UCI Seed dataset: Implementation of hierarchical clustering using KNN classifier

Introduction:

Measurements and geometrical properties are the major part of the dataset. Using the properties form the data either of the 3 varieties of seed type is predicted. The UCI seed dataset consists of independent features as area of seed, perimeter of seed, compactness of seed, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove. The seed type is the required dependent variable to predict. The respective programming approach is hierarchical clustering and KNN classifier based.

Code implementation:

Firstly, the program starts with data reading and analysis. The dataframe consists of 210 rows and 8 columns inclusive of seed type dependent variable. Except, the seed type, all the other variables are of float64 data type.

Secondly, the K-nearest neighbors classifier has implemented to predict the array of seed type with neighbors as 3. Here, 3 is chosen as the value of neighbors as it produced the better accuracy when tried with multiple values.

Thirdly, hierarchical clustering with complete linkage is implemented to predict the array of data points which form the clusters. The dynamic programming approach is used to form clusters. The code returns which data points forms the cluster and the target cluster here the seed type.

Lastly, the accuracy of correctness is predicted by taking the right percentage of correct target variable value produced. The graphs for the clusters are plotted in the program to visualize the clustered data points in the given data frame.

Algorithms implemented:

KNN: The k-nearest neighbors algorithm is a non-parametric supervised learning method used for classification and regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

Hierarchical clustering: In Hierarchical Clustering Technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

Libraries :

Pandas: Used to read the csv data file

Numpy: Used to form the matrices and array

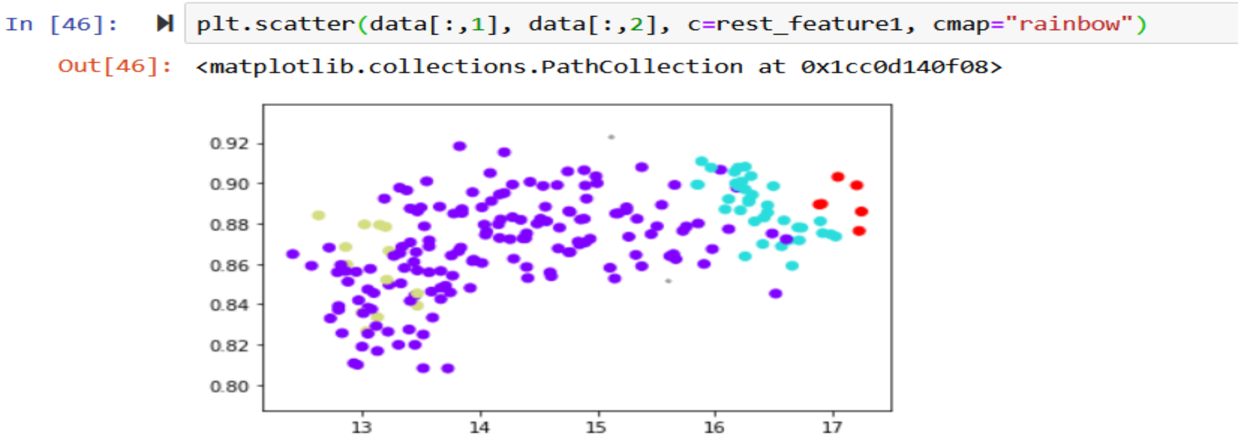
Math: used to implement square root function for the data values.

Matplotlib: Used for data visualization plotting of clusters

Statistics: Used to implement mode function to retrieve data points with high frequency in the predictions list.

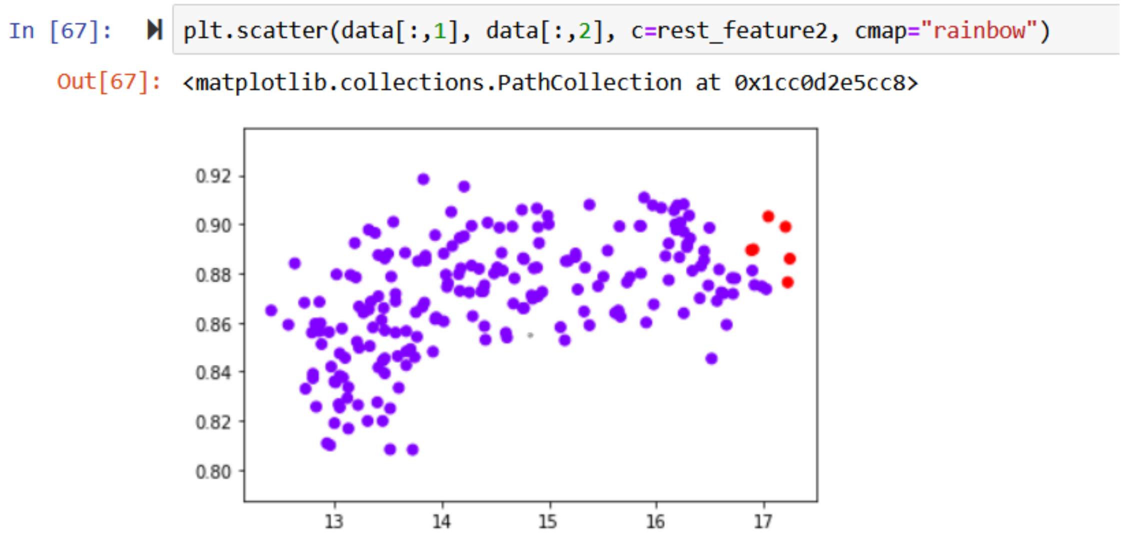
Data Visualization:

The following is the graph when the program produced 4 clusters for the given 205 iterations.



```
In [46]:    plt.scatter(data[:,1], data[:,2], c=rest_feature1, cmap="rainbow")
   Out[46]: <matplotlib.collections.PathCollection at 0x1cc0d140f08>
```

The following is the graph produced 2 clusters.



```
In [67]:  plt.scatter(data[:,1], data[:,2], c=rest_feature2, cmap="rainbow")
Out[67]:  <matplotlib.collections.PathCollection at 0x1cc0d2e5cc8>
```