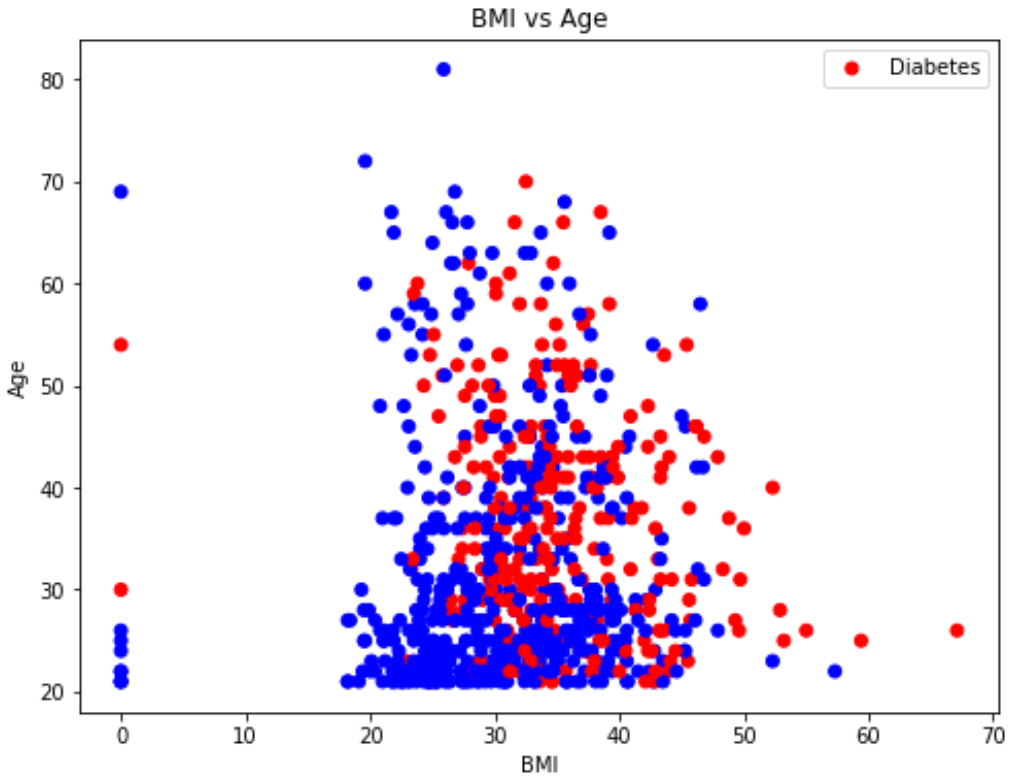
Name : Abhishek Guleri
Roll No. : 185509
Lab : Data Mining Lab
Assignment No. : 5
Branch : CSE DD

---

1. Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the color of the points based on the class of the candidate (0 or 1).

```
import pandas as pd
import matplotlib.pyplot as plt

input_data = pd.read_csv(r'/content/diabetes.csv')
data = input_data[['BMI','Age','Outcome']]
fig, ax = plt.subplots(figsize=(8,6))
ax.scatter(data['BMI'],data['Age'],c=['red' if v == 1 else 'blue' for v in
data['Outcome']])
plt.xlabel('BMI')
plt.ylabel('Age')
plt.title('BMI vs Age')
ax.legend(['Diabetes','No Diabetes'])
```



2. Create a distance matrix for the data.

```
import pandas as pd
from scipy.spatial import distance_matrix
pd.DataFrame(distance_matrix(data[['BMI']], data[['Age']]),
index=data.index, columns=data.index)
```

3. Make a hierarchical clustering analysis using the single linkage method. Then create an object that contains only two clusters.

