Clustering

In [1]:

```
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from kneed import KneeLocator
import pandas as pda
```

```
C:\Users\Dinesh\AppData\Roaming\Python\Python39\site-packages\scipy\__ini
t__.py:177: UserWarning: A NumPy version >=1.18.5 and <1.26.0 is required
for this version of SciPy (detected version 1.26.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

In [2]:

```
df = pd.read_csv('Data/preprocessed_data.csv')
```

In [3]:

```
df
```

Out[3]:

	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	sick	pregna
0	42.0	0.0	0.0	0.0	0.0	0.0	0
1	24.0	0.0	0.0	0.0	0.0	0.0	0
2	47.0	1.0	0.0	0.0	0.0	0.0	0
3	71.0	0.0	1.0	0.0	0.0	0.0	0
4	71.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	
13919	42.0	1.0	0.0	0.0	0.0	0.0	0
13920	47.0	0.0	0.0	0.0	0.0	0.0	0
13921	42.0	1.0	0.0	0.0	0.0	0.0	0
13922	47.0	0.0	0.0	0.0	0.0	0.0	0
13923	47.0	0.0	0.0	0.0	0.0	0.0	0

13924 rows × 21 columns

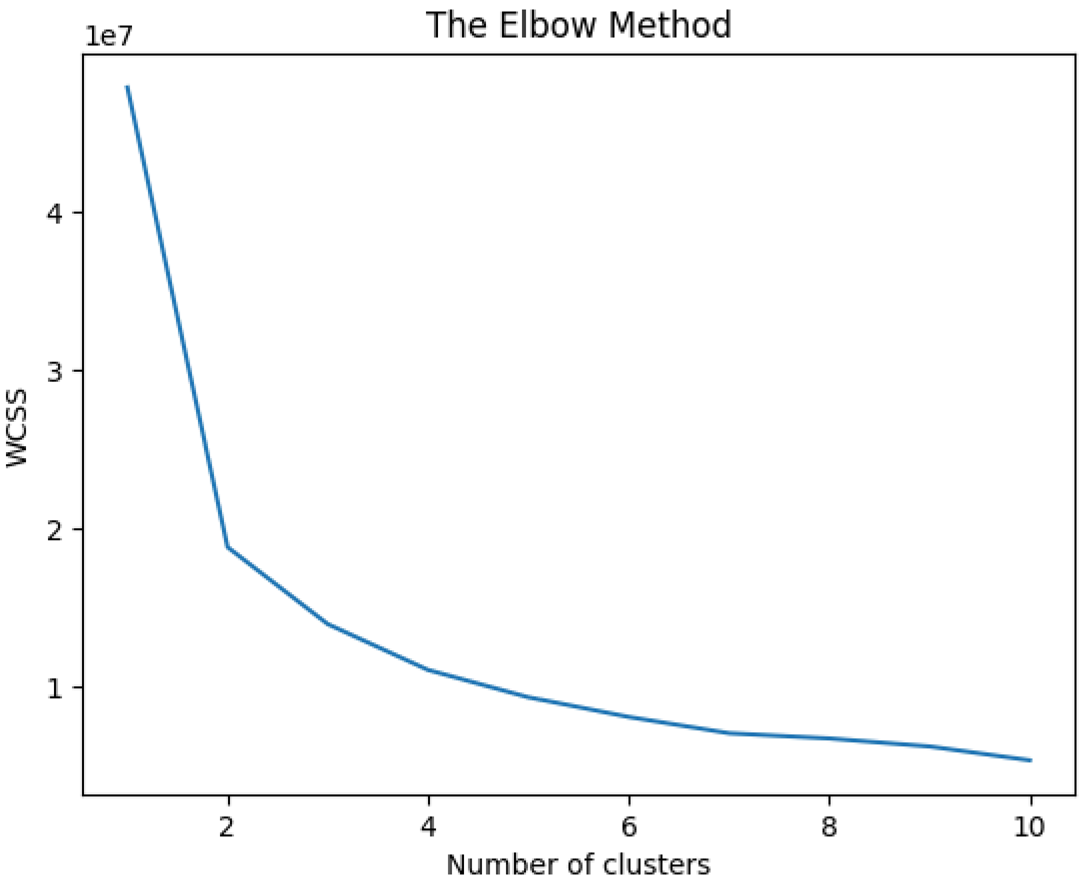
In [4]:

```
X = df.drop(["Class"], axis = 1)
y = df["Class"]
```

In [5]:

```
wcss = [] #within cluster sum of square
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

In [9]:
```
plt.plot(range(1,11),wcss) # creating the graph between WCSS and the number
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



In [11]:
```
kn = KneeLocator(range(1, 11), wcss, curve='convex', direction='decreasing'
num_cluster = kn.knee
```

In [14]:
```
#creating the number of cluster using KMeans++ as the number of cluster is
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=42)
```

In [15]:
```
y_kmeans = kmeans.fit_predict(X)
```

In [16]:
```
# created the cluster and stored the number of cluster
y_kmeans
```

Out[16]:
```
array([2, 0, 2, ..., 1, 1, 1])
```

In [17]:
```
X['Cluster'] = y_kmeans
```

In [18]:
```
X.head()
```

Out[18]:

	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	sick	pregnant	t
0	42.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	24.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	47.0	1.0	0.0	0.0	0.0	0.0	0.0	
3	71.0	0.0	1.0	0.0	0.0	0.0	0.0	
4	71.0	0.0	0.0	0.0	0.0	0.0	0.0	

5 rows × 21 columns

In [19]:
```
X["Label"] = y
```

In [20]:
```
X.head()
```

Out[20]:

	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	sick	pregnant	t
0	42.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	24.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	47.0	1.0	0.0	0.0	0.0	0.0	0.0	
3	71.0	0.0	1.0	0.0	0.0	0.0	0.0	
4	71.0	0.0	0.0	0.0	0.0	0.0	0.0	

5 rows × 22 columns

In [21]:
```
X.to_csv("Data/Cluster_data.csv", index = False)
```

In [22]:
```
import pickle
```

In [24]:
```
with open("Cluster_model/clustering.pkl", 'wb') as f:
    pickle.dump(kmeans, f)
```

In [ ]:
```

```