**Experiment No.06**

**A.1 Aim:** Implementation of K-means clustering using any programming language like JAVA, C++, Python or WEKA Tool.

**PART B**

(PART B: TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per the following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case there is no Blackboard access available)***

|  |  |
| --- | --- |
| **Roll No.** 50 | **Name:** AMEY THAKUR |
| **Class:** Comps TE B | **Batch:** B3 |
| **Date of Experiment:** 15/04/2021 | **Date of Submission:** 15/04/2021 |
| **Grade:** |  |

**B.1 Software Code written by a student:**

***(Paste your problem statement related to your case study completed during the 2 hours of practice in the lab here)***

# importing libraries

import numpy as nm

import matplotlib.pyplot as mtp

import pandas as pd

# Importing the dataset

dataset = pd.read\_csv('diabetes\_csv.csv')

x = dataset.iloc[:, [7, 5]].values

#finding optimal number of clusters using the elbow method

from sklearn.cluster import KMeans

wcss\_list= [] #Initializing the list for the values of WCSS

#Using a loop for iterations from 1 to 10.

for i in range(1, 11):

kmeans = KMeans(n\_clusters=i, init='k-means++', random\_state=42)

kmeans.fit(x)

wcss\_list.append(kmeans.inertia\_)

mtp.plot(range(1, 11), wcss\_list)

mtp.title('The Elbow Method Graph')

mtp.xlabel('Number of clusters(k)')

mtp.ylabel('wcss\_list')

mtp.show()

#training the K-means model on a dataset

kmeans = KMeans(n\_clusters=2, init='k-means++', random\_state= 42)

y\_predict= kmeans.fit\_predict(x)

mtp.scatter(x[y\_predict == 0, 0], x[y\_predict == 0, 1], s = 100, c = 'blue', label = 'Cluster 1') #for first cluster

mtp.scatter(x[y\_predict == 1, 0], x[y\_predict == 1, 1], s = 100, c = 'green', label = 'Cluster 2') #for second cluster

mtp.scatter(kmeans.cluster\_centers\_[:, 0],

kmeans.cluster\_centers\_[:, 1], s = 300, c = 'yellow', label = 'Centroid')

mtp.title('Clusters of patients')

mtp.xlabel('Age(in years)')

mtp.ylabel('BMI(Body Mass Index)')

mtp.legend()

mtp.show()

**B.2 Input and Output:**

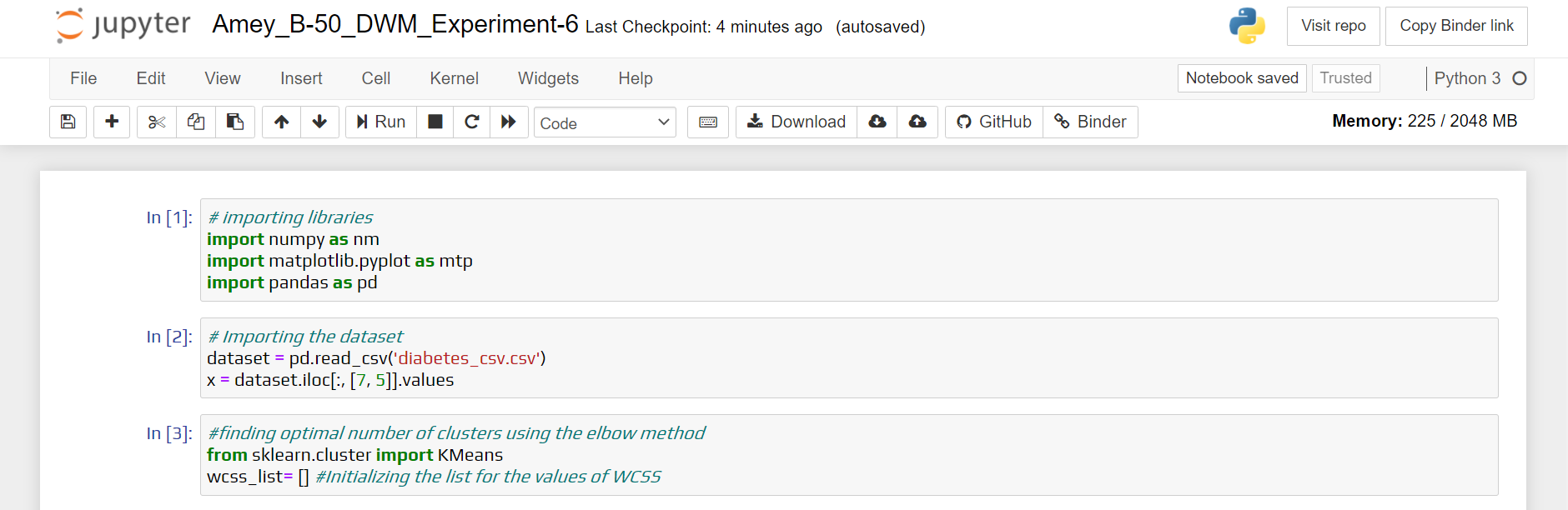
***(Paste your program input and output in the following format, If there is an error then paste the specific error in the output part. In case of an error with the due permission of the faculty, an extension can be given to submit the error-free code with output in due course of time. Students will be graded accordingly.)***

