THE SPARKS FOUNDATION #GRIPMAY21

NAME: BAISHALI GHOSHAL

DATA SCIENCE AND BUSINESS ANALYTICS INTERNSHIP PROGRAM

PROJECT NAME: GIVEN THE ‘IRIS’ DATASET, PREDICTING THE OPTIMUM NUMBER OF CLUSTERS AND TRYING TO VISUALIZE THEM

TOOL USED: PYTHON LANGUAGE

jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

NAME: BAISHALI GHOSHAL

THE SPARKS FOUNDATION- DATA AND BUSINESS ANALYTICS INTERN- MAY 2021

GIVEN THE 'IRIS' DATASET, PREDICTING THE OPTIMUM NUMBER OF CLUSTERS AND TRYING TO VISUALIZE THEM

In [ ]:

```
# importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import datasets
from sklearn.cluster import KMeans
import warnings
warnings.filterwarnings('ignore')
```

jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

In [13]:
```
# loading the dataset
iris = datasets.load_iris()
iris_df = pd.DataFrame(iris.data, columns = iris.feature_names)
iris_df.head() # seeing the first five rows
```

Out[13]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

In [14]:

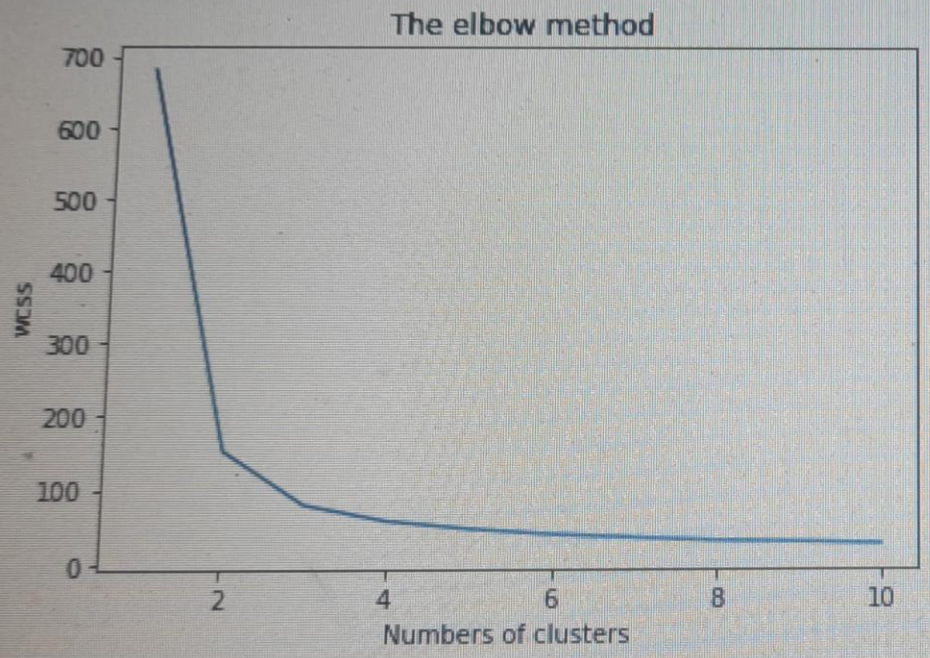
```
# finding the optimum number of clusters for k-means classification

x = iris_df.iloc[:, [0, 1, 2, 3]].values

wcss = []

for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++',
                    max_iter = 300, n_init = 10, random_state = 0)
    kmeans.fit(x)
    wcss.append(kmeans.inertia_)
```

jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted | Pyth

Run Code

In [15]:
```
# plotting the results in a line graph
# allowing us to observe 'The elbow'
plt.plot(range(1, 11), wcss)
plt.title('The elbow method')
plt.xlabel('Numbers of clusters')
plt.ylabel('wcss')
plt.show()
```



'The elbow method'from the above graph, it is seen the optimum clusters is where the elbow ocuurs. This is when the within cluster sum of squares(w does not decrease significantly with every iteration. From this we choose the number of clusters as 3 jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosave

File Edit View Insert Cell Kernel Widgets Help

Trus

Run Code

In [16]:
```
# applying kmeans to the dataset
kmeans = KMeans(n_clusters = 3, init = 'k-means++',
                max_iter = 300, n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(x)
```

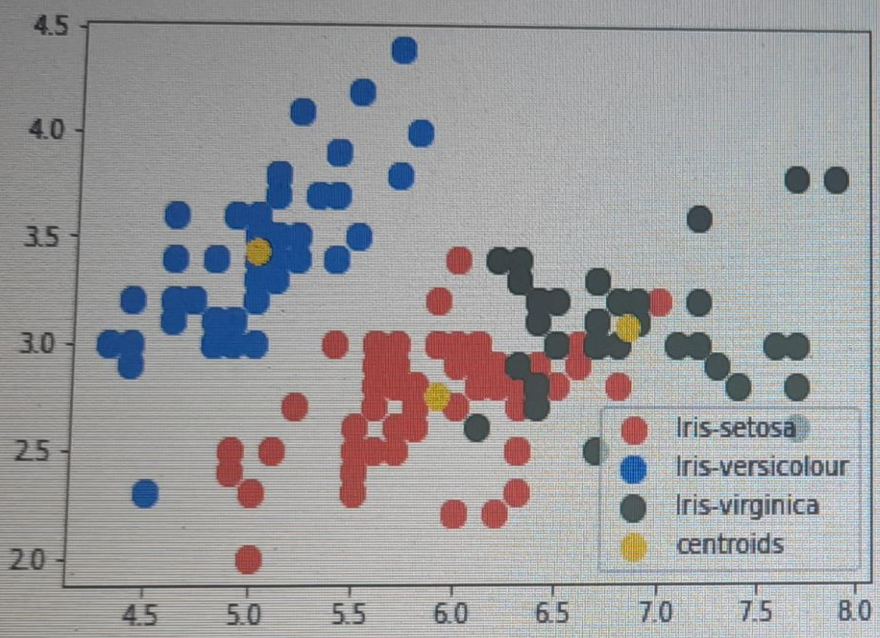
In [17]:
```
# visualizing the clusters on the first two columns
plt.scatter(x[y_kmeans == 0, 0], x[y_kmeans == 0, 1],
            s = 100, c = 'red', label = 'Iris-setosa')
plt.scatter(x[y_kmeans == 1, 0], x[y_kmeans == 1, 1],
            s = 100, c = 'blue', label = 'Iris-versicolour')
plt.scatter(x[y_kmeans == 2, 0], x[y_kmeans == 2, 1],
            s = 100, c = 'green', label = 'Iris-virginica')



# plotting the centroids of the clusters
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
            s = 100, c = 'yellow', label = 'centroids')
plt.legend()
```

jupyter Given the 'Iris' dataset, predicting the optimum number of clusters and tr... Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
Out[17]: <matplotlib.legend.Legend at 0x25d825b0490>
```



We can see that the clustering has done well since most of the reds and blues are seperated and the greens are also very close to eachother. Also the yellow ones represent the center points of each of the Iris species that we have. Thus, we have been able to find the optimum number of clusters and could visualize them.