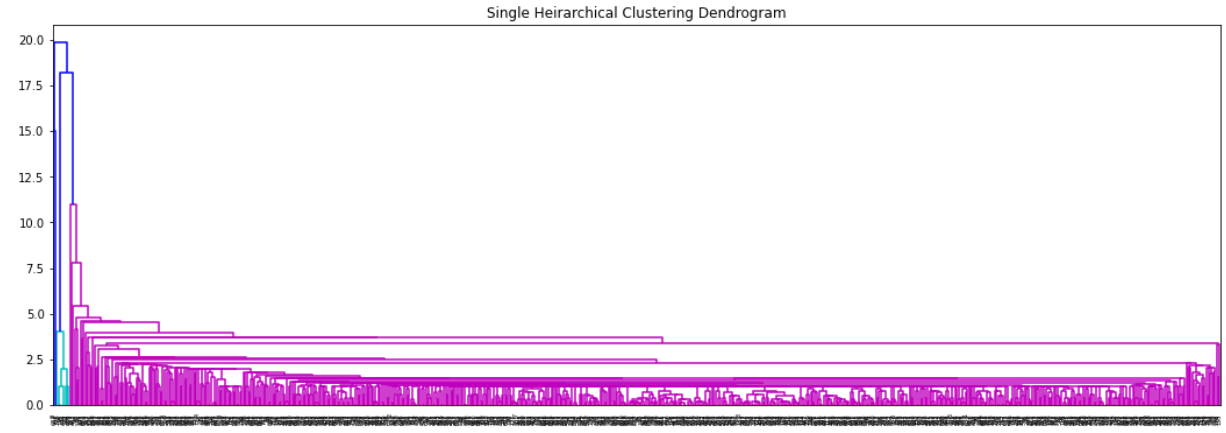
```
from sklearn.cluster import AgglomerativeClustering
import numpy as np
cluster_single = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
linkage='single')
cluster_single.fit_predict(np.array(data[['BMI','Age']]))
```

4. Make a hierarchical clustering analysis using the complete linkage method. Then create an object that contains only two clusters.

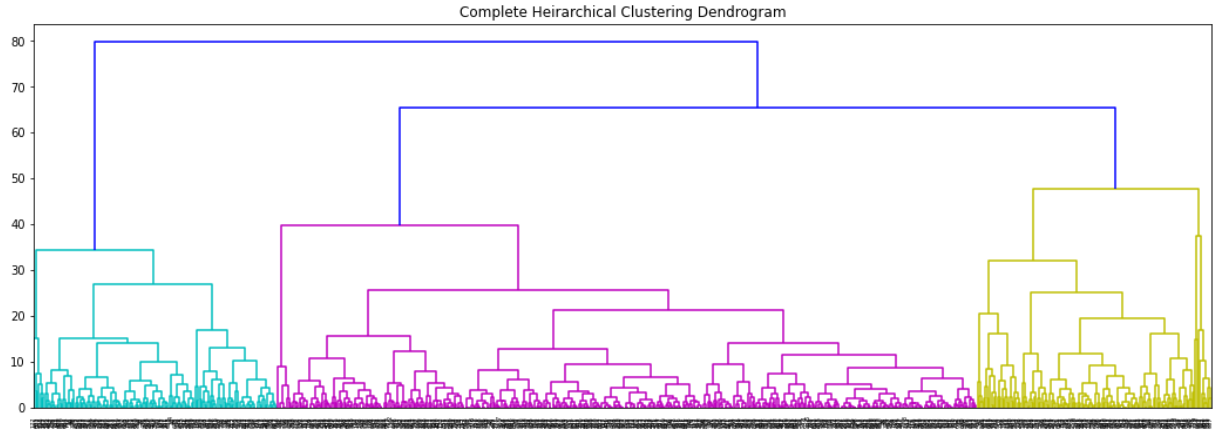
```
from sklearn.cluster import AgglomerativeClustering
import numpy as np
cluster_complete = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
linkage='complete')
cluster_complete.fit_predict(np.array(data[['BMI','Age']]))
```

5. Construct the trees that are produced by Questions 2 and 3 and draw the two clusters (at the plots).

```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(18, 6))
plt.title("Single Hierarchical Clustering Dendrogram")
dend_complete = shc.dendrogram(shc.linkage(data[['BMI','Age']],
method='single'))
```

