PROJECT 3 REPORT MACHINE LEARNING -6363

Name: Sanket Manik Salunke

Student Id: 1001764897

Project Description:

This project is based on implementation of K means cluster Algorithm.

CODE Language Used – Python

Environment- Anaconda (Jupyter Notebook):

Packages:

import pandas as pd

import numpy as np

from pandas.plotting import parallel_coordinates #plot

%matplotlib inline

import sklearn.metrics as sm #accuracy and confusion matrix calculation

from itertools import cycle #for color values of scatter plot

import matplotlib.pyplot as plt #plot

from sklearn.decomposition import PCA #decomposition

Steps:

We have two types of pattern identification methods: 1. Classification 2. Clustering

In this project Kmeans clustering algorithm is used to form 'K' Clusters

Clustering is an Unsupervised Learning algorithm and in this we have only one set of data as a Input data which is not labelled.

1. First, I am importing given Iris data as dataframe and then assigning values for target as per the unique values of target as {'Iris-setosa': 1, 'Iris-versicolor': 2, 'Iris-virginica': 3}

```
]: var1_sorted=sorted(var1)
   var1_sorted

]: ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']

]: var1_mapped= dict([(y,x+1) for x,y in enumerate(var1_sorted)])
   var1_mapped

]: {'Iris-setosa': 1, 'Iris-versicolor': 2, 'Iris-virginica': 3}
```

```
[7]: #Reference: https://stackoverflow.com/questions/60081686/python-how-to-map-numbers-to-unique-items-enumerate-unique-objects-in-a-
df['target']=pd.factorize(df['target'].tolist())[0]
df.tail()
```

[7]:

	sepal_length	sepal_width	petal_length	petal_width	target
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

2. Then I am separating target value and feature values:

t[9]:

	sepal_length	sepal_width	petal_length	petal_width
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
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

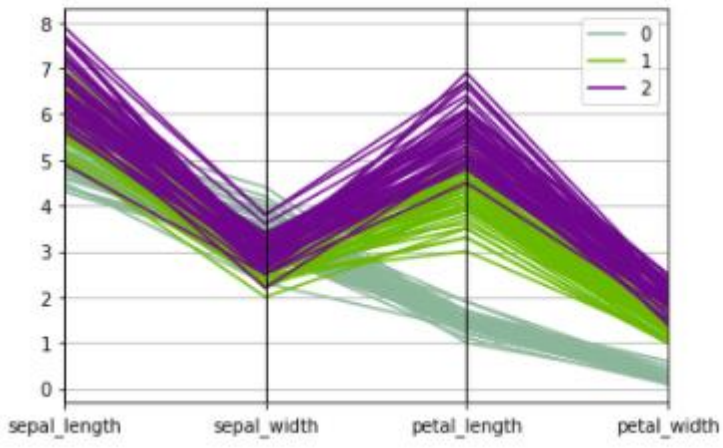
150 rows × 4 columns

3. Just to study iris dataset features, I am trying to plot it visually to see all four features as below. (Target:0,1,2 corresponding to {'Iris-setosa': 1, 'Iris-versicolor': 2, 'Iris-virginica': 3})
This is actual given data and its visualisation.

```
from pandas.plotting import parallel_coordinates
%matplotlib inline

parallel_coordinates(df, 'target')
```

10]: <matplotlib.axes._subplots.AxesSubplot at 0x18cd3d76160>



4. After this I found out given data target value array to compare it after the clustering. So that I can get the accuracy values for the points.

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2], dtype=int64)
```

As we have total 150 rows in the given dataset, array length is 150. (Target values for all the 150 rows).

5. Then I am implementing k means clustering algorithm by giving the value of K for the given iris-dataset.
   While Implementing k means algorithm, I am calculating centroid value first randomly. Then I am iterating over the loop to calculate the Euclidean distance from points to the centroid and then I am finding out the minimum value.
   We will find out all the clusters with given value of 'K'.
   Here by observing Iris data I am giving value of K as 3.

```
clusture_array= Kmeans_Cluster(X,3)
```

```
clusture_array
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2,
       2, 2, 2, 1, 1, 2, 2, 2, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 2, 2,
       2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 1], dtype=int64)
```

6. Now I am calculating the total points which are wrongly assigned with cluster value 3

```
count=0
p=Y.to_numpy()
for i in range(len(clusture_array)):

    if(clusture_array[i]!=p[i]):
        count = count+1
print("number of points wrongly assigned:",count)
# print(count)
```

