Machine Learning Assignment - 2

Report for Q-1

Approach:

1) The dataset is downloaded from the given URL in file “abalone.data”. There are no missing attribute values in the given data.
2) The data in the file is read using pandas dataframe and then we create a 2D numpy array called “features” and a 1D numpy array called “labels”. In features we have 4177 rows, where each row corresponds to a data sample. There are 8 columns where each column represents a particular feature. For our case these 8 features are 'sex', 'length', 'diameter', 'height', 'whole_weight', 'shucked_weight', 'viscera_weight', 'shell_weight'. Only sex is a categorical variable having values ‘M’,’F’ and ‘I’. We encode these variables as: ‘M’ -> 0, ‘F’ -> ‘1’ and ‘I’ -> 2.
3) Before applying PCA we standardize this “features” matrix so that each column vector has mean = 0 and standard deviation = 1. So we compute the mean and standard deviation of the values present in each column. After that we do features[i][j] = (features[i][j] - mean of jth column) / (standard deviation of jth column) . As a result the features matrix is standardized.
4) Next we perform Principle Components Analysis on the features matrix where the number of components are chosen such that 95% of total variance is preserved. We use the inbuilt function of the sklearn library to perform PCA.
5) After performing PCA the 4177*8 sized features matrix reduces to 4177*3 sized matrix. This means that each data point now has 3 features. In order to plot the graph for the PCA, we plot the three components along the three axes. The number of rings of each data point is represented using a color gradient as shown in the results section of this report.
6) On this features matrix obtained in 5) we apply k-means clustering for the cases k=2,3,4,5,6,7,8. In order to initialize the k cluster centers, we choose k data points randomly from the features matrix and take them as the k initial cluster centers. The convergence criterion for the k-means algorithm is as follows:

   $$\sqrt{\sum_{i=1}^{k} (norm(old\ cluster\ center_i\ -\ new\ cluster\ center_i))^2} < \varepsilon$$

   norm(x) denotes the 2-norm of a vector x which is the square root of the sum of squares of its components.

   We update the cluster centers iteratively in the k-means algorithm.

   This means that we stop when the square root of the sum of the squares of distance between the old cluster center and the updated cluster center for the k cluster centers is less than epsilon. Epsilon is chosen as 1e-7. Essentially we stop when the cluster centers do not change much in consecutive iterations of the algorithm.
7) When the k means algorithm converges we know the cluster assignment of each of the data points in the features matrix. In order to compute the Normalized Mutual Information (NMI) we use the following formula: NMI = (2 * I(Y;C)) / (H(Y) + H(C)).

Y denotes the set cluster labels i.e the cluster assignments to each of the data points.
Y[i] = index of the cluster to which the i-th data point belongs

C denotes the set of class labels i.e the number of rings each data point has.
C[i] = number of rings in the i-th data point
H(X) denotes the entropy of the set X.
I(Y:C) denotes the mutual information between Y and C.
I(Y;C) = H(Y) - H(Y|C).
H(Y|C) denotes the entropy of the class labels within each of the clusters.

