**Assignment Problems: Day 13**

The following problems require access to the file named ‘Dataset\_Day13.csv’, provided with this assignment.

This is the Iris dataset. The dataset 150 data points.

It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

Id

SepalLengthCm

SepalWidthCm

PetalLengthCm

PetalWidthCm

Species

Problems to solve –

1. Firstly, treat all outliers and missing values in the dataset.
2. Complete all basic data descriptive statistics by *Species*
3. Use the *Sepal Length* , *Sepal Width*, *Petal Length* and *Petal Width* to find K-Means clusters.
4. Find the optimum cluster number based on, *elbow method, silhouette method* and *Calinski Harabasz Score.*
5. Tabulate the proportion of each *Species* among the clusters found as a result of evaluation in task 4.
6. Share your insights on the data based on the clusters [optional]