```
number of points wrongly assigned: 17
```

I have also calculated accuracy and error for cluster value 3.

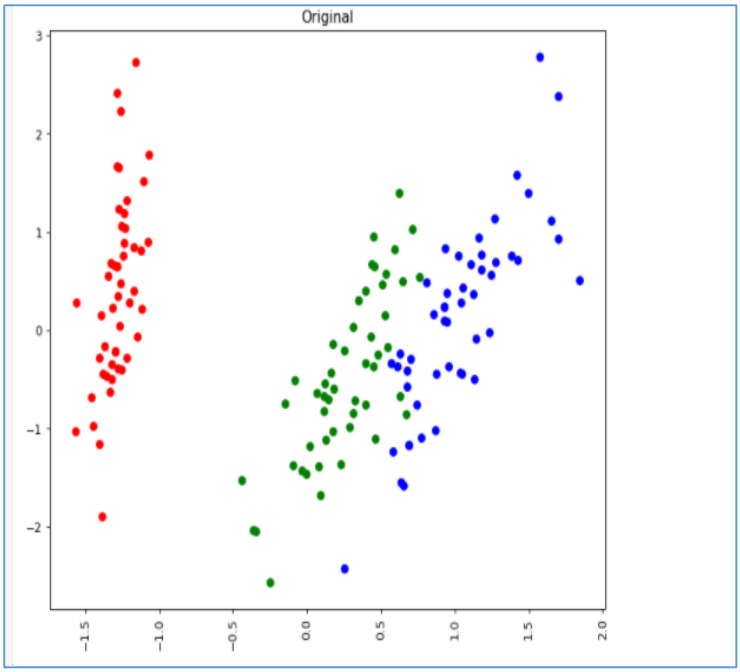
```
In [23]:
#Accuracy Calculation with Cluster Value 3
Accuracy = sm.accuracy_score(clusture_array, Y)
Accuracy= Accuracy*100
print("Accuracy of our method is:",Accuracy,"%")
```

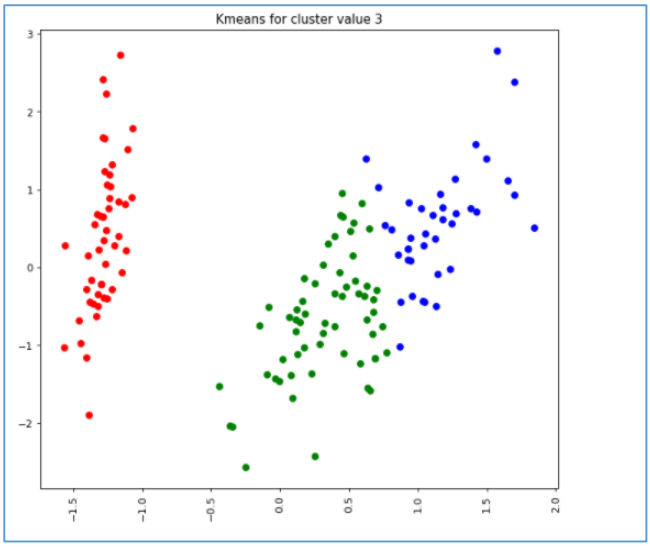
```
Accuracy of our method is: 88.66666666666667 %
```

```
In [24]:
error = 100-Accuracy
print("Error of our method is:",error,"%")
```

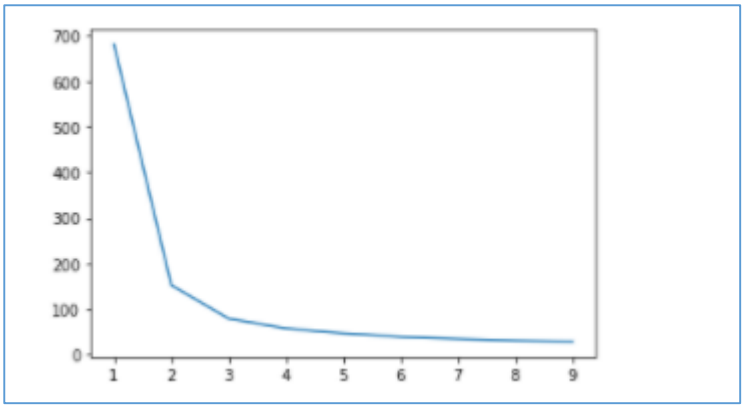
```
Error of our method is: 11.333333333333329 %
```

7. Plotting the scatter graph for original data to the clustered data.
   From the plot we can see that k means algorithm is just creating clusters without taking target values in consideration. You can see from left side first cluster is the same but for 2nd and 3rd cluster there are few changes as its not classifying. It is just creating clusters.

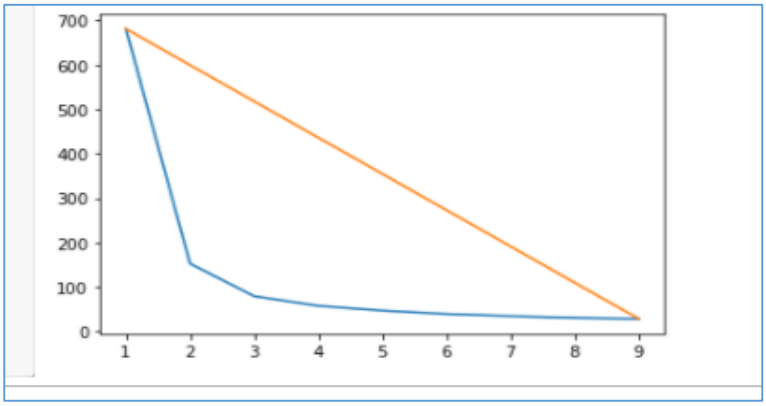




8. Regarding K means value.
   As per my understanding there is no particular method to calculate the K value for k means clustering. Whereas there are few methods which can be used to calculate K value but all of them are subjective.
9. I am trying to implement one method called as elbow method just to get the idea how k values can be predicted using elbow method.
   So based on the values of k I am trying to find out the distance between points with respect to cluster centroids.



In our Iris dataset, number of clusters was easily predictable as we have three type of cluster class values. From above graph, we can see that elbow like structure is in the graph. So I am joining the two points to find out the distance between line and all the clustering points. Then I am calculating Maximum distance value from the line which is cluster 3.



References:

https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation#:~:text=Although%20both%20techniques%20have%20certain,which%20differentiate%20them%20from%20other

https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

https://www.youtube.com/watch?v=W4fSRHeafMo
https://stackoverflow.com/questions/1401712/how-can-the-euclidean-distance-be-calculated-with-numpy
https://www.youtube.com/watch?v=dyH27J_En8M
http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial3/tutorial3.html