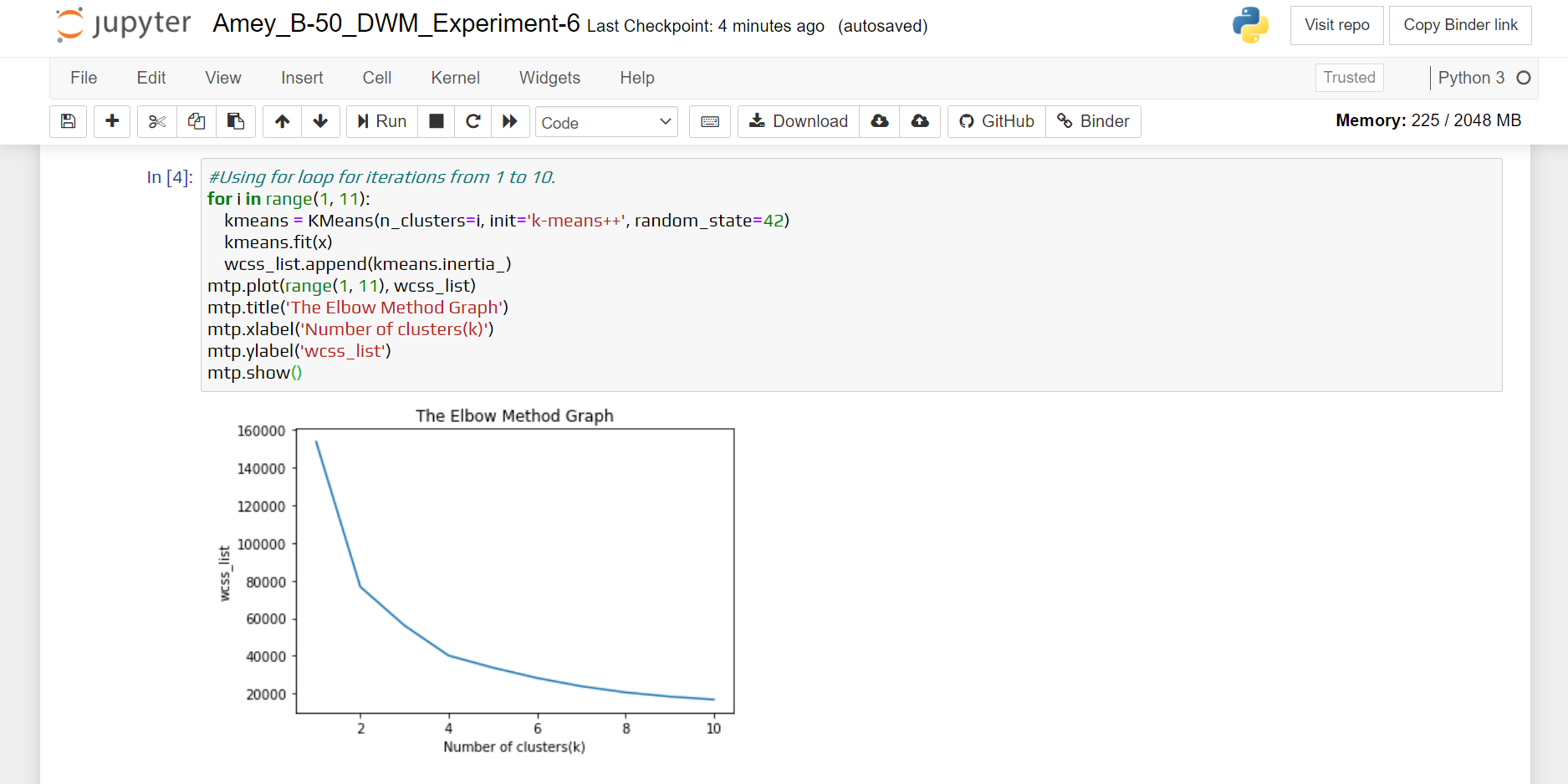
**Jupyter Notebook**

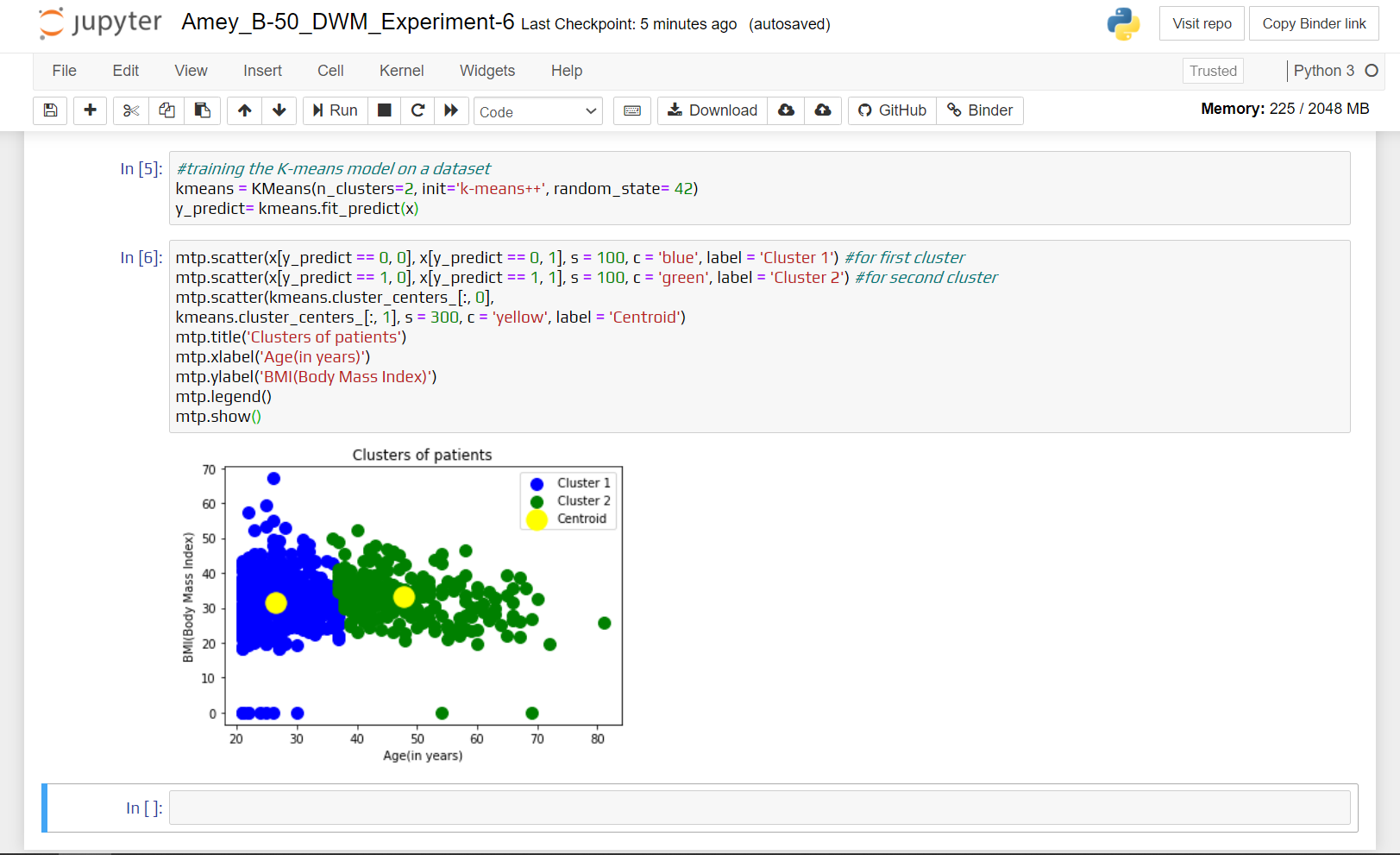
**Jupyter/Binder:**

****

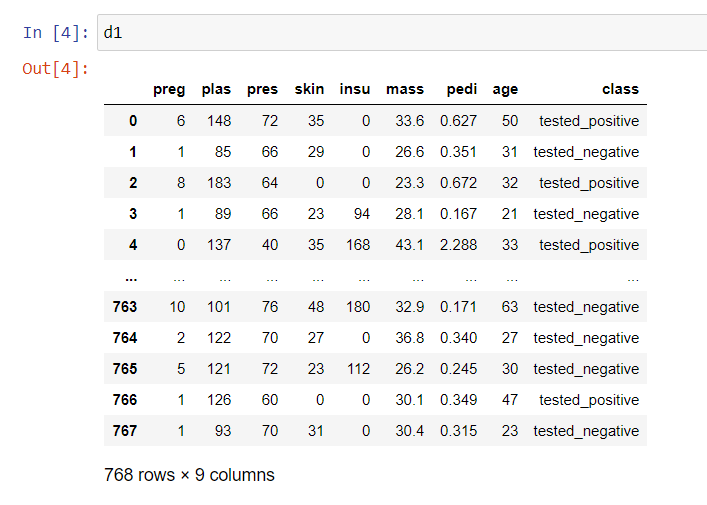
**Jupyter Notebook:**

****

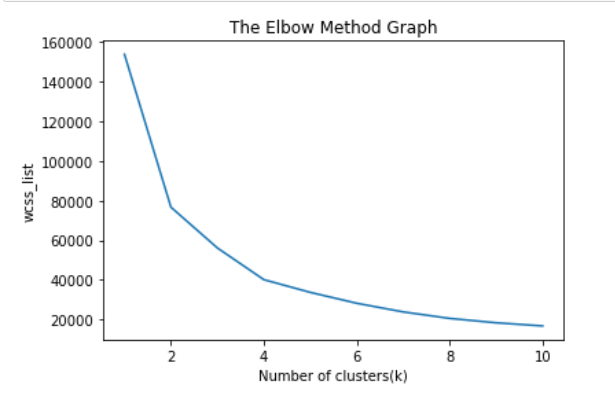
****

****

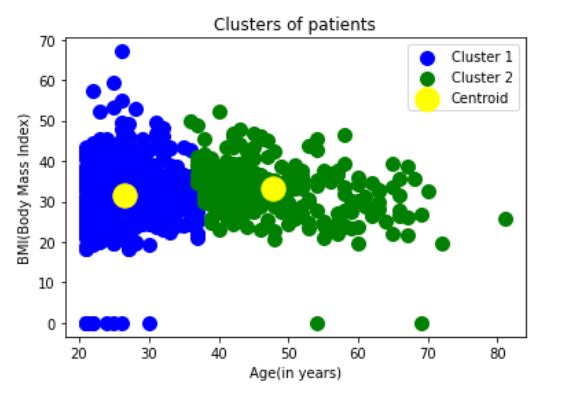
**Sample Dataset:**

****

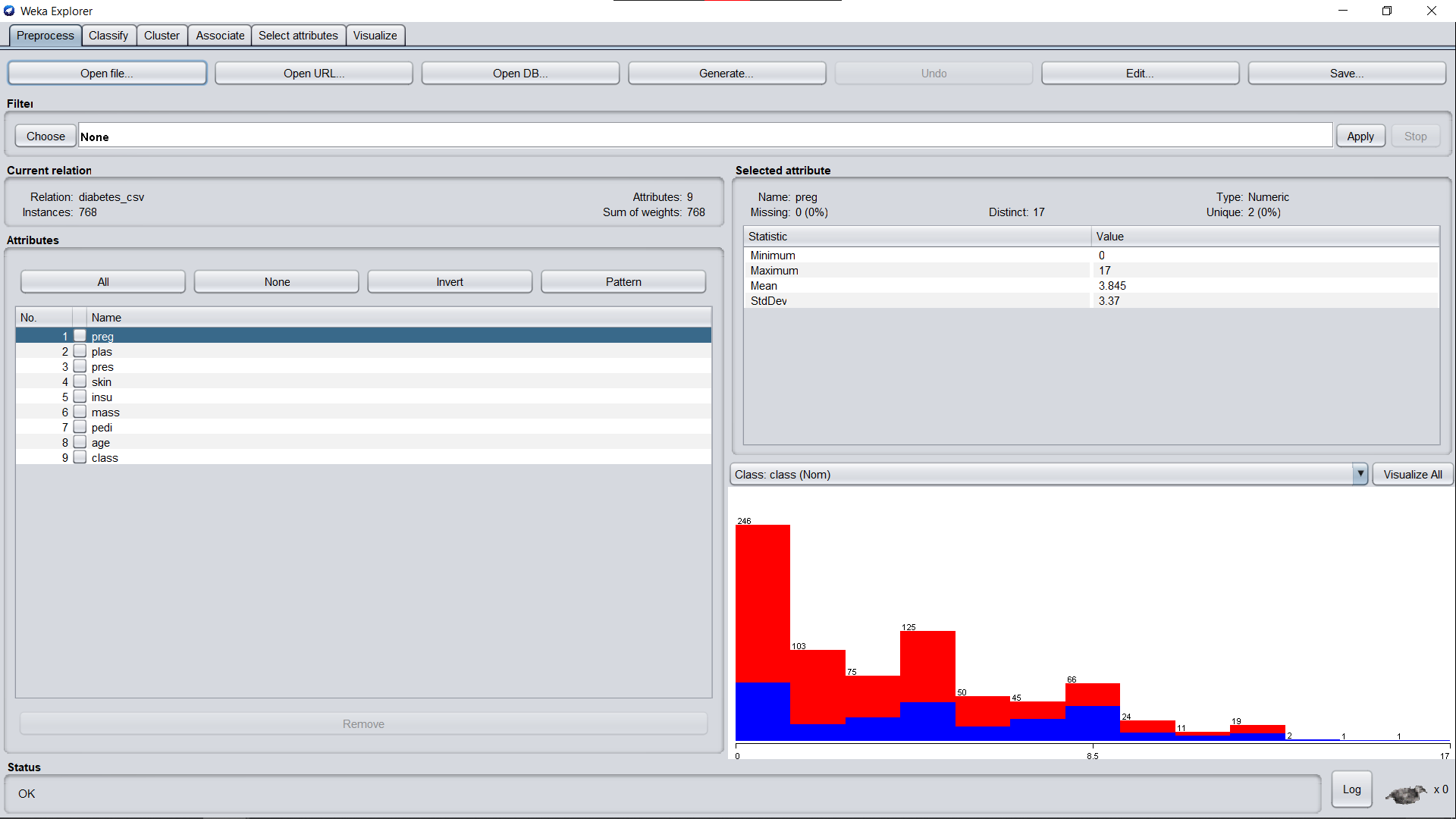
**The Elbow Method Graph:**

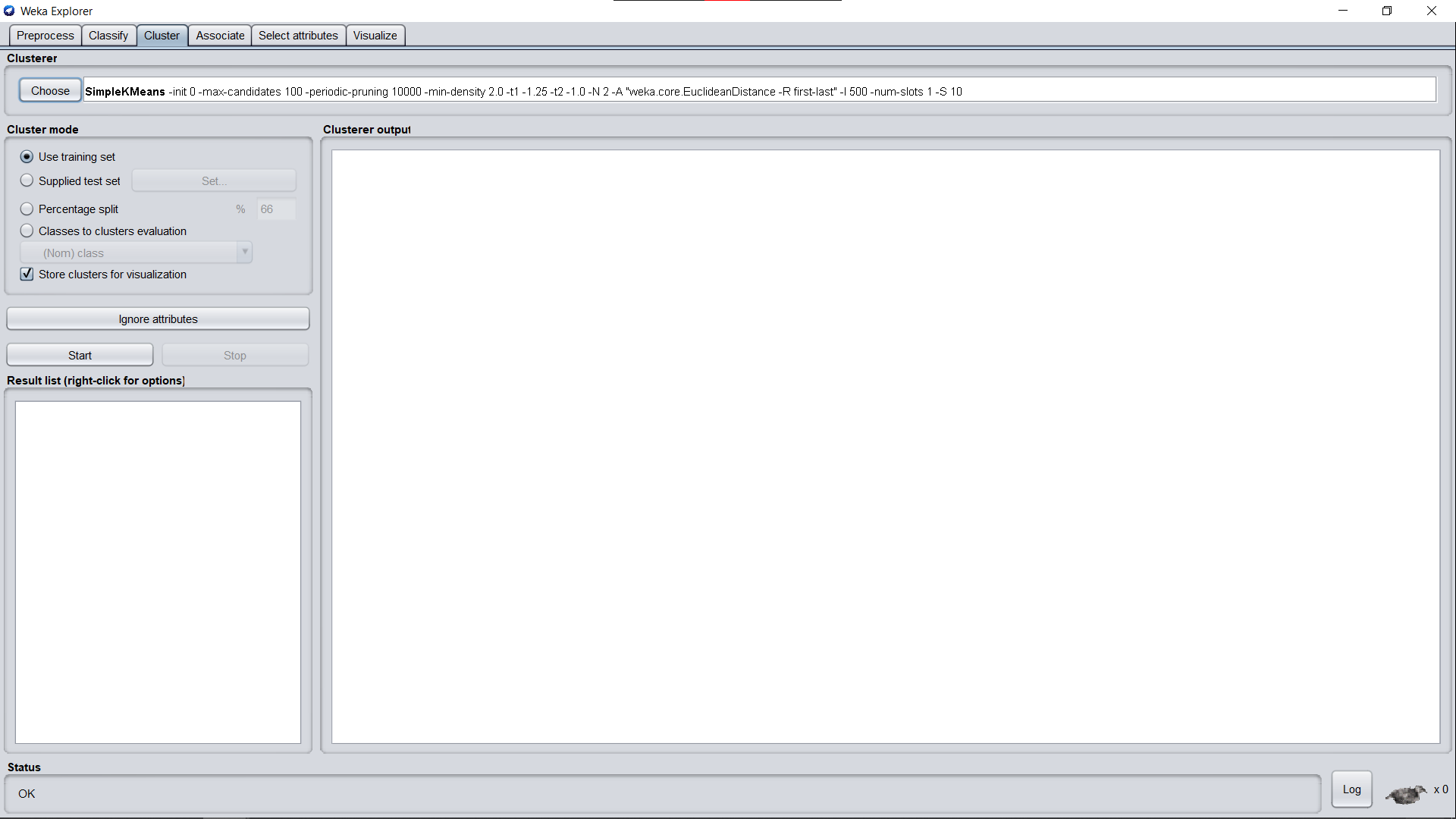
****

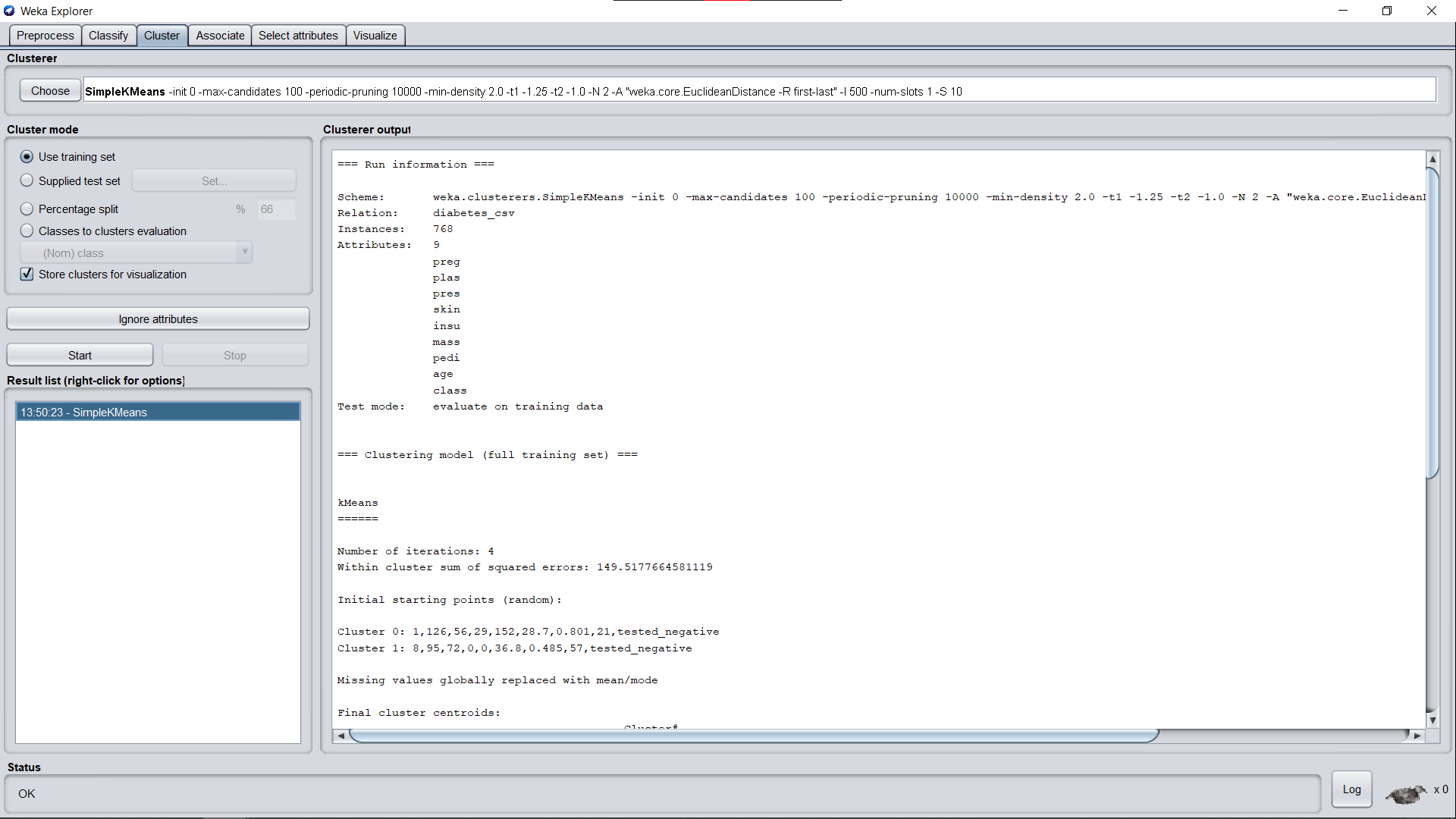
