Artificial Intelligence Project

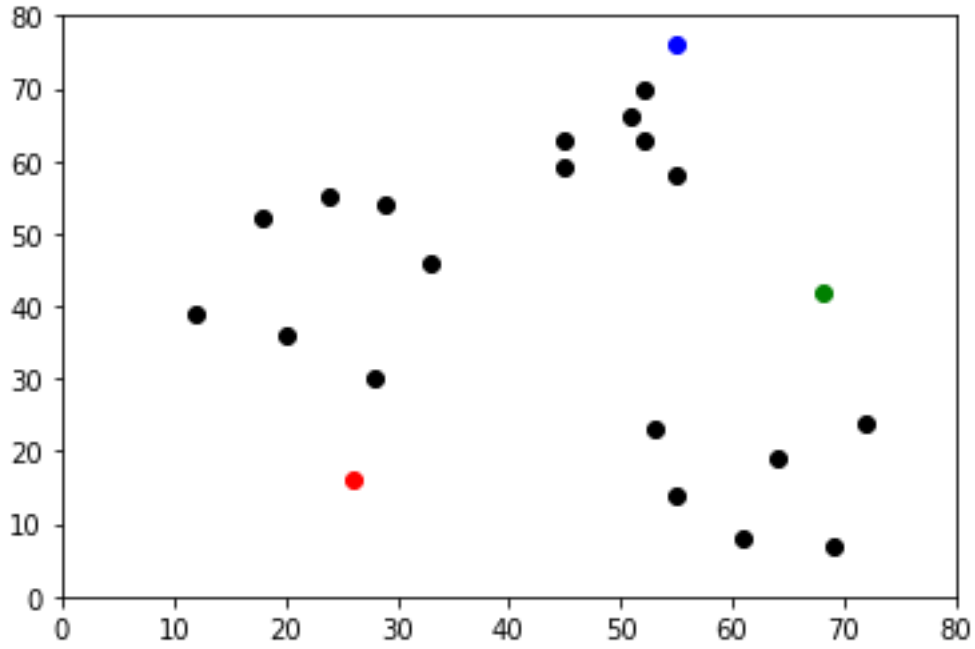
Mentor: Dr Susham Biswas

Step-01: In this Machine Learning project I have used some python library that is more useful to read the CSV file and plot the all relations in data point that I have taken randomly. I am explaining the following python library that I have taken in this project. Firstly '**Pandas**' library, It is used to analyze data. It is an open source python package that is most widely used for data science or data analysis and machine learning tasks. It is built on top of another package named Numpy which provides support for multi-dimensional array. Another one is '**Numpy**' library; it is used for the working with arrays. It also has functions for working in domain of linear algebra, Fourier Transform, and matrices. It is an open source library we can use it freely. Lastly **'Matplotlib'** it is a cross-platform, data visualization and graphical plotting library for python and its numerical extension Numpy. As such, it offers a viable open source alternative to MATLAB.

Step-02: Generally I have taken random data frame in array form and stored in a variable called df.

Step-03: I have written a code using Numpy library. The random seed is numerical value that generates a new set or repeats pseudo-random numbers. I have taken around 200 limits. The value in the Numpy random seed saves the state of randomness. If we call the seed function using value 1 multiple times, the computer displays the same random numbers.

Step-04: Now I plotted a figure using Matplotlib library in which I write their parameter of figure. Now move to scatter plot syntax generally we use to visualize the K-Means data points to see the changes in the centroids or we can call the new mean that will change for every new cluster group until the final classification will not occur. Scatter graph will point out the all centroids that we are taking initially in red, green and blue. It will be easy to visualize the centroids in the three following clusters. I also mentioned the x-axis limit and y-axis limited in the range 0 to 80. And at the last I wrote a code to show the scatter graph. I am attaching the scatter graph that you can see here.

