Python ML Clustering Project

Cancer Data Analysis

Project Description:

In this particular project, we are using a dataset that contains information like quality radius_mean, texture_mean, perimeter_mean,area_mean,

smoothness_mean, compactness_mean, concavity_mean.

However, before you go ahead and make Clustering, it is advised that you first pre-process the data, since it may contain some irregularities and noise.

In addition, try various tricks and techniques in order to gain the best accuracy in your predictions.

Part-1: Data Exploration and Pre-processing

1. Load dataset
2. Find shape of dataset
3. Show basic information of data
4. Check null values
5. Drop unnamed and id columns.
6. Show values counts in diagnosis column
7. Remove Label column diagnosis
8. Create pair plot between two column radius_mean and radius_mean by diagnosis
9. Select only two feature radius_mean & texture_mean for clustering in new dataset
10. Apply scaling on new dataset

Part-2: Working with Models

1. Display hierarchical clustering as a dendrogram using scipy
2. Apply Agglomerative Clustering on dataset with 2 n number of clusters
3. Predict the cluster and create new column for cluster label data
4. Check count of label
5. Plot the label data
6. Check the silhouette score
7. Now apply kmeans clustering no dataset with 2 number of clusters
8. Check wcss score
9. Try different N number from 1 to 10 and plot the result of wcss score

10. Apply kmeans again with different no. of cluster according to best wcss score.
11. Create column for label cluster