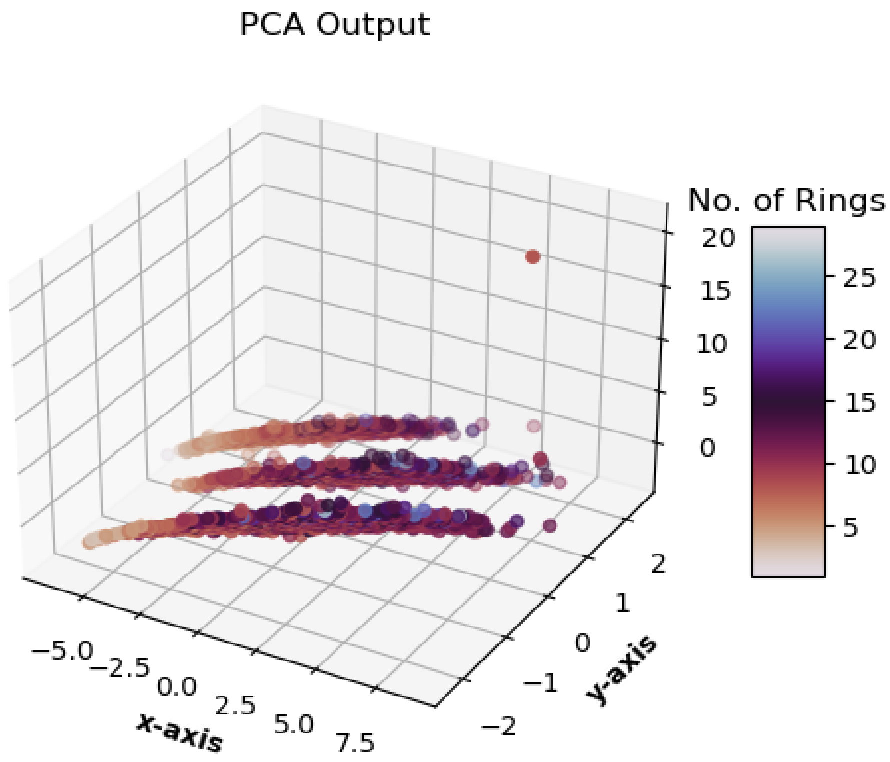
8) We run the k-means algorithm for k = 2,3,4,5,6,7,8 and then we plot the graph of k vs NMI for these cases. We also compute the value of k for which the NMI is maximum.

The results of the code are stated below:

We run the code 3 times and the results of those runs are as follows:

First Run:

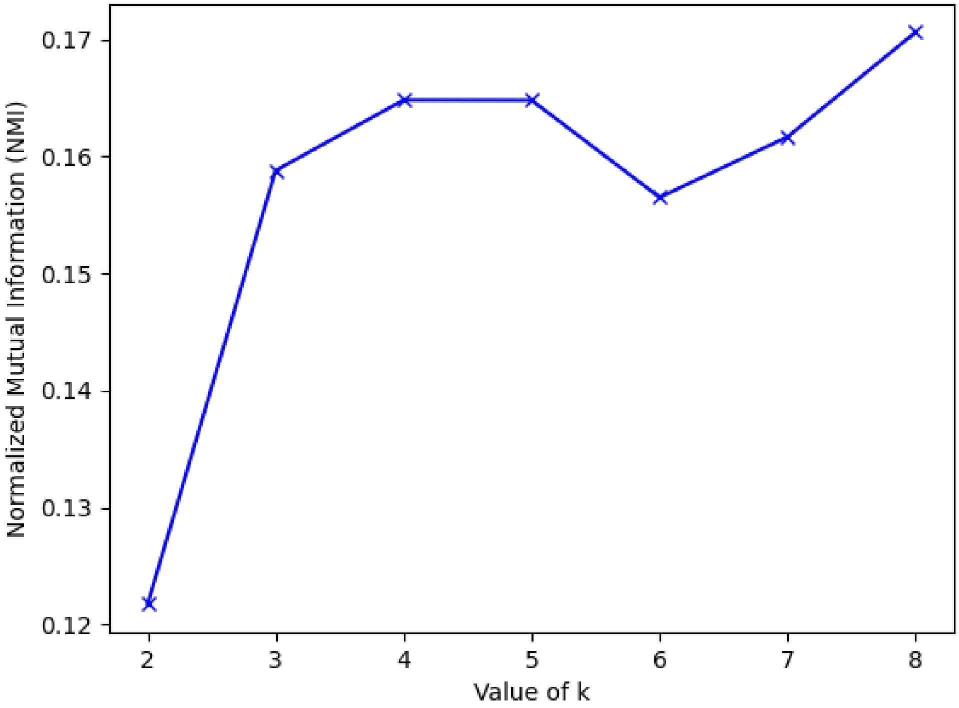
The graph of PCA is as follows:



The output of k-means algorithm is as follows:

