MACHINE LEARNING

Assignment-2

Group - 042

Chukka Jaswanth Kumar (20CS10021)

Sourabh S Das (20CS30051)

CS60050

UNSUPERVISED LEARNING

Principal Component Analysis (PCA):

PCA is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.These new transformed features are called the Principal Components.

- It is one of the popular tools that is used for exploratory data analysis and predictive modeling.
- It is a technique to draw strong patterns from the given dataset by reducing the variances.
- It is also known as a general factor analysis where regression determines a line of best fit.
- PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.
- PCA is mainly used as the dimensionality reduction technique in various AI applications such as computer vision, image compression, etc.

K-Means Clustering:

K-means algorithm is an iterative algorithm that tries to partition the dataset into **K** pre-defined distinct non-overlapping subgroups (**clusters**) where each data point belongs to only one group.
It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

- The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

PROCEDURE:

Since the provided lung-cancer dataset has some missing values denoted by ?, we have replaced those by the mode of the remaining values in the respective columns.

- Then we have performed StandardScaler normalisation on the dataset.
- Then we have applied PCA on the dataset.
- Then the processed dataset is used for KMeans clustering with all values ranging from 2 to 8.
- In the process, the normalised mutual information (**NMI**) is calculated for each iteration of **KMeans**.

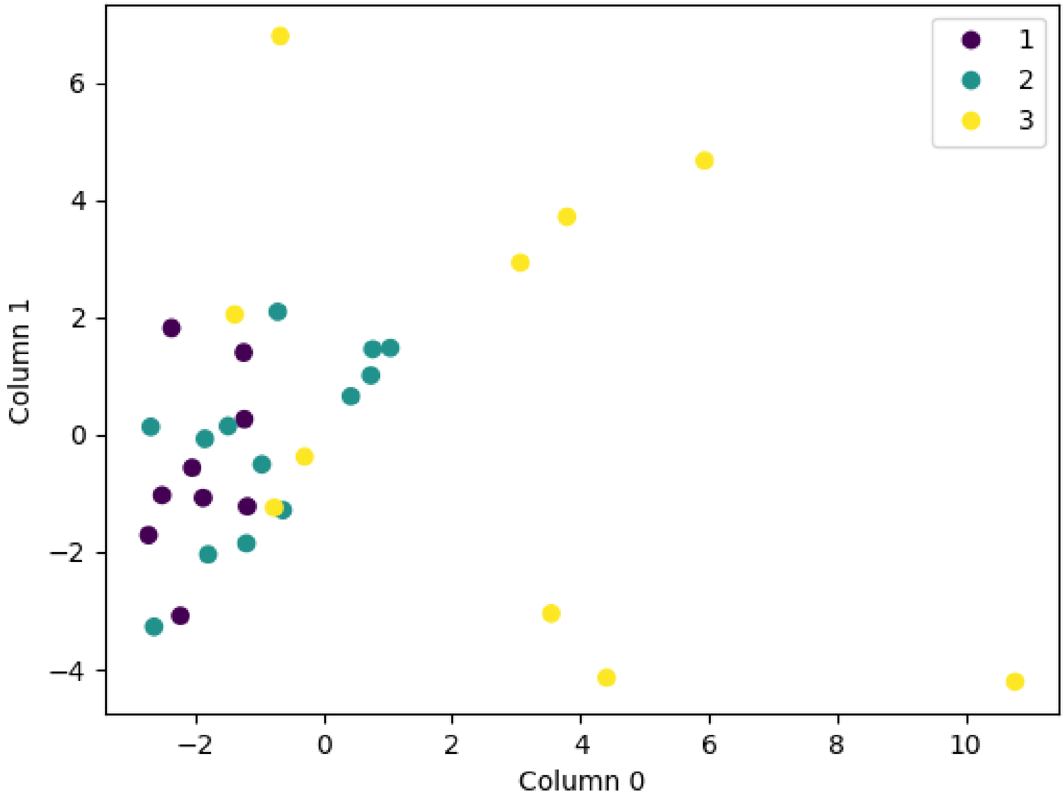
Functions used:

1. **clustering:** It is a function that takes in the dataset, list of centroids and the value of **K**. Then it performs KMeans clustering once and returns the generated new clusters and centroids.

