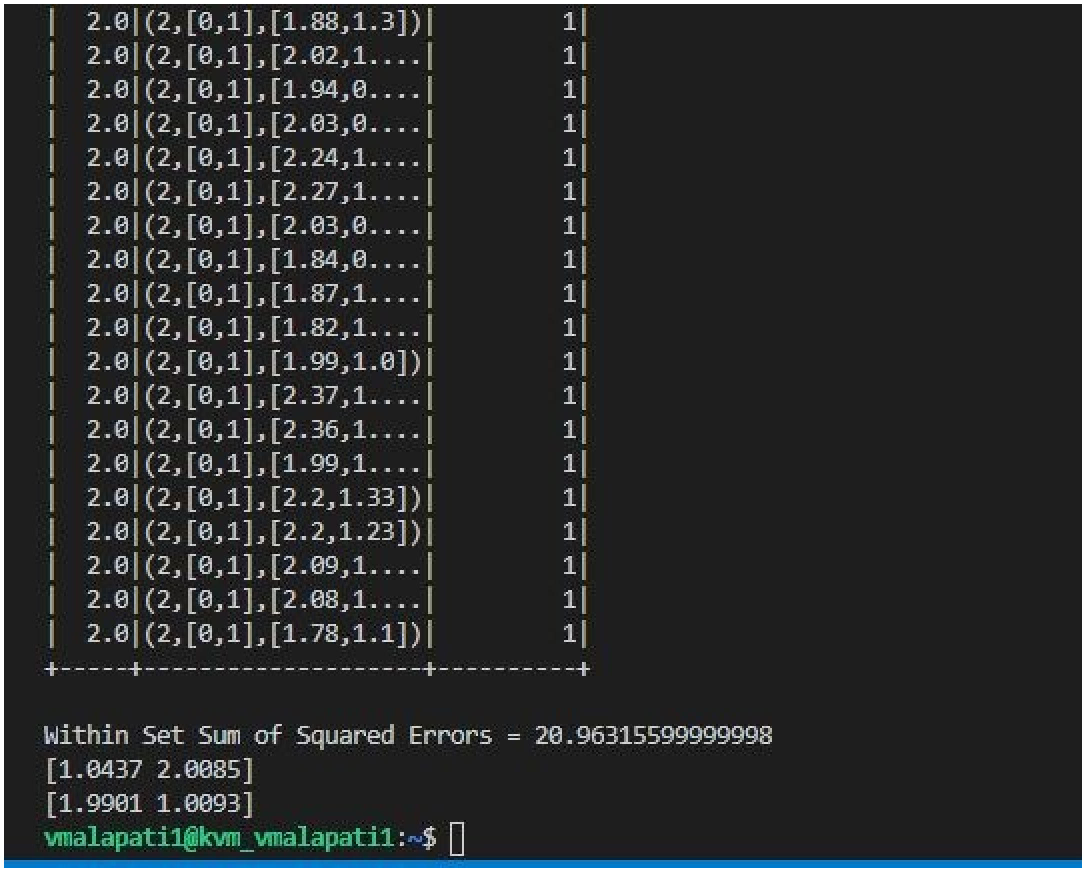


```
|  1.0|(2,[0,1],[0.85,2....|         0|
|  1.0|(2,[0,1],[1.18,1....|         0|
|  1.0|(2,[0,1],[0.96,2....|         0|
|  1.0|(2,[0,1],[1.0,1.99])|         0|
|  1.0|(2,[0,1],[0.91,1....|         0|
|  1.0|(2,[0,1],[0.86,2....|         0|
|  1.0|(2,[0,1],[1.04,2....|         0|
|  1.0|(2,[0,1],[0.91,1....|         0|
|  1.0|(2,[0,1],[1.11,1....|         0|
|  1.0|(2,[0,1],[1.24,2....|         0|
|  1.0|(2,[0,1],[1.13,1....|         0|
|  1.0|(2,[0,1],[1.22,1....|         0|
|  1.0|(2,[0,1],[0.75,1....|         0|
|  1.0|(2,[0,1],[1.15,1.9])|         0|
|  1.0|(2,[0,1],[0.91,2....|         0|
|  1.0|(2,[0,1],[1.19,2....|         0|
|  1.0|(2,[0,1],[1.37,1....|         0|
|  1.0|(2,[0,1],[0.84,1....|         0|
|  2.0|(2,[0,1],[1.99,1....|         1|
|  2.0|(2,[0,1],[1.67,0....|         1|
|  2.0|(2,[0,1],[1.99,0....|         1|
|  2.0|(2,[0,1],[1.79,1....|         1|
|  2.0|(2,[0,1],[2.4,1.02])|         1|
|  2.0|(2,[0,1],[2.15,0.7])|         1|
|  2.0|(2,[0,1],[1.89,0....|         1|
|  2.0|(2,[0,1],[2.13,0.7])|         1|
|  2.0|(2,[0,1],[1.95,1....|         1|
|  2.0|(2,[0,1],[1.77,0....|         1|
|  2.0|(2,[0,1],[1.54,1....|         1|
|  2.0|(2,[0,1],[1.69,0.9])|         1|
|  2.0|(2,[0,1],[2.09,0.5])|         1|
|  2.0|(2,[0,1],[2.08,1....|         1|
|  2.0|(2,[0,1],[2.13,0....|         1|
|  2.0|(2,[0,1],[2.12,0.9])|         1|
|  2.0|(2,[0,1],[2.15,0....|         1|
|  2.0|(2,[0,1],[1.94,1....|         1|
```



```
|  2.0|(2,[0,1],[2.08,1....|         1|
|  2.0|(2,[0,1],[1.78,1.1])|         1|
+-----+--------------------+----------+

Within Set Sum of Squared Errors = 20.963155999999998
[1.0437 2.0085]
[1.9901 1.0093]
vmalapati1@kvm_vmalapati1:~$ 
```



```
|  2.0|(2,[0,1],[1.88,1.3])|         1|
|  2.0|(2,[0,1],[2.02,1....|         1|
|  2.0|(2,[0,1],[1.94,0....|         1|
|  2.0|(2,[0,1],[2.03,0....|         1|
|  2.0|(2,[0,1],[2.24,1....|         1|
|  2.0|(2,[0,1],[2.27,1....|         1|
|  2.0|(2,[0,1],[2.03,0....|         1|
|  2.0|(2,[0,1],[1.84,0....|         1|
|  2.0|(2,[0,1],[1.87,1....|         1|
|  2.0|(2,[0,1],[1.82,1....|         1|
|  2.0|(2,[0,1],[1.99,1.0])|         1|
|  2.0|(2,[0,1],[2.37,1....|         1|
|  2.0|(2,[0,1],[2.36,1....|         1|
|  2.0|(2,[0,1],[1.99,1....|         1|
|  2.0|(2,[0,1],[2.2,1.33])|         1|
|  2.0|(2,[0,1],[2.2,1.23])|         1|
|  2.0|(2,[0,1],[2.09,1....|         1|
|  2.0|(2,[0,1],[2.08,1....|         1|
|  2.0|(2,[0,1],[1.78,1.1])|         1|
+-----+--------------------+----------+

Within Set Sum of Squared Errors = 20.96315599999998
[1.0437 2.0085]
[1.9901 1.0093]
vmalapati1@kvm_vmalapati1:~$ 
```