**Python Output:**

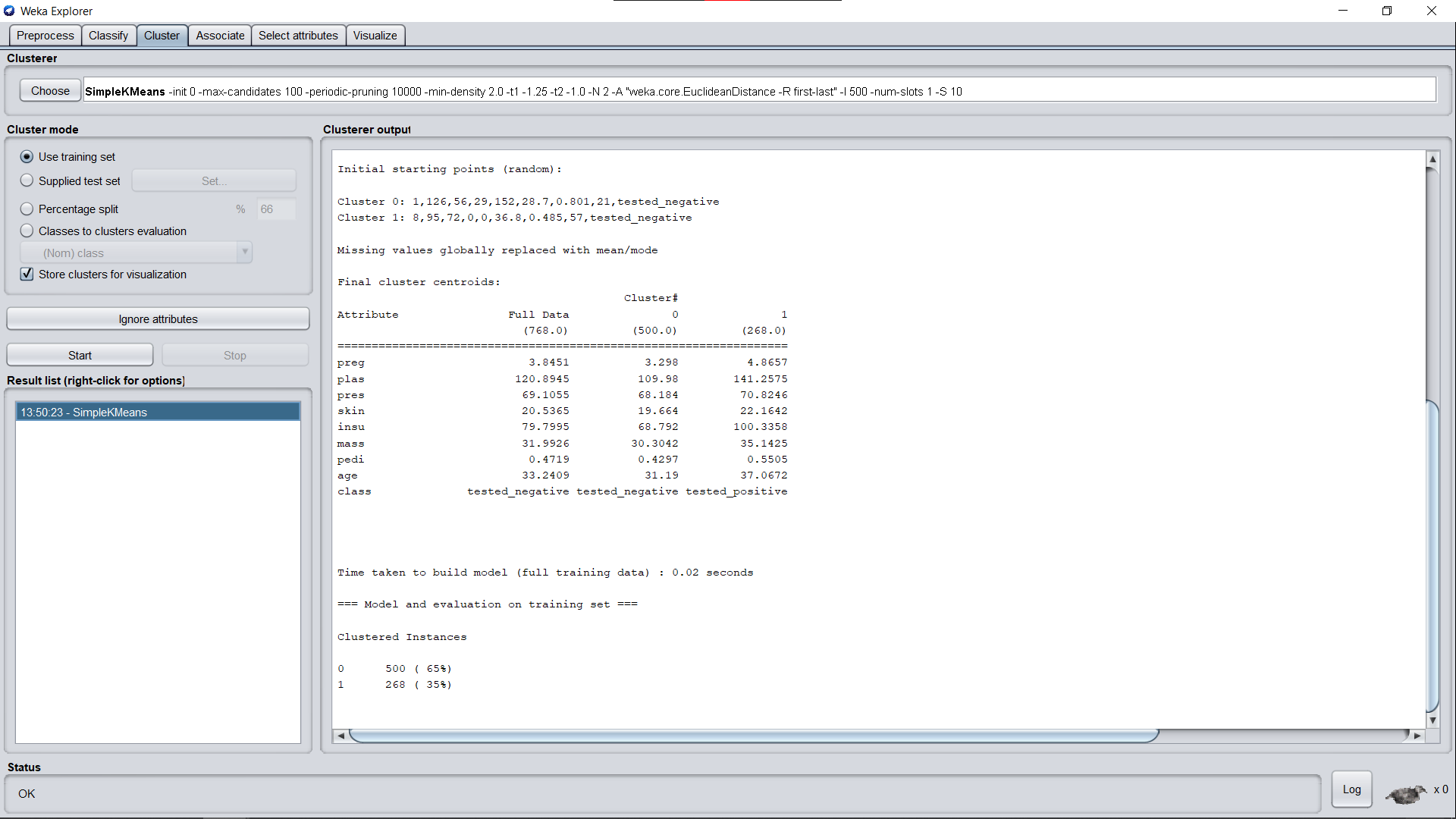
****

**Weka Tool**

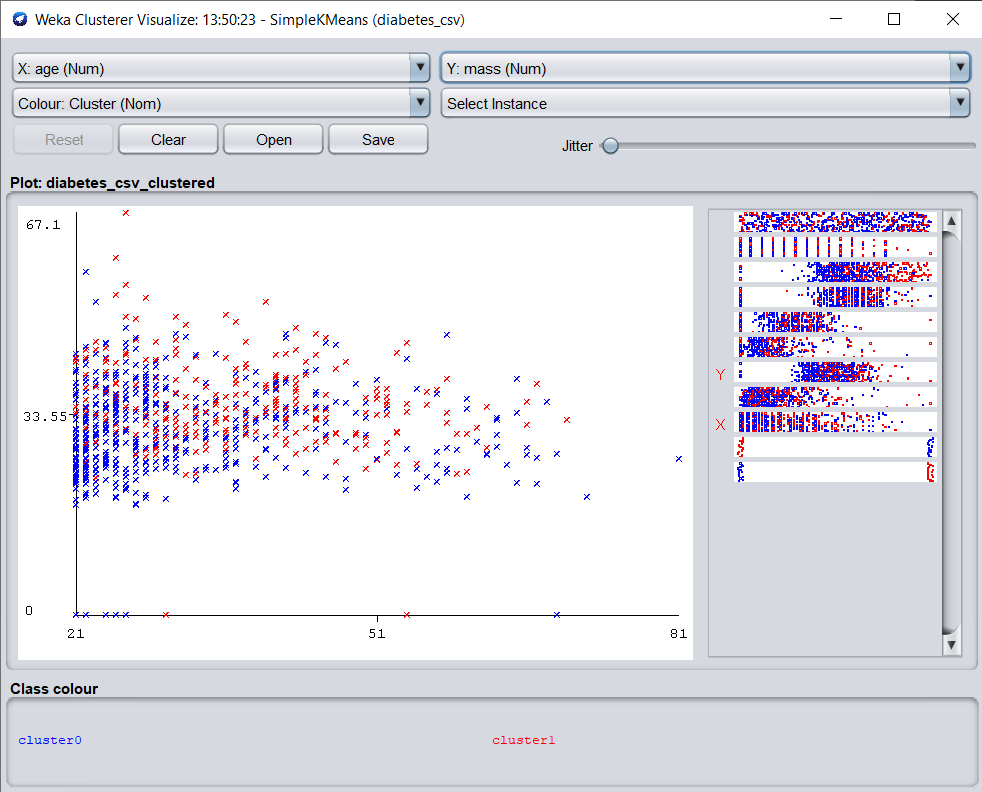
****

****

****

****

**Weka Output:**

****

**B.3 Observations and learning:**

***(Students are expected to comment on the output obtained with clear observations and learning for each task/ subpart assigned)***

K -means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. A cluster refers to a collection of data points aggregated together because of certain similarities. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

**B.4 Conclusion:**

*(****Students must write the conclusion as per the attainment of individual outcome listed above and learning/observation noted in section B.3)***

Hence we’ve successfully implemented K-means clustering through Python as well as Weka Tool.

**B.5 Question of Curiosity**

***(To be answered by the student based on the practical performed and learning/observations)***

1. What is Clustering? Types of clustering? Explain the advantages and disadvantages of clustering.

**Ans:**

**Clustering**

* Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is a collection of objects based on similarity and dissimilarity between them.
* Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

**Types Of Clustering Algorithms**

1. Connectivity-based Clustering (Hierarchical clustering)
2. Centroids-based Clustering (Partitioning methods)
3. Distribution-based Clustering
4. Density-based Clustering (Model-based methods)
5. Fuzzy Clustering
6. Constraint-based (Supervised Clustering)

**Advantages and Disadvantages of Clustering**

* The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention.
* Disadvantages of clustering are complexity and inability to recover from database corruption.

1. Give the advantages and disadvantages of K- means clustering.

**Ans:**

**Advantages of K-means clustering:**

1. Relatively simple to implement.
2. Scales to large data sets.
3. Guarantees convergence.
4. Can warm-start the positions of centroids.
5. Easily adapts to new examples.
6. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

**Disadvantages of K-means clustering:**

1. Choosing manually.
2. Being dependent on initial values.
3. Clustering data of varying sizes and density.
4. Clustering outliers.
5. Scaling with several dimensions.

1. How is the number of clusters chosen?

**Ans:**

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k. Distortion: It is calculated as the average of the squared distances from the cluster centres of the respective clusters. Typically, the Euclidean distance metric is used. Inertia: It is the sum of squared distances of samples to their closest cluster centre. We iterate the values of k from 1 to 9 and calculate the values of distortions for each value of k and calculate the distortion and inertia for each value of k in the given range.