Assignment -3

- **Members (GR-G, 65)**

➔ Name: Surajit Kundu, Roll No: 21MM91R09

➔ Name: Ankur Kumar Jaiswal, Roll No: 21MM62R08

1. **Read the dataset and randomly split it into train, validation, and test parts. The ratio of the train, validation, and test splits should be 70 : 10: 20 respectively.**

   Response:

   We have a total of 768 data points in our diabetes dataset. After splitting the dataset randomly in a 70:10:20 ratio, the number of data points in the training set is 537. The number of data points in the validation set and test set is 77 and 154 respectively.

Train	Validation	Test
537	77	154

Table-1

2. **Reduce the feature dimension of the above data into a two-dimensional feature space using Principle Component Analysis (PCA). Plot the reduced dimensional data of the train split in a 2d plane. In the plot, all data points of a single class should have the same color, and data points from different classes should have different colors.**

   Response :

☐ In the above input data, we have 768 rows and 8 columns. When we have applied the Principal Component Analysis(PCA) with two components on the input data, it gets reduced to a two-dimension feature space with 768 rows and 2 columns.

- After reducing the feature dimension of the above data into a two-dimensional feature space using Principle Component Analysis(PCA), the plot we are getting of the train split in a 2d plane is shown in the Figure-1 below :

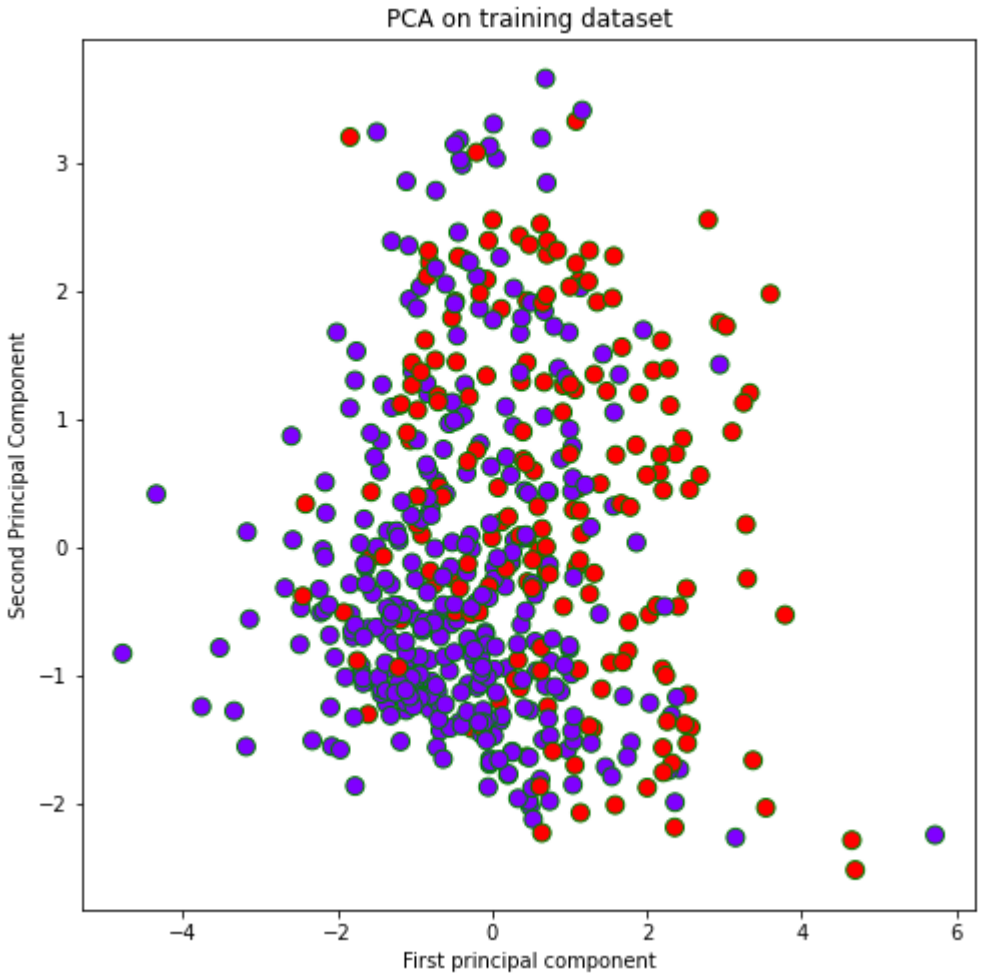

Figure - 1

- In the output column i.e. is Outcome, we have only 2 unique values(0 or 1). Out of 537 data points in the training set, We have 349 patients with no diabetes and 188 patients with diabetes. In the plot, **data points with blue color show the patient no have diabetes and red color shows a patient with diabetes.**

3. **Train an SVM classifier (sklearn.svm.SVC) on the reduced dimensional data generated from step 2. Try different kernel types by varying the appropriate hyperparameters of the classifier and computing the classification accuracy on the validation split. Show the validation accuracy for each combination in a tabular form. Choose the kernel for which the validation accuracy is highest and compute the test accuracy using that kernel. Print the test accuracy.**