2. **KMeans:** It is a function that takes in the dataset and the value of **K** as input. It then does certain initialisation for centroids (selecting k random data points). Then it calls the clustering function upto 1000 max until there is no change in the centroids or the clusters. Then based on the clusters formed, it assigns the cluster representative as the mode of results of each cluster. Based on this, it returns the list of final predicted values.

3. **cmp:** It is a utility function used for comparing the dictionaries of lists. It returns true if and only if every element in the list of the first dictionary matches the corresponding element in the list of the second dictionary by index as well as value and vice versa.

Graph of PCA:

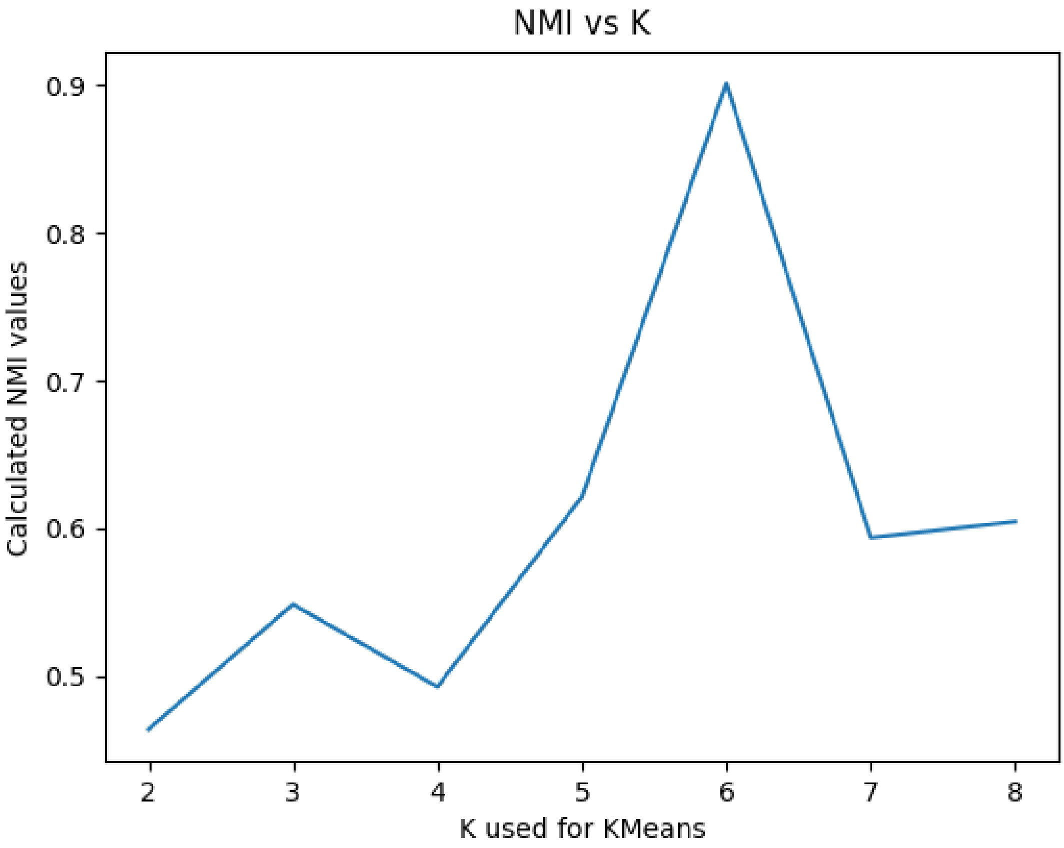


NMI values for respective K value:

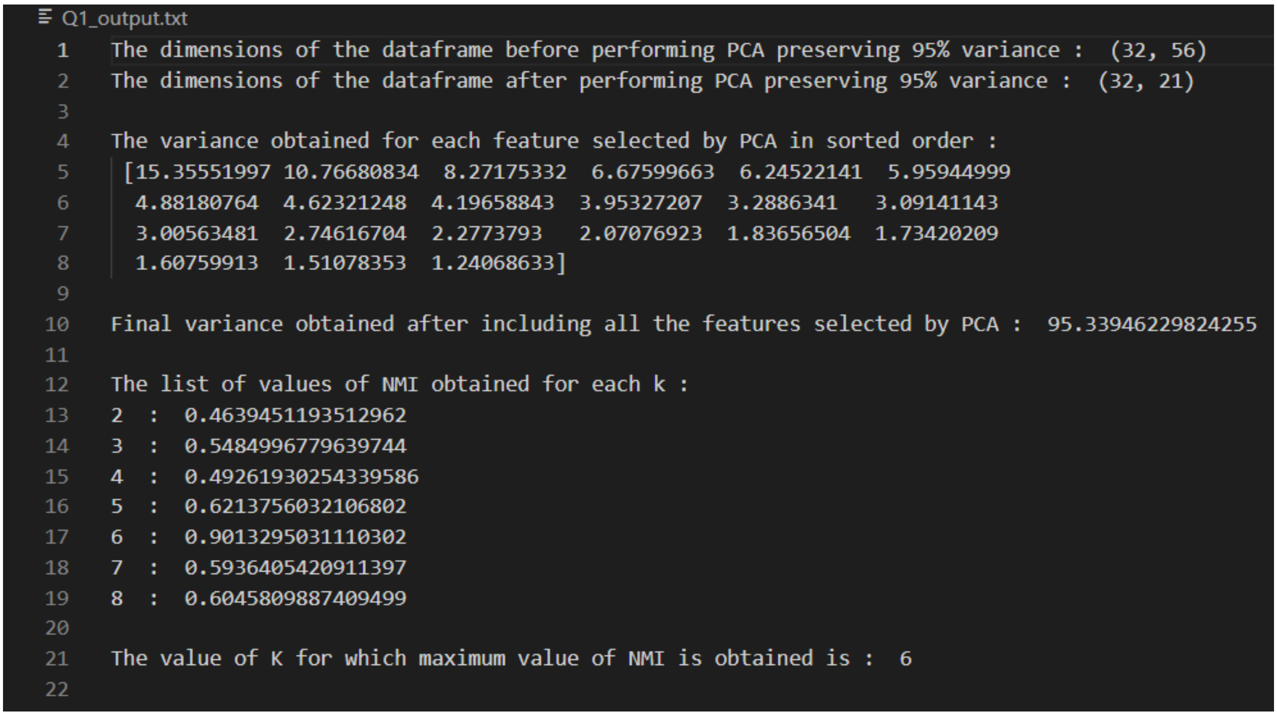
K	NMI
2	0.463945
3	0.548499
4	0.492619
5	0.621375
6	**0.901329**
7	0.593640
8	0.604580

⋆ The maximum value of **NMI** is obtained at **K=6**.

Graph of K v/s NMI:



Output:



```
The dimensions of the dataframe before performing PCA preserving 95% variance :  (32, 56)
The dimensions of the dataframe after performing PCA preserving 95% variance :  (32, 21)

The variance obtained for each feature selected by PCA in sorted order :
 [15.35551997 10.76680834  8.27175332  6.67599663  6.24522141  5.95944999
  4.88180764  4.62321248  4.19658843  3.95327207  3.2886341   3.09141143
  3.00563481  2.74616704  2.2773793   2.07076923  1.83656504  1.73420209
  1.60759913  1.51078353  1.24068633]

Final variance obtained after including all the features selected by PCA :  95.33946229824255

The list of values of NMI obtained for each k :
2  :  0.4639451193512962
3  :  0.5484996779639744
4  :  0.49261930254339586
5  :  0.6213756032106802
6  :  0.9013295031110302
7  :  0.5936405420911397
8  :  0.6045809887409499

The value of K for which maximum value of NMI is obtained is :  6

```

SUPERVISED LEARNING

PROCEDURE USED:

Since the provided lung-cancer dataset has some missing values denoted by **?**, we have replaced those by the mode of the remaining values in the respective columns. We have used the custom scalar function for normalising the data and then we have used the custom sampler function to split the data into **80%** training data and **20%** test data.

- Categorical encoding was not necessary since no string or coded data was involved.
- Then the binary SVM classifier was used with the kernel as linear, quadratic and radial basis function.
- Then the MLP Classifier was used with the given respective parameters:
    1. stochastic gradient descent optimiser
    2. learning rate as 0.001
    3. batch size of 32
    4. for the first classifier : 1 hidden layer with 16 nodes
    5. for the second classifier : 2 hidden layers with 256 and 16 nodes
- Then we selected the MLP Classifier model with best accuracy and used it with learning rates as 0.1, 0.01, 0.001, 0.0001 and 0.00001.
- Then we used forward feature selection, on the best MLP Classifier and listed out the selected features.
- Finally, we used ensemble learning with max voting technique.

