**Data Science and Business Analytics Task**

**TASK:** Prediction Using Unsupervised learning

**GRIP: The Spark Foundation**

**BY**

**NAME:** Utkarsha Solankar

**Dataset: Iris Dataset**

**1. Objective: Predict the optimum number of clusters and represent it.**

**2. Importing Libraries:**

> library("cluster") #import libraries

**3. Importing Dataset:**

> I=read.csv("C:\\Users\\WELLCOME\\Desktop\\Iris.csv") #Read the data

**4. Preprocess Dataset :**

> head(I) #inspecting first lines

ID SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species

1 1 5.1 3.5 1.4 0.2 setosa

2 2 4.9 3.0 1.4 0.2 setosa

3 3 4.7 3.2 1.3 0.2 setosa

4 4 4.6 3.1 1.5 0.2 setosa

5 5 5.0 3.6 1.4 0.2 setosa

6 6 5.4 3.9 1.7 0.4 setosa

> tail(I) #inspecting last lines

ID SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species

145 145 6.7 3.3 5.7 2.5 virginica

146 146 6.7 3.0 5.2 2.3 virginica

147 147 6.3 2.5 5.0 1.9 virginica

148 148 6.5 3.0 5.2 2.0 virginica

149 149 6.2 3.4 5.4 2.3 virginica

150 150 5.9 3.0 5.1 1.8 virginica

> str(I) #structure

'data.frame': 150 obs. of 6 variables:

$ ID : int 1 2 3 4 5 6 7 8 9 10 ...

$ SepalLengthCm: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...

$ SepalWidthCm : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...

$ PetalLengthCm: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...

$ PetalWidthCm : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...

$ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

> dim(I) #check dimensions

[1] 150 6

> names(I) #column names

[1] "ID" "SepalLengthCm" "SepalWidthCm" "PetalLengthCm"

[5] "PetalWidthCm" "Species"

> summary(I) #summary of data

ID SepalLengthCm SepalWidthCm PetalLengthCm

Min. : 1.00 Min. :4.300 Min. :2.000 Min. :1.000

1st Qu.: 38.25 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600

Median : 75.50 Median :5.800 Median :3.000 Median :4.350

Mean : 75.50 Mean :5.843 Mean :3.054 Mean :3.759

3rd Qu.:112.75 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100

Max. :150.00 Max. :7.900 Max. :4.400 Max. :6.900

PetalWidthCm Species

Min. :0.100 setosa :50

1st Qu.:0.300 versicolor:50

Median :1.300 virginica :50

Mean :1.199

3rd Qu.:1.800

Max. :2.500

**5. Apply K-means clustering algorithm**

> set.seed(30)

> km=kmeans(I[,-6],3,nstart=30)

> km

K-means clustering with 3 clusters of sizes 50, 50, 50

Cluster means:

ID SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm

1 25.5 5.006 3.418 1.464 0.244

2 125.5 6.588 2.974 5.552 2.026

3 75.5 5.936 2.770 4.260 1.326

Clustering vector:

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[38] 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

[75] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2

[112] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

[149] 2 2

Within cluster sum of squares by cluster:

[1] 10427.74 10456.03 10443.12

(between\_SS / total\_SS = 88.9 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"

[6] "betweenss" "size" "iter" "ifault"

**The algorithm has grouped the data into three clusters with 30 different random assignments.**

**Let’s now compare this with the species column.**

> t=table(I$Species,km$cluster);t #compare with Species column

1 2 3

setosa 50 0 0

versicolor 0 0 50

virginica 0 50 0

**From the table, we can see that the algorithm has clustered the three species i.e. Setosa, Versicolor and Verginica correctly.**

**By plotting the clusters we would be able to distinguish them.**

|  |
| --- |
| > clusplot(I,km$cluster,color=TRUE,shade=TRUE,lines=0) #plotting clusters with their centers |
|  |
| |  | | --- | |  | |



The above plot provides a clear understanding of the three clusters in different colors.

**6. Conclusion:** I was successfully able tooptimum number of clusters and presented it.