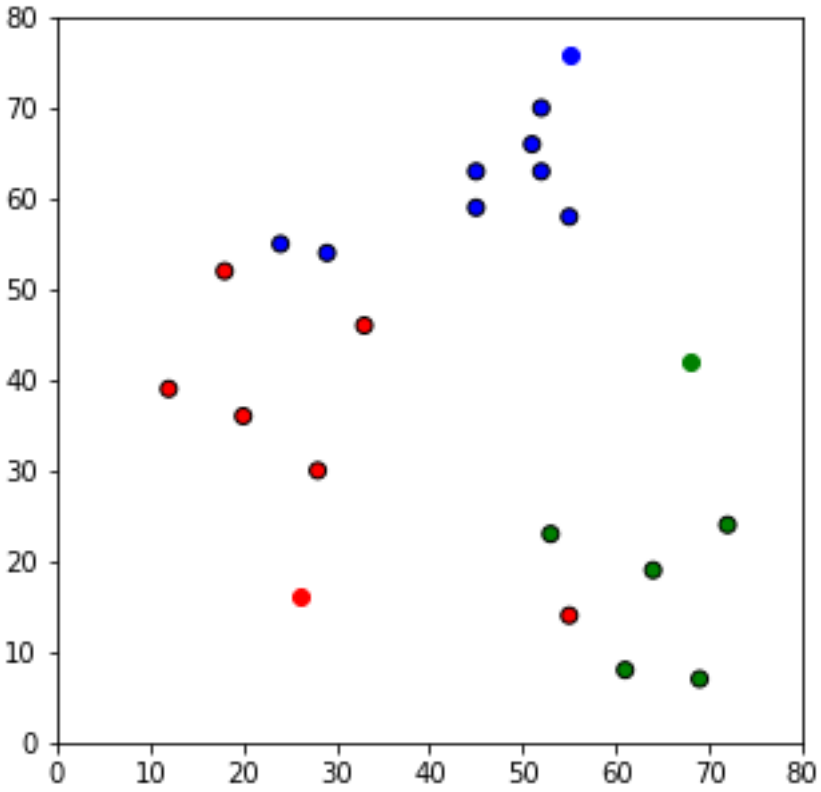


So, you can see in this plot that initially I have taken a random mean value that is mentioned in the red, blue and green color.

Step-05: I defined **assignment** variable that is storing my dataset (df) and centroids and then I am using **'for'** loop to calculate distance from each data points to others. That is a distance equation to calculate the distance between two data points in which I used Numpy library to calculate the distance between two points that loop will calculate every distance from first assumed mean value and It will compare the distance between all points where as it will extract that all value that consist closest distance from one point to another point. In that code I have used lambda function because I have not created a new variable to store the closest distance between the data points.

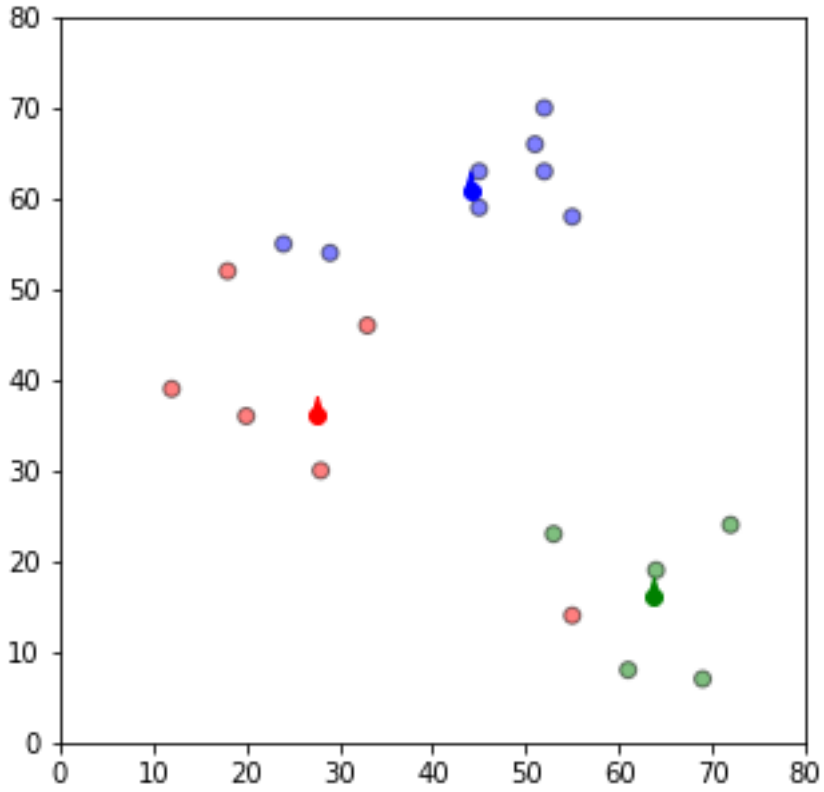
Step-06: Now I printed the all value that is occurring from the closest distance loop and all x values and y values. In which we can see the closest distance from point 1, distance from point 2 and distance from the point 3.