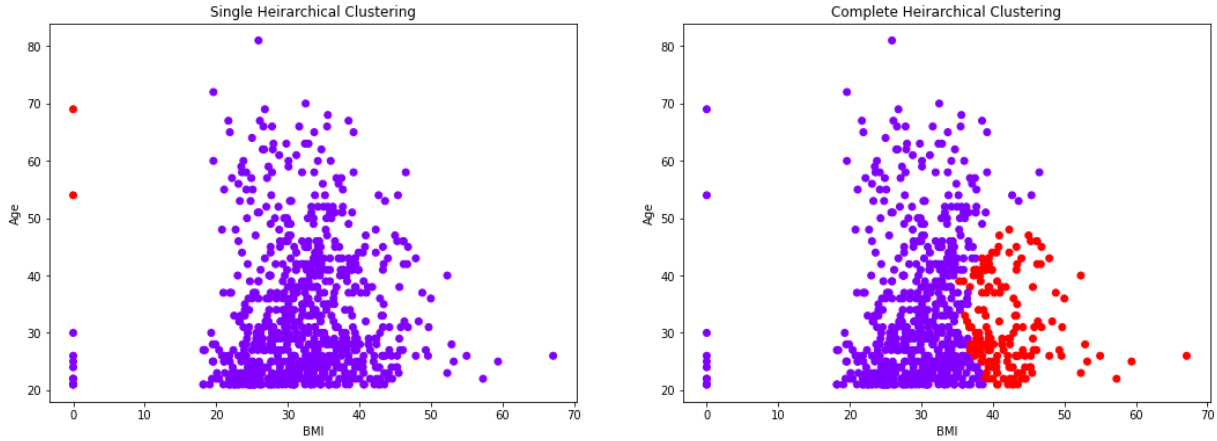


```
plt.figure(figsize=(18, 6))
plt.title("Complete Hierarchical Clustering Dendrogram")
dend_complete = shc.dendrogram(shc.linkage(data[['BMI','Age']],
method='complete'))
```



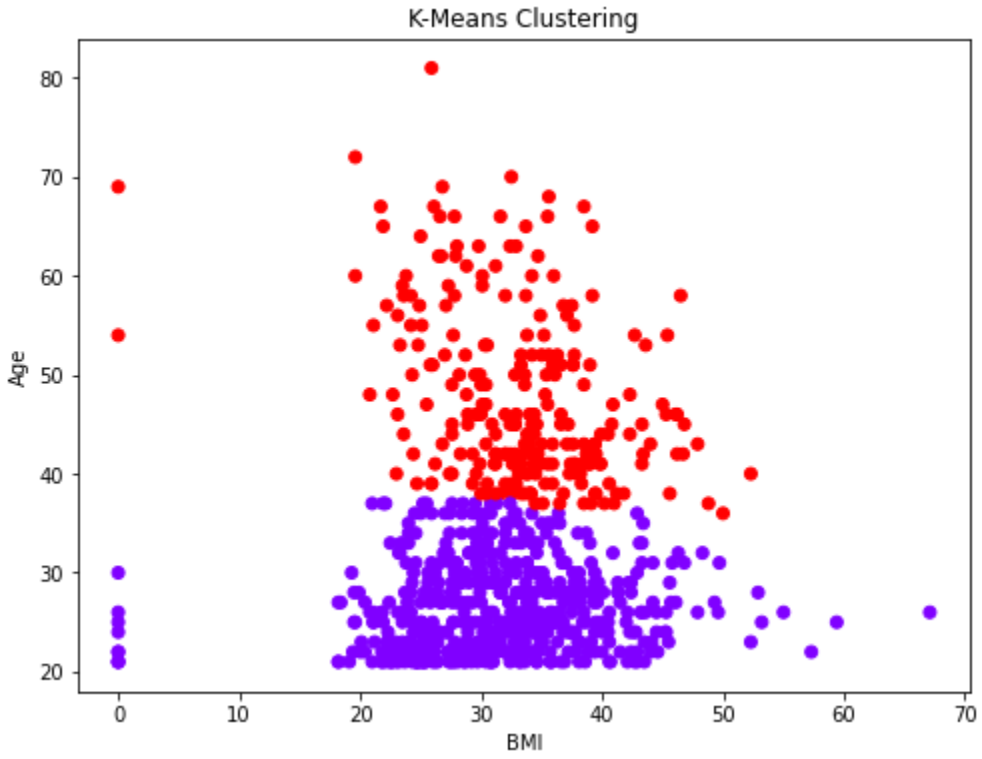
6. Construct two scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the color of the points based on the cluster that those points belong to. Each scatterplot is for a different clustering method.

```
fig, ax = plt.subplots(1,2,figsize=(18,6))
ax[0].set_xlabel('BMI')
ax[0].set_ylabel('Age')
ax[0].set_title('Single Hierarchical Clustering')
ax[0].scatter(data['BMI'],data['Age'], c=cluster_single.labels_,
cmap='rainbow')
ax[1].set_xlabel('BMI')
ax[1].set_ylabel('Age')
ax[1].set_title('Complete Hierarchical Clustering')
ax[1].scatter(data['BMI'],data['Age'], c=cluster_complete.labels_,
cmap='rainbow')
```