Response:

☐ We have trained the two-dimensional data generated using PCA with the SVM classifier(sklearn.svm.SVC). We have also tried different kernels like linear, polynomial, sigmoid, radial basis function, etc by varying the appropriate hyperparameters like gamma, C, degree, etc of the classifier. We got different validation accuracy for different kernels and different hyperparameters. The list of accuracy is shown below in the tabular form:

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 0.1	70.12%
Sigmoid	Gamma = scale, C = 0.1	67.53%
Polynomial	Gamma = scale,C = 0.1	66.88%
Radial Basis Function	Gamma = scale,C = 0.1	68.83%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 10	70.12%
Sigmoid	Gamma = scale, C = 10	60.38%
Polynomial	Gamma = scale,C = 10	67.53%
Radial Basis Function	Gamma = scale,C = 10	69.48%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 100	70.12%
Sigmoid	Gamma = scale, C = 100	60.38%
Polynomial	Gamma = scale,C = 100	67.53%
Radial Basis Function	Gamma = scale,C = 100	70.12%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 1000	70.12%
Sigmoid	Gamma = scale, C = 1000	60.38%
Polynomial	Gamma = scale,C = 1000	67.53%
Radial Basis Function	Gamma = scale,C = 1000	67.53%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 0.1	70.12%
Sigmoid	Gamma = auto, C = 0.1	62.98%
Polynomial	Gamma = auto, C = 0.1	67.53%
Radial Basis Function	Gamma = auto, C = 0.1	70.12%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 10	70.12%
Sigmoid	Gamma = auto, C = 10	55.84%
Polynomial	Gamma = auto, C = 10	67.53%
Radial Basis Function	Gamma = auto, C = 10	68.83%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 100	70.12%
Sigmoid	Gamma = auto, C = 100	55.19%
Polynomial	Gamma = auto, C = 100	67.53%
Radial Basis Function	Gamma = auto, C = 100	67.53%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 1000	70.12%
Sigmoid	Gamma = auto, C = 1000	55.19%
Polynomial	Gamma = auto, C = 1000	66.88%
Radial Basis Function	Gamma = auto, C = 1000	66.23%

- ☐ We can conclude from the above results that the linear kernel has the highest validation accuracy above all kernels. So, the test accuracy using a linear kernel is shown below :

Kernel	Hyperparameter	Validation Accuracy
Linear	Gamma = scale or auto, C = [0.1,10,100,1000]	70.12%

4. **Reduce the feature dimension of the above data into a one-dimensional feature space using Linear Discriminant Analysis (LDA). Plot the reduced dimensional data of the train split. In the plot, all data points of a single class should have the same color, and data points from different classes should have different colors.**

Response:

- In the input data, we have 768 rows and 8 columns. When we have applied the Linear Discriminant Analysis(LDA) with one component on the input data, it gets reduced to a one-dimension feature space with 768 rows and 1 column.
- After reducing the feature dimension of the above data into a one-dimensional feature space using Linear Discriminant Analysis (LDA), the plot we are getting of the train split in a 2d plane is shown in the below figure-2:

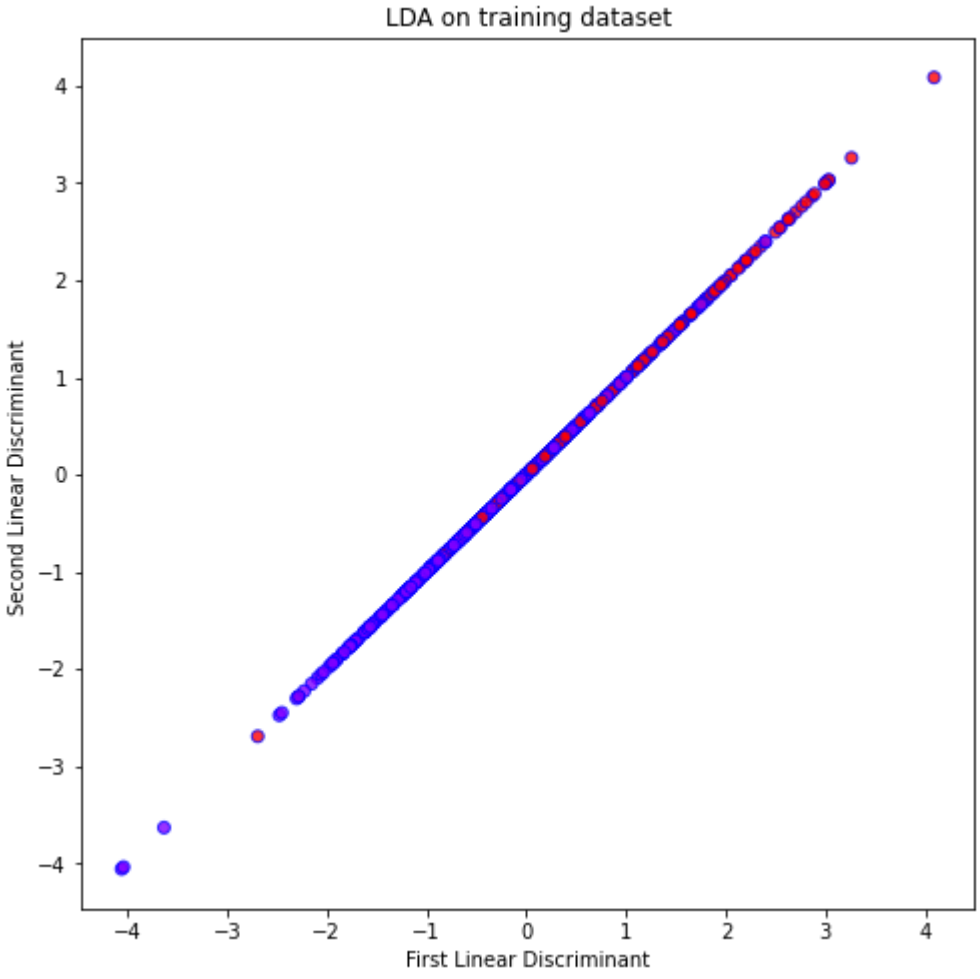

Figure-2

- In the output column i.e. is Outcome, we have only 2 unique values(0 or 1). Out of 537 data points in the training dataset, We have 349 patients with no diabetes and 188 patients with diabetes. In the plot,

data points with blue color show the patient no have diabetes and red color shows a patient with diabetes.

5. **Repeat Step 3 on the data obtained from Step 4.**

Response:

- ☐ We have trained the two-dimensional data generated using PCA with the SVM classifier(sklearn.svm.SVC). We have also tried different kernels like linear, polynomial, sigmoid, radial basis function, etc by varying the appropriate hyperparameters like gamma, C, degree, etc of the classifier. We got different validation accuracy for different kernels and different hyperparameters. The list of accuracy is shown below in the tabular form:

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 0.1	78.57%
Sigmoid	Gamma = auto, C = 0.1	66.23%
Polynomial	Gamma = auto, C = 0.1	75.97%
Radial Basis Function	Gamma = auto, C = 0.1	77.92%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 10	79.22%
Sigmoid	Gamma = auto, C = 10	57.79%
Polynomial	Gamma = auto, C = 10	75.97%
Radial Basis Function	Gamma = auto, C = 10	77.27%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 100	79.22%
Sigmoid	Gamma = auto, C = 100	57.79%
Polynomial	Gamma = auto, C = 100	75.97%
Radial Basis Function	Gamma = auto, C = 100	75.97%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = auto, C = 1000	79.22%
Sigmoid	Gamma = auto, C = 1000	57.97%
Polynomial	Gamma = auto, C = 1000	75.97%
Radial Basis Function	Gamma = auto, C = 1000	74.67%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 0.1	78.57%
Sigmoid	Gamma = scale, C = 0.1	77.27%
Polynomial	Gamma = scale, C = 0.1	74.67%
Radial Basis Function	Gamma = scale, C = 0.1	79.22%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma =scale, C = 10	79.22%
Sigmoid	Gamma =scale, C = 10	75.32%
Polynomial	Gamma =scale, C = 10	75.97%
Radial Basis Function	Gamma =scale, C = 10	77.27%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 100	79.22%
Sigmoid	Gamma = scale, C = 100	74.67%
Polynomial	Gamma = scale, C = 100	75.97%
Radial Basis Function	Gamma = scale, C = 100	77.27%

Kernel	Hyper-parameter	Validation Accuracy
Linear	Gamma = scale, C = 1000	79.22%
Sigmoid	Gamma = scale, C = 1000	74.67%
Polynomial	Gamma = scale, C = 1000	75.97%
Radial Basis Function	Gamma = scale, C = 1000	75.32%

☐ We can conclude from the above results that the linear **kernel** has the highest validation accuracy above all kernels. So, the test accuracy using a linear kernel is shown below :

Kernel	Hyperparameter	Validation Accuracy
Linear	Gamma = scale or auto, C = [0.1,10,100,1000]	76.62%

6. **Is there any significant difference between the final test accuracy obtained from Step 3 and Step 5 ? If so, justify the results with the proper reason?**

   Response:

- ☐ As we can see from the results that test accuracy with the LDA dimensionality reduction is higher than the test accuracy with the PCA dimensionality reduction. Using PCA, they can lose some spatial information which is important for classification, so the classification accuracy decreases. PCA performs better in cases where the number of samples per class is less. Whereas LDA works better with large datasets having multiple classes; class separability is an important factor while reducing dimensionality.

7. **Prepare a report clearly describing the process followed and showing the results of the above steps.**

Response: Please refer to the document “Procedure and Results.pdf ” for procedure and results.