Step-07: I plotted all new points that are mentioned in the new scatter plot.

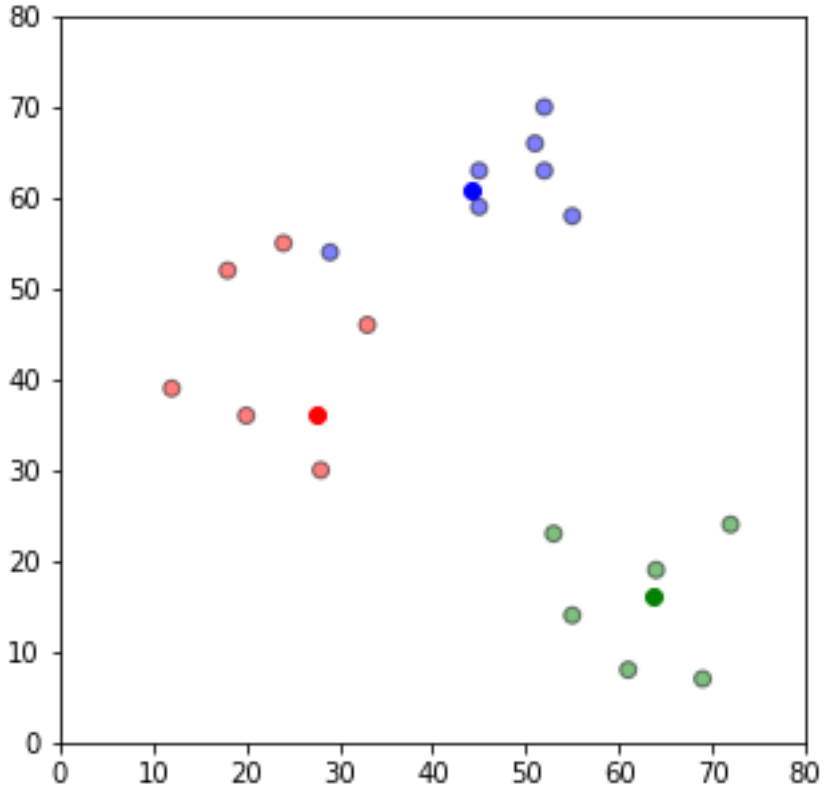


Now, you can see in this plot that it has pointed all nearest point that is in the range of three mean data points.

Step-08: In this step firstly I imported the **'copy'** because a deep copy creates a new compound object before inserting copies of the items found in the original into it in a recursive manner. It means first constructing a new collection object and then recursively populating it with copies of the child objects found in the original. In the case of deep copy, a copy of the object is copied into another object. It means that **any changes** made to a copy of the object **do not reflect in the original object.** And then I defined a update that will update all the new centroids after the calculation of all distances from one data points to other data points. Again I plotted the scatter plot for the analysis of new centroids and newest cluster that has been updated. I am attaching the scatter plot below.



So, you can see in this graph the changes between clusters are not stopped that is why we are updating again and again all mean value and new clusters.



In the final graph as I don’t know that what will be the last step in which the clustering will not change for that I used while loop in the last step to reach the final output of clustering.

Step-09: In this last step I used while loop that is **‘while True’** It will not proceed further when the centroids does not change. So I added **‘If’** condition that is for if the centroids will not change or the clustering will there then it will **‘break’** the loop for the specific step. And here is the final clustering group that is attached below.