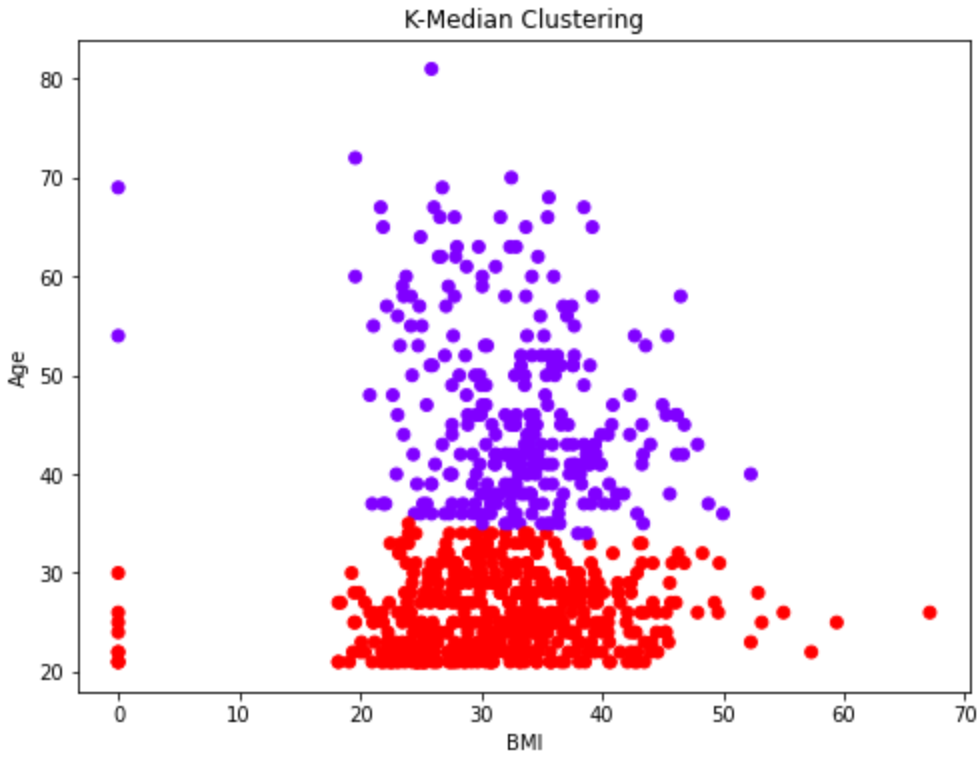
7. Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the color of the points based on the cluster (retrieved from k-mean method) that those points belong to.

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=2, random_state=0).fit(data[['BMI','Age']])
kmeans.predict(data[['BMI','Age']])
fig, ax = plt.subplots(figsize=(8,6))
ax.set_xlabel('BMI')
ax.set_ylabel('Age')
ax.set_title('K-Means Clustering')
ax.scatter(data['BMI'],data['Age'], c=kmeans.labels_, cmap='rainbow')
```



8. Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable.Moreover, determine the color of the points based on the cluster (retrieved from k-median method) that those points belong to.

```
from sklearn_extra.cluster import KMedoids
kmediods = KMedoids(n_clusters=2, random_state=0).fit(data[['BMI','Age']])
kmediods.predict(data[['BMI','Age']])
fig, ax = plt.subplots(figsize=(8,6))
ax.set_xlabel('BMI')
ax.set_ylabel('Age')
ax.set_title('K-Median Clustering')
ax.scatter(data['BMI'],data['Age'], c=kmediods.labels_, cmap='rainbow')
```


K-Median Clustering
Age
BMI
80
70
60
50
40
30
20
0
10
20
30
40
50
60
70