Three models were considered :

1. **Quadratic SVM**
2. **Radial basis function SVM**
3. **Best MLP Classifier**

For each datapoint, the mode of results was considered. For all such models, the accuracy was calculated.

Functions used:

1. **scaler:** It takes the dataset as input. Then it calculates for each column their mean and standard deviation. Then it normalises the data by using the formula:

$$\text{new\_data} = \frac{\text{(old data - mean)}}{\text{standard deviation}}$$

2. **sampler:** It takes the dataset and the sampling fraction as input. Then it calculates the no. of points needed to be considered using the dataset size and the input sampling fraction. Then a random set of required no. of points are considered for output.

3. **forward_feature_selection:** It takes the training dataset, training result values, testing dataset and testing result values as input. Then it takes each feature one by one. Then it selects the feature which gives maximum increase in accuracy. Then it iterates until any further addition of any feature results in reduction in accuracy. In the process, it stores the index of the feature column which is selected permanently. Then it returns the list of such indexes.

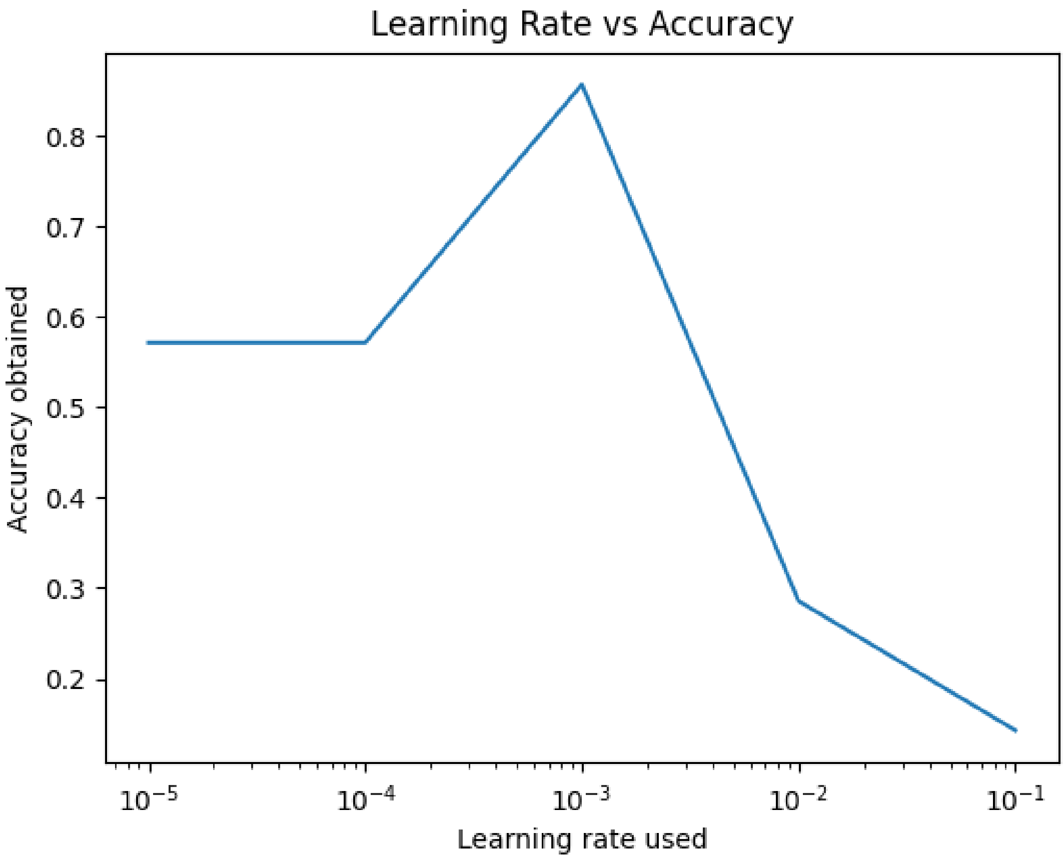
Accuracy obtained using different models:

- Accuracy for binary SVM classifier using linear kernel: 0.285714
- Accuracy for binary SVM classifier using quadratic kernel: 0.428571
- Accuracy for binary SVM classifier using radial basis function kernel: 0.285714

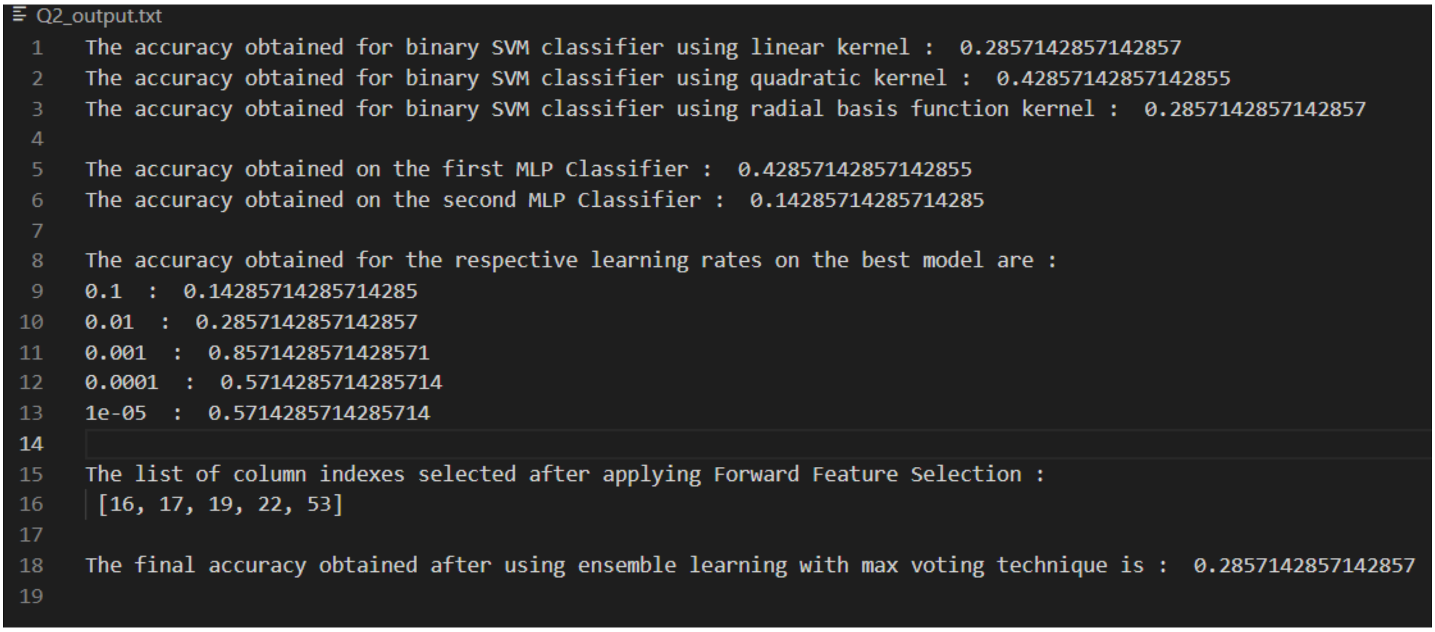
Accuracy obtained using Learning rates on best model:

Learning rate	Accuracy
0.1	0.14285714
0.01	0.28571428
0.001	0.85714285
0.0001	0.57142857
0.00001	0.57142857

Graph of Learning Rate v/s Accuracy:



Output:



```
The accuracy obtained for binary SVM classifier using linear kernel :  0.2857142857142857
The accuracy obtained for binary SVM classifier using quadratic kernel :  0.42857142857142855
The accuracy obtained for binary SVM classifier using radial basis function kernel :  0.2857142857142857

The accuracy obtained on the first MLP Classifier :  0.42857142857142855
The accuracy obtained on the second MLP Classifier :  0.14285714285714285

The accuracy obtained for the respective learning rates on the best model are :
0.1  :  0.14285714285714285
0.01  :  0.2857142857142857
0.001  :  0.8571428571428571
0.0001  :  0.5714285714285714
1e-05  :  0.5714285714285714

The list of column indexes selected after applying Forward Feature Selection :
 [16, 17, 19, 22, 53]

The final accuracy obtained after using ensemble learning with max voting technique is :  0.2857142857142857

```