# RUMAISA SHAHAB 2211-024-DEG-KHI AND USMAN ZAMAN 2211-032-DEG-KHI

In [21]:

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.metrics import calinski_harabasz_score
from sklearn.metrics import davies_bouldin_score
```

In [22]:

```
iris = datasets.load_iris()
x = iris.data
y = iris.target
```

## MODEL_1 (Kmeans)

In [23]:

```
model_1 = KMeans(n_clusters=3, n_init=4, max_iter=300)
labels = model_1.fit_predict(x)

davies_bouldin_score_1=davies_bouldin_score(x, labels)
calinski_harabasz_score_1 = calinski_harabasz_score(x,labels)

print("davies_bouldin_score_1 = " , davies_bouldin_score_1)
print("calinski_harabasz_score_1 = ", calinski_harabasz_score_1)
```

```
davies_bouldin_score_1 =  0.6619715465007465
calinski_harabasz_score_1 =  561.62775662962
```

## MODEL_2 (Kmeans with PCA)

In [27]:

```
pca = PCA(n_components=2)
x_reduced = pca.fit_transform(x)

model_2 = KMeans(n_clusters=3, n_init=4, max_iter=300)
```

```
labels_2 = model_2.fit_predict(x_reduced)

davies_bouldin_score_2=davies_bouldin_score(x_reduced, labels_2)
calinski_harabasz_score_2 = calinski_harabasz_score(x_reduced,labels_2)

print("davies_bouldin_score_2 = " , davies_bouldin_score_2)
print("calinski_harabasz_score_2 = ", calinski_harabasz_score_2)
```

```
davies_bouldin_score_2 =  0.5648157434964133
calinski_harabasz_score_2 =  693.708433418847
```

## EXPLANATION

## The Calinski-Harabasz (to check Cohesion):

We used The Calinski-Harabasz to evaluate cohesion. It also known as the Variance Ratio Criterion, is calculated as a ratio of the sum of inter-cluster dispersion and the sum of intra-cluster dispersion for all clusters (where the dispersion is the sum of squared distances).

## davies_bouldin_score (to check separation):

The score (DBI) is calculated as the average similarity of each cluster with a cluster most similar to it. The lower the average similarity is, the better the clusters are separated and the better is the result of the clustering performed.

## CONCLUSION

Since the cohesion and separation validation proved that Model_2 had better clustering in contrast with Model_1.PCA dimension reduction proved to be a better approach to enchance the model's behaviour.

In [29]:

```
if davies_bouldin_score_2<davies_bouldin_score_1 and calinski_harabasz_score_2>calinski_harabasz_score_1:
    print("Model 2 has better clustering")
else:
    print("Model 1 has better clustering")
```

```
Model 2 has better clustering
```