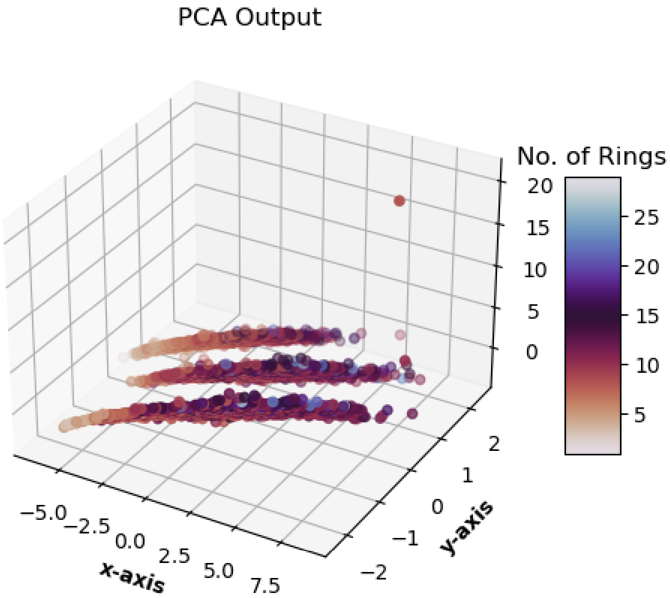
```
Value of k = 2
NMI value = 0.1217315098678864

Value of k = 3
NMI value = 0.1587998722269132

Value of k = 4
NMI value = 0.16481761867824635

Value of k = 5
NMI value = 0.16480002457430412

Value of k = 6
NMI value = 0.1565030582275382

Value of k = 7
NMI value = 0.16167149425344315

Value of k = 8
NMI value = 0.1706280321504191

The value of k for which NMI is maximum = 8
Corresponding maximum NMI value = 0.1706280321504191
```

The k vs NMI graph is as follows:



Second Run:

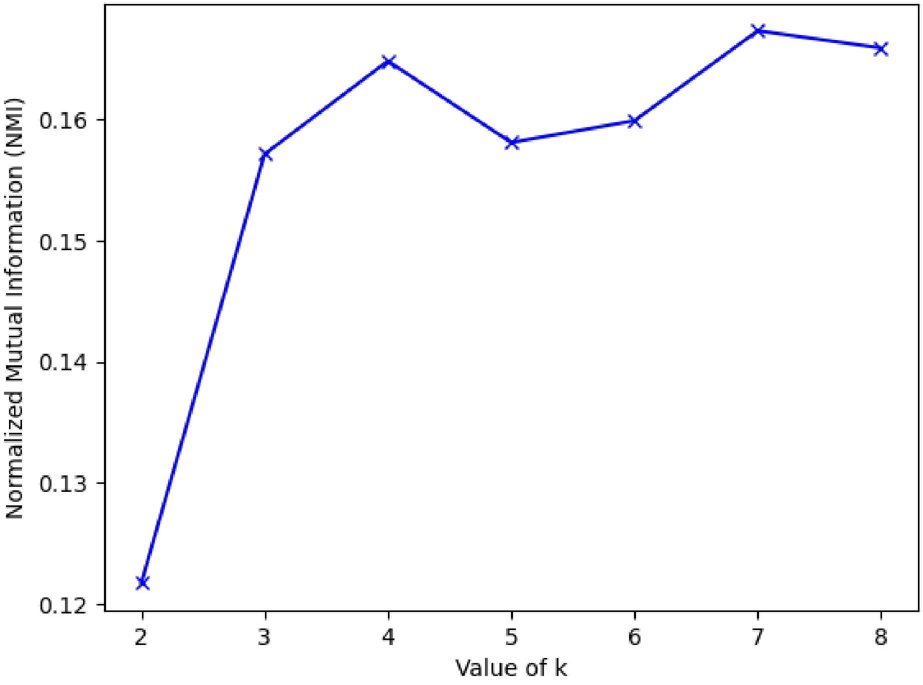
The graph of PCA is as follows:



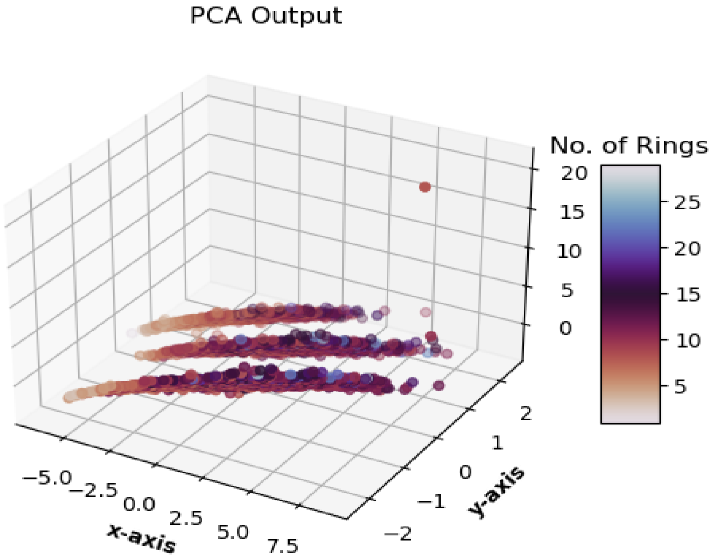
The output of k-means algorithm is as follows:

```
Value of k = 2
NMI value = 0.1217315098678864

Value of k = 3
NMI value = 0.1571277640376759

Value of k = 4
NMI value = 0.1648176186782465

Value of k = 5
NMI value = 0.1580924565104648

Value of k = 6
NMI value = 0.15989045724529938

Value of k = 7
NMI value = 0.167311548105901

Value of k = 8
NMI value = 0.16589118758204519

The value of k for which NMI is maximum = 7
Corresponding maximum NMI value = 0.167311548105901
```

The k vs NMI graph is as follows:



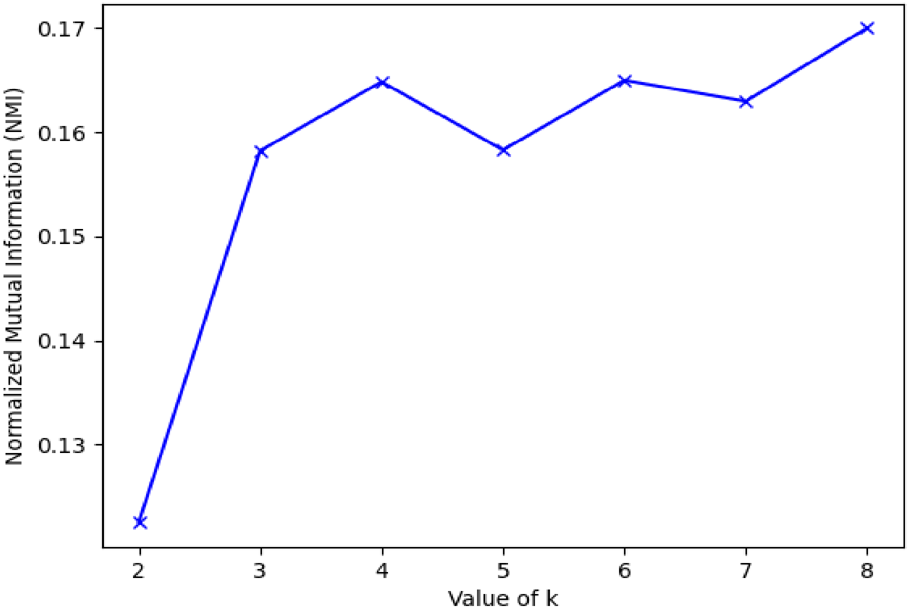
Third Run: The graph for PCA is as follows:



The output of k-means algorithm is as follows:

```
Value of k = 2
NMI value = 0.12251276586640608

Value of k = 3
NMI value = 0.15820353484841687

Value of k = 4
NMI value = 0.1648176186782465

Value of k = 5
NMI value = 0.1582805339524841

Value of k = 6
NMI value = 0.16496614350773786

Value of k = 7
NMI value = 0.16295307440824544

Value of k = 8
NMI value = 0.1699832463214506

The value of k for which NMI is maximum = 8
Corresponding maximum NMI value = 0.1699832463214506
```

The k vs NMI graph is as follows:


Normalized Mutual Information (NMI)
Value of k